A Computational Memory and Processing Model for Prosody

by

Janet Elizabeth Cahn

B.A. in Computer Science, Mills College (1983)
S.M., Visual Studies, Massachusetts Institute of Technology (1989)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1999

Author .....
Program in Media Arts and Sciences, School of Architecture and Planning
October 30, 1998

Certified by .....
Kenneth Haase
Visiting Professor, Program in Media Arts and Sciences
Academic Advisor

Accepted by .....
Stephen A. Benton
Chairman, Departmental Committee on Graduate Students

A Computational Memory and Processing Model for Prosody

by

Janet Elizabeth Cahn

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
on October 30, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis links processing in working memory to prosody in speech, and links different working memory capacities to different prosodic styles. It provides a causal account of prosodic differences and an architecture for reproducing them in synthesized speech. The implemented system mediates text-based information through a model of attention and working memory.

The main simulation parameter of the memory model quantifies recall. Changing its value changes what counts as given and new information in a text, and therefore determines the intonation with which the text is uttered. Other aspects of search and storage in the memory model are mapped to the remainder of the continuous and categorical features of pitch and timing, producing prosody in three different styles: for small recall values, the exaggerated and sing-song melodies of children's speech; for mid-range values, an adult expressive style; for the largest values, the prosody of a speaker who is familiar with the text, and at times sounds bored or irritated. In addition, because the storage procedure is stochastic, the prosody from simulation to simulation varies, even for identical control parameters. As with with human speech, no two renditions are alike.

Informal feedback indicates that the stylistic differences are recognizable and that the prosody is improved over current offerings. A comparison with natural data shows clear and predictable trends although not at significance. However, a comparison within the natural data also did not produce results at significance.

One practical contribution of this work is a text mark-up schema consisting of relational annotations to grammatical structures. Another is the product – varied and plausible prosody in synthesized speech. The main theoretical contribution is to show that resource-bound cognitive activity has prosodic correlates, thus providing a rationale for the individual and stylistic differences in melody and rhythm that are ubiquitous in human speech.

Academic Advisor: Kenneth Haase
Title: Visiting Professor, Program in Media Arts and Sciences

Thesis Committee

Committee Chair

Kenneth Haase, Ph.D.
Visiting Professor
Program in Media Arts and Sciences

Research Advisor

Stan Sclaroff, Ph.D.
Assistant Professor, Computer Science
Boston University

Member

Julia Hirschberg, Ph.D.
Head, Human-Computer Interface Research Department
A.T. & T. Labs – Research

Member

Pattie Maes, Ph.D.
Associate Professor
Program in Media Arts and Sciences
Massachusetts Institute of Technology

Acknowledgments

Thanks first to my committee. Ken Haase has encouraged both creativity and community spirit in his students and has been an enthusiastic supporter of my research. Julia Hirschberg has provided expert commentary and in addition, has always pointed me to exactly the right people, papers and data. Her insight and standards of scholarship have made this a far better piece of work. Pattie Maes has the rare ability to identify the key strengths and weakness of any work – her advice and encouragement have been both practical and deep. Stan Sclaroff has been indispensable as a friend and an advisor. He provided encouragement and feedback at key points and was instrumental in helping me defend and finish. It could not have happened without him.

This work would not have begun without the help of Dave Barrett, who generously gave of his time and skill to help me machine my chair to ergonomic standards. It would not have proceeded without the brilliant practical and theoretical suggestions of Marilyn Walker and Susan Brennan. It would have taken even longer were it not for the generosity of: Marilyn Walker, who gave me her code for the memory model; Anil Chakravarthy, who gave me his WordNet code; Trevell Perkins and Erin Panttaja and Hannes Vilhjalmsson, who provided technical assistance with the speech synthesizer software and related hardware; Mari Ostendorf and Stephanie Shattuck-Hufnagel, both of whom gave me many useful papers and provided me with the early versions of their annotated speech data.

Philip Resnik gave me the word collocation data from his dissertation and Lisa Stifelman gave me the timing data from hers. While my explorations of their data did not find their way into this work, I am grateful for their trust and generosity.

Many other friends and colleagues have patiently and expertly answered my many technical questions, both mundane and complex, especially, Deb Roy, Sumit Basu, Baback Moghaddam, Kris Popat, Lee Campbell, Raissa d'Souza, Bill Butera, Josh Smith, Ravi Pappu, Dan Ellis, Louis Weitzman, Warren Sack, Mike Best, Anil Chakravarthy, Tim Chklovski and Bernd Schoner.

Mike Best's organizational and content suggestions have greatly improved this document. My father's editing expertise has greatly improved the first two chapters.

The writing has also benefited from feedback from Ken Haase, Pushpinder Singh and Wendy Plesniak.

Thanks also to faculty, students and staff outside my research group, especially to: Sandy Pentland for sensible and insightful advice on both practical and research matters; Aaron Bobick, Neil Gershenfeld, Bruce Blumberg and Justine Cassell for useful feedback at various stages of the work; Fred Martin, Steve Strassman and Lenny Foner for helpful technical advice on all kinds of questions over the years; Henry Holtzman and Ben Lowengard for sublimely effective social and technical solutions; Will Glesnes, Viet Anh and Tomas Revesz for patient and expert hardware and software support; Linda Peterson, Santina Tonelli and Laurie Ward, for clear and effective counsel on academic matters; Greg Tucker and Dennis Irving for timely and wise logistical help. Rebecca Prendergast, Glen Sherman, Felice Gardner and Florence Williams for providing administrative support with unmatched professional and human competence.

Many other people have (also) provided support and friendship, all of which have been incalculably wonderful and necessary. Aside from those friends I have already thanked, I am also excruciatingly grateful to Laura Teodosio, Tomoko Koda, Sara Elo, David Dean, Babette Wils, Jennifer Gonzalez, Steve Whittaker, Candy Sidner, Kathy Blocher, Michelle Fineblum, Michael Gevelber, Maribeth Back, Tara Rosenberger, Marina Bers, Josh Bers, Paula Hooper, Ingeborg Endter, Joey Berzowska, Flavia Sparacino, Dave Rosenthal, Martin Hadis, Mike Travers, Soo-jin Chung, Chris Verplaetse, Lorin Wilde, John Underkoffler, Teresa Marrin, Judith Donath, Nuria Oliver, Kristin Hall, Kim Binsted, Kim Foley, Judy Bornstein, Nancy Etcoff, Amy Pearl, Stephanie Graham, Kaitilyn Riley, Nancy Blachman, Steve Gray, Vicki O'Day, Anna Henderson, Lisa Borden, Lydia Mann, Barbara LiSanti, Cathleen Craviotto, Pat Lordan, John Wilkes, Sally Stuffin, Alex Sherstinsky, to name a few...

Final thanks go to my family, immediate and extended. My parents, Arno and Blossom, provided financial support for my ergonomic research. They also provided strategic professional advice and unconditional support and encouragement. It is a debt only partially repaid by graduating. My sister, Nancy, and brother-in-law, Craig, have been supportive despite their complete bewilderment at my career path. My brother Jonathan has held to his unwavering belief that I would graduate before the millennium, and recognizes it for the Sign that it is. I also thank my cousins Joanne, Roberta, Dan, Zach, Doron and Ilan for providing welcome and warmth.

This research was generously supported by funding from the News in the Future consortium.

In memory of Diane McEntyre, Gena Tan, Zvi Erez and Celia Sack, and dedicated to the three furies, the three graces — Arianna, Gabrielle and Nicole.

Contents

List of Figures

List of Tables

Part I

Foundations

Chapter 1

Introduction

Few utterances are alike. Whether the same words are repeated many times by one speaker, or once by many speakers, each rendition will have a unique prosody – that is, the combination of melody and rhythm will vary with the speaker and the speaking situation. The nature of the variation is predictable when the speakers are socially or professionally distinct. Children's prosody is animated while that of adults is more subdued. Sports announcers shout while classical music announcers soothe. Intra-speaker, inter-speaker and stylistic variations abound and are ubiquitous in both public and private life, in readings of prepared text and in daily conversation.

The study of prosodic variation is a study of the actual instead of the ideal, of deviations instead of the norm. It is a study of linguistic *performance*, which is one half of the distinction made famous by Chomsky [Cho65]. The other half is *competence*, a term that Chomsky employs to distinguish a grammar that generates well-formed sentences from the actual sentence structures that a speaker produces on the fly and in real time. Chomsky draws this distinction to support his contention that by focusing on competence, grammar can be studied as a formal system. As he explains, ([Cho65], pp.9):

> To avoid what has been a continuing misunderstanding, it is perhaps worth while to reiterate that a generative grammar is not a model for a speaker or a hearer. It attempts to characterize in the most neutral possible terms the knowledge of the language that provides the basis for actual use of language by a speaker-hearer.

That is, for the purpose of advancing linguistic theory, the grammar of a language should be abstracted away from a particular user of that grammar. Beyond recommending a method of study and analysis, Chomsky uses this distinction to support his claim that there exists a knowledge of grammar which is *universal* among humans and whose expression in a particular language is the result of (universal) parameters

that are turned on or off [Coo88]. For example, some grammars (Japanese) require explicit lexical markers for attentional focus, while others (English) do not. It is the feature of abstracting away from a particular speaker or hearer that identifies most theories of prosody as theories of prosodic competence. However, there is no further claim that prosodic abstractions are innate. They may have been learned. Similarly, it is not necessary to claim that the prosodic abstractions are universal[1], but rather, that they depict a recognized standard for a language, which may be either an ideal or a norm.

Theories of prosodic competence present formal and generative accounts. In such accounts, intonation (the melody) conveys the attentional and logical propositions expressed by the structure and semantics of the text ([WH85, PH90, Bir91, Oeh91, Ste90]). Timing features – syllable stress within a word and word stress within a phrase – are the consequence of phonetic, morphological and syntactic forms ([CH68, LP77, Sel84]). Because these theories link prosody to the structure and semantics of text, they have little to say about why prosody varies within and among speakers.

The performance explanation I propose does not reject theories of competence but adds that the prosodic elucidation of a text also depends on a particular speaker's ability to understand it. More specifically, my proposal links prosody to the attentional capacities of the speaker, and prosodic differences to differences in the attentional capacities within and among speakers. This is not a traditional explanation and so raises several questions.

The first question concerns the meaning of attentional capacity. I use a definition that is simple and quantitative – attentional capacity is the number of items that a speaker can access in her working memory. A large capacity speaker can access many items; a small capacity speaker can access only a few. Attentional capacity is separate from storage capacity. Even though working memory may be large and full, a speaker with a small attentional capacity can access very little of its contents.

The second question concerns the source of differences in attentional capacity. One claim is that differences in attentional capacity are innate. Indeed, some memory pathologies are explained by extremely limited attention. However, the exploration of pathologies is not the focus of this work. The explanation I prefer is that differences in attentional capacity (within a normal range) are task related and that attention is limited with respect to a particular cognitive task or is situational. For example, a child (to whom many things are new) is likely to be performing many computations in working memory at the same time. Consequently, only a small part of her[2] attentional capacity is available for listening and speaking. The opposite case is represented by a person who is knowledgeable or skilled with regard to a particular task. By

[1] Although Bolinger [Bol58] states that the use of a rising pitch to mark new information is universal across languages.

[2] Following [Bre90] and Heeman [Hee91] the speaker is denoted by female pronouns, the hearer by male. This imperfect usage replaces the more imperfect usage offered by current American English.

definition, this person is in possession of numerous and salient facts about the task. By the same token, she has a large attentional capacity, as defined by the number of accessible items. Accordingly, skill at a particular task may be depicted by differences in attentional capacity. This explanation does not depend on pathology but rather accounts for the influence of both the current situation and acquired skills.

The next question deals with the rationale for linking attentional capacity to prosodic variation. A partial answer is that the link between attention and prosody is already established for intonation. Speakers typically mark salient information with rising pitch if they believe that it is *new* to the hearer or no longer salient to him [Wal92, Wal93, Cru93]. Conversely, when they believe that the hearer is already aware of the information, it is typically unmarked [Bro83]. Intonational variance is therefore explained by the speaker's presumptions about the hearer's attentional capacities and needs.

The last question touches on the reason for linking prosody to the attentional capacities of the speaker. The main reason is that even though a speaker typically adapts her prosody for the hearer, her own attentional capacity (for the task) constitutes the upper bound on her performance. Thus, while the ideal prosody is optimized for the hearer, the prosody produced by the limited capacity speaker is often a compromise between optimizing the information content for the listener and optimizing the ease of production for herself. In a situation where the text or its content are familiar, a speaker can most afford to adapt to the needs of the hearer. For example, when an adult instructs a child, she will typically use exaggerated intonation and speak slowly and clearly as well [FS84, AK96]. Her prosody mainly reflects her assumptions about the child's limited attentional capacity. However, when production itself is difficult, her prosody is likely to reflect these difficulties, for example, in pauses that interrupt the flow of text [GE58].

The main goal of this work is to demonstrate a link between prosodic variation and attentional and working memory capacity. A supporting goal is to develop a model that predicts and generates prosodic quantities as well as qualities. The final goal is to apply it to synthesized speech to improve its expressive and stylistic range. The remainder of this section describes the terms of the discussion and the plan of the thesis in support of these goals.

1.1 Terms

In this section, I define the speech terms that are used throughout this work. The first is **prosody** itself. It is defined as the melody and rhythm of speech. The melody is described by fluctuations in fundamental frequency over the course of an utterance. The **fundamental frequency** is determined by the rate of vibration of the vocal cords. It is the lowest harmonic in the speech signal. The other frequencies contribute to

the perception of both voice quality and the actual phoneme type. Phonemes are the unique sounds of a language that are combined into words. They are functionally defined such that exchanging one for another produces a different word. Sounds that are separate phonemes in one language may be a single phoneme in another. For example, *r* and *l* are unique phonemes for English, but count as one phoneme for Japanese.

The fundamental frequency is referred to in the speech literature as F0 ("F-zero" or "F-nought") to distinguish it from F1, F2, etc., which are the formants – high energy regions in the speech spectrum – that characterize a phoneme. F0 is a function of vocal cord vibration, while a formant is a function of a vocal tract configuration, which boosts some frequencies and dampens others.

Pitch denotes the perception of fundamental frequency. Intonation describes the communicative aspects of melody in speech, especially pitch fluctuations at the word level and at the end of a phrase. The change in pitch over time is called a pitch contour; a distinctive pitch contour applied to a single word is a pitch accent [Bol58]. Speakers use pitch accents to highlight salient words and thus, the concepts to which they refer. They use distinctive phrase final contours to group words into phrases, distinguish sentence types such as a declarative or a question, and to convey discourse structure – for example, phrase final rises convey continuation and phrase final falls convey completion.

The prominence of a pitch accent or phrase final contour is the magnitude of its rise or fall. The distance between the highest and lowest F0 values of a phrase describes the pitch range. The pitch range often undergoes change over the course of a phrase. It is typically widest at the beginning and shrinks thereafter.

Pitch accents are acoustical devices that convey emphasis at the sentence level. Changing the location or the type of the accent changes the meaning of the sentence (whereas changing a phoneme changes the meaning of the word). Over-articulating or increasing the duration of a word are two other means for conveying sentential stress. In general, the location of sentential stress is variable. However, the location of lexical stress is restricted by the citation form of a word, which describes its most typical phoneme and stress pattern. The lexically stressed syllable for the word bears primary stress. For example, the first syllable carries primary stress for the noun, "*INsult*",' and no stress for the verb, "*inSULT*".' Longer words show more gradations of stress. For example, in "CON*dem*NA*tion*", "NA" carries primary stress and "CON" carries secondary stress.

The rhythmic component of prosody is partially characterized by patterns of stress across the phrase (such as iambic pentameter for poetry) and partially by the duration of words and pauses. Pauses may be expressed by vocalization, such as "um" or "uh", or by silence. The first is called a filled pause, the second is called an unfilled pause. Pauses are also distinguished by their location in the structure of a phrase or sentence. Hesitation pauses occur within a syntactic and intonational unit, while fluent

pauses occur between syntactic units. The actual size or make-up of these units is unspecified in both definitions. Typically, hesitation pauses occur before **content words** – those words within a language whose form, content and members change relatively quickly over the life of the language. Usually, these are nouns, verbs, adjectives and adverbs. **Function words**, on the other hand, describe the small set of words whose membership is relatively stable over time and whose use and meaning are slow to change. Such words have little intrinsic meaning but are essential to the syntax. Pronouns, prepositions and conjunctions are examples of function words.

The final set of definitions distinguishes the two main types of speech synthesis. **Text-to-speech** systems take unrestricted text and assign phrasing and intonation according to punctuation, the part of speech for a word and perhaps a syntactic parse. **Concept-to-speech** systems (formerly, "message-to-speech") are text generation systems with a speech interface and a task model. The text of an utterance is derived from both the system's representation of domain knowledge and its model of the current state of the domain or communicative task.

1.2 Plan of the thesis

This is document in four parts: Foundations (chapters 1 through 4), Approach (chapters 5 through 10), Results and Evaluation (chapters 11 and 12) and Conclusion (chapter 13).

Chapter 2 discusses the principle impetus for this work: the need for a cognitive model that will generate prosodic quantities as well as qualities, and for improved prosody in synthetic speech.

Chapter 3 reviews the theoretical and empirical foundations and divides them according to whether they describe competence or performance. This is followed by a discussion of prosodic style and genre. Finally, the computational theories of prosody are reviewed. The intonational theory developed by Pierrehumbert and colleagues is reviewed in depth because it is widely accepted and is one of the foundations of this work.

Chapter 4 discusses two areas of related work: the work on prosody for synthetic speech, and the few computer simulations that focus on the effects of cognitive resource bounds on linguistic performance. The related synthesis work includes both text-to-speech and concept-to-speech efforts. The most directly related computer simulations model the effects of limited attention on learning, comprehension and interaction. However, a few other simulations are reviewed, including connectionist models of language processing.

Chapter 5 provides an overview of the work and its components: the memory model,

the text analysis and the mapping of search and storage in working memory to pitch and timing in speech. Each component is discussed in more detail in the chapters that follow.

Chapter 6 introduces the computational model of attention and working memory developed by Landauer [Lan75] to simulate the effects of limited search and retrieval on learning and recall. I use it in LOQ, the implemented system, to demonstrate the effects on prosody of limited attention and storage in working memory. The model is critiqued with respect to insufficiently defined properties, in order to introduce changes that are more in line with limited capacity assumptions.

Chapters 7, 8 and 9 describe the operation of LOQ, the system that maps the processing of text in working memory to pitch and timing in speech. Chapter 7 describes how a reader's knowledge of language use and meaning is approximated by the combination of text mark-up, automated text analysis and the incorporation of online linguistic databases. Chapter 8 describes the operations of attention and working memory that are relevant to storage and retrieval. The essence of retrieval is the search for a *match* in memory to the current stimulus. A match is determined by domain-dependent criteria. For speech and language, they include pragmatic, semantic, syntactic, acoustic and orthographic criteria. Chapter 9 presents the algorithms that map search and storage in the memory model to both categorical and quantitative features of prosody. They produce specifications for pitch accents, phrase final contours, intonational prominence, pitch range, the duration of a word and the duration and location of fluent pauses.

Chapter 10 describes the integration of linguistic databases, linguistic input, memory operations and the mapping algorithms in LOQ, the computer-based simulation system for generating prosody in speech.

Chapter 11 presents the results of the simulations. These show the effects on pitch and timing of three simulation parameters: (1) the size of the working memory model; (2) the attentional capacity within the model; (3) the speed at which the focus of attention changes.

Chapter 12 evaluates the intonation generated by LOQ by comparing it to human readings of the same text – a news story originally broadcast on WBUR in Boston. To compare the two, intonation is treated as a text mark-up that is performed by a speaker. This motivates the use of Cohen's kappa as an evaluation tool. It is a statistic originally developed to determine how closely evaluators agree when classifying the same data [Coh60]. The variation within the natural and LOQ speech is first examined separately, and then the two are compared.

Chapter 13 discusses the contributions and future directions and concludes the work. The key contributions are the limited resource explanation of prosody and the implementation that produces varied prosody as a consequence of limited attention and working memory. The future directions propose work in both speech synthesis and

cognitive modeling.

Chapter 2

Motivations

My aim is to demonstrate that prosody is not predictable from text alone and that claims about default prosody encode implicit assumptions about the processing capabilities of the default speaker. More strongly, prosody is not speaker independent and cannot be. It is, after all, delivered by a speaker and not by a text.

If this claim is plausible, it gives rise to the challenge of devising the best characterization for differences in processing abilities. The few extant analyses of speaker-dependent prosody locate the source of differences in the emotional state of the speaker. These analyses have their own problems, since emotion labels are often subjective. At best, they form a qualitative taxonomy, if not several taxonomies. Moreover, none of the taxonomies so far has been generative – they do not in and of themselves account for the central causes and effects of the phenomena they describe. Probably as a reflection of the fact that there is not yet a generative theory of emotion, efforts to produce synthetic emotional speech have opted for a descriptive approach of the output, and one whose acoustical descriptors are selected beforehand.

The processing approach I adopt is generative, quantitative and does not fall prey to the problems of a descriptive taxonomy. It links prosody to the simplest mechanisms of working memory, namely, to those that affect attention and recall. It further claims that quantifiable differences in attentional and recall capacities can be linked to characteristic differences in prosody.

The advantage of this perspective is that a focus on attention and recall neither implies nor forbids any particular emotional or affective state. Furthermore, they are common to all forms of cognitive impression and expression. In addition, the link between attention and prosody is theoretically and empirically supported. Intonation conveys salience and grouping and therefore, is an attentional device by definition. Similarly, duration has been shown to respond to differences in retrievability – salient words are uttered more quickly [FH87, Lev89, FLB97] and longer pauses occur before rare words [GL83].

A second and related motivation is to expand the expressive range of synthetic speech. While synthetic speech has attained significant intelligibility and a reasonable (male) voice quality [Kla87], it remains under-employed in all areas where its use should be obvious: in information applications (such as one that gives driving directions [DH88, DT87, Dav89]), as a tool for dramatic prototyping [Cah90, Cah], as the voice of an autonomous agent [WCW97], as a reader to the blind [Ram94] or as a voice for the voice-impaired [Kla87, Cah90, MAN88, MA93]. The main problem is that synthesized speech is still acoustically and perceptually impoverished. The voicing source is often too regular, the phoneme transitions awkward and the prosody inappropriate.

In general, synthesized speech contains less information per speech feature, less redundancy across features, and information that is often distorted in comparison to natural speech. Luce *et al.* [LFP83] describe it as perceptually equivalent to natural speech in a noisy environment because the acoustics of both are degraded. Consequently, it requires the listener to direct more of his processing toward recovering basic forms, such as phonemes and words, rather than toward encoding and manipulating content. Not surprisingly, experimental subjects show impaired performance in memory and comprehension tasks when synthetic speech is used to present information [LFP83, WT85, RPL+91].

The quest for better prosody is one avenue by which synthetic speech can be improved. It promises to produce prosody that is both more accurate (it expresses plausible propositions about the text) and more natural (it produces informative prosodic patterns that are also easily processed). For text-to-speech synthesis, the quest has mainly been for an algorithm that will generate the best possible *discourse-neutral* [BF90] prosody. This is a necessary goal for unrestricted and minimally understood text. However, it produces prosody that is repetitive and often inappropriate to the content of the text. Concept-to-speech synthesis is better able to produce prosody that conveys both task semantics and discourse structure because the generating system has access to both. However, its prosody is only as good as its knowledge representations and the algorithms that map them to prosody.

Moreover, both approaches fail to convey other information for which hearers listen, namely, the pitch and timing features that reveal a speaker's characteristic physiology (e.g., young, old), momentary physiology (e.g., tired, energetic), affective state (e.g., calm, annoyed) and attitude towards the text or the interaction (e.g., sincere, ironic, grudging). From this, hearers can then deduce the speaker's likely intention, and therefore, the context in which to interpret her utterances. The search for better algorithms for emulating linguistic competence ultimately requires not a default competence, but one that belongs to a specific speaker, with a specific physiology and history, all of which impinge on how she stores and accesses information and thus, on her lexical and acoustical choices.

In summary, although processing in working memory is not the only influence on prosody, it is an important one and one that has so far been missing from both

theory and implementation. I suggest that it is only by accounting for speaker specific influences in a model of prosody that we can hope to duplicate in synthetic speech the stylistic and individual variation that is a matter of course in human speech.

Chapter 3

Theoretical and Empirical Foundations

By conveying emphasis and grouping, prosody serves many purposes. Speakers use it to disambiguate syntax, distinguish questions from statements, draw attention to salient semantics, coordinate turn-taking and structure a discourse into topic segments. This information helps the hearer locate referents, identify topic structure and participate in the interaction.

A cooperative speaker will try to provide this information as a matter of course. However, like hearers, speakers have finite cognitive capacities. If they are engaged in spontaneous speaking situations, they are simultaneously constructing an utterance, producing it and attending to the hearer's response – and all in real time. At times, their prosody may well reflect the difficulty of managing these tasks as much as the structure and content of the information they deliver. However, the possibilities are not exclusive. A focus on communication emphasizes the speaker's accommodation to the limited attention and working memory of the hearer. It treats prosody as instructions that help the hearer process both text and interaction. A focus on production emphasizes the limited resources of the speaker. It treats prosody as information about the speaker's online processing as she constructs and organizes the text. Both address aspects of resource-bound processing in conversation. The first focuses on how prosodic form reflects the speaker's accommodation to the information and processing needs of the hearer. The second focuses on prosodic form as a reflection of the speaker's own limited resources. Alternatively, these differences illustrate competence and performance explanations, respectively.

In the first two sections of this chapter, I organize the accounts of prosody according to whether they emphasize the speaker's competence or performance. As a reflection of competence, prosody conveys information about the form and content of a text. As a reflection of performance, it demonstrates the effects of the speaker's processing and production capacities. This is not a traditional distinction, mainly because most

studies and models are of the competence variety. However, discussing prosody as an elucidator of both text and processing clarifies the reason for attempting an approach based on the limited cognitive resources of the speaker.

Following that, I discuss the research on stylistic and individual differences in prosody, including the scant work that links prosodic differences to individual differences of cognitive capacities and mental state. Finally, I review computational approaches to prosody. I discuss the intonational theory of Pierrehumbert and colleagues in depth because it provides the main framework for research on understanding and synthesizing intonation.

3.1 Competence explanations of prosody

The competence explanations treat prosody as a communicative and interactive resource. They propose that speakers use intonation to group and to emphasize, and thereby structure the discourse into topics, subtopics and turns [PH90, GH92, HG92, SG94]. and direct the hearer's attention to the areas where processing effort is most profitably applied [Cut84].Applied to an individual word, intonation communicates its information status such as whether it is familiar, in context, or novel [Pri81, Bro83]. It also conveys specific propositions about the word or phrase to which it belongs, such as contrast [WH85], similarity, or set membership [LP84, PH90]. The different phrase final contours (rising, falling, level) distinguish among sentence types, such as interrogative or declarative, and speech acts ([Aus62, Sea69]), such as requesting or informative. They also convey topic structure, coordinate turn-taking [Sch82, Yng70, SSJ74] and affirm social connection [McL91]. In essence, the interactive and information structure interpretations identify prosody as a resource used by the speaker to facilitate the hearer's comprehension and cooperation.

Studies in this vein have uncovered useful correlations between prosodic features and discourse structure, information status and speech acts. For example, speakers typically signal topic shifts by expanding their pitch range [MB82, HG92]. Ayers [Aye94] found correlations between the size of the pitch range and the position of a discourse segment within the hierarchy of segments that make up a conversation. She notes that the relation of the pitch range to the discourse hierarchy is more pronounced for read speech than for spontaneous speech. Hirschberg and Grosz [HG92, GH92] found that a speaker's pitch range tends to expand for quotations and to compress for parentheticals. They also report a slower rate of speech and an increased amplitude at the beginning of new discourse segments, and a faster rate of speech for parentheticals.

Pause duration is often a cue to the strength of a phrase boundary, and more so for read speech. Ayers [Aye94], *inter alia*, found long pauses at the boundaries between segments of read speech, but found no clear pattern of pause duration in spontaneous speech. However, when the durations of filled pauses ("um", "uh") are taken into

account for spontaneous speech, there is evidence that pause duration reflects topic structure. Swerts *et al.* [SWB96] found that filled pauses that initiate a new segment have longer durations and a higher average pitch than those that are phrase internal.

Phrase final contours are often correlated with speech acts – e.g., request, inform, command – and the corresponding sentence types – e.g., question, declarative, imperative [LS74, MB82, EC86, Cam95, GvH95]. The prototypical yes-no question in English has a high phrase final rise, and both the prototypical wh-question and the declarative statement are produced with a phrase final fall. As with other intonational features, Hirschberg [Hir95] found that these trends were more consistent in read speech than in spontaneous speech.

An important function of pitch accents is to convey information status such as *theme* (the current topic) and *rheme* (additional information about the topic) [Hal67, PS94], and especially the distinction between *given* and *new*. Prince [Pri81] defines given information as information that the speaker believes is available to the hearer. She enumerates the criteria under which a speaker is entitled to this presumption: (1) if the information has just been presented; or (2) if it is *inferable*; or (3) if it has been *evoked* textually by prior mention, or situationally. Her taxonomy predicts that new information is accented (with high pitch) and that given information is de-accented.

Brown [Bro83] investigated this claim and found that whether an item was previously mentioned was not a sufficient predictor of givenness. In her studies of elicited spontaneous speech, referents of the currently evoked item (the item in focus) tended to be de-accented, and pronominalized as well. In contrast, if an item had been previously mentioned but was not the current focus of attention, it was likely to be accented with high pitch if the speaker believed that the hearer did not expect it. This corroborates with the converse [NT82], that a speaker will tend to de-accent if she believes that an interpretation is already available to the hearer.

Accent is also correlated with the textual features that convey information status, such as grammatical role, sentence position (for word order languages) and pronominalization. It has been found that items about to become the current focus are often in object position, while items already in focus tend to occur in subject position [Sid79, Pri81, BFP87]. Such items are given, and therefore tend to be de-accented. In SVO languages such as English and Dutch, subject position is usually sentence initial. Terken, *et al.* [NT82, Ter84, TH94] have found that de-accenting is most probable when these two features coincide. This often occurs for referring expressions of greatest specificity,[1] such as pronouns, since they are indicators of continued salience [Sid86, WW90].

[1]Indefinite reference (*a book*) is less specific than definite full-NP reference (*the book*), which is less referentially specific than *it*. The importance of context increases along with increased specificity. Pronouns depend almost entirely on context to select their referent.

3.1.1 Discussion

Competence accounts mainly focus on how prosody is shaped by the propositions in the text, but less so on how it is shaped by the individual speaker who produces the text. The main linkage of prosody to attentional capacity focuses on the hearer rather than the speaker. In the next section, I review research that explicitly links prosody to the attentional and memory resources of the speaker.

3.2 Performance explanations of prosody

Psychological studies attribute shorter completion times, for both production and recognition tasks, to the greater accessibility of task-related information in working memory. In studies of speech performance, both the duration of the word and the pause that precedes it appear to correlate with word frequency. Geffen and Luscsz [GL83] have found that speakers produce significantly longer pauses before rare words and that rare words take marginally longer to articulate. Hesitation pauses occur most often before content words (whose members are numerous and varied) than function words (whose members are ubiquitous and few). Pauses in general are far more frequent in spontaneous than read speech [Cut96] – the result of constructing the content and form of the utterance in real time. Goldman-Eisler [GE61] observes that longer pauses precede the expression of more complex and interesting thoughts, a result (it is hypothesized) of a search for not only the right word, but the right thought.

Recency is also a well-known predictor of retrievability in that the more recent the mention, the more quickly the item will be recalled. This is suggested by both Levelt's [Lev89] and Fowler and Housum's [FH87] findings that word durations are shorter at second mention than at first mention. Levelt found that words at second mention are also lower-pitched and softer. Fowler and Housum found that they are less intelligible. It appears that both retrieval time and production effort are reduced at second mention. Elaborating on these findings, Fowler *et al.* [FLB97] found that the duration of repeated words was reduced within a (read) discourse segment but blocked at boundaries marked by "meta-narrative statements" rather than narrative discontinuities. This suggests that a meta-narrative statement effects a shift in perspective that reduces the accessibility of previously salient items.

Intonation is most often linked to the meaning and structure of a text. However, accenting has also been shown to respond to *frequency of occurrence.* Function words, the most ubiquitous of a language, tend to be de-accented [Hir90, Hir93] and content words tend to be accented [Ter84], reflecting the relative frequency of use for the language and the corresponding retrievability for an individual.

In addition, accenting is also responsive to *recency.* Referring expressions tend to be

accented upon first mention but de-accented subsequently [NT82, Ter84]. This trend appears to be strongest for topical items. Recently mentioned items that are not topical are more likely to receive accents upon subsequent mention, indicating their less salient status.

Evidence from other areas of language points to the influence and sometimes the primacy of the speaker's own limited capacity. Wasow proposes that the tendency of speakers to put complex constituents at the ends of sentences comes from their need to simplify planning and production, especially because such constructions are more difficult for a hearer to parse [Was97]. In his examples, production expediency outweighs communicative efficiency. In a similar vein, the lexical choice experiment conducted by Brown and Dell [BD87] shows that speakers were more likely to explicitly mention items typical to a scenario than to tailor their description to the particular information needs of the hearer. Such findings imply that linguistic behavior is determined by expediency in the face of limited production resources, as well as by accommodation to the hearer's processing and information needs.

In the remainder of this section I review the (scant) research that links variation in prosody to individual differences among and within speakers, whether they are differences of skill, affective state or idiosyncratic preferences.

3.2.1 Prosodic differences among speakers

In this section I discuss studies of prosodic differences between and within groups of speakers. Differences among groups of speakers are often characterized as differences of speaking style, particularly those that are influenced by social and professional genre. Differences between the sports announcer and the news announcer are differences of style, as are differences between the speech of an adult to an infant *versus* the prosody that adults employ when talking to each other.

Standard performance within a style has been the main subject of empirical speech research. Studies of the differences among speakers of the same style have been far rarer.

3.2.2 The influence of style and genre

In the speech research literature, the term "speaking style" usually refers to the distinction between read speech and spontaneous speech. Although each are in fact composed of many styles, the read—spontaneous distinction captures a categorical distinction about how much improvisation is either necessary or allowed for both text and prosody. For example, the text is fixed for the typical read speech scenario, although the prosody is not. In contrast, neither the text nor the prosody are fixed for

the typical spontaneous speech scenario. However, there are many variations – both the prosody and the text of ritual speech (chants, prayers) are often fixed; the prosody of sorority speech is often highly constrained by social status [McL91] although the text is more fluid; and the prosody of parent-infant speech is typically exaggerated [FS84] and while the text is not fixed, it is likely to be simple and repetitive.

Nonetheless, it will be useful to consider the prototypical situational and processing distinctions in more detail. The text of read speech is determined beforehand and is either read aloud or delivered from memory. When the reading is a monologue, the listener need not be physically present (as in a radio broadcast) although the speaker is likely to have a target audience in mind. When both speaker and audience are co-present, the audience's responses are often restricted in content and timing, for example, to applauding at the end of a presentation or only when explicitly solicited by the speaker.

Spontaneous speech usually occurs in social interactions in which all parties are present. Today's technology modifies the requirement that co-presence be physical or even temporal. Voice mail messages are an example of asynchronous spontaneous speech.

These differences have cognitive implications. Planning and production tasks tend to be easiest for read speech because readers are not interrupted as frequently (if at all) and their main task is to interpret a text. Therefore, they can plan over longer chunks of text [SSH96], and with the reasonable expectation of executing their plans.

Spontaneous speech also requires planning ahead, but over shorter chunks of text. In synchronous communication, a speaker's plans are updated and changed according to the listener's response [FK92]. Given both the availability and unpredictability of feedback, it makes no sense to devise long range plans. Moreover, it is not feasible because the speaker has only limited attention and working memory to allocate over many real time tasks, such as planning, producing and evaluating her speech [Lev89].

The cognitive and interactive differences between the read and spontaneous styles show up prosodically in characteristic ways. Phrase final rises are considered more typical of interactive spontaneous speech because of the multiple functions they perform. As in read speech, they indicate discourse structure, namely that there is more to come. However, they are also used to solicit feedback from the hearer [Dun72] or to affirm social connection [McL91]. In addition, as a result of planning load, spontaneous speech is comprised of shorter phrases and frequent pauses [Cut96]. It also evinces more disfluencies [Fro71] such as false starts [O'S92], repairs [Lev83], mispronunciations [SH86, FC77] and filled pauses [MO59]. These features have received increased attention in recent years because speech recognizers must be able to distinguish disfluencies from the intended words and phrases, and if possible, use them to help recognition [BDSP93, SBD92, ZDG+89]. Finally, spontaneous speech tends to be less clearly articulated, no doubt as a result of performing many cognitive tasks simultaneously [Koh95].

The consensus has been that read speech exhibits more phrase final falls and a steeper pitch range declination. This is partially an artifact of collecting readings of isolated sentences ([SC80]) in the lab. With no actual audience, readers tend to compress their pitch range and employ a falling intonation at the end of the phrase [Ume82]. However, the general trends tend to hold for longer samples. Silverman *et al.*, compared the prosody of directory assistance calls to readings of their transcripts by the original callers. The spontaneous samples contained many more phrase final rises [SBSP92] than the read versions. In their comparison of a telegram read aloud and a spontaneous retelling of its context, Swerts *et al.* found significantly greater fundamental frequency (F0) declination in the read samples and likewise, greater F0 resets (pitch range expansion) at phrase and utterance boundaries [SSH96]. Zue and Daly compared the spontaneous speech elicited in human-machine interactions with readings by the original (human) speaker of their transcribed utterance. They too found more phrase final rises in the spontaneous data. However, because the data came from human–machine interaction, their finding of a significantly higher mean fundamental frequency for the spontaneous speech differs from most other studies.

Although the distinction between spontaneous and read speech seems to be the defining one in speech research, it is actually one of several dimensions that together describe social and professional speech genres. As Hirschberg [Hir95] observes, an important dimension is the degree of planning. She proposes others such as the presence, type and number of interlocutors. Similarly, Clark and Brennan [CB91] enumerate conditions of production and comprehension in dialog, often imposed by technological intermediaries, that determine the communicative strategies of both listener and hearer. In their taxonomy, the least precision of form and content is required when the interaction is face-to-face and synchronous. For speech research, it is important to remember that the conditions under which speech is collected influence its acoustical and lexical form [Ume82]. Touati [Tou95] proposes that researchers distinguish between "spontaneous speech" (*in situ*) and "spontaneous lab speech".

Sociolinguistic research contributes much of the data about how prosody conveys social status and affiliation. For example, studies of American and British English speakers show that women tend to employ greater enunciation and longer pauses than men [Hen95, Whi95]. Workplace speech shows characteristic prosody as well. Douglas-Cowie and Cowie [DCC95] found that the intonation of the phone speech used by secretaries varied according to whether they were answering or initiating a call, and whether they were transferring a call or handling it themselves. Goodwin [Goo96] reports distinctive differences in the intonation used by airport control tower personnel to distinguish ordinary from out-of-the-ordinary events. In all, a useful set of descriptive or predictive factors has yet to be enumerated, let alone related in a functional or causal manner. This remains an intriguing and interesting challenge for speech research in many fields.

The effect of skill

The evidence that differences in working memory capacity (and contents) might be related to prosodic differences has received little direct attention. A suggestive exception is the work of O'Connell [O'C88] and colleagues on the stylistic differences between expert and amateur readers. He reports on two experiments correlating temporal aspects of prosody with skill level in the dramatic arts. In the first, three groups of readers – professionals, amateurs, and untrained college students – read a dramatic biblical passage. While all readers used the same percentage of pause time to total time (41%), the differences in skill and training were reflected in the *distribution* and *duration* of pauses. The untrained readers produced the fewest words per phrase and used the slowest articulation rates. The most skilled and experienced readers used the greatest number of long duration pauses (greater than 1.2 seconds), showing the deliberate use of pausing as a dramatic and expressive device. The suggestion is that with a greater number of novel tasks to perform, the unskilled readers had fewer cognitive resources available for lookahead and planning. The evidence comes from both structure and timing – their phrases were composed of the fewest words, but nevertheless took the longest to articulate.

The second study compared readings by three poets[2] of their own work to readings of the same poems by other adults. Half of the adult readers were English professors, the other half, educated adults but not expert. Despite their greater expertise, the English professors did not produce more expressive readings than the other adults. All the adult readers did not read the poems expressively, according to O'Connell, while the poets who were intimately familiar with their own work, did. However, this greater expressiveness manifested itself in ways opposite from the previous study. The articulation rate was slowest for the poets and the phrase lengths were shortest. Part of the slowing was due to exaggerated syllable lengths rather than unfamiliarity with the task and the text. O'Connell proposes that exaggeration, slowness and shorter phrases are expressive devices that are more appropriate to poetry than to prose. By the same token, he notes that hesitations classed as disfluencies for spontaneous speech are deliberately employed as rhetorical devices in public oratory. An interesting result in this study is that the greater background knowledge of the professors did not produce any better readings. Rather, the familiarity of the poets with their own work was key, most likely because the content and form of the work were already salient for them.

The effect of idiosyncratic biases

Speakers often have a choice of strategies for indicating grouping and emphasis. For example, they may indicate phrase boundaries by pausing, lengthening the last word,

[2] e e cummings, Randall Jarrell and Robert Frost

adding extra melodic elements to the end, or any combination thereof [vDvB96, O'C88]. Similarly, they can indicate word prominence by increased duration, a pitch accent or both. The conditions under which speakers choose to de-accent also show individual differences [TH94]. These may indicate the effects of situation on limited resources or they may simply reflect the longstanding expressive biases of the speaker.

3.2.3 Prosodic variation for one speaker

Intra-speaker variation in general is often situational. In a high stress situation, a speaker is likely to apportion her processing capacities over a variety of tasks, of which speaking may be only one [OC74]. Gable *et al.* [GNH92] found that as task demands increased, so did the number of non-word disfluencies (pauses, interjections) and planning errors. Anxiety and increased arousal have been correlated with a decrease in the use of qualifiers (adjectives and adverbs) and an increase in the use of nouns and verbs [Col85]. They have also been correlated with a wider pitch range and faster speech [FP39, Fai40, Dav64, WS72].

Repeated practice also influences intra-speaker variation. Under affectively neutral lab conditions, Atal reports durational differences for the same words across multiple readings of the same text by the same speaker [Ata76]. This is likely a reflection of both recency frequency results – by repetition, the words have become more salient to the speaker.

3.3 Computational theories of prosody

In this section, I discuss theories of prosody that apply across style and genre. Their main focus is on stress and intonation, primarily as reflections of the lexical, syntactic and information structures in the text. Stress and intonation are related – stress patterns in a phrase or text (such as iambic pentameter) usually arise from the distribution of pitch accented words. Lexical stress and pitch accents are also related – a pitch accent typically falls on the syllable of a word that carries primary stress.

Theories of prosody are distinguished by whether they propose a structure that is strictly layered or recursive. The hierarchical but non-recursive Strict Layer hypotheses identify distinguished and named levels ([BP86, Hay89, NV83, NV89, BE91, WSHOP92, GG83, Sel86]). In contrast, the recursive and unrestricted hierarchies build structures analogous to the trees of transformational grammar ([LP77, Hal87, Lad86, Lad88, Mar87, Gib84, Ris87]).

Ladd [Lad83] makes another distinction for intonation, between the *contour interaction* theories, which treat pitch accents on words as local perturbations of a global

contour for the phrase, and the *tonal sequence* approaches, which treat phrasal tune as compositional from a sequence of elements associated with the word (e.g., pitch targets [LS74] or pitch contours [Cru86]). These approaches require minimal lookahead, which Levelt [Lev89] claims is a requirement for theories of human speech production. Tonal sequence theories also lend themselves to compact description as a finite state automaton, a familiar structure in computation.

I will group the intonational theories by whether they belong to the Strict Layer or recursive hypotheses,

3.3.1 The Strict Layer hypotheses

The Strict Layer hierarchies are part of a larger hierarchy of scales in conversation and speaking that range from the discourse or conversation as the largest, to vocal cord vibration as the smallest (see Figure 3-1). The exact constituents of the phonological levels vary across languages, but those above and below are relevant to most languages. At the top of the hierarchy are the scales whose structures are roughly analogous (for illustrative purposes) to textual structures such as the chapter, the paragraph and the sentence. The *conversation* or discourse is the largest scale. It is composed of one or more *discourse segments*, each defined by a main topic. In dialog, a discourse segment is composed of *speaking turns* that alternate between speakers; in a monologue, the entire segment consists of one extended turn. Next is the *utterance* – a speaking turn may be composed of one or more. An utterance, in turn, may be composed of one phrase or many.

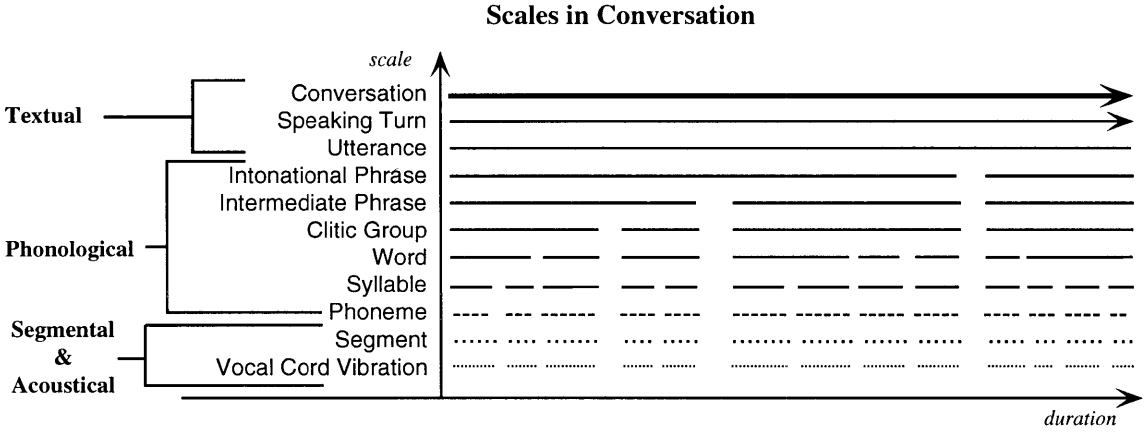

Figure 3-1: A hierarchy of textual, phonological and acoustical scales in conversation and speech.

At the level of the *intonational phrase*, the mapping between textual and phonological scales is no longer isomorphic because intonational and major syntactic clause boundaries often do not coincide. Theories about intonational structure also vary at this

point. The theory developed by Pierrehumbert and colleagues [Pie80, BP86, PH90] defines the intonational phrase as composed of at least one *pitch accent*, and a *phrase accent* and a *boundary tone* which together describe the phrase final contour. The intonational phrase is composed of *intermediate phrases* (alternatively, *phonological phrases* [GG87]) whose phrase contours after the *nuclear* (phrase final) pitch accent are defined by a rising, falling or level pitch contour but not a final boundary tone.

A pitch accented word and the unaccented words that surround it have been termed a *phonological word* [GG87, BF90] or *clitic group* [Hay89], identifiable because the phonemes at the boundaries of the unaccented words are often reduced. For example, because infinitival *"to"* is almost never accented, the cliticization, *"wanna"*, for *"want to"*, is possible. A second type of clitic group contains lexical forms that do not stand on their own, such as the possessive particle *"'s"*. At about this level, Beckman and Pierrehumbert identify the *accentual phrase* in Japanese, which consists of at most one pitch accented word.

Words are composed of syllables, which are composed of zero or more consonants and one vowel or diphthong. The consonants and vowels are acoustical rather than orthographic. For example, *"sh"* is a single phonetic consonant even though it is two consonants orthographically. Each syllable bears stress of some kind. It if bears primary stress, a pitch accent will fall on this syllable. However, it will fall on a syllable with secondary stress if phrasal constraints mandate a stress shift [LP77, SHOR93]. A syllable may also be unstressed, in which case its vowel is often pronounced as a schwa[3] and its consonants reduced or deleted.

Below the syllable are the segmental scales, whose features directly reflect the mechanics of speaking. Phonemes – the sounds of a language – may be steady state – such as vowels and fricatives – or composed of a series of transitions – such as plosives and diphthongs. Below the phoneme segment is phonation itself, the vibration of the vocal folds. Physical properties of the vocal folds such as thickness, length and closure patterns are responsible for a speaker's characteristic pitch range and voice quality.

3.3.2 The recursive accounts

Text centered theories of prosody tend to produce only the categorical components of intonation, namely, pitch accents and phrase final contours. However, there are two exceptions, notable because they generate (relative) quantities from categorical phenomena. The first is the *metrical grid* approach to predicting lexical and phrasal stress [LP77]. A metrical grid describes a binary constituent tree whose branches at each node are labeled as either strong or weak. The strong and weak assignments propagate upward and are additive – the constituent with the most strong nodes above

[3]**schwa**: the vowel produced when the vocal tract is in its most neutral position. The first and last syllables of "America" are schwas.

it receives the main stress of the word or phrase. Therefore, it is equally applicable to the syllables of a word ([CH68] and the words of a phrase ([LP77]).

In essence, the accumulated strong and weak assignments describe the relative prominence of a syllable within a word and a word within a phrase. It also predicts stress shift. If the original grid contains a node with two contiguous strong branches, the stress will shift so that two consecutive strong syllables are not contiguous. For example, *Tennessee* shows two different stress patterns in *"In Tennessee they speak a Tennessee dialect."* In the second occurrence, *"Tennessee"* is part of a complex nominal. Its main stress has moved leftward to avoid conflict with the first syllable of *"dialect"*, which bears the main stress for the word and the complex nominal phrase.

The recursive accounts of intonation developed by Ladd [Lad88, Lad86], Gibbon [Gib87] and Marek [Mar87], *inter alia*, are inspired by the metrical grid approach. They build intonational trees in an interactive fashion, using the strong/weak distinction at each node to describe the relative intonational prominence of two successive branches. This gives theoretical representation to the *magnitude* of intonational features as well as their type. At higher levels, the strong/weak distinctions become differences in the pitch range of phrases; at the lowest level, they define the prominence of a pitch accent.

Ladd argues that prominence is categorical. In early work [Lad83], he attributes the location of the peak prominence to the feature $[+/-delayed]$ and prominence itself to the feature $[+/-raised]$. Both determinations are binary and local. However, his later work derives these features from a recursive analysis [Lad86, Lad93].

The recursive accounts integrate pitch range expansion and compression (both are outside the Strict Layer hypotheses). However, because they are recursive, they derive intonational tones and levels from fully constructed intonational trees. This is unsatisfactory for a limited capacity account, since the mapping of structure to intonation requires more lookahead than most speakers possess. Moreover, deep and recursive representations are difficult to maintain in working memory, which is where they presumably reside until they are realized as speech.

3.3.3 Discussion

The recursive hypotheses are appealing because they generate gradient features from the local and recursive application of the binary distinction between strong and weak nodes. However, their local generativity is offset by the recursive structures they create and the large amount of lookahead they require. Therefore, they are not as cognitively feasible as the simpler structures of the Strict Layer hypotheses. The difference is in what each seeks to explain. Recursive accounts link the emergent features of prosody to locally determined categorical distinctions in the underlying structure. The Strict Layer hypotheses describe only surface phenomena.

The intonational grammar developed by Pierrehumbert and colleagues is a both a Strict Layer and tonal sequence theory. As such, it requires minimal lookahead. Therefore, it does not exceed the limits of human computation and is easily implemented on a machine. In addition, its expression as a finite state grammar is the basis for the compositional interpretation of the semantics of intonational tune proposed by Pierrehumbert and Hirschberg [PH90].

3.4 Pierrehumbert's intonational grammar

Pierrehumbert's framework provides a compositional and generative description of intonational contours. Its introduction in 1980 ([Pie80]) decisively shifted the balance in several debates in phonetics and linguistics to favor compositional contours over global ones, and continuous tone levels over categorical ones. In addition, it addressed the question about whether intonation was best described as discrete pitch targets ([LS74]) or continuous pitch contours, ([Cru86]) by combining elements of both in the bitonal accents.

Since its introduction, it has become the standard intonational framework for both theoretical and speech synthesis work. It is applicable to languages as dissimilar as Standard American English and Japanese [BP86] (with some adaptation). It has been combined with theories of propositional logic ([Bir91, Oeh91, Ste90]), discourse structure ([HP86, HLPW87, HL87, PH90]) and centering ([Kam94, Cah95, Nak95]). It has also advanced the practical work on prosodic algorithms for synthesized speech and provided the basis for later theoretical work on the atomic and compositional meaning of intonational features.

Its significant innovations are: (1) the simplification of the intonational tone inventory into two tones – H ("high," for rising pitch) and L ("low," for falling pitch); (2) a finite-state grammar of intonational structure; (3) an account of the meaning of intonational contours.

3.4.1 The tonal inventory

One of the main contributions of this framework is to simplify the tonal inventory for English intonation. It describes both pitch accents and phrase final contours with combinations of only two tones, H for a high pitch target and L for a low one. Differences among heights and depths of these tones are ascribed to an independent *prominence* variable. This provides a cleaner description than proposals that mix quantitative and categorical information in one category. For example, Pierrehumbert's system represents both the High and Mid-High tones of Liberman and Sag's inventory [LS74] as High tones of differing prominence.

Pierrehumbert constructs the full pitch accent inventory from high (H) and low (L) pitch targets. These are relative designations and denote a pitch contour that either rises (H) or falls (L). She combines these tones to produce the *bitonal* pitch accents, whose contours are defined by two pitch targets. They are further distinguished according to which of their two tones fall on the lexically stressed syllable of a word. The two possibilities for each bitonal are denoted with an asterisk to mark the main tone. Thus, the L+H bitonal form is the basis of the L*+H and L+H* accents, and the H+L form is the basis of the H*+L and H+L* accents.

Beckman and Pierrehumbert [BP86] propose an inventory of six pitch accents altogether: H*, L*, H*+L, H+L*, L+H*, L*+H. Variants include the seven categories of Pierrehumbert's original proposal [Pie80] and even more in the ToBI[4] standard for prosodic mark-up [SBP+92, WSHOP92].

3.4.2 A context-free grammar of intonation

Pierrehumbert *et al.*'s taxonomy of pitch accents are incorporated into a finite state grammar of intonation. The other tokens of the grammar depict the intonational phenomena that occur at the end of a phrase. These are the *phrase accent*, which describes the shape of the pitch contour after the *nuclear* (last) pitch accent of the phrase, and the *boundary tone*, which specifies the phrase final fundamental frequency for the *intonational phrase*. The full grammar is shown in Figure 3-2.

$$
\begin{array}{rcl}
\langle \text{IntonationalPhrase} \rangle & ::= & \langle \text{IntermediatePhrase} \rangle^{+} \ \langle \text{BoundaryTone} \rangle \\
\langle \text{IntermediatePhrase} \rangle & ::= & \langle \text{PitchAccent} \rangle^{+} \ \langle \text{PhraseAccent} \rangle \\
\langle \text{PitchAccent} \rangle & ::= & \text{H*} \mid \text{L*} \mid \text{H*+L} \mid \text{H+L*} \mid \text{L*+H} \mid \text{L+H*} \\
\langle \text{PhraseAccent} \rangle & ::= & \text{H} \mid \text{L} \\
\langle \text{BoundaryTone} \rangle & ::= & \text{H\%} \mid \text{L\%}
\end{array}
$$

Figure 3-2: Pierrehumbert *et al.'s* context-free grammar of intonation. ("+" denotes "one or more", "|" denotes "or", and nonterminals are in angle brackets.)

As with any finite-state grammar, this syntax is generative as well as descriptive. It describes how intonation combines on a formal level, but also generates the contours that are found in the empirical data. For this reason, it has been incorporated into algorithms that generate intonation in synthetic speech [APL84, Pie81].

[4]ToBI stands for "Tone and Break Indices", a prosodic mark-up standard for English. The (human) coders identify pitch accents, phrase final tones, and annotate the degree of phrase break between words, such that 0 denotes a reduced boundary between words, 1 indicates that there is no break but no phoneme reduction at the boundaries and 4 denotes an intermediate phrase break. The main pitch accents are: H*, L*, H+L*, L+H*, and L*+H. H*+!H replaces H*+L in Pierrehumbert *et al.*'s account. The downstepped variants apply to the H phrase accent and to pitch accents that contain H tones. They are denoted with "!" before the downstepped tone.

The height or depth of a pitch accent or phrase tone is produced from a combination of prominence, pitch range declination and downstep [Pie80]. *Downstep* describes the pitch range compression that occurs after uttering a bitonally-accented word [Lad83, LP84]. It is distinct from *declination*, which is continuous and is a property of the intonational phrase as a whole. *Prominence* describes the magnitude of the height or depth of a pitch accent or phrase tone. Like downstep, it is applied locally. However, like declination, its values are continuous. Therefore, both declination and prominence are outside the finite state grammar.

Although downstep is also excluded from the grammar, it is a categorical function of pitch accent category [LP84, Lad88]. Thus, it would appear to belong in the finite state automaton. However, its inclusion would complicate the automaton, and require either the explicit representation of all the possible sequences that produce downstep, or the addition of a global register to track whether a bitonal has just been produced. The ToBI mark-up chooses the first option, and marks downstepped tones with the downstep symbol, "!"

3.4.3 The meaning of intonational contours

To explain when and why speakers choose one intonational contour over another, Pierrehumbert and Hirschberg [PH90] propose a semantics for both pitch accents and phrase final intonation. They base their proposal on the observation that pitch accents mark *salient* information and phrase final contours indicate *discourse structure.*

They interpret the meaning of the pitch accents in light of the speaker's conceptions about *mutual beliefs*, the collection of propositions that each conversant believes and in addition, believes her conversational partner to also believe [CM81]. In this context, they ascribe to H* the function of instructing the hearer to *predicate* the marked information as a mutually believed proposition, especially because the information is new and has not yet been so predicated. L* on the other hand, is interpreted as *failing to mark* information as mutually believed. However, the meanings are not in opposition. While H* marks information as a new addition to the set of mutual beliefs, L* marks information that should not be instantiated as a mutual belief, but does not explicitly comment on its status as given or new. That is, a speaker may wish to exclude a proposition from predication because she believes that it is already part of the mutual beliefs, or because she believes that it is false and should not be believed.

The H+L bitonal is taken as a commentary on *inference paths*, that is, on whether the accented information is inferable even if not explicitly mentioned or previously predicated as a mutual belief. The other bitonal, L+H, invokes a *salient scale* from which a referent is selected (the L+H* accent is the contrastive stress contour). In all, the meanings of the bitonals are compositional from the meanings of the main tones, H* or L*, and the bitonal structures, H+L and L+H. Any accent that contains

a H* component affirms the speakers commitment to predication of a proposition as mutually believed, while any accent that contains L* fails to affirm this. The interpretations of all six accents are summarized in Table 3.1.

Accent	*Predicate as mutually believed?*	*Salient scale?*	*Inference path?*
H*	yes	–	–
L*	no	–	–
L+H*	yes	yes	–
L*+H	no	yes	–
H*+L	yes	–	yes
H+L*	no	–	yes

Table 3.1: Pierrehumbert and Hirschberg's (1990) account of pitch accent meanings.

To construct the meaning of phrase tones, Pierrehumbert and Hirschberg appeal to Grosz and Sidner's theory of the hierarchical structure of discourse [GS86], and in particular, to the local relations between successive discourse segments. They propose that H phrase tones convey forward reference and therefore signal continuation, and that L tones fail to do so.

They construct phrase final contours from a phrase accent – either H or L – and a final boundary tone – either H% or L%. Each contour is often correlated with a sentence type. For example, the low-rise contour (LH%) typifies continuation rises, which speakers use to solicit backchannel feedback or to indicate that they intend to continue speaking. The high rise contour (HH%) is said to be typical of yes-no questions in English. Pierrehumbert and Hirschberg interpret both contours as making forward reference to upcoming material. They also propose that the level contour (HL%) typically ends statements which add supporting details to previous statements, and that a falling contour (LL%) simply fails to make forward reference. This is consonant with its typical usage – it is usually found at the end of a declarative sentence or a discourse segment.

3.5 Summary

In this section I have divided the approaches to prosody into those that stress competence and those that stress performance. Most are theories of competence because they link prosody to the logical propositions that illuminate information in the text. In contrast, psychologically based work emphasizes performance and focuses on the capabilities of an individual speaker, which may be innate or situational. Pierrehumbert 's theory is stated mainly as a theory of competence. However, in Chapter 5

and beyond, I will describe a performance approach that produces pitch accent and boundary tone types as the consequence of processing in working memory.

Chapter 4

Related work

The traditional distinction between text-to-speech and concept-to-speech synthesis hinges upon whether the text is given (text-to-speech) or generated by the system itself (concept-to-speech). Text-to-speech synthesis is general purpose and takes unrestricted text as input. Context-to-speech synthesis is currently special purpose and is typically built around interactive applications that possess some form of domain-specific expertise.

By these accounts, the paradigm for text-to-speech synthesis is read speech. Yet, as read speech goes, text-to-speech systems are poor approximations of the reading process. Unlike human readers, they lack both the attentional and semantic understanding of the words they speak. In contrast, concept-to-speech systems are usually equipped with both. Therefore, their counterpart in human behavior is spontaneous speech. However, unlike human speakers, they are limited to conversing only about one domain of expertise. Moreover, their prosody is only as good as the algorithms that map task semantics and pragmatics to text and prosody. Both text-to-speech and concept-to-speech efforts strive to increase linguistic competence – text-to-speech, because it has so little, and concept-to-speech, because the aim is to model a speaker who is well-informed about both interaction and the particular task domain.

With rare exceptions (found mainly in the work on emotional and stylistic variation) neither approach to synthesis considers the effect of limited cognitive resources on prosodic or lexical choice. The awareness that resource bounds are relevant and causative is found instead in computer simulations of other aspects of linguistic behavior such as interaction strategies and reading comprehension. In this chapter I describe the related work in text-to-speech and concept-to-speech synthesis, and the few extant simulations of the effects of resource bounds on linguistic performance.

4.1 Text-to-speech synthesis

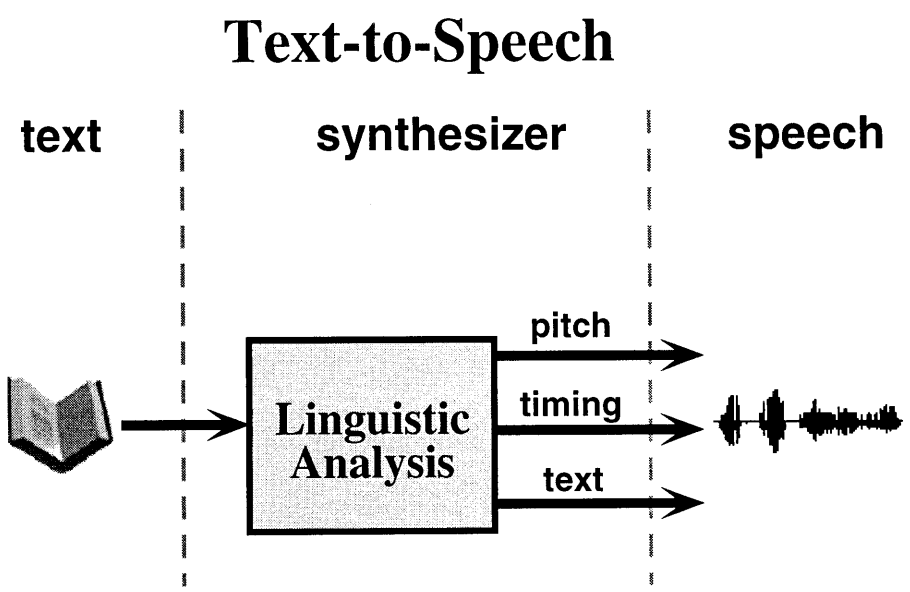

Figure 4-1: Prototypical text-to-speech system.

The prototypical text-to-speech system (Figure 4-1) takes unrestricted text and computes prosody based on the information it can deduce from surface form. In the earliest implementations, this often amounted to computing phrasing from punctuation, paragraph boundaries and part of speech information, and computing emphasis from part of speech distinctions alone. The simplest algorithms accent content words with a rising pitch accent [Kla87], de-accent function words and use punctuation as a guide to phrase final prosody. More sophisticated synthesizers also use syntactic parse information to determine both phrasing and emphasis.

In this tradition, improving prosody is a question of producing better text-to-prosody rules. Because current natural language processing methods cannot reliably infer semantics and pragmatics from text, such efforts are characterized by the search for better *discourse-neutral* prosody. Bachenko and Fitzgerald [BF90] describe an algorithm that calculates accent from part of speech information and performs recursive balancing of phrases around the main verb in order to arrive at the final phrasing. Similarly, Monaghan [Mon90]'s algorithm calculates sentential rhythm and lexical stress-shift from syntactic and phonological information.

Corpus-based techniques have been shown to produce improved prosody over rule-based approaches. At their essence is the application of automatic learning algorithms to prosodically annotated corpora in order to identify co-occurrence patterns for words and their prosodic realization [SHY92, RO96]. These are used in combination with a decision algorithm to identify the best match of the current text to the learned patterns and thereby, choose the most appropriate pronunciation [Yar94] or prosody [WH92, SHY92, HP94, PM98].

The corpus-based approaches add semantic and usage information to text-to-speech systems, which typically have none. While they do not directly provide semantics, they do identify patterns of use that implicitly express semantic relations and constraints. Unlike hand-crafted heuristics, their automatically generated rule sets and decision trees are fine-grained and context-sensitive. Practically, these approaches shorten the time needed to find the regularities of prosodic use. By being grounded in empirical data, they are more accurate than methods that depend on human coders (depending, of course, on the quality and size of the data set). Moreover, they are general purpose. Corpus-based techniques have been applied to symbolic [SHY92, PM97], stochastic [OV94, WO92, SHOR93, MTKI96] and neural net driven synthesis [CHW98] of prosody.

Another problem is that most text-to-speech systems lack a discourse model for tracking the focus of attention. Therefore, they cannot determine whether information is given or new, let alone deduce whether it is contrastive or inferable. And therefore, they cannot apply the full set of pitch accents in accord with logical and attentional propositions. Addressing this, Hirschberg [Hir90, Hir93] has devised a method for capturing the effects of prior discourse history. It is based on the content words of the topic sentence of the paragraph, which are retained for use in future comparisons. A match of subsequent text in the paragraph to these words or their uninflected roots identifies given information, and conversely, the lack of a match identifies information that is new. This method has since been incorporated in other text-to-speech systems ([BT94]).

Horne *et al.* [HFLL93] describe a related technique for Swedish intonation that also assigns given and new status based on previous mention or a match to a root form. Instead of a paragraph they use a moving attentional window of sixty words. In addition to exact or root-based identity, they also assign given and new status according to semantic associations derived from an online thesaurus of Swedish. Both synonymy and hyponymy[1] count as evidence that the current instance is given. The given and new assignments are then mapped to their respective focal and non-focal accents. Informal listener tests confirmed improved naturalness over the default output of the synthesizer. This technique is interesting because it uses both attentional and semantic information to produce text-to-speech intonation.

As discussed so far, the overarching problem with text-to-speech synthesizers is that they have little information about the structure of a text and even less about its semantics or pragmatics. Efforts to increase the linguistic competence of text-to-speech synthesis provide two main techniques: (1) improving the mimicry of the results of linguistic competence; (2) obtaining the competence directly from online semantic databases or indirectly from corpus analysis; (3) employing a reasonable attentional unit such as the paragraph or a large moving window of text.

[1] In their system, the current word or compound nominal is the super-ordinate category, and the prior word is the hyponym.

All the efforts described so far are text-based and aim for better text-appropriate prosody. However, not all of human prosody is determined by the text. In addition, the discourse and affectively neutral prosody aimed for by most applications is neither appropriate nor sufficient for many human interactions. In the remainder of this section I will review the text-to-speech work that recognizes that there are different styles and different speakers and that even one speaker can produce varying prosody, depending on her affective state or her situation.

4.1.1 Stylistic variation

As previously discussed, the main distinction in speech research has been between read and spontaneous speech. Johnson [Joh96] reports success in modeling the differences by using a feedback model such that a short-term closed-loop models the shorter lookahead ascribed to spontaneous speech, and a long-term loop models the greater lookahead ascribed to read speech. The read speech simulations exhibit more downstepped contours and steeper final lowering than spontaneous speech, and the prosody overall is reported to be acceptably natural. Most significant is that the prosodic differences for the two styles accord with the style-based differences found in human speech, and do so as a direct consequence of simulated attentional differences.

Abe [Abe96] uses statistical analysis techniques to isolate the acoustical properties of three genres of Japanese read speech: advertisement, fiction (an excerpt from a novel), and non-fiction (an encyclopedia entry). The analysis identifies characteristic formant frequencies, phoneme durations, power spectra and phrase final contours. The characteristic differences are then incorporated as rules into a synthesis system to produce the different styles, and are also used to convert a human sample in the encyclopedia style into the novel and advertisement styles. Listener tests using both the converted and synthesized samples showed reliable identification for the novel and advertisement styles.

Of these efforts, only Johnson's is truly a production model. It attributes performance differences to the influence of lookahead as a limited resource. Abe's work takes the more standard approach to synthesizing speaker variation, which is to imitate the output rather than the cognitive (or physiological) mechanisms that produce it.

4.1.2 Individual variation: Emotional and expressive synthesized speech

The pursuit of individual prosodic variation occurs mainly in the work on emotional and expressive synthesized speech. Thus far, all the text-to-speech efforts have attempted to imitate the prosodic output rather than generate it from underlying causes. HAMLET [MAN88] simulates six emotions by rule, using unrestricted text.

The Affect Editor [Cah88, Cah89] applies a parameterized model to syntactically analyzed text. Varying its acoustical and linguistic parameters produces many kinds and shadings of vocal affect.

The emotion rules in HAMLET and the parameters in the Affect Editor describe the acoustical evidence of the effects of emotion on cognitive processing. For example, the Affect Editor's *hesitation pause* parameter controls the points at which a speaker pauses within an intonational phrase, thus mimicking the result of a speaker's difficulties in retrieving an item from memory. More recent work continues in the descriptive and imitative tradition. Carlson, *et al.* [CGN92] report work on developing a synthesizer that reproduces the acoustical characteristics of basic emotions. Hinton and Edelman [HE95] describe a system that manipulates the acoustical controls of an existing synthesizer. Higuchi *et al.* [HHS96] manipulate the fundamental frequency and segmental durations of pre-recorded samples to explore the conversion among samples of unmarked, hurried, angry and gentle speech.

The main problem with these methods is that they are not explanatory or predictive. In addition, extending them often means developing a finer-grained taxonomy, whether of the emotions themselves or their acoustical correlates. Generally, such taxonomies are problematic. While researchers tend to agree on the basic emotions of anger, sadness, fear, surprise, disgust and gladness, they often disagree about the finer distinctions. More fundamentally, emotions are complex. One taxonomy is not likely to capture all their salient features nor place them in their proper relation.

Ultimately, a production model is the better path. Instead of imitating the acoustical features of emotional speech, it will predict, explain and produce them as the emergent consequence of the physiological and cognitive biases and capacities of an individual speaker [Cah89].

4.2 Concept-to-speech synthesis

While text-to-speech systems must perform a reverse-engineering feat from text back to concept, concept-to-speech systems start with task knowledge and some form of a discourse model (Figure 4-2). Their main challenge is to map attentional and logical propositions about the state of the task and the discourse to words and prosody. Because they create the communicative and propositional intentions, they determine thematic role,[2] speech act,[3] focus and word sense, which they can then use in their algorithms for generating prosody. And because they construct the text, they have access to its structure on many levels – grammatical role, syntactic constituency and part of speech. Interactive tutorial or advising systems have been the main application

[2]Such as Agent, Patient and Instrument.

[3]Such as Request, Inform and Command.

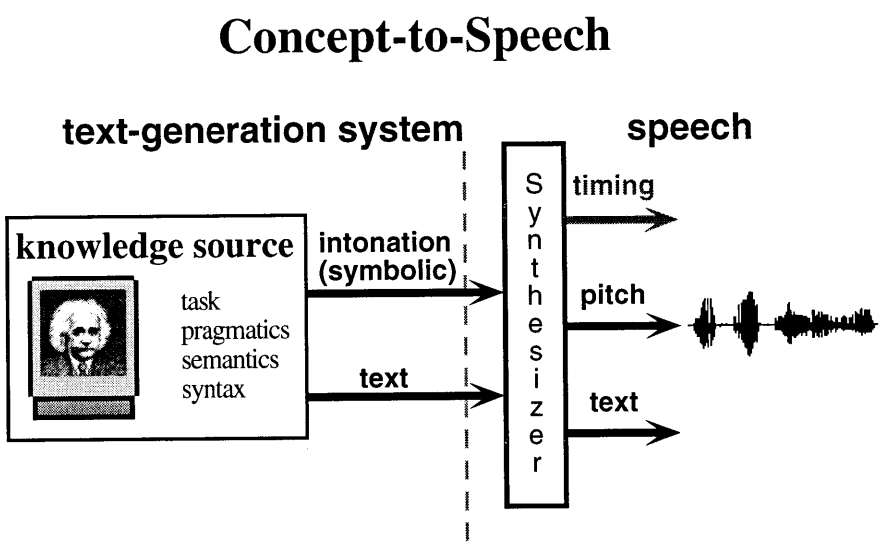

Figure 4-2: Prototypical concept-to-speech system.

platforms for concept-to-speech systems. Recently, they have been joined by sports-casting demonstrations and lifelike computer character simulations. Currently, there is no general purpose concept-to-speech system. The base of facts and often the mapping heuristics as well, are specific to the application [PM98].

4.2.1 Tutorial and advising systems

In this section, I discuss the salient concept-to-speech work on generating prosody that is appropriate to the topic structure and the current focus of attention.

Hirschberg and Pierrehumbert [HP86] use a tutoring system to explore the relationship between intonational and discourse phenomena. Their system correlates intonational tune with propositional attitude, accenting with given and new status, and pitch range with the discourse segment hierarchy, such that wider pitch ranges introduce top-level segments.

Davis and Hirschberg's [DH88] direction-giving system integrates discourse information with lexical and intonational decisions. It pronominalizes the current focus of attention and de-accents both pronouns and the second mention of a word within the same discourse segment. This work was included and extended in Davis's Back Seat Driver navigation instruction program [Dav89].

Using an airline reservation system as the underlying application, Youd and House [You92, YH92] aim for contextually-appropriate prosody, where context is mainly the semantic and pragmatic contents of the system's discourse model. The system is rich

in linguistic information, and uses speech acts, semantic content and surface form to calculate syllable features (focus, emphasis and accent) and therefore the location and type of phrase boundaries and pitch accents.

Prevost and Steedman [PS94] apply combinatory categorial grammar to distinguish between *theme* (topic, usually in focus) and *rheme* (comment) [Hal67]. Typically, the theme is unaccented while the rheme, which contains non-focal material, is accented. This approach is flexible and sensitive to prior thematic context – depending on what has been in focus, it can generate more than one contour for a sentence. However, it is mainly demonstrated for *contrastive* theme and for contrast in general. As Theune [The97] points out, the items that Prevost [Pre96] identifies as contrastive because of membership in a common set, are usually de-accented because of givenness.

4.2.2 Data-to-speech

In most concept-to-speech systems, the task expertise is frozen beforehand. Recently, Theune *et al.* [TOKdP97, The97] and Spyns *et al.* [SDvTC92] have experimented with hybrid techniques that apply task expertise to live data. In both cases, the domain is sports casting and the technique requires formatted data. However, because the domain roles are limited (player, spectator, referee, etc.) and the format fixed by reporting tradition, no manual formatting is required. Theune *et al.* deduce focus, givenness and contrast from the text feeds, which are then mapped to prosody according to the generally accepted view that a new item receives a high pitch accent (H*), a contrasting item receives a contrastive accent (L+H*), and a given item is de-accented.

4.2.3 Lifelike computer characters

Concept-to-speech approaches are a natural platform for expressive synthesized speech, especially for conveying the character and personality of a synthetic actor. Walker *et al.* [WCW97] report on an improvisation system for synthetic characters that derives lexical form and content from social variables as well as from the degree to which the speech act is an imposition on the addressee (a command is ranked higher than a request). The acoustical form comes directly from the independent specification of the character's affective state. Varying the social and affective specifications varies the lexical form and its prosodic delivery, often to comic effect.

Most recently, Binsted [Bin98] reports on an implemented architecture for a simulated "talking head" soccer commentator. The motivation is to explore the interplay between believability and consistency in synthetic characters. Her system generates facial expression and vocal affect according to the simulated soccer match data and the speaker's attitude toward the play or the team.

4.2.4 Summary

Concept-to-speech algorithms take advantage of application and discourse information to generate intonation that expresses propositional, structural and attentional relations according to the task and discourse model. Their success depends on the accuracy and specificity of the mapping algorithms. One problem is that assembling the application knowledge has been costly in terms of human effort. Thus, the use of automated sources of knowledge is explored in recent data-to-speech efforts. Another problem is that the prosodic algorithms are often intertwined with the specifics of the application. Pan and McKeown [PM98] report ongoing work that applies machine learning techniques to prosody, and links these results with the output of a text-generation system, with the overall aim of modularizing components of concept-to-speech systems. The key challenges for concept-to-speech synthesis are to simplify the encoding of task-based expertise and to generalize the mapping from state of the task and the discourse to prosody.

4.3 Discussion

Originally, whether the text was constructed by the system (concept-to-speech) or fed to it (text-to-speech) could be reliably associated with whether the system was also knowledge-rich or knowledge-poor. Today that association no longer holds. In the main, it is the text-to-speech systems that have been augmented with statistical and database methods. However, hybrid concept-to-speech techniques have been developed for applying application expertise to live data and to producing prosodically acceptable full-text reports as a result. It appears that the approaches are merging and that techniques developed for one approach are also applicable to the other.

4.4 Simulations of processing effects on linguistic behavior

Very few systems in the AI tradition explicitly model cognitive processing constraints, let alone link them to linguistic behavior. One exception is Landauer's work [Lan75] on the effects of resource bounds on learning and recall. The limited resource in his system is *attention* and therefore, retrieval capacity. As attentional capacity is increased, so is retrieval and therefore learning. His system also shows that both the frequency and recency of a stimulus will increase its chances of being recalled. It is described in detail in Chapter 6.

Using Landauer's model, Walker [Wal93, Wal96b] shows that resource bounds are significant predictors of lexical choice, utterance semantics and interaction strategy. Her

system is the only multi-agent simulation so far that links resource-bound processing to interactive behavior.

The other notable simulation is Corriveau's work on reader-based comprehension [Cor96]. Whereas Landauer and Walker focus on spatial limits, Corriveau identifies *time* as the limited resource. Items decay and must respond to a retrieval signal within a pre-set time span. Age and activation determine how long a linguistic item stays in working memory before it is transfered to short-term memory and then finally to long term memory.

Corriveau's is an ambitious system that performs many language processing tasks, such as referent identification, prepositional attachment and inference. It does this with a combination of rules and thesaural, orthographic and acoustical annotations that are provided interactively by the user.[4] Its main simulation parameters are the richness of the knowledge base (poor, average, rich) and the amount of time allowed for responding to the retrieval signal (short, medium, long) ([Cor96], pp. 179). Variations affect the amount and kinds of inference in predictable ways – the richest knowledge base and longest decay times produces the most complete interpretations of the text. The poorest knowledge base and shortest decay times produced the most superficial interpretations, independently and together.

The system's many rules and dependencies make it difficult to isolate the effect of its timing variables. In addition, its performance is limited to a three sentence paragraph at most. However, it is the first (to my knowledge) of recent efforts in the symbolic AI tradition to incorporate a performance perspective. Moreover, it provides a wide-ranging description of performance, defining it as a product of both attentional and knowledge limitations.

Using the exemplar of garden path sentences,[5] Just and Carpenter show how individual differences in storage and computational capacities predict individual differences in text comprehension [JC92]. The limited resource in their simulations is an *activation capacity* that is apportioned over computation and storage tasks. Simulated readers with higher activation can maintain more simultaneous (partial) interpretations of a text, and for longer. Therefore, they are less likely to have difficulty with garden path sentences. Besides modeling human performance, the results show that processing can be interactive for high activation readers, but highly modularized for low activation readers. Although the model was developed for sentence comprehension (as many psycholinguistic models are), its methods and insights are useful to any simulation of resource bounds on linguistic behavior.

Most other simulations are connectionist. With few exceptions, they use topology

[4]However, the user selects from fixed choices, as provided by online semantic, spelling and pronunciation databases.

[5]The canonical example is *"The horse raced past the barn fell"*. Garden path processing that comes from limited capacity will not recognize "raced past the barn" as an embedded relative clause.

and thresholds as implicit resource bounds and focus on language learning rather than on language use. Notable efforts include systems that learn letter-to-sound rules [SR87], coarticulation [GL90] and syllable stress patterns within words [GT91]. In an interesting departure from most connectionist work, Dell [Del85, Del88] addresses a performance issue directly. His simulations produce word pronunciation errors as a consequence of activations of phonemes but also as a consequence of speech rate – higher speech rates produces more phonemic disfluencies.

4.5 Summary

Both text-to-speech and concept-to-speech synthesis efforts have traditionally aimed for increased linguistic competence in the form of more knowledge about the text and application, and better mappings of text structure and propositional content to prosody. Some have explored the effect of individual speaker performance influences such as affect and style. Of these efforts, only Johnson's work [Joh96] simulates the influence of limited capacity directly.

One problem shared by most synthesis systems is the lack of a sufficiently flexible discourse model. This is a large problem for text-to-speech systems, whose identification of topic and focus is minimal at best. In contrast, concept-to-speech systems include discourse models for tracking the focus of attention and overall structure. By default, such systems have perfect recall – at least for the items in the current global focus space. Therefore, they do not predict the linguistic behavior (or its prosodic subset) determined by resource bounds.

For these reasons, I employ Landauer's model of attention and working memory. It varies attentional capacity and therefore, varies retrieval. It is described briefly in the next chapter (Overview) and more fully in Chapter 6.

Part II

Approach

Chapter 5

Overview

My aim is to develop a model that generates prosodic variation from cognitive causes, and to incorporate the model into a speech synthesis system. So far, variation has not been a priority for most synthesis applications, whose goal is to produce better prosodic correlates of text structure and semantics. The main exceptions are found in the work on emotional and expressive synthesized speech, and in the more recent work on stylistic variation. Almost all the work in this area employs a descriptive model of the acoustical correlates of emotion or style. Such approaches are neither explanatory nor predictive.

In my approach, the productive cause of prosodic variation is the speaker's ability to recall previously encountered items. The software system that demonstrates this is called LOQ, variously, for *Loquacious*, *L-O-Q* (as in "elocution") and because *"loq"* itself comes from the Latin root, *loqui*, *"to speak"*. Its main function is to map search and storage in working memory to pitch and timing in speech. Its simulations show that variations in attentional capacity have the most influence on prosodic style, while variations in the pattern of storage produce variation within a style. Currently, LOQ produces three styles likely to be associated with attentional and memory differences: an animated and child-like prosody when attentional capacities are small, an expressive adult style for mid-range capacities, and a knowledgeable style for the greatest capacities.

As shown in Figure 5-1, LOQ takes text as input and produces speech as output. Thus, its equivalent human activity is read speech. Although a spontaneous speech application would appear to be the natural platform for simulating the effects of resource bounds on prosody, its implementation requires both text generation and text comprehension capabilities. A read speech application has the practical advantage of requiring only text comprehension and the theoretical advantage of requiring fewer simulation parameters.

Currently, LOQ simulations assume that the acts of reading (input) and speaking

(output) are problem-free. Of course, this is the ideal case, and has been chosen to keep the focus on the effects of limited attention and memory. In Section 13.3, I discuss the addition of reading and speaking as resource-bound components.

The basic function of LOQ is to map search and storage in the memory model to prosody in speech. The linguistically analyzed text is key to the matching algorithms and the matching algorithms are key to the search operation. The acts of storage and search are mapped to tone type, tone prominence, pitch range and duration.

The main components of the LOQ system are: the text, the linguistic analysis, the memory model, the algorithms that map search and storage to pitch and timing, and the synthesizer itself. Each is reviewed in the remainder of this chapter.

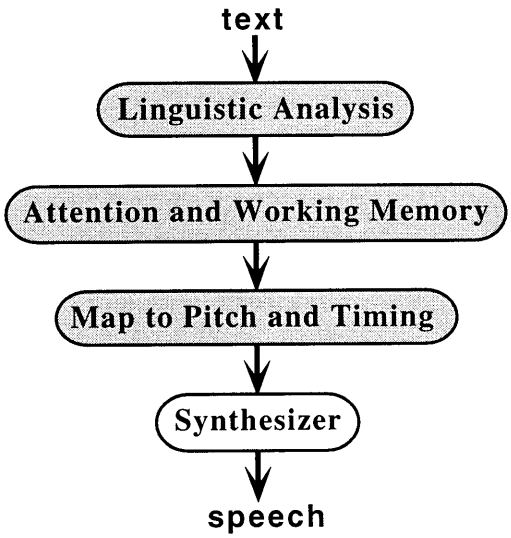

Figure 5-1: The LOQ system.

5.1 The text

Currently, the LOQ simulations use text from three genres: (1) fiction, represented by an excerpt from *One Hundred Years of Solitude* by Gabriel Garcia-Marquez; (2) non-fiction, represented by a news report originally delivered on National Public Radio; (3) rhymed poetry, represented by Lewis Carroll's poem, *Jabberwocky*. The same analysis techniques are applied to each text.

5.2 The linguistic analysis

The main body of LOQ's linguistic competence resides in the analyzed text. As currently implemented, a text is analyzed once, offline, for all simulations. Therefore, the initial linguistic information is the same for any simulation. The analysis (Figure 5-2) combines three methods: (1) manual annotation; (2) the use of natural language processing software; (3) the use of online linguistic databases. The manual annotations provide information such as grammatical role and anaphoric reference. Both are necessary for comparison and retrieval but cannot be reliably provided by current software. In general, because the hand-entered information is generic to language and text, it is highly probable that natural language processing software will be developed that can identify and tag these features. Currently, the natural language processing tools used by LOQ are: a tagger, a stemmer (which provides the uninflected form of a word) and a noun phrase identifier.[1]

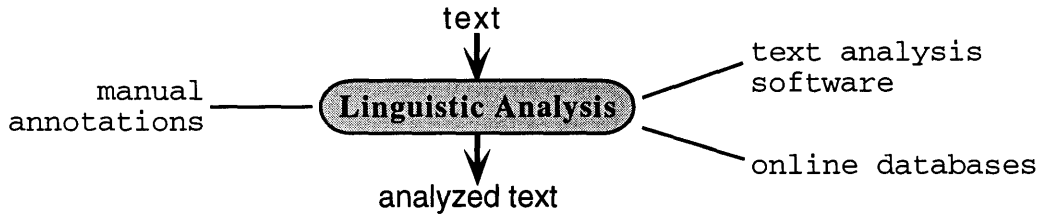

Figure 5-2: Sources and types of text analysis in LOQ.

The use of online databases is a technique that has been already used in various guises in other synthesis systems. Commercial synthesizers use table lookup to guide pronunciation and stress. Researchers at AT&T Bell Labs have pioneered the corpus-based approach to codifying pronunciation and phrase stress regularities [Spr98]. Horne *et al.* [HFLL93] use an online thesaurus to inform decisions about given or new status, in order to assign focal (for new information) or non-focal accents. LOQ also uses an online thesaurus, although in a manner different from Horne *et al.* So far, its collection of databases provides semantic, word frequency and pronunciation information. Their use illustrates a technique that is easily extensible as more kinds of linguistic data collections become available online.

5.3 The memory model

As used in LOQ, the working memory model shows how attentional capacity determines whether information is given or new, and thereby determines its prosody. The

[1]The parsing programs I tested produced parse trees that were too deep for a model of limited cognitive capacity, and too complicated to use in phrasing decisions. Therefore, the parsing into grammatical and syntactic clauses occurs by hand.

original model was developed and implemented by Thomas Landauer [Lan75]. It was later used by Walker [Wal93] to model interaction strategies and lexical choice as the consequence of limited attention.

It depicts *working memory* as a three dimensional periodic Cartesian space. Its contents are stored at integer-addressable locations in the space. Attentional capacity is depicted as a region within the space. The larger the region, the more items it is likely to contain and the more likely it is that recall will occur. The search for a match to the current stimulus proceeds outward from the center of the region and stops at the edge. Recall (i.e., retrieval) is the case in which a match is found.

Incoming stimuli are stored via a *search and storage* pointer that moves in a random walk across the space. The pointer's current location is the equivalent of the *focus of attention.* As a consequence of its random walk, incoming stimuli are stored in a spatial pattern that is locally random but globally coherent. That is, temporal proximity in the stimuli is represented by spatial proximity in the model.

This model contrasts with stack models of memory, which are strictly chronological, and semantic spaces in which distance is conceptual rather than temporal. Moreover, its operations reproduce the basic findings of learning studies which show that subjects are likely to recall items presented most recently and most frequently [Lan75]. The recency effect occurs because the most recently encountered items are stored closest to the pointer. The frequency effect occurs because items that have been frequently presented will be stored throughout the space and therefore, are likely to be retrieved from any number of locations.

In the work of both Landauer and Walker using this model, the main simulation parameter is the search radius, which defines the size of the search region and therefore quantifies attentional capacity. In Section 6.5, I motivate changes to the original model that add two more control parameters to the simulations: (1) the *stability* of the attentional focus; (2) the total *storage capacity* of the memory space.

As used by LOQ, the input to the working memory component is linguistically analyzed text, which is processed word by word and phrase by phrase. First, the incoming stimulus is stored. This is followed by the search for a match within the search region. The features that are identified in the text analysis are used by the matching process to find a match and determine its strength. The matching process measures the mutual information between the current stimulus and those that have been previously stored. Alternatively, it measures the amount of memory priming for the current stimulus. It is described in detail in Chapter 8.

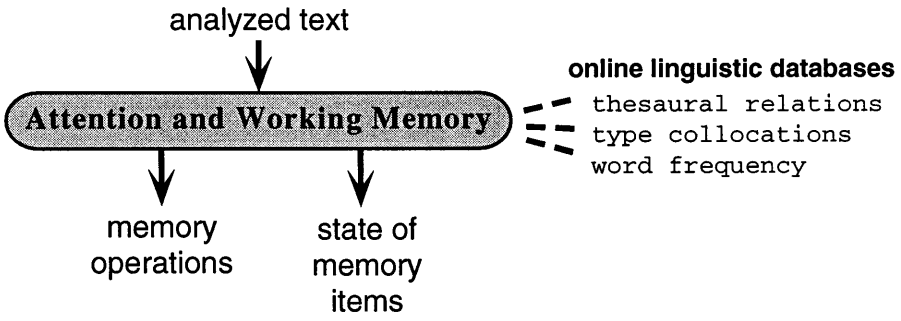

Figure 5-3: The input, output and sources of linguistic knowledge that are used in the matching process.

5.4 Mapping cognitive processing to prosody

As shown in Figure 5-4, the ongoing sequence of memory operations and their results are mapped to categorical and continuous features of prosody. Currently, the *duration* of a word is proportional to the total time it takes the system to carry out its main tasks of storing the word and searching for a match. The results of the search are mapped to *tone type* and *tone prominence*. The occupation density in the search region determines the *pitch range*. The motivations for the mapping are discussed in Chapter 9.

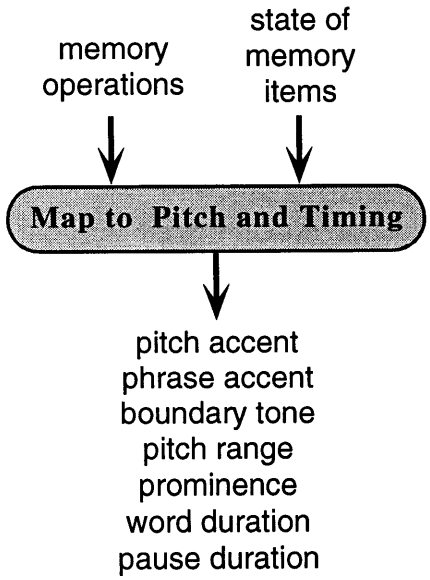

Figure 5-4: Memory operations and their results are mapped to continuous and categorical features of prosody.

5.5 The synthesizer

The prosodic specifications produced by the mapping must be translated for the synthesizer. Thus, the final product of LOQ's computation is a text annotated with prosodic instructions that the synthesizer can understand. Currently, LOQ transmits instructions to the TrueTalk synthesizer, which was developed at Bell Labs and is now distributed by Entropic Systems.

5.6 Summary

The limited resource explanation of prosodic variation is embodied in LOQ, a program that simulates the effect of different attentional capacities on prosody using Landauer's model of attention and working memory. The simulations test the hypothesis that prosody is significantly influenced by attentional capacity. They also test hypotheses that are implicit in Landauer's original model, namely, that the stability of the attentional focus and the total storage capacity of working memory may also have an effect.

This approach emphasizes the effects of processing in working memory. However, it cannot work without some background knowledge of language, the nominal contents of long term memory. In the current implementation, the text analysis that occurs offline provides the bulk of the speaker's *a priori* knowledge of language. Additional knowledge is incorporated from the databases as needed during the matching process.

Detailed descriptions of the memory model, the linguistic analysis, the matching process and the mapping to pitch and timing are described in Chapters 6, 7, 8 and 9, respectively.

Chapter 6

A model of attention and working memory

The claim that prosody reflects the capacities of the speaker is not controversial for the timing features of speech, which have long been correlated with individual processing and production capacities. However, it is unusual for intonation, which is typically explained without reference to an individual speaker. My reason for suspecting that intonational variation can also be linked to individual cognitive capacities is based on the established correlations between pitch accents and the given-new distinction. Typically, given information is de-accented and new information is accented. However, as I will show, Landauer's model demonstrates that given and new designations are not fixed but instead, vary with attentional capacity. Therefore, accenting should likewise vary.

In this section, I develop the basic claims of Landauer's model of attention and working memory from even simpler models. I then motivate some changes to the original model that more precisely quantify its resource bounds.

6.1 Limited attention and the given-new distinction

Most studies of intonation propose that a speaker accents a word if she believes it is not yet salient for the hearer, and de-accents it otherwise. Prince [Pri81] ties this to the given–new distinction, observing that whether information is given (salient) or new (not salient) depends on context. Most clearly, it is given if it has been previously mentioned in the current discourse segment [Bro83, Hir93]. Under looser criteria, the speaker is justified in assuming givenness for words and concepts with no explicit prior mention, for example, if they are closely associated with the topic or the speaking

situation [Pri81].

Implicit in these proposals is the idea that both the speaker and the hearer possess the information necessary for determining given and new status. However, as Walker [Wal96b] demonstrates, the availability of propositions for inference is not only a function of shared context and interaction history, but of the storage and retrieval capacities of each individual agent. The case in which each agent is able to retrieve all the relevant contents from their respective working memories is only one of many.

The mapping from recall in the memory model to pitch and timing in speech starts with the correlation between information status and accenting ([Pri81, Bro83, TH94]). To show how memory limitations vary the given or new status of a recently-mentioned item, I will start with one of the simplest memory models – a last-in-first-out queue. When used as a model of attentional state ([GJW95]), the item at the head of the queue is the focus of attention.[1] The model classifies the stimulus as given or new according to the outcome of a sequential search of the queue. If a previously stored instance is found, the current stimulus is deemed to be already part of the salient context. If not, it is new.

Finding a matching item in the memory queue is a consequence of the maximum search distance within the queue. Both are variable. When the search distance equals the queue size, all previously encountered items will be accessible, and only stimuli that are entirely new will be classified as new. When the stimuli are linguistic data, the corresponding intonation will contain mostly de-accented words. At the other extreme, a search limit of zero retrieves nothing, and the current stimulus is always treated as new information (even with unlimited storage). The corresponding intonation is composed entirely of pitch accented words.

Doubtless, the storage and recall capacities of most speakers lie within these two extremes. For the LIFO memory in Figure 6-1, a speaker whose retrieval capacity is small (represented by a search limit of two) will not retrieve a previously stored instance of *blue*. For this speaker, *blue* is new information, which she denotes with a high pitch accent. In contrast, a speaker with greater retrieval capacity (represented by a search limit of five) will retrieve a previously stored instance. Because *blue* is already salient, she utters it with either a low pitch accent or none at all.

To give this model more flexibility, consider a series of modifications that starts by moving the focus of attention to the center of the queue, as shown in Figure 6-2. This allows bi-directional search and storage outward from the center. For the same search limit, twice as many locations are accessible as in the original LIFO queue. The items closest to the focus in both directions will tend to be those stored most recently. However, this depends on the storage algorithm. If storage alternates direction on successive turns, spatial proximity to the focus of attention will map to temporal proximity. However, a global register is needed to keep track of the alterna-

[1] In a pushdown implementation, it is the item the top of the stack [GS86].

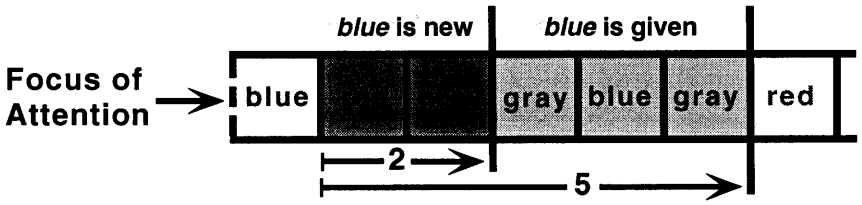

Figure 6-1: Working memory as a last-in-first-out queue. The speaker's attention is focused on *blue*, at the head of the queue. Items within the shaded regions are accessible for a search distance of five. Items within the darker region are accessible for a search distance of two.

tions. Choosing the direction randomly at each turn has the advantage of requiring no memory augmentations and will achieve roughly the same temporal/spatial equivalence.

The next modification links the ends of the queue to make it circular, as shown in Figure 6-3. As a result, items stored in distant time may become available serendipitously because an item stored outside the search limit in one direction may be found within it from the other. In addition, the circular configuration is a closed one. Therefore, the storage capacity of the entire memory becomes explicit.

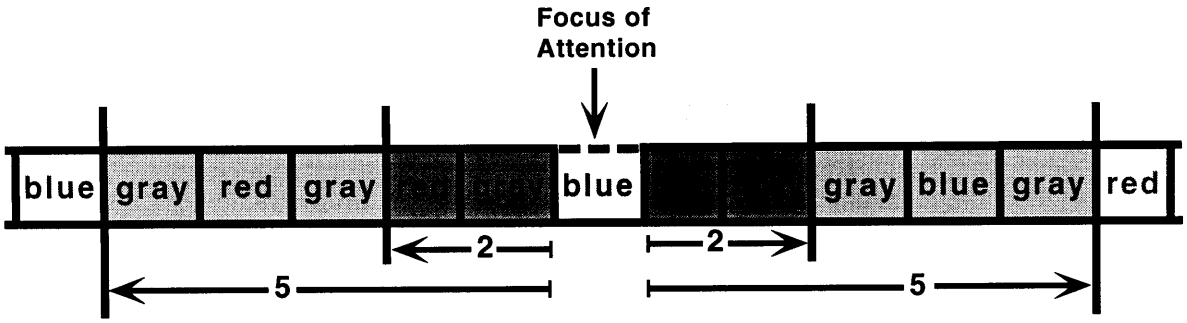

Figure 6-2: Working memory as a bi-directional queue. Search extends in two directions outward from the current focus of attention at the center.

The next modification maps the circular queue onto two dimensions. While this does not change the model's basic search and storage properties, the storage locations are no longer points on a circle, but integer-addressable locations on the surface of a torus. And instead of a bi-directional search in one dimension, a search is conducted within a region of a plane. The search distance in one dimension becomes the radius of a search region in two dimensions. All items stored beyond the radius are not retrievable and are therefore forgotten. As in the one dimensional case, the wraparound topology allows the retrieval of items stored in distant as well as recent time. One difference is that by allowing more comparisons for any distance, more items in the two dimensional memory are retrievable, and therefore are more likely to be classified as (representations of) given information.

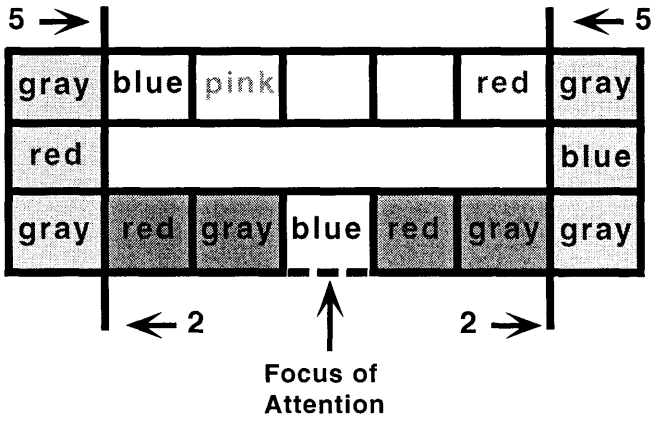

Figure 6-3: Working memory as a circular queue. Search extends outward from the focus of attention.

However, as given, the storage algorithm, mapped to two dimensions, will make displace previously stored items (randomly) in four directions instead of two, along the axes that intersect with the focus location. This is a poor use of the memory space. An alternative is to keep all the items in memory to remain stationary and allow only the focus of attention to move. The random choice of location in the one dimensional memory becomes a random walk of a moving focus of attention in two dimensions. The effect of these changes is shown in Figure 6-4.

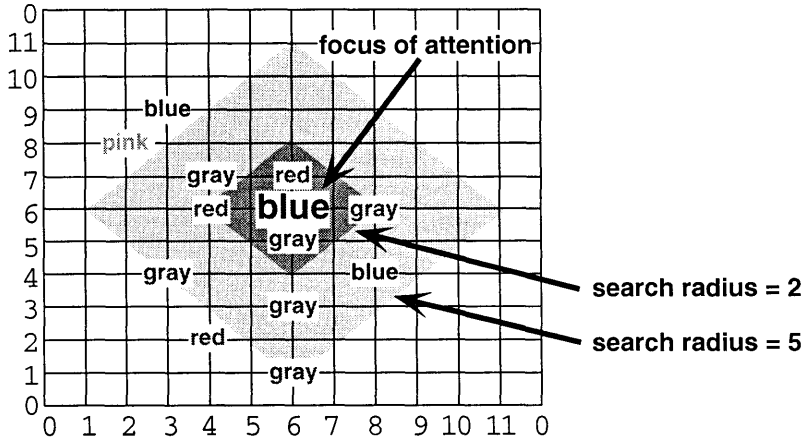

Figure 6-4: A section of a two-dimensional representation of working memory. *blue* is new information for a search radius of 2, and given information for a search radius of 5. The focus of attention is at the center.

We are now left with a model of limited attention and memory that is close to the one developed by Landauer [Lan75]. It is linked to prosody by its ability to calculate the given or new status of the current stimulus, which is mapped in turn to de-accenting and accenting, respectively. Because the search radius is a variable simulation parameter, its size will affect the determination of given and new. This

in turn affects the proportion of de-accented to accented words, and illustrates how attentional capacity produces prosodic variation.

6.2 Landauer's computational model of attention and working memory

Like the examples in Figures 6-3 and 6-4, Landauer's model of attention and working memory (hereafter, AWM, following [Wal93]) is a Cartesian space whose edges wrap around and whose contents are stored in locations addressable by integer coordinates. As implied by the previous examples, the axes of AWM are merely spatial – retrieval depends on the absolute distance from the focus of attention. An item outside the search region is not retrieved. Therefore, relative location has attentional consequences, but absolute location has no semantic interpretation.

Landauer instantiates AWM in three dimensions. Operationally, it shares the same properties with its two-dimensional counterpart – it assumes a distinguished location for the focus of attention, allows search within a region of the space as defined by the search radius, and stores its input via the random walk of a moving pointer. It operates with a simple update rule centered on the pointer – MOVE–STORE–SEARCH. This rule is applied at each time step as follows:

- MOVE the pointer. The pointer may either stay where it is or move one city block[2] in any direction. In a three dimensional Cartesian space, there are seven possible locations from which to choose, counting the current one. AWM implements this as a random choice at each step. Over time, this local algorithm describes a random walk across the space. In this way, the AWM pointer simulates changes in focus and context.

- STORE the current stimulus. The pointer may move in the absence of a stimulus. However, if there is a stimulus, its representation will be stored at the pointer's new location. This stimulus is an idea, event or entity in the world to which the simulated agent is currently attending.

- SEARCH for a match using the newly stored stimulus as a retrieval cue. The search proceeds outward from the pointer up to the radius that is fixed for the simulation. Data stored within the radius are *salient.* Therefore, the larger the radius, the greater the likelihood of recall. However, because the search proceeds in *constant time* (a search of all items at the same distance occurs in parallel), a greater search radius may result in longer search times.[3]

[2]Landauer points out that short term memory pathologies can be modeled by speeding up the pointer, or, by introducing bounded local storage and restricting the pointer to such a small region that it will begin to overwrite recently entered data.

[3]This also depends on whether the search strategy is exhaustive or self-terminating [Mac87].

RECALL, or retrieval happens when the search yields a match to the current input.[4] The simplest case occurs when the cue, A, retrieves an identical copy, A. However, more flexible match criteria may allow A to retrieve the similar item, A', or the related item, B [Lan75, RM88, RM94]. This depends on the application – retrieval criteria are task specific.

Figure 6-5 shows the schematics of the model. The stimuli for this memory are elementary – either filled or unfilled circles.

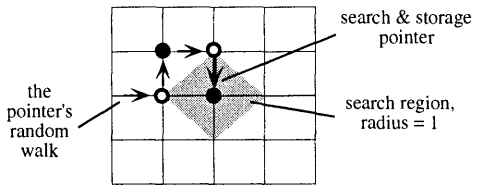

Figure 6-5: Search radius, storage and the path of the pointer's random walk for Landauer's model of attention and working memory.

Given or new status is determined by whether a matching item lies within the search region. As shown in Figure 6-6, prior history is only a partial determinant of given and new. The search radius determines the accessibility of previously stored items, and therefore, determines the classification of the current stimulus as given or new.

Landauer's simulations with AWM have reproduced human experimental results such as the *recency* effect, in which subjects are likely to recall the items presented most recently because they are stored closest to the pointer in AWM, and therefore, remain salient and retrievable. They have also reproduced the *frequency* effect in which subjects recall items presented most frequently. In AWM, frequently encountered items are stored throughout the space. The more even the distribution, the more likely that one of these items will be retrieved during any search operation.

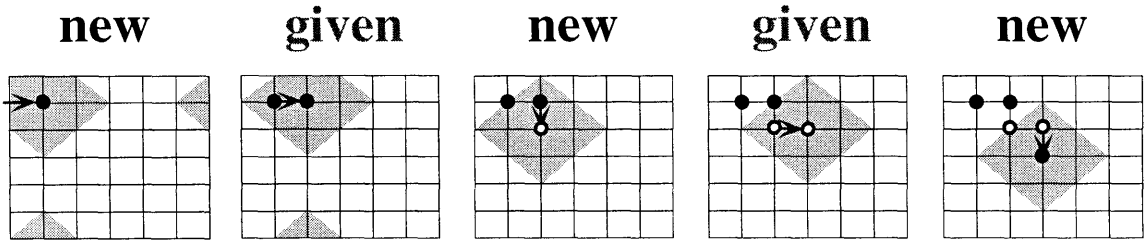

Figure 6-6: Given and new determinations as a consequence of the random walk, temporal sequencing of stimuli, and the search radius. The memory is periodic (wraparound) as shown by the area covered by the search region.

[4]In Walker's simulations, items that are retrieved are then re-stored by the pointer. This adaptation is consistent with AWM and in addition, illustrates how items come to be in (re)focus.

6.3 Operational and spatial effects of the AWM design

In both Landauer's and Walker's implementations, the main simulation parameter is the search radius. It determines two important aspects of recall: (1) roughly, how far back in time retrieval can go; (2) how many items can be examined in one search operation. Both would appear to be the sole consequence of the magnitude of the search radius. However, the number of items is also determined by local topology, namely, the number of intersections at each node. For example, a two dimensional Cartesian space has two axes intersecting per node and four nodes at unit distance from any intersection. If the regular space has more than two intersections per node (e.g., three for a triangular lattice, four for an octagonal one) the number of nodes at unit distance will be greater.

The effect of the pointer's slow random walk is to map temporal proximity in the world to spatial proximity in the model. Stimuli encountered closely together in time will be stored closely together in the model, although in a cluster rather than in strict temporal order.[5]

The effect of the periodic topology is that items stored in distant time may be serendipitously retrieved. This is because, in a periodic space, not all the nodes are uniquely at the edges of one search radius. Because the space wraps around, nodes accessible at small distances will be again accessible at larger ones. As Figure 6-7 shows, the number of new locations is the derivative of the total number of locations in the search region (as defined by the search radius). In a two dimensional $n \times n$ space, the maximum distance required to access all locations once during a search, is the length, n, in any one dimension. For a Cartesian space, the equation for calculating this distance is:

$$(1) \quad distance_{max} = \tfrac{1}{2}(n \times d).$$

Therefore, for three dimensions and higher, the search radius that covers the entire space is larger than the length in each dimension.

6.4 Discussion

The AWM model is agnostic about the data it processes. Therefore, it is a general purpose model of attention and memory. Its particular strength as a model of cog-

[5]If instead, the pointer movement were linear (the simplest case is a traversal along one axis) the temporal ordering would be strict along this linear path but unpredictable away from it.

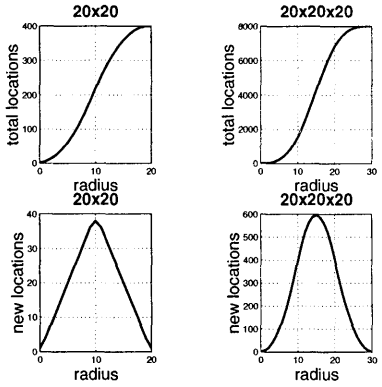

Figure 6-7: The number of nodes in a region of periodic Cartesian space increases in an S-curve as the radius of the region increases. Likewise, the number of new locations is the derivative of the total number of locations. The smallest (0) and largest distances contribute only one node each.

nitive processing is that it reproduces experimental data on the effects of frequency and recency on recall. This makes it especially well-suited for simulating the effects of attentional capacity on prosody because both accenting and timing features respond to frequency and recency. It is also well-suited for producing individual differences because of the stochastic motion of the pointer, which assures that each individual simulation will follow a different course, even if overall averages are the same.

6.5 Critique

AWM has one feature that is inconsistent with its basic premise of limited capacity – it allows each location to store more than one item. Landauer implements this as a pushdown stack [Lan75]. This is problematic. With unbounded storage at each node, the storage capacity becomes theoretically infinite. The model's dimensionality is effectively increased by one but the actual size of the added dimension is unknown beforehand. Moreover, while the dimensionality is increased for storage, it is not necessarily increased for search because the contents of each locus are outside the AWM coordinate system. Therefore, the random walk will not explicitly access them. Instead, a separate search procedure must be designed for traversing the local stack. If stack search is not also limited, the effect of unlimited access at each node will override the effects of a limited search in the AWM space.

An alternative is to restrict storage to one item per node so that the total storage capacity is finite and predictable beforehand. This dispenses with stack search as an AWM operation and motivates the addition of memory operations that approximate summary in human processing. They are discussed in the next section.

6.6 Revision: AWM with restricted local storage

To show more clearly the effects of limited capacity, storage in AWM is restricted to one item per node. This makes the total storage capacity more precise – it is simply and precisely the number of nodes in the space. The few ways to increase storage capacity are to increase the length in any one dimension, the number of dimensions or both. I will refer to AWM with restricted local storage as AWM′.

A direct consequence of restricting local storage is the addition of new memory operations and the alteration of existing ones. For example, when nodes can store only one item, the MOVE operation must include *a search for unoccupied space.* For the same reason, the act of *freeing up space in memory* becomes essential. More space may be obtained by *swapping* out (all or part of) the current contents of memory. Swapping is standard for computer memory. Walker [Wal96a] suggests that the pauses that occur before topic changes in spoken dialogue reflect the time it takes for a speaker to displace old referents and swap in new ones ([Cah92]).

Another strategy is to allow the model to *overwrite* old data, or to simply prevent it from storing new data. As Landauer points out, both alternatives correspond to different short term memory pathologies [Lan75]. In the first case, old data are lost; in the second, new data are not stored. In contrast to either option, data that are swapped out into long term storage are not lost because they can be swapped back in. The problem with adding swap to AWM is that it requires the design of long term storage. Currently, the organization and operation of long term storage (memory) are outside the AWM model. Therefore, I introduce *compression* as an alternative that frees space but does not (yet) remove items from working memory. It performs a simple form of summary and models the fact that people tend to remember the gist but not the exact surface form of spoken or written text [Sac67].

6.6.1 Compression of items in working memory

What I have termed "compression" denotes the creation of a more compact representation. Ideally, it is implemented by using the same number of bits to store information of varying complexity, as in the information theory definition. In the connectionist tradition, Pollack proposes a compression and expansion scheme to explain how recursive structures [Pol90b] might be stored and retrieved from long term memory. Within this framework, Plate [Pla94] proposes circular convolution to store (compress) and retrieve (expand) linguistic components and their case frame roles from one representation. Both aim for an efficient representation of complex relations, in order to show how finite long term memory stores the vast amount of information that it does.

By comparison, the compression I propose for AWM′ is simplistic – data formerly

stored at separate addresses are combined into one complex symbolic representation that is stored at a single location. Thus, it is most similar to the chunking procedure in the Marcus parser [Mar80], which recursively groups smaller syntactic units into larger ones. To avoid the obvious problem, that compression itself might remove salient items from memory and thus undo the search radius predictions, the items in memory have a simple form of state that includes whether or not they are compressible. This depends on whether the item itself is a predictor of upcoming events. In LOQ, prediction is based mainly on syntax, grammatical role and *verb subcategorization.*[6] For instance, a determiner predicts an upcoming noun. Therefore, until the noun itself is in memory, the determiner is incompressible and cannot be removed. Likewise, the identification of a subject clause predicts the imminent construction of a full grammatical clause. However, until the parent clause is constructed, the subject clause is incompressible. This schema is explained more fully in Section 8.3.

6.6.2 Revised update rule

As a consequence of limiting storage to one item per location, the steps in the update rule for AWM′ becomes more complex. For example, the revised MOVE step also includes compression if no free space is found. Similarly, when compression is included as a precursor to storage, its inverse, *expansion*, becomes part of retrieval. Therefore, expansion of a compressed element is one possible outcome of retrieval and as such, one possible effect of SEARCH. It is not implemented *per se* in LOQ for reasons discussed in Chapter 10. However, LOQ includes procedures for retrieving an item from a compressed representation, for example, a word from a noun phrase, or a noun phrase from a grammatical clause.

Because each node stores only one item, storage in AWM′ cannot occur without moving the pointer. AWM′ adds the reverse constraint, which is that the pointer cannot move without storing something. Items are therefore stored in the model at a constant rate. Overall, the link between moving and storage complicates the internal structure of each step, but simplifies the update rule as a whole, to become STORE–SEARCH.

The differences between the update rules for AWM and AWM′ are:

- In AWM, storage at a node is, in theory, unlimited. In AWM′, storage can only occur at an unoccupied node, because each node can only store one item.

[6] **Verb subcategorization** describes the thematic (θ) roles of the syntactic clauses that are linked to a verb. The roles vary according to the kind and number of arguments for the verb – usually, according whether the verb is transitive, intransitive or bi-transitive verbs (among others). Examples of such roles include Agent, (the one who acts), Patient (the one acted upon) , Instrument (used by the agent to act upon a Patient), Goal (a place or state). In theory, the subcategorization or *case grammar* properties of a verb are represented in a language-user's **mental lexicon** [Bre82].

- For AWM, the random selection of the next location includes the current location. Therefore, the pointer may remain where it is at any time step. In AWM′, the pointer changes its location at each time step because storage and pointer movement coincide, and because storage is limited to one item per node.
- In AWM, the pointer may move even when there is no incoming stimulus. In AWM′, it moves only when a new stimulus is encountered and therefore, stored.
- In AWM′, compression is a possible outcome of attempting to move the pointer to an unoccupied node.
- Likewise, expansion of a compressed item is a possible outcome of retrieval in AWM′.[7]

The differences are summarized in Table 6.1.

Operation	AWM	AWM′
MOVE	Select location. Move pointer.	—
STORE	Store stimulus.	Select location ⇔ If occupied, compress. Move pointer to unoccupied location. Store stimulus.
SEARCH	Compare stimulus to memory contents. Retrieve matching items.	Compare stimulus to memory contents. Retrieve matching items ⇔ Expand.

Table 6.1: Comparison of the update rules for AWM and AWM′.

6.7 Summary

In this section, I have motivated the use of Landauer's model of attention and working memory (AWM) as a cognitive model that is appropriate for producing prosody. The key is that given and new determinations are a consequence of attentional capacity in the model. Moreover, the model itself reproduces the effects of recency and frequency on learning. Both are also relevant to intonation and duration.

As originally conceived, AWM allows infinite storage at each location. However, this appears to contradict its philosophy of limited resources. It also complicates both

[7]Expansion is currently implemented as a retrieval procedure only, rather than one that also re-stores the retrieved item in an independent location.

attentional predictions and procedures. Therefore, the revised model, AWM′, limits storage to one item per node. This gives some predictive power to the overall size of the space. In addition, it adds compression and expansion to the update rule and in accord with results that show that humans generally have poor working memory for verbatim content. Compression and expansion also set the stage for further extensions to AWM′, such as the swapping of compressed representations into long term memory.

Chapter 7

Input: Processed Text

Most linguistic comparisons are far more complex than the boolean criterion used in the AWM examples. They may compare many features instead of one and allow partial matches instead of requiring an exact match. Indeed, to use a linguistic item as a retrieval cue, a speaker must know about the structure, meaning and use of language. LOQ applies this knowledge via the match criteria, which distill from the literature the most relevant and prevalent ways that linguistic items *prime* for the current stimulus, and by the same token, the ways that the current stimulus functions as a *retrieval cue*. Comparisons on these criteria gauge the mutual information between items in memory and the current stimulus.

In this chapter, I discuss the linguistic information that is required by the matching process, and the methods for imparting it to LOQ. For a phrase, the required information is its syntactic constituency, its clausal constituency, its relation to other phrases and its meaning. For a word, the required information is its part of speech, its uninflected root, its frequency of occurrence for English, its pronunciation and its meaning. For a punctuation mark, the required information is mainly the function it performs in segmenting text.

LOQ obtains this information from three sources: (1) a manual text mark-up; (2) automatic natural language analyses; (3) online linguistic databases. The text mark-up provides the structural and relational information that is not yet reliably available from either online databases or text analysis software. The text analysis tools perform part of speech tagging, word stemming and noun phrase identification. The online databases provide pronunciation, word frequencies and the semantics of words and phrases.

The phrase, word and punctuation information comprise the LOQ speaker's background knowledge of spoken and written English, that is, the *a priori* contents of long term memory. Because the effects of long term memory on linguistic performance are outside the scope of this thesis, its contents are the same for all LOQ speakers. More-

over, its contents are assumed to be error-free – only the correct parses count and only the most appropriate meanings apply. Currently, to ensure this, a human editor must review the output from the automatic analyses.

7.1 Manual text mark-up

The manual text mark-up substitutes for automated text comprehension methods that do not yet exist, or are in the early stages of development. It identifies grammatical and syntactic clause boundaries, simple verb subcategorization relations such as argument and modifier, pragmatic relations such as anchoring and co-reference, non-lexicalized constituents such as traces and deletions, and the function of punctuation and layout on the page. The mark-up provides three kinds of information: (1) clausal constituency; (2) clausal relations; (3) the types of text segmentation indicated by punctuation and layout.

7.1.1 Clausal constituency

The standard means for grouping words into coherent units is the (context-free) syntactic parse, which builds a tree structure for each syntactic sentence. Another approach is to divide a sentence according to the thematic roles for the verb, such as Source, Goal, etc. [Lyo77] or more simply, according to grammatical roles, such as Subject, Verb and Object.

Of these methods, text analysis software exists only for producing sentence-based syntactic parses. However, such parses are not necessarily the best bases of intonational phrasing. For example, Chomsky [Cho65] (pp.13) observes that when people read:

(2) [[$_{NP}$ this] [$_{VP}$ is [$_{NP}$ the cat that caught [$_{NP}$ the rat that stole the cheese]]],

they normally place intonational breaks after *"cat"* and *"rat"*, instead of before the main syntactic breaks. In contrast, in a parse based on grammatical role, the ends of each clause coincide with the typical phrase break locations:

(3) [[$_S$ this] [$_V$ is] [$_O$ the cat]]
(4) [[$_S$ that] [$_V$ caught] [$_O$ the rat]]
(5) [[$_S$ that] [$_V$ stole] [$_O$ the cheese]].

The thematic roles based on verb subcategorization will also locate these breaks. However, because they are semantic, their taxonomies are sometimes ambiguous. While this is not a reason to reject an organization based on thematic role, LOQ

currently opts for a parse based on the simpler taxonomy of grammatical roles. They are also verb-based (the subject and object are arguments of the verb) but are fewer and unambiguous. In addition, their use in the mark-up is particularly relevant to prosodic applications, since grammatical role and accenting are often correlated [Ter84, TH94].

The structure of a full grammatical clause

The core of an English grammatical clause is the SV (Subject Verb) sequence. More complex clauses are constructed from constituents that are either required or optional, depending on the verb. For example, some verbs take no objects, others take two – e.g., an object and an indirect object, and still others, like the verbs of cognition, take full sentences (denoted here by *SS* for "Syntactic Sentence"):[1]

(6) [$_{SS}$ [$_{NP}$ He] [$_{VP}$ claims [$_{SS}$ [$_{NP}$ the butler] [$_{VP}$ did [$_{NP}$ it]]]].

An analysis according to grammatical role says that every verb belongs in its own grammatical clause (denoted by *Cl*), as does every subject. In sentence (6), the SV construction occurs twice. Therefore, it becomes two grammatical clauses – the SV clause,

(7) [$_{Cl}$ [$_{S}$ He] [$_{V}$ claims]],

and the SVO clause,

(8) [$_{Cl}$ [$_{S}$ The butler] [$_{V}$ did] [$_{O}$ it]]].

The full syntax for the grammatical clause includes not only subjects, verbs and objects, but complements and indirect objects as well. A complement is the "object" of the predicative verbs such as (i.e. *is, feels*). These verbs link a subject and predicate in a relation that is often *equative*, as in, *"She is a butcher."* or *attributive*, as in, *"She is skilled."* An indirect object (IO) occurs for a bi-transitive verb, as in, *"She gave them a present."* The context-free specification for a full grammatical clause is:[2]

(9) Cl ::= S V {{C} | {O} | {IO O}}.

[1]Technically, according to Chomskyan theory, the root node of a syntactic (deep structure) sentence is composed of a complementizer (*what, where, if, whether, that*, etc.) followed by a simple Sentence, which is in turn composed of a noun phrase (NP) and verb phrase (VP).

[2]Optional constituents are denoted within brackets, *"*"* is glossed as *"zero or more"* and "|" denotes logical OR.

That is, a full grammatical clause contains a mandatory subject and verb, which may be followed by a complement, an object, or an indirect object–object sequence.

Some verbs take arguments that are neither objects, complements nor indirect objects but instead fill a thematic role. Often, these are prepositional phrases. For example, the verb, *"put"*, takes two arguments – a noun phrase object and a prepositional phrase that specifies a location, as in:

(10) I put [$_{NP}$ the vase] [$_{PP}$ on the table].

To capture this relation in the syntax, but without going so far as to specify thematic roles, I augment the syntax with two generic types: the `Post` clause, which applies to prepositional and adverbial phrases that follow the verb and which may fill the functions of argument, modifier or adjunct; the `Pre` clause, which applies to constituents that precede the subject, such as adverbials that modify the whole sentence, and pre-posed modifiers of the verb, as in, *"In the chair, he slept."* The main difference between the two types is their position in the full clause. They are specified as independent types mainly to highlight unusual uses. For example, arguments typically follow the verb (in English), while adverbials may either precede or follow the verb.

The final addition to the syntax is the `Br` ("bridging") clause, which applies to connectives between clauses that do not count as adjuncts. Mainly, they are conjunctions, such as *"and"* and *"but"*, and relativizers, such as *"that"* and *"which"*. These additions produce the following syntax for the grammatical clauses that LOQ processes:

```
(11)  Cl  ::=  {Br} Pre* S V {{C} | {O} | {IO O}} Post*.
```

7.1.2 Representational and computational consequences

Building a parse around grammatical roles has a number of advantages and consequences. It produces shallow and minimally recursive parse trees, supports the bottom-up processing of read text, and motivates both the use of empty category constituents and the inclusion of information about the kinds of relations among phrases and clauses.

Shallow parse trees

A parse based on grammatical role produces shallow trees that are minimally recursive and also contain fewer words per tree than their syntactic parse equivalents. This is consistent with a limited resource approach. Most people are unable to maintain

either recursive representations [Cho65, Pin94] or long lists of verbatim content in their working memory [Mil56].

The representational advantage of a parse based on grammatical roles is illustrated in Figure 7-1. It shows a syntactic sentence-based parse from the Penn Treebank[3] and, to the right, its representation as two grammatical clauses. The syntactic tree has a depth of thirteen. Its grammatical clause equivalents have depths of four and six for the first and second clauses, respectively.

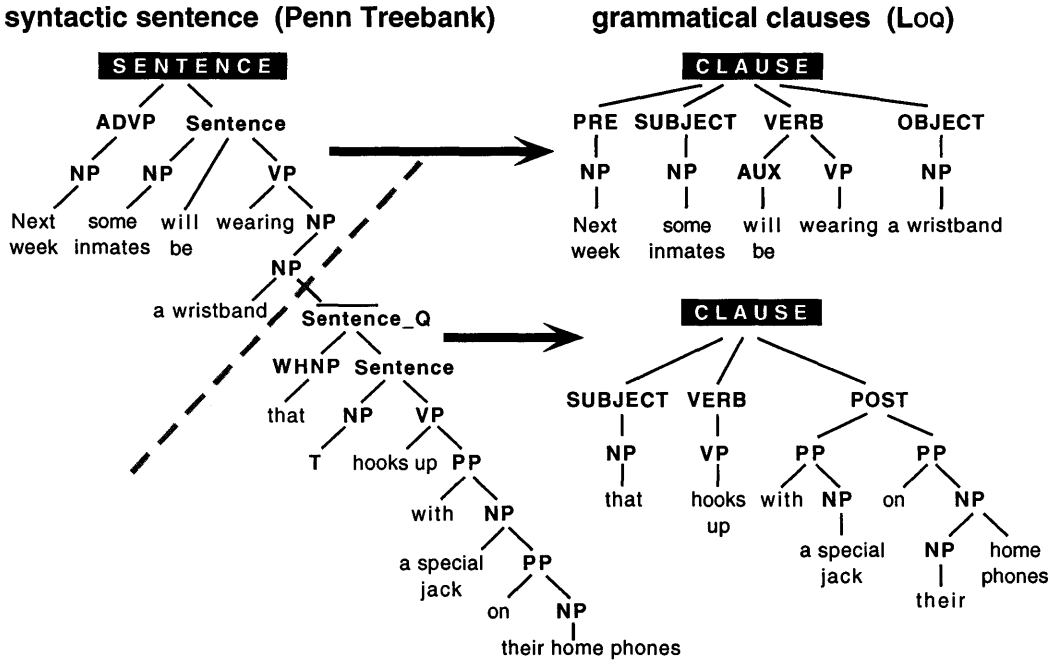

Figure 7-1: A deep syntactic parse from the Penn Treebank and its representation as two shallower and minimally recursive trees, which are organized around grammatical roles and fewer syntactic categories.

Sequential processing

Because the grammatical clause categories replace some of the syntactic sentence categories such as Sentence and WHNP, only a subset of all syntactic categories is needed for the LOQ parse trees. They are: the adverbial phrase (AdvP); the adjectival phrase (AdjP); the noun phrase (NP); the possessive (genitive) noun phrase (GenNP); the prepositional phrase (PP); the relativizer phrase (RelP); the auxiliary verb phrase (Aux); and the verb phrase based on the main verb (VP). In the mark-up, the verb phrase labeled "VP" is actually a truncated form of the verb phrase that is found

[3]The condition of using this parse or any other from the Penn Treebank is that the user must proclaim its provisional and preliminary nature, which I do, hereby. This parse is from the 1991 version of the Treebank.

in the sentence-based syntactic parses. It contains the main verb and any optional adverbial modifiers. However, it is separated from its arguments or modifiers, which appear as object, indirect object or adjunct clauses.

As much as possible, the syntactic constituents are grouped into a linear sequence of phrases within the grammatical clause. Generally, this produces trees that are hierarchical but non-recursive. For example, the first clause in Figure 7-1 (top right) is strictly hierarchical; the second (bottom right) contains only two embedded syntactic phrases (*"a special jack"*, *"their home phones"*) and only one recursive construction (*"their home phones"*).[4]

A left-to-right and bottom-to-top traversal of the grammatical parse tree mimics the bottom up processing of read text by which words are assembled into phrases and phrases are assembled into clauses. The traversal is easily converted into a sequence of tokens. Figure 7-2 shows the sequential decomposition of the first grammatical clause in Figure 7-1. This is the input stream that feeds into the AWM′ component in LOQ.

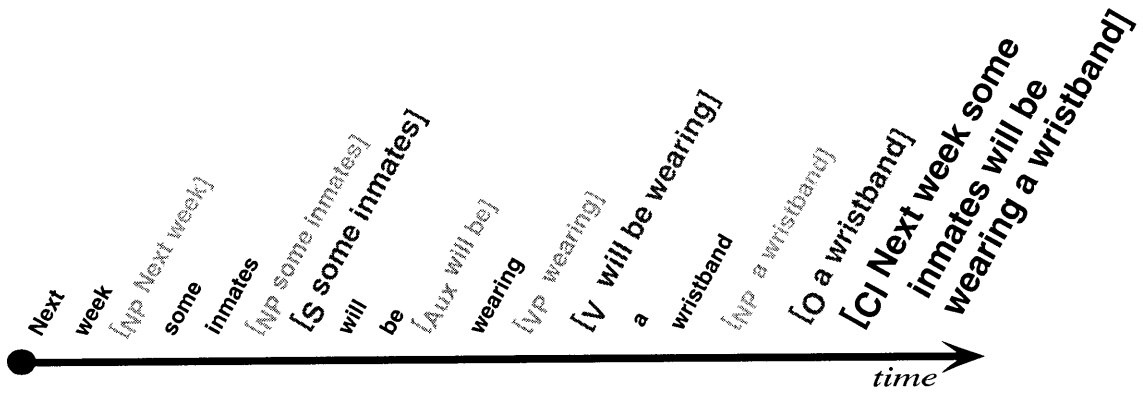

Figure 7-2: Bottom up sequential processing as the decomposition of a parse based on grammatical role.

Empty categories

The minimal full grammatical clause requires both a subject and a verb. This motivates the recognition of unlexicalized subjects and verbs. For example, the sentence,

(12) `Winter came and went.`

contains two verbs and therefore, the bases of two full grammatical clauses. However, it contains only one subject, *"Winter"*, which is the lexicalized subject of the first clause, and the unlexicalized subject of the second clause:

[4]In contrast, the sentential parse contains six embeddings and four recursive constructions.

(13) $[_{Cl}$ Winter$_1$ came]
(14) $[_{Cl}$ and e_1 went].

The recognition of unlexicalized constituents assists in the identification and construction of full grammatical clauses. As the constituents of a clause, they are part of the stream of tokens that LOQ processes. Their main effect during processing is to reinforce the salience of the entities to which they refer. Formally, they are the *empty categories* [Cho82]. Chomsky identifies three types: the *trace*, the *parasitic gap* (deletions) and the *PRO subject* of an infinitival clause. All three are used in the LOQ representations and comparisons.

The trace Chomsky [Cho82] analyzes the *trace* as an artifact of a syntactic transformation from underlying to surface form. His explanation relies on the *Move Alpha* rule, which says that if a grammar allows any constituent, α, to move, it will leave a trace in the syntax tree at the location from which the constituent was moved [Coo88]. The trace will persist in the surface form as well. For example, in:

(15) the book$_1$ that I read tr_1

a trace marks the original position of *"the book"*. Note that the trace and its lexicalization are *co-indexical.*[5] Both constituents refer to the same entity. With a slight shift in emphasis, Chomsky's explanation becomes a theory of cognition as well as textual relations. It says that the verb, *"read"*, takes an object but because of α movement (a cognitive operation), the object position contains only an unlexicalized trace. A competence explanation says that this fulfills a subcategorization requirement of the mental lexicon; the performance interpretation proposes that at the point at which *"read"* is uttered, *"book"* continues to be salient to the speaker.

The parasitic gap Another type of non-lexicalized constituent is the *parasitic gap*, or ellipsis, as it is commonly known. It marks the place from which a constituent has been deleted. However, unlike the trace, a lexicalized form will not be found elsewhere in the tree. The clause in (14) contains a parasitic gap.

The PRO constituent Because every verb must have a subject, Chomsky's Universal Grammar also includes the PRO constituent for subjects that can not be lexicalized.[6] These are the subjects of infinitives, which are co-indexical with a lexicalized subject. For example, English permits a lexicalized subject in:

[5]Two lexical items are co-indexical if they both refer to the same entity. This is also termed *co-reference* or *co-specification.*

[6]And then, there is the opposite case, in which no referent exists for a subject, but the verb requires that it be lexicalized. These are the *pleonastic* subjects, as in, *"It is raining."* or *"There was a long pause."* Such pronouns have a grammatical role but no actual referent.

(16) [$_{Cl}$ Rose wanted [$_{Cl}$ John to go]

but not in:

(17) [$_{Cl}$ Rose$_1$ wanted [$_{Cl}$PRO$_1$ to go].

Like the other empty categories, the PRO constituent provides evidence that a referent is still salient. However, because it is required by the syntax of a grammatical clause, it provides very little new information. Probably for this reason, cliticization often occurs across PRO boundaries. For example, a clitic between the main verb and *"to"* occurs far more often when *"to"* precedes an infinitive, than when it heads a prepositional phrase. Thus, *"want to"* may undergo cliticization (*"I wanna go."*) but *"went to"* will not (*"I went to the beach."*).

7.1.3 Relational annotations

Currently, the relational annotations for phrases and clauses provide information about verb subcategorization and cohesive relations.

Verb subcategorization

The attachment relations that reflect verb subcategorization are explicit in the sentence-based parse trees but often absent from the grammatical parse trees. However, the type of relation that is the basis of the attachment is not conveyed by either structure. For verbs, it must be found in the mental lexicon, which contains information about arguments and modifiers and their thematic roles. The LOQ annotations indicate whether an argument or modifier relation is responsible for the structural linkage, thereby both restoring and augmenting the information that is explicit in the sentence-based parse. By naming the relation as either argument or modifier, the annotation provides some of the verb subcategorization information that is attributed to the mental lexicon. For the clauses X and Y, argument and modifier relations are defined as follows:

- If X is an *argument* of Y, then X is *required* according to verb subcategorization, as in:

  (18) [$_{Cl}$ [$_S$ I] [$_{V_Y}$ closed] [$_{O_X}$ the door]];

- If X *modifies* Y, then X is *optional* according to verb subcategorization, as in:

  (19) [$_{Cl}$ [$_S$ I] [$_{V_Y}$ closed] [$_O$ the door] [$_{Post_X}$ silently]].

Cohesive relations

Cohesive relations describe the many textual devices by which authors or speakers relate current material to previous or upcoming material [HH76]. Some kinds of cohesion may be inferred from syntactic structure. For example, in *"Dora surprised herself."*, the antecedent of *"herself"* must be *"Dora"*, according to Chomsky's government-binding theory [Cho82]. However, many cohesive relations occur across sentences and clauses. The LOQ annotations mark two: co-reference and anchoring.

Co-reference Determining co-reference is currently one of the most difficult tasks for natural language processing. Yet locating the salient referents is necessary for text comprehension and for accenting and de-accenting appropriately. Useful textual clues to salience are often provided by syntactic form. The typical progression that reflects increasing salience goes from an indefinite noun phrase, to a fully specified definite one, to a pronoun and possibly to ellipsis. However, pronominalization and ellipsis are not restricted to noun phrase referents. The referents of verbs and even whole sentences may be denoted by *"do so"* (for verbs only) and *"it"* [Web84], and ellipsis can occur for most constituent types [HH76].

The simplest methods for computing co-reference are based on recency or selectional restrictions, as when *"he"* is taken to select the most recently mentioned male. However, because this method is neither reliable nor robust and because more sophisticated methods are still under development (e.g., [AHG98]), the annotations for co-reference are added by hand to the LOQ input.

Anchoring Prince [Pri81] identifies anchoring as a pragmatic device that allows a speaker to introduce new items to the discourse without marking them as fully new. In this way, the main focus of attention is not displaced by new information, because anchoring marks it as background. As a technique, it provides a shorter alternative to the traditional sequence for introducing new information, yet allows this information to be treated as given at first mention. According to Prince, the referent of a noun phrase is anchored if the noun phrase contains within it another noun phrase whose referent is not brand new to the discourse. Thus, $[_{NP}$ `you]` is the anchor in:

(20) $[_{NP}$ `the book` $[_{SS}$ `you recommended]],`

and $[_{NP}$ `the Beauty]` is the anchor in the appositive noun phrase,

(21) $[_{NP}$ `Remedios` $[_{NP}$ `the Beauty]].`

LOQ also treats as an anchor both possessives and any properly contained constituent that constrains the search for the referent of a noun phrase. Thus, [$_{NP}$ her] is the anchor in:

(22) [$_{NP}$ her [$_{NP}$ unscheduled meals]],

and [$_{PP}$ of [$_{NP}$ solitude]] is an anchor in:

(23) [$_{NP}$ [$_{NP}$ the desert] [$_{PP}$ of [$_{NP}$ solitude]]].

Noun phrase modifiers

Constituents are annotated as modifiers if they provide information that is not essential to locating referents. Often, they are prepositional phrases, as in:

(24) [$_{NP}$ her dreams] [$_{PP}$ without [$_{NP}$ nightmares]]].

Adjectives, by definition, modify the head noun of the noun phrase that contains them both. LOQ make this relation explicit via automatic annotation.

An example

Figure 7-3 shows how the argument and cohesive annotations to the grammatical clause restore attachment information that is explicit in the structure of the sentential parse shown in Figure 7-1.

7.1.4 Punctuation and layout tokens

Read text is segmented and grouped by both punctuation and layout. Punctuation is a visible orthographic token, whereas layout is implicit from the blank portions of the page. For example, paragraph boundaries are signaled by the blank space that follows the last sentence of a paragraph and by the indentation or blank line that precedes the next. The LOQ mark-up represents layout information as explicit tokens. This approximates the reader's recognition of the end of a paragraph or, for poetry, the end of a line.

As much as possible, the mark-up places most layout and punctuation tokens outside any grammatical clause. This expresses the fact that punctuation is often optional after a grammatical clause. To assume that it is an underlying part of a clause and

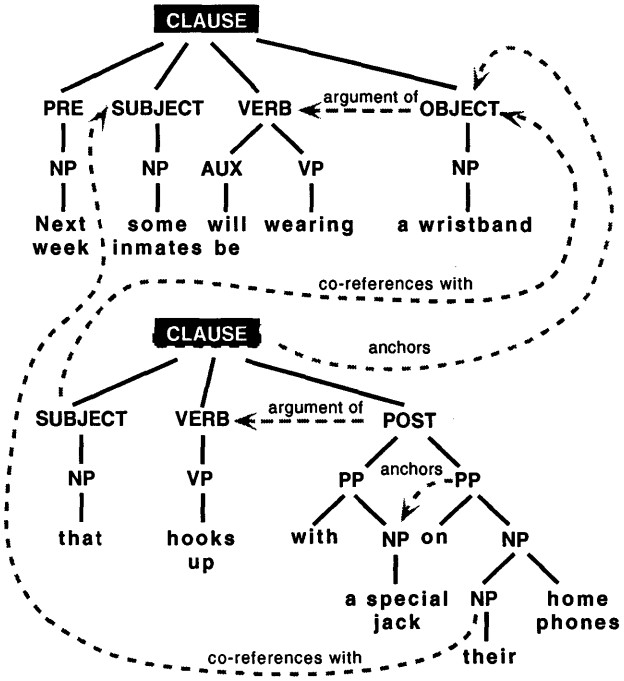

Figure 7-3: Annotations to the clauses preserve the attachment information that is explicit in the deeper structure of a syntactic parse.

therefore a token in the context free grammar of clauses is not necessarily problematic, even if it is superfluous. The deeper problem is that such an addition would make grammatical clauses different for text and speech. In speech, the only "punctuation" is the phrase final intonation. Because there is no compelling reason to propose different syntaxes for spoken and written text, the punctuation and layout tokens are (in the main) placed outside the clause in the LOQ mark-up.

The punctuation and layout tokens are themselves annotated to approximate the reader's correct interpretation of each token. For example, the annotations distinguish among commas that separate items in a series, delimit an intervening phrase or clause (often, a relative, adverbial or appositive clause), or separate a phrase initial or phrase final adverbial clause from the main grammatical clause.[7] Currently, this information is used for very little except as grist for the matching operation. For instance, two commas will always match based on orthographic identity, but will match even more closely if their functional annotations are also identical. Table 7.1.4 summarizes the annotations and the kinds of punctuation to which each applies.

Because clause-final punctuation is outside the clause in the mark-up, the functional annotations for punctuation and layout tokens are duplicated for the clause or phrase that precedes them. For example, if a full grammatical clause is followed by a period, and the period ends a paragraph, then both the period and the clause (tokens) receive

[7]Adverbial clauses in these positions fall into neither of the previous categories, hence they have their own functional designation.

the end annotation. The annotation to the clause expresses the fact that readers scan (and plan) ahead, especially for phrasing decisions. This is one of the few explicit provisions that LOQ makes for lookahead.

Annotation	Applied to	Function
series	Comma; semi-colon.	Separates items in a series of like items, e.g., *1, 2, 3.*
push	Left double quote; left single quote.	Starts quoted text.
	Left parenthesis; comma; colon; dash.	Marks the start of an interjection, interruption or intervening relative clause.
pop	Right double quote; right single quote.	Marks the end of quoted text.
	Right parenthesis; comma; dash.	Marks the end of an interjection, interruption, or intervening relative clause (including an intervening adverbial).
break (supersedes *pop* if both apply)	All end-of-sentence punctuation: period, question mark, exclamation point; line break token (e.g., in poetry and other formatted texts).	Marks the end of a clause or series of grammatical clauses (e.g., a sentence).
end (supersedes *pop* or *break*)	All end-of-paragraph punctuation: period, question mark, exclamation point; paragraph token.	Marks the end of a paragraph final sentence
more	Comma.	Separates adjuncts and modifier clauses that do not meet any of the other conditions, as in, *"Finally, he went home."*, and *"He went home, finally."*

Table 7.1: LOQ annotations that denote the function of punctuation and layout tokens.

7.1.5 Semantic annotations

LOQ obtains its knowledge of semantic relations from WordNet [MBF+90], an on-line database of thesaural relations. The automatic methods for retrieving the most

appropriate meaning of a word or phrase from WordNet are often limited to nouns and noun phrases (e.g., [Res95, RA95]), and in addition, are not yet 100% accurate. Therefore, the most appropriate word or phrase *sense*[8] is currently assigned by hand. If the word or phrase has no WordNet entry, then a synonym that does is assigned instead. LOQ uses the word sense as the starting point for automatic semantic calculations using WordNet.

7.1.6 Other annotations

The remaining annotations deal with special cases. One occurs when a retrieval cue cannot be automatically computed; another deals with special forms such as compounds, contractions and interrupted phrases and clauses.

Retrieval cues

LOQ uses both the word and its uninflected form as retrieval cues, both for operations in AWM′ and for accessing linguistic databases. For the most part, LOQ computes retrieval cues for phrases and clauses using a simple algorithm: the verb is the main retrieval cue for the clause; the head noun is the retrieval cue for the subject, object or indirect object (which are always noun phrases); the head noun is also the retrieval cue for a prepositional phrase (thus treating the prepositional phrase as an inflected noun phrase).

When this algorithm fails, the appropriate cue is entered by hand. For example, in the National Public Radio text (Appendix B.2), the subject clause, *"the J.N.C."*, refers to the Judicial Nominating Committee. Current text analysis methods would have a hard time deducing this. Therefore, the subject clause is manually annotated with the retrieval cue, *"committee"*:

(25) [$_S$ (*committee*) the J.N.C.].

Compound structures

The main compound structures in English are contractions and multi-word forms. Contractions are problematic for the grammatical clause representation when they combine more than one grammatical role in one word. LOQ represents such constructions by clauses that contain either a null subject or a null verb, depending on which

[8]**word sense:** When a word (or phrase) has more than one meaning, each meaning is a different sense. For example, each numbered entry in a dictionary definition represents a different sense.

one is contracted. For example, the full grammatical clause that starts with *"'Twas"* contains a null subject, because the subject, *"it"*, is reduced:

(26) $[_{Cl}$ $[_{S}$ $tr_1]$ $[_{V}$ 'Twas$_1]$ $[_{C}$ Brillig]].

Conversely, the clause that starts with *"It's"* contains a null verb because *"is"* is reduced.

(27) $[_{Cl}$ $[_{S}$ It's$_1]$ $[_{V}$ $tr_1]$ $[_{C}$ brillig]].

The multi-word form may be a verb and its participle, or a sequence of words that together take on a new meaning. When a verb and its participle are not contiguous, the constituent that appears latest in the phrase is treated as a trace earlier. In addition, the retrieval cue for the verb and its participle is provided by hand, as in:

(28) $[_{V}$ (*open up*) open $tr_1]$ $[_{O}$ them] $[_{Post}$ up$_1$ wide]

The retrieval cue for a compound noun is the head noun. LOQ computes this automatically. However, when an adverbial phrase is a compound, it is hand-annotated with the compound as the retrieval cue, for example:

(29) $[_{AdvP}$ *(all over)* all over].

Interrupted constituents Sometimes, the construction of one clause is interrupted by the construction of another, as in:

(30) The dinner you cooked was delicious.

The LOQ representation for the interrupted clause includes an unlexicalized verb clause that makes forward reference, as in:

(31) $[_{Cl}$ $[_{S}$ The dinner$_1]$ $[[_{V}$ $e_2]]$.

This conveys the speaker's expectation that the delayed clause will at some point be constructed and the delayed reference resolved. The clause that contains the realized information will also contain a non-lexicalized subject that co-indexes with the subject in the initial clause, as in:

(32) $[_{Cl}$ $[_{S}$ $e_1]$ $[_{V}$ was$_2]$ $[_{C}$ delicious]].

7.1.7 Discussion

The text that LOQ processes is organized into grammatical clauses rather than full sentential syntactic parses. The grammatical parse trees it produces are shallow, narrow and generally non-recursive. This is consistent with the limited representational capacities of working memory. In addition, the trees are easily mapped to a linear sequence of input tokens, which represents the reader's serial recognition of words and the larger syntactical and grammatical structures to which they belong.

Currently, no software or algorithm is sophisticated enough to accurately identify full grammatical clauses or their subject, verb, object, (etc.) constituents. Were such tools available, they would apply to text comprehension for any mode of speech (read or spontaneous) or text (written or spoken) since theories of textual relatedness ([HH76, Sid79]) depend on the grammatical role as well as syntactic category.

7.2 Automatic analyses

If this dissertation were completed ten years from now, the reportage in this section would be voluminous, and the previous section would be correspondingly much smaller. At present, the most reliable natural language processing tools are the word stemmers and part of speech taggers. Therefore, they provide the bulk of the automated text analyses for the unmarked text. LOQ performs some automated mark-up after incorporating information from all the other analyses. The analyses from both sources are described in this section.

7.2.1 Parts of speech, word stems and noun phrases

The English Constraint Grammar (ENGCG) software developed by Voutilainen, *et al.* [Vou92, VHA88, VJ95]) provides word stemming, part of speech tagging and noun phrase identification. The word stem is the uninflected root of the word, and is used as the main retrieval cue in memory comparisons and database access. The part of speech tags are used in feature comparisons during the matching process. In addition to providing instantaneous information, the tags also carry expectations about upcoming syntactic structures (e.g., a preposition carries the expectation of a noun phrase) or the resolution of such expectations (the appearance of a noun phrase, as predicted by the preposition).

The ENGCG noun phrase extractor (*NPtool*) locates noun phrases, and always chooses the largest (e.g., *"hound dog"* instead of *"hound"* and *"dog"*). Therefore, it identifies possible subjects, objects, compliments and even prepositional phrases and helps to shorten the process of hand-tagging the grammatical clauses.

The ENGCG software was chosen because of its many capabilities and in addition, because its part of speech tags are compositional. Instead of the eighty-nine or so unique categories employed by the main alternative, the Xerox Parc tagger [CKPS91], it uses thirteen major syntactic categories[9] and tags approximately one-hundred and forty features. The feature tags often apply to more than one syntactic type. For example, the *plural* feature applies to nouns, pronouns and verbs. For the purposes of matching items in memory, this decomposition is helpful because it allows feature-based comparisons across syntactic categories. The advantage this provides will become clearer in Chapter 8, where the matching procedure is described.

Although the text tagging is automatic, it is not always accurate. Therefore, a human reviewer must check for errors, superfluous tags and idiosyncratic representations. The errors often arise from lookahead that is too limited. For example, *"chairs"* is incorrectly marked as a noun in: *"Attorney Haskell Kassler chairs the judicial nominating committee.* To obtain the correct tags, the sentence is re-submitted with the syntactically unambiguous *"governs"* as the main verb, and the results are then incorporated by hand into the tag file.

In general, ENGCG deals with garden path parses by providing more than one tag set. This would appear to be helpful for modeling cognitive operations, since multiple possibilities indicate points at which garden path interpretations might occur. However, because it is not a focus of this work, the incorrect tags are currently unused (not even to note a critical processing juncture). In other words, LOQ presumes a reader with sufficient lookahead to produce the correct parses.

One problem, mainly of standardization, is that ENGCG occasionally makes one word from two, especially if one of the words is a preposition. For example, it tags *"out of"* as one prepositional word. Because the linguistic databases are based on the orthographic word, the words that ENGCG has joined are separated by hand and each is annotated for membership in a compound. Because the ENGCG software already tags words that are members of a compound, undoing the compounds it creates is straightforward.

7.2.2 Automated analysis in LOQ

Using the grammatical and syntactic information provided by the manual and natural language processing techniques, LOQ performs additional computations that assign syntactic category to each grammatical clause, determine the main retrieval cues for phrases and clauses, and derive the inverse subcategorization and cohesive relations from the explicit annotations.

[9]These are: adjective, abbreviation, adverb, conjunction, determiner, infinitive marker (usually *"to"*), interjection, noun, negative-particle (*"not"*, *"n't"*), number, preposition, pronoun, verb.

Syntactic category

The algorithm for deriving the syntactic category of a grammatical clause is straightforward. Subjects, objects and indirect objects are automatically tagged as noun phrases. Verb clauses are automatically tagged as verb phrases. The syntactic phrase category for a `Pre` or `Post` clause is the category of the phrase it contains, if there is only one, and if the clause is not a prepositional phrase that functions as an adverbial. Otherwise its syntactic category is provided by hand.

Retrieval cues

The retrieval cue annotations are used in comparisons during the search for a match to the stimulus. The cues denote the main concept of a phrase via the head constituent of its syntactic type. Thus, the head noun is the retrieval cue for a noun phrase,[10] and the main verb is the retrieval cue for a verb clause. Both the head noun and main verb are retrieval cues for a full clause, although the verb is treated as the main cue.

Retrieval cues are provided by hand when the default algorithm fails. This is always the case for referring expressions. Therefore, each is annotated with the retrieval cue that selects its referent. For example, *"Remedios"* is the referent and retrieval cue for *"her"*, in:

(33) $[_{NP}$ $[_{GenNP}$ *(refers-to "Remedios")* `her]` $[_{AdjP}$ `unscheduled] meals]].`

It is also the referent and retrieval cue for the deleted subject in:

(34) $[_{Cl}$ $[_{S}$ *e (refers-to "Remedios")*`]` $[_{V}$ $[_{VP}$ `wandering]]].`

Co-reference relations are resolved during the search for a match to the stimulus.

Relation

LOQ explicitly annotates relations that are unambiguous in the structure. For example, an object or indirect object is always the argument of a verb (and not a modifier) and an adjective always modifies the head noun of the noun phrase that contains

[10]This noun phrase may be the grammatical clause itself, or it may be contained within a grammatical clause.

them both. LOQ also uses the manual annotations for subcategorization and cohesion, which are one-way relations, to calculate the inverse relation. For example, if X is annotated as a modifier of Y, LOQ annotates Y to indicate that it is modified by X. This occurs for argument, modifier and anchoring relations.

7.3 Online linguistic databases

The databases used by LOQ provide some of the knowledge that a speaker must have about her language. Currently, LOQ makes use of pronunciation, word familiarity and semantic databases. The word pronunciations[11] are available for all the words found in the Brown Corpus [KF67], a corpus that has been a standard test suite for statistical analyses of text. The information on word familiarity uses the word frequency counts from the Brown Corpus, as provided in the online Oxford Psycholinguistic Database [Qui92]. Semantic relations and verb subcategorization information comes from the WordNet semantic database [MBF+90, MF91], an online compendium of thesaural relations among words and phrases.

The use of databases as a knowledge source for speech and language is a viable technique because of the development of large online acoustical and psycholinguistic (e.g., WordNet, the Oxford Psycholinguistic Database) and common-sense (e.g., CYC) databases. The use of general purpose databases is an appropriate strategy because humans are not simply experts in one area but possess wide-ranging general knowledge. This approach differs from that used by concept-to-speech systems, whose knowledge is specific to the application and often encoded by hand.

7.3.1 Pronunciation

The pronunciation database provides pronunciations for all the words in the Brown Corpus. Each entry identifies the syllables, their phonemes and their stress markings – primary, secondary and unstressed. These features are used in acoustical comparisons among words and phrases. One problem with this database is that it lacks pronunciations for both inflected forms and proper names. This may be computed algorithmically (e.g., [Spi85, Dav89]). However, the missing information is currently added by hand to the database.

[11] Translated into ARPAbet notation by James Raymond Davis at the M.I.T. Media Lab.

7.3.2 Word frequencies

The word frequency information is part of the background knowledge that refines the predictions of the knowledge-free attentional model. It is used in LOQ calculations to balance newness to the discourse with prior expectedness for the language. The word frequencies (expressed as percentages) are the starting probabilities for the word. For example, because *"the"* has the highest word count and a relative frequency of 1, LOQ will not treat it as new information, even if it is the first word of a text.

The frequencies come from the Brown Corpus, five hundred text samples[12] taken from fifteen popular adult text genres and collected in the 1960s. Although it is small by today's standards, its strength is that it is a well-balanced corpus. In addition, it was the first online corpus. Therefore, it has been a standard test suite for many computer-based applications. Although its word frequency counts may not precisely reflect the most current trends, what is important to the LOQ computations is the overall order and not the exact counts.

Because the Brown Corpus data comes from written text, it is especially appropriate for LOQ, which simulates read text. However, because its sources are from many genres, its counts are not differentiated by part of speech or by word sense. It would be useful to have this information so that rare uses of a common word would be treated as novel. Another desirable extension is a database of collocation (co-occurrence) frequencies, since word likelihoods are often context-dependent.

7.3.3 Semantics

LOQ's grasp of English semantics comes from WordNet, an online semantic net developed at Princeton by George Miller and colleagues [MBF+90]. It contains thesaural relations for nouns, verbs, adjectives and adverbs and for both words and phrases. It is intended as a psycholinguistic database, that is, as the plausible contents of long term memory. Therefore, it is especially appropriate for simulations using a cognitive model.

It is organized around the *synset*, a set of synonyms. Each synset represents a different sense of a word or phrase. The whole database is a semantic net, organized from general to specific categories, for four parts of speech – the noun, the verb, the adjective and the adverb. It provides thesaural relations among synsets, such as antonymy, hyponymy (subset) and hypernymy (superset). It is particularly well-suited as an aid for extracting from text the relational propositions that pitch accents convey. For example, membership in a common WordNet parent category has a propositional counterpart in the pitch accents (L+H* and L*+H) that mark a lexical

[12]These samples contain 69971 unique words, and over one million words altogether.

item as an alternative selection from a salient set. Likewise, the part/whole relations assembled in WordNet have a counterpart in the pitch accents (H*+L and H+L*) that convey the existence of an inference path. Although LOQ does not directly map these relations to pitch accents, it does use them for comparison in the retrieval process.

As Resnik [Res95] points out, WordNet's coverage is dense for some concepts but sparse for others. Therefore, the number of nodes between two synsets is not a reliable indicator of semantic distance.[13] One useful extension to WordNet is the addition of function words which also have different senses and participate in semantic relations, such as synonymy and antonymy.

7.4 Summary

The components described in this chapter are combined in LOQ to approximate just some of the knowledge of language that most human speakers possess. The LOQ speaker has an idiosyncratic sampling: the semantic knowledge of WordNet; the word frequency expectations of a literate adult from the 1960s; the part of speech and word stem knowledge contributed by ENGCG; the syllable structure and pronunciation from the Brown Corpus; and (with human assistance) an understanding of grammatical clause and syntactic phrase and their structural and cohesive links. The grammatical clause is particularly important because it can be decomposed into a sequence of text and grammatical tokens. This approximates the reader's recognition, over time, of the structural and pragmatic links between constituents.

This information is richer than the linguistic information that is currently available to most text-to-speech systems. The biggest drawback is that the clause analysis is not automated. One alternative is to design a concept-to-speech system that provides the grammatical role classifications and links as part of generating the text. However, the more generalized solution for both speech synthesis and text comprehension is to automate the analyses that are currently performed by a human coder. The desiderata include software that would automatically identify grammatical roles and in addition, the attachment, subcategorization and cohesion relations among the clauses. On the semantic front, continued work on the automatic extraction of word and phrases senses from semantic databases ([Cha94, HPC97]) will eventually produce methods to replace the manual selection of retrieval cues.

[13] Agirre and Rigau [AR86] present a method that normalizes for the uneven density, and therefore provides a meaningful measure of semantic distance.

Chapter 8

Memory operations

In Chapter 6, I reviewed Landauer's model of attention and working memory, and motivated alterations and extensions to its operations. In Chapter 7, I described the linguistic knowledge that LOQ acquires from previously analyzed text. In this chapter, I focus on how the memory operations and linguistic knowledge combine during matching. I also describe the compression and expansion operations that distinguish AWM′ from the original model and discuss how time progresses for memory operations.

8.1 Matching: The main operation

Identifying a match is central to classification and retrieval tasks in many fields, from gene-sequencing in molecular biology [SOWH96] to object identification in computer vision [SP95]. In the simulations using AWM′, the matching process characterizes the SEARCH step, in which a text-based stimulus is compared to the contents of a region in memory, as defined by attentional limits.

For any pattern matching task, a positive result is variously termed recognition, recall or retrieval. In the AWM′ matching operation, the stimulus serves as a *cue* that *retrieves* an identical or associated item from working memory. However, the search for a match also reveals the converse relation, in which the salient (or activated [And83]) items in memory *prime* for a stimulus, possibly the current one. These items comprise the attentional context that classifies the current stimulus as expected or unexpected, given or new. The matching process measures the mutual information between the stimulus and the memory items to which it is compared.

To capture the many ways in which a textual or linguistic stimulus may serve as a retrieval cue, LOQ compares features of many types – pragmatic, semantic, grammatical, structural, syntactic, orthographic and acoustical. Because this work emphasizes

attention more strongly than the quality of the information that is attended to, a reasonable collection of match predicates is currently assembled by hand. The guiding questions for this phase are:

1. What are the typical retrieval cues for linguistic constituents?
2. What features of text and language are typically associated with prosody?

8.1.1 Retrieval cues for linguistic constituents

The research on speech errors provides the main body of evidence about how linguistic constituents are encoded in long term memory. It points to phonological encoding by stress and phoneme patterns [SH86], structural encoding by syntactic and grammatical category, and semantic encoding by thematic role, similarity and association [FC77]. The LOQ match criteria include predicates that compare features from these categories.

Retrieval in any category can be differentiated by the generic kinds of structural relations that can occur between the stimulus (or *cue*) and an item in memory (or *target*). LOQ makes three distinctions:

1. The identity relation, in which the stimulus retrieves an identical or similar copy, e.g., *"search"* retrieves *"search"* or *"searching"*. There is no indirection between cue and target for this relation.
2. The association relation, in which the stimulus retrieves an associated item e.g., *"search"* retrieves *"warrant"* [RM88]. This describes retrieval with one level of indirection between cue and target.
3. The compound relation, in which a combined cue retrieves a related item, e.g., *"search"* and *"warrant"* retrieve *"legal"*. This is also an association relation, but with a more complex inference step. There are two levels of indirection between cue and target. The current stimulus may itself be a combined cue [RM94] or alternatively, retrieval could proceed in two stages – the stimulus first retrieves a companion cue, and then their combination retrieves a third item.

These three distinctions describe mutual information in the formal sense: the identity relation describes the mutual information between the random variable X and itself (self-information); the association relation describes the mutual information between X and Y (the reduction of the uncertainty of X due to knowledge of Y); the compound relation describes the conditional mutual information between X and Y, given Z (the reduction of the uncertainty of X, due to knowledge of Y when Z is given) [CT91].

LOQ uses these distinctions to evaluate the certainty of the association between the stimulus and a memory item. Usually, LOQ treats the results of identity matches as

the most certain. However, LOQ also includes a separate "intensifier" category for identity predicates that are conclusive but insignificant. They are special in that LOQ only considers positive outcomes when there is a match for one of the other predicate types. For example, both *"govern"* and *"exist"* are verbs, but if that is all they have in common, LOQ will not register the match. However, if they share the same subject as an argument (as association match), the fact that they are also verbs increases the certainty that there is a match.

Discussion

Viewed in this framework, the match criterion in Landauer's simulations is a boolean *identity* predicate because the stimuli are either ones or zeroes. Walker's system also applies a boolean criterion, but because it includes inference procedures, it retrieves a logical proposition using identity, association and compound cues. In both Landauer's and Walker's simulations, retrieval occurs based on one feature – whether the proposition may be retrieved from the AWM space, or constructed its contents. There are no partial or gradient matches in these implementations.

8.1.2 Textual and acoustical correlates of prosody

Textual features that indicate salience and discourse structure are also correlated with prosody, which also conveys the same information. The main findings of research in this area (Chapter 3) are incorporated in LOQ as match predicates. For example, grammatical role and sentence position indicate focus status and are correlated with accenting and de-accenting; therefore LOQ tests for matches based on grammatical role and sentence position. Text layout and punctuation convey discourse structure and are correlated with phrase final contours; therefore, LOQ tests for matches based on discourse function. The repetition of intonational contours often correlates with the repetition of linguistic types, for example, items in a list are frequently recited with the same bitonal accent, albeit downstepped [LP84]; therefore, LOQ tests for matches based on parallel structure and analogous function.

Information status has also been correlated with prosody. However, because my claim is that information status is speaker based as well as text based, it is not explicitly included in the match criteria. Indeed, the goal of the simulations is to show how information status varies according to attentional capacity. Therefore, designations of given or new should emerge from processing in the memory model.

8.1.3 The predicate categories

The matching operation compares the current stimulus to items in AWM′ according to predicates that test for a match on the many features of language, text and speech. The collection and ordering of these predicates attempts to capture the many ways that a target may prime for a stimulus, and the many feature combinatorics that contribute to retrieval. It is somewhat the inverse of Yarowsky *et al.*'s [SHY92, Yar96, Yar94] decision list algorithm for identifying homographs. In their work, the test that yields the highest score determines the pronunciation of the word or phrase. It is a best match approach. For LOQ comparisons, the particular features that match are not as important as whether the cue and target match on any combination of criteria strong enough (in aggregate) to induce retrieval.

In all, LOQ applies seventy-four predicates. They are organized into twenty-four minor categories, which in turn are distributed across nine major categories:

1. **Co-reference**: Predicates in this category will report a match if the cue and target refer to the same entity. Notable examples are *anaphora* – in which pronominal reference follows the full syntactic specification, and *cataphora* – in which pronominal reference precedes full specification [HH76]. Empty category constituents may also co-refer.

2. **Acoustical and orthographic identity**: Predicates in this category test for homographs and homophones, as well as for similar or identical rhymes and rhythms. For example, *"loan"* and *"lone"* are exact matches on acoustical criteria but not on orthographic criteria.

3. **Morphological and syntactic identity**: Predicates in this category look for identity matches based on inflection or syntactic category.

4. **Structural parallelism**: Matches on parallel structure are matches of form. For example, the noun phrases *"the red barn"* and *"the large ocean"* have the same internal structure – a determiner followed by an adjective followed by a noun.

5. **Semantic identity**: Predicates in this category encompass synonyms, antonyms, hypernyms and hyponyms – that is, the sibling, parent and child relations in the (WordNet) semantic net.

6. **Semantic association**: Predicates in this category include attribute and part-whole relations, as well as the verb relations of cause and entailment.

7. **Collocation patterns within the text**: Collocations are co-occurrences of linguistic constituents (generally words). They are typically specific to a genre, or even to a single author. For example, *"periodic table"* is a likely collocation in a chemistry text, whereas *"table a motion"* is most likely in a text on parliamentary procedure. As AWM′ processes a text, collocations that are repeated

become distributed throughout the space and each part is stored near the other. Therefore, if the current stimulus belongs to one of these collocations, it is likely to retrieve a previous whole collocation (e.g., *"pitch"* retrieves *"pitch accent"*) or one of its elements (e.g., *"accent"* retrieves *"pitch"*). In this way, the association among tokens within a collocation increases as the text is processed.

8. **Collocation patterns within the language**: Thus far, no LOQ predicates test for word collocations that are general to English, mainly because an appropriate collocation database is not available. However, they do test for syntactic and grammatical patterns, such as the fact that a determiner and noun co-occur in a noun phrase, and a verb clause co-occurs with its case grammar complements in a full grammatical clause.

9. **Structural association**: The predicates in this category identify associations with the verb clause, such as argument, modifier and case grammar role. They also include the predicates that test for anchoring associations between noun phrase constituents.

8.1.4 Scoring the predicate outcomes

In the LOQ comparisons, the result of applying any single predicate is a boolean value – true or false. However, the quality of the test result depends on the test itself. For example, a match on co-reference is a more definitive indicator of identity than one on syntactic category. Therefore, the result of applying a predicate, which is boolean, is assigned a numerical value, which indicates its relative informativity, or conversely, the amount of mutual information it reveals between cue and target, relative to the other predicates. Each value is derived from the partial ordering of the matching criteria, according to these principles:

1. The more precisely a target predicts a feature of the stimulus, the higher the ranking of the predicate that tests for the feature. Mutual information is typically highest for identity predicates. The ranking of predicates usually follows this order:

   (35) *identity* > *association* > *compound* > *intensifier.*

2. Predicates that reveal local patterns within a text are ranked higher than those that are generic to English.

The ranking of the major categories reflects these principles, as shown in Table 8.1. So too, does the ranking of minor categories within a major category, and the ranking of predicates within a minor category. The LOQ predicate hierarchy and the relative rankings of its elements are presented in detail in Appendices A.1 through A.4.

The numerical value for each of the seventy-four predicates is the product of the weighted rankings of the predicate, its minor and major category. A weighted ranking

Category	Predicate types within the category
Reference	Identity
Acoustical and orthographic identity	Identity
Collocation patterns within the text	Association
Semantic identity	Identity
Semantic association	Association, Compound Cue
Collocation patterns within the language	Association
Structural association	Association, Compound Cue
Morphological and syntactic identity	Intensifier
Structural parallelism	Intensifier

Table 8.1: Rankings for the major predicate categories. Matches on co-reference criteria contribute the highest mutual information.

is calculated as follows: If n is the number of elements at any level of the predicate hierarchy, and x is the rank of an element at this level, such that $0 < x \leq n$, then R_{group}, the weighted ranking of x within its group, is:

(36)
$$R_{group} = \frac{x}{\sum_{i=1}^{n} i} .$$

Thus, if a group has three members, the weighted rankings it contributes for the first, second and third members are 1/2, 1/3 and 1/6, respectively.

The final value for each predicate, R_{final}, is the product of the local rankings at each level:

(37) $$R_{final} = R_{predicate} \times R_{minor} \times R_{major}.$$

With seventy-four predicates, twenty-four minor categories and nine major categories, the values produced by this method are small. Currently, the highest score is for co-reference, at .0444 – the product of $5/15 \times 3/6 \times 9/45$. This method produces an absolute ordering within a group and a partial ordering among all the predicate values. In addition, a high-scoring predicate in a lower-ranked category will often have a higher value than a low-scoring predicate in a category that is more highly ranked.

8.1.5 Discussion

The match predicates encode features of language and text that apply across genres.[1] The strongest claim about their breadth and their ordering is that both are plausible according to the background literature. However, when more precision is needed, a statistical analysis of empirical data is called for to select and rank the predicates. Depending on the application, the data may need to be genre-specific as well.

8.1.6 Combining the scores

To determine whether the stimulus has matched an item in the search radius, the values for the *successful* predicates are summed to produce the total score for the comparison. Scores above a pre-set *threshold* confirm a match. Note that once the score is calculated, it is no longer possible to tease out which of the predicates have contributed to the match. For example, one comparison might score above threshold because the cue and target fill the same syntactic and grammatical roles and for the same subject, as in, *"Tara sang."* and *"Tara laughed."* Another comparison might score above threshold because the cue and target rhyme, as in, *"delegation"* and *"jubilation"* and because they are both nouns.[2] What counts is whether the aggregate score exceeds the threshold. If it does, then there is a match between the stimulus and the memory item to which it is compared.

Adjusting for prior expectations

Currently, the AWM′ space is completely empty at the start of a simulation. Therefore, even ubiquitous words, such as *"the"*, will be designated as new information if they are the first item stored in an empty memory. To correct for this and thereby, to approximate the prior availability of ubiquitous information, LOQ adjusts the match score by the prior probability for the stimulus. So far, this is implemented only for words, because other kinds of frequency data (e.g., word collocations) are not available. The prior probability for a word is calculated from $freq$, its rank in the Brown Corpus, and the log of 69971, the maximum rank in the corpus (for *"the"*):

$$(38)\quad Pr(word) = \log_{69971} freq.$$

This value is added to the total match score for a comparison. Because the current match threshold is small, at .035, the values for the most ubiquitous words will always

[1] And possibly across languages as well.

[2] See [Lev89] (pp. 224) for an example relating to vision.

exceed it and induce a match. To make this adjustment schema less rigid, the prior probability is scaled by the inverse of the occupation density of the search region. This is a simple way of stating that with more items in the attentional focus, there is a greater uncertainty that prior expectations will hold.

For a low frequency word, the addition of its miniscule prior probability as calculated in (38) will not make the difference between matching and not matching. Whether there is a match remains an effect of its local probability for the text, as revealed via the match predicates.

Adjusting for relevance

Some comparisons between the stimulus and a memory item are simply irrelevant. For example, verb phrases do not participate in anchoring relations, and punctuation can not rhyme. Therefore, LOQ suppresses tests on irrelevant criteria. As a practical note, when data are sparsely distributed in the AWM′ space, a test may be triggered with minimal justification. However, there is a higher computational cost for more densely distributed data simply because more items and features will be compared. Therefore, to limit computation, the standards for applying a predicate may need to be more strict. Because the data tend to be sparsely distributed in the AWM′ space, most of the match predicates are applied for each cue–target comparison.

There are two other conditions under which testing is blocked. One is a consequence of the ordering within a major or minor category – if the highest-scoring predicate yields a positive response, no weaker test is applied. For example, if exact acoustical identity is recognized, the weaker rhyme and stress predicates need not be applied. On the other hand, sometimes a negative result for one feature mitigates against positive results for any others in the category. For example, if the stimulus and memory item do not co-refer, the more specific tests on the form[3] of the co-referential expressions are irrelevant.

8.1.7 Detecting recognition: The role of the match threshold

The comparison between the stimulus and any one memory item takes the results of many boolean tests and converts them into a score, which is expressed as a numerical value. This value is then converted back into a boolean to answer the simple question: has a match occurred? This is determined by the *match threshold* – the minimal score that a comparison must achieve to be considered a match, or conversely, to classify the stimulus as primed-for, and therefore, as given. A score greater than the threshold

[3]E.g., full noun phrase, pronoun, ellipsis.

has a cascading effect: if it exceeds the threshold, there is a match. And if there is a match, the search stops, even if the maximum search radius has not yet been reached. As shown in Figure 8-1, the main function of the match threshold is to stop the search.

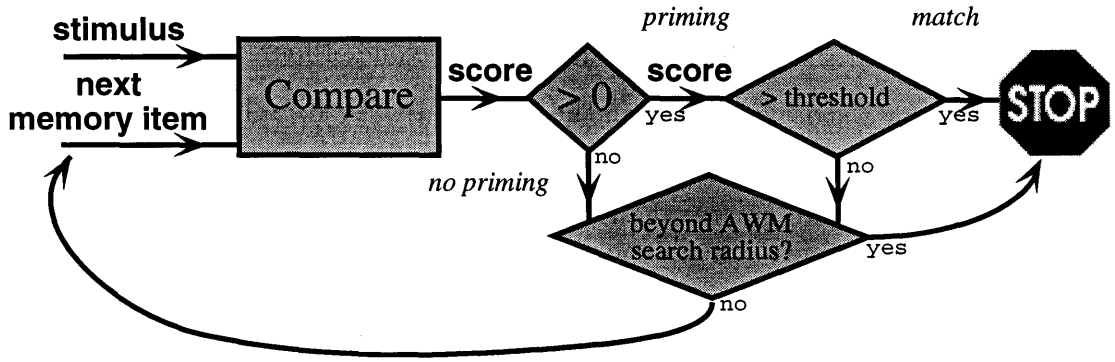

Figure 8-1: The matching algorithm for LOQ.

Currently, the threshold is fixed at .035, a value that permits only strong identity predicates to match in isolation. All other predicates must match collectively to bolster the determination that there has been a match.

Usually, the choice of matching strategy for any application is a variant of either a *first match* or *best match* strategy [Mac87]. A first match strategy stops when the minimum criteria have been met; a best match strategy seeks the best fit rather than one that is merely adequate, and pretty much guarantees an exhaustive search every time. Overall, the choice of strategy depends on the application. A first match strategy is consistent with a limited resource perspective. It implies that both time and computational power are limited. Therefore, LOQ uses this strategy. However, because matching is implemented using a numerical threshold, raising the threshold will prolong the search. Raising it high enough will eventually transform a first match into a best match strategy.

Discussion

The match threshold is necessary for detecting a match. However, because it is not yet clear what specific performance claims should be made for this parameter, it is fixed in LOQ to allow only strong identity matches to succeed independently. Although its current value appears to work for the three text genres used in the current simulations, the it may turn out to be genre or situation dependent. On the other hand, it may turn out that a fixed threshold is not desirable no matter what the value or the genre, and that it should instead be implemented as an adaptive parameter. The effect of varying the match threshold is a good starting point for investigations beyond this dissertation.

8.2 The results of a comparison

In the previous sections, I discussed how values are assigned to individual predicates and then how they are summed to produce one score for a comparison between the stimulus and a memory item. As may be deduced from these discussions, each attempt at individual comparison with the stimulus will have one of four outcomes:

1. Retrieval, for a score greater than the threshold – the stimulus and the memory item match.
2. Priming, for a score greater than zero but less than or equal to the threshold – there is some affinity but no match.
3. Inhibition, for a score of zero – the match predicates apply but none yield a positive outcome. Therefore, there is neither a match nor priming.
4. No response, because the items are so dissimilar that to test for a match would be vacuous and therefore, no matching process is triggered.

A match stops the search, and according to the interpretation developed so far, confirms the given status of the stimulus. However, we are also interested in the distribution of responses throughout the search region. For example, suppose that no item matches above the threshold but all the match scores are greater than zero (case 2). Even though no match is found, there is a uniformly positive response to the stimulus, which indicates some priming and therefore some degree of givenness. This differs from a failure to match because of uniformly negative responses (case 3), or because all comparisons are vacuous (case 4). It also differs from non-uniform distributions of responses that may occur for some or all of the four cases.

As illustrated previously in Figure 6-6, the mapping proposed so far only distinguishes between given (a match) and new (no match). This is insufficient for a speech application because it will only map to two outcomes – accenting new stimuli with H*, and de-accenting given stimuli. Therefore, I construct a mapping that reflects the different kinds and distributions of responses in the search region, and in addition, a mapping that will generate the remainder of the pitch accents.

It relies on an annotation to each memory item that summarizes the result of the most recent comparison in which the memory item participated. I will term this kind of annotation the "polarity" of the memory item because it indicates a binary distinction.[4] It works as follows:

1. An annotation of L denotes a score greater than zero, and therefore, a *positive* response to the stimulus (alternatively, activation). The polarity, being qualitative, does not record the strength of the response.[5] Instead, the difference

[4] And also to reserve the use of "annotation" for the text annotations described in Chapter 7.

[5] However, the distribution of responses in the search region provides a quantitative measure.

between a positive response and a match is that any positive response is recorded as L, no matter what its value, while a match is functional in the model – a score above the threshold stops the search.

2. An annotation of H denotes a score of zero, and therefore, a *negative* response to the stimulus (alternatively, inhibition).

3. When no comparison is triggered, the previous polarity will hold.[6]

The items in the search region comprise the context in which the current stimulus is evaluated. Their polarities reflect their priming in response to prior stimuli. If the current stimulus is sufficiently similar to what has been primed for, it will be treated as highly predictable (i.e., given) by the model. In this way, each stimulus affects the state of the items within the search radius, and conversely, the state of these items records the effect of the stimulus. This distinguishes between finding a match and the effects of searching for it, that is, between retrieval and priming. Moreover, it registers the magnitude of the priming via the aggregate response of the items in the search region.

What makes this new to any use of AWM is that the results of the matching process feed back into the model continually and in ways that are both categorical (polarity) and quantitative (the distribution of polarities in the search region). In Landauer's simulations, only storage affects the model. In Walker's, the results of retrieval affect the occupancy pattern because the retrieved item is re-stored close to the pointer, thereby depicting its re-established salience. This extension is consonant with the attentional claims of the model. However, when a match fails, there is no effect – only positive outcomes affect the model. In AWM′, all outcomes have an effect on the model's contents, although not on its pattern of occupancy.[7]

The consequence of recording the results of the matching process in this way is that prior history persists and, as they should, prior biases affect current processing. Moreover, the various configurations of activation or inhibition are the basis of the mapping that produces all the intonational types and continuous prominence as well, as will be described in the following chapter.

8.3 Compression: An augmentation to storage

Storage in AWM′ is limited to one item per node. Therefore, when a new item is stored, it must be stored at a node that is unoccupied. Without a means to free

[6] In the current implementation, the initial polarity is always null. Another option is to randomly assign an initial polarity of "L" or "H".

[7] However, compression is the one AWM′ operation (besides storage) that does affect occupancy patterns via the deletion of memory items.

up space on an incremental basis, the pointer must keep searching for an unoccupied node. Because its current search pattern is a random walk, there is nothing to prevent it from re-visiting the same occupied node several times. However, even with a less random search pattern, the eventual problem is that memory fills up.

As discussed in Section 6.6, external swap, deletion and compression are three options for re-claiming storage space. However, the current model does not include an external swap space, nor does it contain provisions for what to do when items are not recoverable (as is the case with deletion). Compression, however, is an operation that frees up space in memory while preserving information. Therefore, it is included as an option to STORAGE in AWM′.

The global option performs compression whenever the occupation density exceeds a threshold for the region or the entire space. This addresses the problem of space filling up, but not the problem of excessively long searches for an unoccupied node. Therefore, LOQ employs compression in an incremental fashion, as the consequence of the effect of local density on attempts at storage.

One obvious problem with incremental compression is that it will tend to compress items that are nearest to the pointer and therefore the items that are most likely to be salient. Although LOQ includes procedures for retrieving information from a compressed representation, the compression of items that are salient is counter-intuitive given the attentional claims of the memory model. Therefore, compression is limited to items that are likely to be less salient, and also likely to have been in memory longer, as determined by the following criteria:

1. **The stimuli for which the memory item primes have already been encountered and stored in memory.** For example, given the regularities of English syntax, a determiner primes for a noun, a noun primes for the completion of a noun phrase, and a subject clause primes for a verb clause. The condition of *waiting for* the completion of a sequence or hierarchy is boolean. It will be false if a memory item does not prime for any future stimuli, or if such stimuli have been encountered and stored.

2. **A larger structure to which the memory item belongs is also in memory.** This follows from the hierarchy within a clause. For example, a noun is deletable if its parent noun phrase is in memory; a noun phrase is deletable if its parent subject clause is in memory; a subject clause is deletable if its parent full grammatical clause is in memory.[8] In the current scheme, a full grammatical clause cannot be deleted unless it is an argument of a verb clause, as is the case for clauses that contain a cognitive verb. For example, the clause, *"it will work"*, can be deleted, since it is still retrievable as the argument to *"think"*, in *"They think it will work"*. Currently, punctuation and layout tokens are not compressible.

[8]This is possible because each memory item has, as part of its state, links to its parent and child tokens in the grammatical parse.

3. **One compression attempt has already been made.** This necessitates the addition of a binary state variable for each memory item. It is false initially, and becomes true after the first compression attempt has been made.

These restrictions on compression affect what happens when the pointer moves to an occupied location. There are two possibilities:

1. **If the stored item is compressible** – that is, if conditions (1), (2) and (3), above, are true – then it is deleted and the current stimulus is stored in its place.

2. **If the stored item is not compressible**, then it is marked as the object of a compression attempt. This fulfills condition (3) above. If it is again encountered by the pointer and conditions (1) and (2) hold as well, it will be compressed. In the meantime, the pointer continues its random walk in search of a location that is either unoccupied or contains a compressible item.

In actuality, compression is implemented as deletion from the AWM′ space. As such, the compression described here is only an analog of true data compression. However, it retains the essential (analogous) features which are that less space is used to store the same information, and information that is "compressed" is not irretrievably lost, but is instead recoverable from another representation, as will be described in the next section. The compression and expansion methods should be seen as a placeholder for methods such as those developed by Pollack [Pol90b] and Plate [Pla94], which represent linguistic structure and relations as bit strings and therefore, whose parts are accessible by data expansion techniques.

8.4 Retrieving a compressed item

The opposite of compression is expansion. The effect of expansion is first of all, to retrieve a part from a whole. In LOQ, this occurs via tree descent because all LOQ tokens belong to a hierarchy.[9] For the AWM′ model, a second effect of expansion is to re-store the matching part in the memory space, thus undoing the original deletion. However, this step is not implemented because the most likely heuristics for doing so have problems and the cognitive claims are not strong enough to identify a convincing option.

For instance, a plausible option is to implement Walker's [Wal93] method and store the retrieved item near the pointer, thus asserting its re-established salience. There is then no reason not to implement this in a general fashion (as Walker does) for

[9]The LOQ hierarchy is flat at the top. At this level it will contain the tokens for full grammatical clauses, punctuation and layout.

any item that is retrieved, regardless of whether the retrieval counts as expansion. However, space in AWM′ is most limited nearest to the pointer.[10] Moreover, because storage at an AWM′ node is limited to one item, the search for free space during re-storing will consume much of the model's processing time.

Another option is to store the retrieved item near the item from which it was retrieved. However, the principles that would guide this are unclear. (How near? for example.) In addition, this option requires a search for free space near the parent representation, rather than near the pointer.

Because the obvious solution is problematic for AWM′ and the cognitive claims are not compelling for either option, comparison occurs for a memory item that has been "compressed" within another one, but without inducing re-storage. The compressed item is located within its parent (or ancestor) structure using tree descent and the results of the comparison affect the polarity of the parent.

8.5 Special tokens

As much as possible, AWM′ processing and its outcomes are emergent from search and storage. In line with its attentional claims, the cognitive operations generally occur within the search region. The only exceptions occur at the end of a paragraph or verse, as indicated by the end-of-segment punctuation and layout tokens. The final punctuation triggers the *global resetting* of the polarities of all memory items, to L. The final layout token triggers *compression* for all the (compressible) items in memory. Both global reset and compression approximate the summary that occurs at the end of a topic and, via the mapping of density to pitch range, produce the pitch range expansion that occurs at the beginning of the next discourse segment [Lad88].

8.6 The state of a memory item

One result of introducing polarity and compression is that the items in memory begin to have dynamic state and moreover, a state that is not a reflection of their static linguistic features but of their participation in AWM′ operations. Altogether, AWM′ memory items have three kinds of state. The first is composed of the *static* linguistic information that is associated with the item – its semantics, its grammatical and syntactic type, its place within a clause structure, its subcategorization and cohesive links, its pronunciation, its orthography and its frequency of occurrence in English.

[10] For example, in a two dimensional Cartesian lattice, a maximum of four locations are available at a distance of one.

The second kind of state describes values and information that *change once* during a simulation. Some of these are included to increase the computational efficiency of the program. For example, WordNet information is accumulated only as needed. Other values of this type reflect the accrual of co-reference information as a result of the matching process. A third is represented by the two kinds of compression-related information: (1) whether the stimuli for which the item primes have been encountered; (2) whether one compression attempt has already been made.

The third kind of state is *dynamic*. It describes values that may change many times over the course of the item's tenure in working memory. So far, only two values have this property. One is the calculation of the predictability of a word,[11] in which its prior probability is adjusted by the number of items in the search region (as described in Section 8.1.6). This value is recalculated each time the word is involved in a comparison, and produces a continuous value.

The main example of dynamic state is the binary-valued polarity of a memory item in response to the stimulus. This value changes for each comparison in which the item is involved. As will be shown in the next chapter, it is used to determine both the intonational category and prominence of the current word stimulus.

8.7 Model time: Incrementing the clock

In Landauer's model, the clock is incremented each time the pointer chooses its next location. The random walk algorithm includes the current location as a possibility for the next randomly chosen location. Therefore, the clock is incremented for every location choice, rather than for every location change. The clock is also incremented during search. Landauer proposes that search time expands outward from the radius in constant time, one time step per integer distance from the pointer (up to the search radius maximum).

In AWM′, because storage has just occurred at the current location, it is not available as an option for the next location choice. Therefore, choosing a location is synonymous with pointer movement. Currently, each step taken by the pointer takes place in unit time. Compression is also assigned unit time, as is the act of recognizing the next unit of text – word, punctuation or layout – or constituent structure – a syntactic phrase, a grammatical clause or a full grammatical clause. These augmentations propose that there is a time cost to these operations. However, because the operations may in fact occur instantaneously, the time costs ascribed to pointer movement, reading and compression are conservatively set to unit time. Expansion is currently assumed to occur instantaneously – that is, retrieval occurs for any cue, regardless of whether

[11] So far, this is limited to words because frequency data is only available for words. Collocation frequencies would allow the entropy calculations to apply to other constituents.

it entails tree descent. This is not a strong assumption and awaits a better answer to the questions raised in Section 8.4.

One source of variable time in AWM′ is the search for free space – when space is tight, the pointer may move several times for each attempt at storage. The other source is the matching process, which constitutes the main example of high-level cognitive activity in the model. Landauer implements constant time for the search outward from the pointer. Each city block distance that is covered increments the clock by one. This scheme is modified in AWM′ so that the clock is incremented only if a comparison occurs at the current search distance.

8.8 Summary

In this section, I have described the main memory operations and their effects on the model. In general, the matching process (by which SEARCH is conducted) applies mode and domain specific criteria to the comparison between a stimulus and a memory item. Because the LOQ domain is language and speech, its match criteria are fine-grained. However, they are both genre and speaker independent criteria.

The results of the matching process become part of the state of the items to which the current stimulus is compared. This develops a notion of state for the contents of the memory model, such that the state of a memory item reflects attentional data as well as linguistic knowledge. It is a notion of dynamic state that is mapped to prosody, which is also dynamic.

Compression is another activity in the model that is mandated by the restriction of storage to one item per node. It is currently implemented as deletion and as a place-holder for bit string representations of linguistic information such as those developed by Plate [Pla94] and Pollack [Pol90a] which implement true data compression and expansion. Currently, expansion – the retrieval of a component from a whole – is implemented as tree descent.

All memory operations take time. Expansion is currently treated as a kind of retrieval and so is already accounted for in the search that expands outward from the pointer. Pointer movement, compression and the recognition of structural and orthographic tokens are fixed at unit time. However, searching for free space takes variable time and as does searching for a match, once free space is found.

The effect of storage and retrieval in the memory model determines prosody. In the next section, I explain how memory operations and their outcomes are mapped to pitch and timing in speech.

Chapter 9

Output: Prosodic correlates of search and storage

In this chapter I develop the mapping from processing in AWM′ to the key components of prosody: (1) pitch accent type and prominence; (2) phrase tone type and prominence; (3) pitch range; (4) the duration of words; (5) the location and duration of pauses. My position is that a mapping exists. My claims for the accuracy of the mapping I present are less emphatic. This is in part because the data that links limited attention and working memory to prosody is mainly available for timing. The mapping I employ is rooted in the performance data when possible. Otherwise, it aims for computational consistency with the more supported claims. In this chapter I explain how it works and develop the reasoning behind each component.

9.1 Pitch accents

I have described how AWM′ determines the given or new status of a stimulus by whether or not it matches an item in the search region. The intonational expression of these results is either de-accenting for given stimuli, and the H* accent for new stimuli. However, this algorithm will not produce the other five accents identified by Pierrehumbert *et al.* [BP86, PH90]. Therefore, obtaining them requires an alternative approach, one that can derive seven categories from what appear to be only two.

The mapping I propose distinguishes between accenting and de-accenting based on prominence, and determines the pitch accent type based on the distribution of responses to the current stimulus. In this section, I describe the derivation of pitch accent type and the calculation of pitch accent prominence.

9.1.1 Pitch accent type

The mapping treats the H or L tones of a pitch accent as representing the dominant outcome for all the comparisons between the stimulus and the items in the search region. The result of each individual comparison is denoted either as L for priming or H for inhibition. If most of the comparisons yield L polarities, the tone is L and the L* accent is assigned *to the stimulus.* Since bias is unavoidable in determinations based on inequality, the calculation of the dominant tone currently has a H bias:

$$(39)\ tone = \begin{cases} H & \text{if } \sum H \geq \sum L, \\ L & \text{if } \sum L > \sum H. \end{cases}$$

This demonstrates one of the uses of the polarity annotations described in the last chapter. However, so far, this derivation produces only the simple accents.

To produce bitonals, it is extended to consider the polarities both *before* and *after* the comparisons between the stimulus and the items in the search region. The polarities before the comparisons describes the *context* in which the comparisons occur. The configuration after the comparisons shows the *effect* of the stimulus on the region. The defining tone for each stage is calculated as described in (39). The denotation of BEFORE+AFTER (or CONTEXT+EFFECT) derives all the pitch accent forms. For example, if the defining tone both before and after the comparisons is L, the denotation is L+L, which is interpreted as the simple pitch accent, L*. However, if the defining tones differ, the result is a bitonal. For example, if L dominance changes to H dominance, its representation is the bitonal form, L+H. The derivation of the simple accents is shown in Table 9.1, and of the bitonal forms in Table 9.2.

The interpretation of L+L is, roughly, that a familiar item was both expected and also provided. Likewise, the interpretation for L+H is that a familiar item was expected but a new (or unfamiliar) one was provided and for H+L, that a new item was expected and a familiar one provided.

	context		effect	context		effect
tone counts	$\sum L > \sum H$	$\Rightarrow$	$\sum L > \sum H$	$\sum H \geq \sum L$	$\Rightarrow$	$\sum H \geq \sum L$
denotation	L	+	L	H	+	H
accent	**L***			**H***		

Table 9.1: Derivation of the simple tones.

To complete the bitonal derivation, LOQ treats the location of the main tone as a categorical reflection of *how much change* the stimulus has induced in the polarities of the items in the search region. If it has induced very little, then the influences of

	context		effect	context		effect
tone counts	$\sum L > \sum H$	$\Rightarrow$	$\sum H \geq \sum L$	$\sum H \geq \sum L$	$\Rightarrow$	$\sum L > \sum H$
denotation	L	+	H	H	+	L
accent	**L+H**			**H+L**		

Table 9.2: Derivation of the bitonal forms.

previous stimuli are still represented in the model and the main tone is the first tone. However, if it has induced many changes, overriding previous effects, the main tone is the second tone. The mapping is currently biased towards accents whose main tone is the second tone, as follows:

$$\text{(40) } bitonal = \begin{cases} T_1 * + T_2 & \text{if polarity retention} > \text{polarity change,} \\ T_1 + T_2 * & \text{if polarity change} \geq \text{polarity retention.} \end{cases}$$

Figure 9-1 shows how the location of the main tone reflects the (relative) number of items whose polarities were changed by the current stimulus.

This schema produces the six pitch accents identified by Pierrehumbert *et al.* However, it also produces some accents that are currently not part of the taxonomy (and perhaps need not be), namely, L+L*, L*+L, H*+H and H+H*. These are all treated as simple pitch accents, for which the location of the main tone (i.e., the magnitude of the effect of the stimulus) is currently irrelevant.

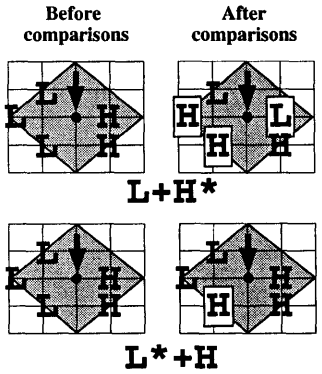

Figure 9-1: Bitonal mapping.

9.1.2 Pitch accent prominence

In the mapping, comparisons with any type of stimulus affect the items in the search region, and therefore, any stimulus can receive a pitch accent, even the most ubiquitous of function words. In effect, five pitch accents are gained but de-accenting is lost.

To restore it in the context of the proposed mapping, the prominence value is used as a *threshold* to distinguish accented from unaccented words. Only if the prominence of a word exceeds this value is the word accented.[1] If it is accented, its height or depth are scaled by its prominence. In general, prominence serves as an adverbial modifier to the pitch accent "adjective". For example, the H* accent marks new information, while its prominence indicates just how new it might be.

The motivation for the prominence mapping comes mainly from timing studies, which link shorter word and pause durations to greatest salience (e.g., [FLB97, Bre95, FH87, GL83]). In AWM′, the shortest retrieval times occur when a match is found close to the pointer. Therefore, prominence is tied to retrieval distance so that prominence below the accenting threshold is correlated with the most salient items and produces de-accenting. As in the empirical studies of human prosody, words representing the most salient concepts will be unaccented and will also have the shortest durations.

In the calculation, the *distance* at which a match is found, $d_{retrieval}$, is scaled by the inverse of the search radius, d_{max}. The result is then adjusted by $1 - Pr(word)$, where $Pr(word)$ is the probability of the word, as calculated in (38). For example, because *"the"* has a prior probability of 1 in English, the prominence of its first occurrence is scaled by zero, effectively nullifying any pitch accent that may have been assigned it. In addition, to allow local probabilities to eventually overwhelm prior ones, the prominence increases with n, the number of items to which the stimulus is compared. This number tends to increase with the number of items in memory as a whole.

LOQ calculates pitch accent prominence as follows:

(41)

$$n \times \frac{d_{retrieval}}{d_{max}} \times (1 - Pr(word)).$$

9.2 Phrase tones

The current mapping for phrase tones reflects the model's response to a phrase final item. In LOQ, these are mainly the non-word tokens – the syntactic phrase and the grammatical clauses. If the word is followed by at least two such tokens, the response to the first token after the word becomes a phrase tone and the response to the final one becomes the boundary tone. The boundary tone mapping is easiest to justify – it encapsulates the results of attempting to match the last token of the phrase. The phrase accent mapping is more *ad hoc* because intonational theory claims that it applies to the region of the pitch contour that follows the nuclear accent and stops just short of the boundary tone. Therefore, an average response over all the tokens between the nuclear accent and phrase final boundary tone is just as plausible as the

[1] Currently, the prominence threshold is set low, to .05.

current proposal and bears further investigation.

9.2.1 Tone

L polarities indicate that familiar items have been encountered. As a (coherent) text is processed by LOQ, the L polarities should increase. Indeed, at the end of a topic, there is no text left to be processed and the search for referents is resolved *de facto.* Extrapolating from these trends, LOQ takes an increase in L polarities to indicate that a reduction has occurred in the amount of processing that is left; and therefore, takes an increase in H polarities to indicate that a reduction has not occurred, and in fact, that further processing is anticipated. This functional interpretation is consonant with the relations among intonational and intermediate phrases that Pierrehumbert and Hirschberg [PH90] ascribe to discourse structure.

It allows a very simple mapping to phrase tone type and prominence. The tone for the phrase accent or boundary tone is defined by the dominant polarity in the search region after the comparison to the associated stimulus. The sampling is for the whole region rather than only for the items to which the current stimulus has been compared. The purpose is to include the influence of all the currently salient items. Unlike the pitch accent calculation, this determination is slightly biased toward L tones:

$$(42)\ tone = \begin{cases} H & \text{if } \sum H > \sum L, \\ L & \text{if } \sum L \geq \sum H. \end{cases}$$

Because phrase tones are interpreted as expectations about future processing, LOQ provides an alternative means of computing their properties based on the compressibility status of an item in memory. A predominance of items that are part of an ongoing structure (and therefore, incompressible) is mapped to a H tone, and a predominance of items that are part of completed structures is mapped to a L tone, as is the case at the end of a topic. The current schema is only preferred because it calculates all intonational features from one property, polarity.

9.2.2 Phrase tone prominence

For a L phrase tone, prominence answers the question, how low? For H, how high? Phrase tone prominence is calculated as the ratio of the count of the defining tone to the total count for the search region:

(43)

$$prominence = \begin{cases} \frac{L}{total} & \text{for a L phrase tone,} \\ \frac{H}{total} & \text{for a H phrase tone.} \end{cases}$$

At the end of a discourse or discourse segment, the polarities of all the items in memory will be L. Therefore, the final boundary tone will be a L tone uttered with the maximum prominence. This will tend to produce a steep final fall, depending, of course, on the prominence and type of the preceding phrase accent.

Unlike the pitch accents, phrase tone prominence is not directly subject to a threshold. However, it is indirectly subject to the prominence threshold for accents because no phrase final intonation is realized until a pitch accent has been assigned. This condition is a filter on the mapping of AWM′ activity to prosody. It ensures that the intonation conforms to the syntax of an intonational phrase: at least one pitch accent, followed by a phrase accent and then a boundary tone.

9.3 Word duration

In AWM′, salience is spatial – the most salient items are closest to the pointer. Because search time accrues monotonically[2] with distance from the pointer, the retrieval times for the most salient items are also the shortest. Finding a match at the very edge of the search region will take longer than finding one closer to the pointer. Failing to find a match within the search region will also take the maximum amount of time, since the entire region is traversed. This directly expresses the results of the timing studies that link longer durations for words and pauses to reduced accessibility, and shorter durations to greater accessibility.

The calculation of word duration starts with the time spent searching for a match, as measured by ticks of the AWM′ clock. Because time accrues for other operations as well, the total processing time for the word token is the sum of all the processing times associated with it – the time it takes to read it, store it and conduct the matching operation. As described in Section 8.7, expansion is instantaneous, and reading, pointer movement and compression are assigned unit time. Variable times occur for the search for an unoccupied node (which may include several attempts at compression and therefore several moves) and for the search for a match to the stimulus.

Storage time is the sum of all the clock ticks associated with the act of reading (recognition), all pointer moves and all the compression operations that occur prior to actual storage. Because unit time is ascribed to reading, pointer movement and compression, the total storage time for any one token is:

$$(44)\quad Store = Read + \sum Compress + \sum Move.$$

[2]In Landauer's model, it accrues linearly with distance.

Although Landauer proposes that searching expands outward from the pointer in constant time, the calculation of search time is modified in AWM′ so that time time accrues for each successive distance from the pointer *at which a comparison is performed* (see Section 8.7). The maximum search time is never greater than the search radius for the simulation:

$$(45)\ 0 < Search \leq SearchRadius.$$

The total time it takes to process any token is:

$$(46)\ TokenProcessingTime = Store + Search.$$

The total duration of a word also includes the time it takes to process the syntactic and grammatical constituents to which it belongs – the syntactic phrase, the grammatical role clause, and possibly the full grammatical clause as well. Currently, word duration also includes the processing time for a preceding empty category if it is not realized as a pause.[3] Thus, the total duration for the word is the sum of the processing times for all the tokens associated with it:

(47)

$$WordTotal = EmptyCategory + Word + Phrase + GrammaticalRoleClause + Clause.$$

By this accounting, if a word is the last word in a phrase, its duration will include the processing time for all the phrase, clause and punctuation tokens that directly follow it. All other things being equal, a word in the middle of a phrase will be spoken at a faster rate than one that is simultaneously the last word of a syntactic phrase, a grammatical role clause and a full grammatical clause. This is how LOQ produces the phrase final lengthening noted by Klatt [Kla75] and others.

9.3.1 Adjustments

As is, this algorithm will produce the longest search times and therefore the longest word durations for the largest search radii. In practice, this produces a linear distribution of word durations, such that the smallest radii produce the fastest speech, and the largest produce the slowest. Restricting the upper bounds of the speech rate range speeds up the slow speech so that it is no longer excruciatingly slow. However,

[3]This is an artifact of the current clause representation, as will be discussed in Section 9.4. It should probably accrue to the previous word or punctuation.

the more vexing problem is that there is very little durational range for any one radius. Therefore, to achieve the varied duration that is found in human speech, the durations are adjusted (inversely) for the search radius as follows:[4]

(48)

$$AdjustedTotalProcessingTime = \frac{TotalProcessingTime}{min(\log_2 SearchRadius \ \ .99)} .$$

The curves it produces are shown in Figure 9-2. Thus, a search of the entire region, regardless of its actual size, is eventually mapped to the longest possible duration of the word. This tends to produce the slowest speech for the smallest radii. It also enforces the result that no matter what the radius size, newest words will take the longest to utter. Its further advantage is that, unlike a strictly linear scaling, it allows for a variety of word durations for any radius.

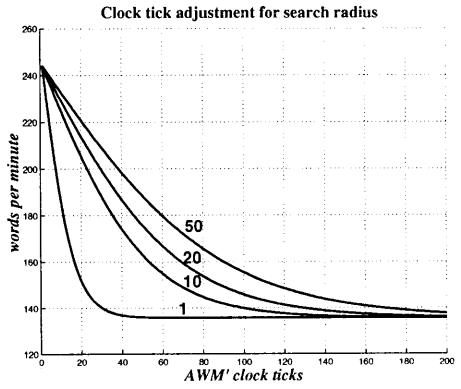

Figure 9-2: Inverse log scaling of duration for search radius. The shortest durations are mapped to the slowest speech for the smallest radii. Shown for the radii of 1, 10, 20 and 50.

9.4 Pause duration and location

As currently implemented, the duration of a pause directly reflects the duration of its processing by AWM′ – the more clock ticks, the longer the pause. To say that pauses have duration presupposes a means for selecting their locations. In LOQ, they occur for only two kinds of tokens – punctuation and the empty category. Pauses are mandatory for phrase and sentence-final punctuation. They are optional and highly restricted for the empty categories, mainly because the cognitive claims are not strong. PRO is currently not realized by a pause because, unlike the other empty categories, it does not mark a site that once held a lexicalized word. The remaining empty

[4]The minimum condition is included to prevent division by zero for the radius of one ($\log_2 1 = 0$).

category constituents are realized as a pause only if: (1) the calculations produce a H main tone for a deleted word, or a H phrase tone for other deleted constituent types; (2) its prominence is above a pre-set threshold, currently set to .2.

These restrictions are included to limit pausing mainly to punctuation, and otherwise to cases where the stimulus is new and significantly so. Another reason to restrict pausing for empty categories is because the current LOQ mark-up and mapping do not predict the correct pause location for the parasitic gaps when they follow a conjunctive. For example, the typical pause location is before *"and"* in:

(49) Winter$_1$ came [PAUSE] and e_1 went.

While LOQ does not aim for typical pausing only, its sole means for realizing a parasitic gap as a pause will only produce a pause after *"and"*, but not before:

(50) [$_{Cl}$ [$_S$ Winter$_1$] [$_V$ came]]
[$_{Cl}$ [$_{Br}$ and] [$_S$ e_1] [$_V$ went]].

This kind of pausing is mainly appropriate when *"and"* receives contrastive stress, as in *"They sell coffee and tea."* Therefore, the pausing restrictions should be viewed as a placeholder for a method founded on a theoretical treatment of variable pausing strategies in read speech and perhaps a reconsideration of how connectives are represented in the text mark-up.

Van Donzel and Koopmans–van Beinum [vDvB96] observe that pause strategies in Dutch spontaneous speech are speaker-specific – they may be filled, unfilled or instead occur as the increased duration of a word. Currently, LOQ produces unfilled pauses and lengthened words. In addition, as is appropriate for simulations of read text, LOQ generates only fluent pauses. However, cognitively based simulations of spontaneous speech will need to explain and generate both hesitation and filled pauses.

9.5 Pitch range

Pitch range is correlated with the hierarchy of segments within a discourse [HP86, HG92, Aye94, SSH96]. It tends to be widest at topic beginnings and then reduced over the course of an utterance [Pie80, Lad88]. It also has been found to be reduced for parentheticals [HG92, GH92].

The corresponding spatial correlates of topic structure are, first, that the AWM′ space is emptiest at the beginning of a reading[5] and also at the beginning of any

[5] AWM′ does not yet model the effects of items that may already be in working memory due to previous processing.

new paragraph due to the global compression that is triggered by the boundary tokens. Conversely, as the topic end approaches, memory occupancy is likely to be at its highest because the salient tokens are not yet compressible, and most of the (incompressible) full clauses are stored in memory.

Because they occur as interruptions, parentheticals are also associated with a more densely packed memory. At the point at which they occur, AWM′ contains the tokens for the interrupted clause, which are incompressible until the remainder of the clause is constructed. In addition, as the parenthetical is processed, it is adding tokens to the memory space. Many of these tokens can not be compressed until the entire parenthetical is processed. Therefore, the occupation density will be greatest for parentheticals and other interruptions.

These correlations motivate the mapping of density in AWM′ to pitch range in speech. In the mapping, the width of the pitch range varies inversely with the occupation density of the search region. Density is calculated as the ratio of the current to the maximum occupancy. Recall that in a periodic and Cartesian space, the maximum occupancy does not increase linearly with the search radius but instead follows an S-curve distribution, as previously shown in Figure 6-7.

Because a pitch range is delineated by a topline and a baseline, its reduction based on density raises the baseline:

$$\texttt{(51)}\quad Baseline = MinBaseline + (BaselineRange \times Density)$$

and lowers the topline:

(52)

$$Topline = \frac{1}{MaxTopline \times Density}.$$

The pre-set value ranges for the baseline and topline ensure that the baseline rises very little, while the topline falls more steeply.

9.6 Final filters: Speaker biases

The mapping algorithms produce abstractions of categorical and continuous features of prosody. However, the final forms and values depend on the syntax of the synthesizer commands. Because the TrueTalk synthesizer uses the ToBI annotation, the only translation required for the categorical designations is to change the H*+L accent into H* L – a H* pitch accent followed by a L phrase accent.

In contrast, the continuous features are expressed in AWM′ units and may need further translation into absolute numbers. For example, the mapping to pitch range

is a function of density. However, its final expression must be in Hz. This is a function of a pre-set baseline and topline, as shown in (51) and (52).

Because the TrueTalk synthesizer also treats prominence as a percentage of the pitch range within which the accent is realized, the LOQ prominence results may be used as is. However, because LOQ calculates, at most, a prominence of 1, even though the TrueTalk allows greater values, prominence is currently scaled to exceed 1. Thus, its lowest whose lowest value is always 0 but its highest value may be larger than 1.

Word and pause durations in AWM′ are counted in clock ticks. Therefore, it is easy to attach a unit of time (centiseconds, for example) to these values. Pause duration is assigned in this way. It is also straightforward to map the clock ticks to a speech rate. The TrueTalk synthesizer expresses speech rate as the inverse percentage of the default rate, which it defines as 190 words per minute. Therefore, the mapping is from AWM′ clock ticks to a range of percentages. The current values are fixed at .55 for the minimum (and quickest speech rate) and 1.4 for the maximum (and slowest) rate. 1, of course, is the norm. A filter[6] function is applied to suppress values above the pre-set maximum, as shown in Figure 9-2. Therefore,

(53) $$WordDuration = filter(ClockTicks \times (MaxRate - MinRate)) + MinRate.$$

This typically produces values in the range of .75 to 1.35, between about 253 words per minute and 141 words per minute, respectively.

The pre-set range and threshold values effect the final mapping of the AWM′ quantities. They portray influences that are currently outside the model, such as the transitory effects of emotion on physiology, or the more stable characteristic physiological and expressive biases of the speaker. They are currently fixed for all the simulations at the values shown in Table 9.3. However, they are implemented in the program as variables, to allow the future exploration of different settings.

9.7 Discussion

The aim of the mapping is, first of all, to be accurate or in the absence of conclusive data, to be plausible. For some features, this is an exercise in reverse engineering from acoustical features back to a property of the model. In the case of word duration, the straightforward mapping from duration in the model produces two unnatural features: very long word durations for the largest radii (longer than is typical even for slow speech), and very short durations for the smallest radii. Therefore, a filtering is imposed to allow slow speech for the smallest radii and fast speech for larger radii. This produces a range of durations for any one search radius.

[6] Alternatively, a sigmoid function could be applied.

Parameter	Value	Unit	Applies to
Maximum H* prominence	100	% of pitch range	Words: H*, H*+L, L+H*.
Maximum L* prominence	100	% of pitch range	Words: L*, L*+H, H+L*.
Minimum prominence	5	% of pitch range	All tokens, all intonation.
Maximum phrase tone prominence	100	% of pitch range	Phrases and clausal tokens: L, H, L%, H%.
Minimum prominence for silences	20	% of pitch range	Empty categories expressible as pauses (currently, deletions and traces).
Minimum baseline	75	Hz	Pitch range.
Default baseline	75	Hz	
Maximum baseline	125	Hz	
Minimum reference line	85	Hz	Pitch range.
Default reference line	96	Hz	
Maximum reference line	155	Hz	
Minimum topline F0	108	Hz	Pitch range.
Default topline F0	116	Hz	
Maximum topline F0	215	Hz	
Minimum word duration	.55	inverse proportion of speech rate	Word.
Default word duration	1.0	default speech rate	
Maximum word duration	1.4	inverse proportion of of speech rate	

Table 9.3: Fixed output parameters and ranges.

Table 9.4 summarizes the mappings between pitch and timing features and the properties and effects of storage and search in the model.

Prosodic feature	**Function of**	**Interpretation**
Pitch Accent	Context/effect ratio for the items in the search region that have been compared to the stimulus.	Expected *vs.* actual salience or familiarity of the stimulus.
Pitch Accent Prominence	Number of items compared, stop radius, and word frequency (adjusted for search radius).	Magnitude of the expectations.
Phrase Tone Type	Polarity of all items in the search region.	Expectations for upcoming processing.
Phrase Tone Prominence	Polarity of all items in the search region.	Magnitude of the expectations.
Word Duration	Storage + Search (adjusted for search radius).	Processing time for word and associated tokens.
Pause Duration	Storage + Search.	Processing time for punctuation, layout and empty category tokens.
Pitch Range	Density of items in the search region.	Number of simultaneously salient items.

Table 9.4: The mapping between pitch and timing features and properties of the model, and their interpretations.

9.8 Summary

The algorithms presented in this chapter map AWM′ operations and their results to the pitch and timing of a word and the duration of a pause. The mapping assigns intonation based on the results of comparisons between the stimulus and items in a region of working memory. Pitch accents reflect the response of items to the current stimulus. Phrase tones summarize the response of all items in the search region, regardless of whether they have been compared with the current stimulus.

Processing time comes from the number of clock ticks associated with storage and search. However, because words are the only lexicalized tokens, the processing times for phrasal and clausal constituents accrue to the word. Currently, punctuation pro-

duces pausing and, by triggering global resets, produces phrase final intonation as well. The trace and the parasitic gap are also mapped to pausing if the information they represent is sufficiently novel or if they are preceded by punctuation.

The aim of the mapping is to be plausible, if not accurate, with regard to attention and memory processes and at the same time, to produce more varied and more natural prosody. The results of this effort are reported in Chapters 11 (Results) and 12 (Evaluation).

Chapter 10

System integration

In this chapter I describe how the linguistic analysis, memory operations and their mapping to pitch and timing are integrated in the LOQ system. The basic design is shown in Figure 10-1.

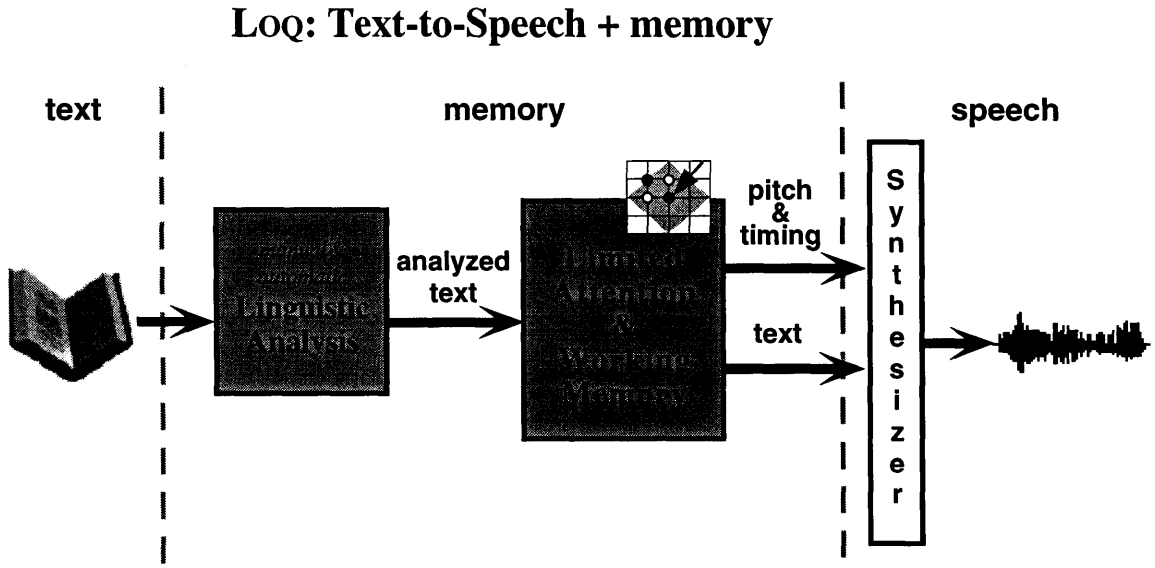

Figure 10-1: The LOQ system. Text is analyzed and then processed by a dynamic model of limited attention and working memory.

10.1 The text

The process starts with the text. So far, three texts representing three different genres have been prepared for LOQ processing. The fiction genre is represented by an excerpt from *One Hundred Years of Solitude*, by Gabriel Garcia-Marquez. Non-

fiction is represented by a news story, originally broadcast on National Public Radio. Rhymed poetry is represented by the nonsense poem, *Jabberwocky*, from *Alice in Wonderland*, by Lewis Carroll. The word and segment counts for each text are shown in Table 10.1.

Text	Genre	Words	Sentences	Segments
100 Years of Solitude	fiction	278	8	6 paragraphs
NPR News report	non-fiction	442	24	6 paragraphs
Jabberwocky	rhymed poetry	167	21	title + 7 verses

Table 10.1: The three texts prepared for processing by LOQ.

10.2 Manual and automatic text analysis

As currently implemented, most of the text analyses occurs offline and with human input. Some of the human input takes the form of a manual mark-up; the rest requires running the data files through programs that convert their contents into a LOQ-readable form. The result is a collection of files, one for each type of analysis and one per database. The contents of these files are combined in LOQ data structures and subject to a final analysis that computes cross-references from one-way annotations, and assigns missing retrieval cues. This information constitutes the bulk of the LOQ speaker's knowledge of the properties and use of language and text.

10.2.1 Components of the analysis

To use the information in the online databases, the relevant data must be retrieved, parsed and then converted into LOQ-readable form. The clause structure and relational analysis reflect the efforts of both human and machine, which together parse the text, identify relations among the constituents of the parse and provide retrieval cues. Table 10.2 lays out the sequence of tasks for each information source and indicates whether they are performed by human or machine. The process of readying the information for use by LOQ is described in the remainder of this section.

Online databases

The databases are provided as ASCII files. Because on the fly file access slows down the LOQ simulations, the current design requires the *a priori* gathering of the database

(manual) **input**		(automatic) **database**		**output**		(manual) **edit**	(automatic) **re-format**
word list	⇒	Brown Corpus pronunciations	⇒	pronunciations	⇒	Add missing pronunciations (inflected forms and proper names)	—
word list	⇒	Oxford Psycholinguistic Database	⇒	database entries	⇒	—	⇒ re-format
word & phrase list	⇒	WordNet	⇒	synsets	⇒	Select appropriate synsets and case frame descriptors for verbs.	—
text	⇒	ENGCG	⇒	tagged text	⇒	Select correct tags from multiple tags; correct incorrect tags.	⇒ re-format
			⇒	noun phrases	⇒	Group phrases into grammatical clauses; annotate relations between constituents; add co-reference and retrieval cues.	—

Table 10.2: Manual and automatic processing of linguistic information for a text.

entries that are germane to the current text. These entries are then read into program memory when the program starts up.

Gathering the relevant data occurs automatically, by iterating through a list of words and phrases from the text and using them as data retrieval keys. Most linguistic databases only store information on the uninflected form of a word. Therefore, the original list is augmented so that it includes both the inflected forms as found in the text, and their uninflected roots.

Once retrieved, the data entries are assembled into a file. These files are then processed by hand or machine to convert them to LOQ-readable form. For example, although the Oxford Psycholinguistic Database contains linguistic data from many sources, LOQ only uses the word frequency data from the Brown Corpus, which informs about the word count, the word rank, and the number of samples and genres in which each word was found. The original entry is automatically parsed and the frequency information is written out as a LOQ data structure file.

The pronunciation database contains pronunciations for uninflected words. Although it is possible to compute the missing pronunciations for inflected forms and proper names, they are currently added by hand, since the work required is minimal.[1]

The WordNet database is indexed by the individual words, compound words (e.g., *"open up"*) and phrases (e.g., *"come to an end"*) of the text. This produces a file of indices that retrieve a synset (a set of synonyms) from WordNet. Because each synset represents a different word sense, the correct sense(s) must be identified. Verb subcategorization information must also be identified. For expediency and accuracy, this currently occurs by hand.

Automatic tagging and stemming

The ENGCG (English Constraint Grammar) software is the main source of automated natural language processing for LOQ. It provides word stemming, part of speech tagging and noun phrase identification. Although it is usually accurate, the output of the tagger must be reviewed by a human editor, who selects the correct set of tags when more than one are given, and re-submits parts of the text when the tags are incorrect (as described in Section 7.2.1). Once an acceptable tagging is achieved, the file is fed into a conversion program, which organizes the many ENGCG tags into twenty-five tag categories, such as number (plural, singular), person (first, second, third) and scalar (absolute, comparative, superlative, cardinal). These categories are used in the matching process.

[1] Morphological information for each syllable, such as whether it is a root or an inflection particle, has been manually added. However, it is not yet used by any of the match criteria.

Manual clause analysis

The clause analysis is the key to LOQ processing. It provides the basic computational structures with which all other information is associated.

Although most of the clause analysis is manual, the output of the noun phrase extractor tool (*NPtool*) reduces the initial work by identifying the noun phrases that are most likely to be subjects, objects, indirect objects, or the objects of a prepositional phrase. This aids in the construction of a grammatical parse. Once it is constructed, other manual annotations may be added, such as those that identify the empty categories, provide referents, and denote subcategorization and anchoring relations.

Figure 10-2 presents an example of the result of a clause analysis. It shows a full grammatical clause that is composed of a subject, verb and a post-posed adjunct clause. Each grammatical role clause contains one or more syntactic phrases. Its structure is hierarchical but not recursive.

```
(Clause (S    (NP "REMEDIOS")
              (NP (anchors "Remedios")
                  "THE" "BEAUTY"))
        (V    (VP "STAYED"))
        (Post (arg-of "stayed")
              (ADVP   "THERE")))
```

Figure 10-2: Example of the clausal analysis input to LOQ

For example, the recursive syntactic parse of *"Remedios the Beauty"* –

(54) $[_{NP}$ Remedios $[_{NP}$ the Beauty]]

– is represented by a sequence of two noun phrases. The annotation to the second noun phrase shows an anchoring relation between the two phrases. This presents the functional basis of the original recursive structure. It identifies $[_{NP}$ `the Beauty]` as the anchor for the phrase whose retrieval cue is *"Remedios"*. Likewise, an annotation to the `Post` clause identifies it as the argument of the Verb clause whose retrieval cue is *"stayed"*.

10.2.2 Consolidation

The information in the database and text analysis files is consolidated in LOQ data structures as follows: LOQ first reads in the grammatical parse file. Because the

parses are hierarchical, they are easily expressed as a sequence of tokens, which represent the reader's serial recognition of words, syntactic phrases and grammatical roles (Section 7.1.2). Hence, the parse in Figure 10-2 becomes the sequence of tokens shown in Figure 10-3.

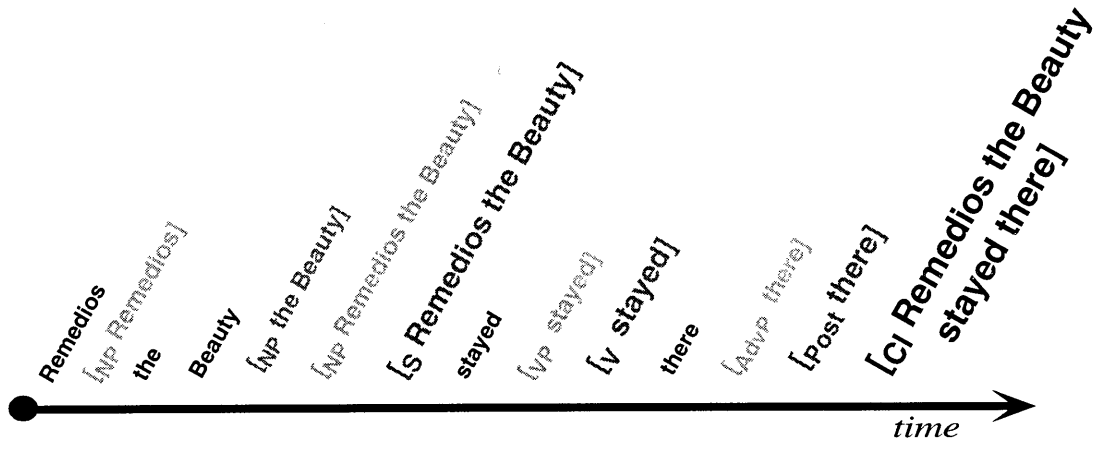

Figure 10-3: Tokenization of clause structure analysis.

The information in the remainder of the files is read in and added as annotations to the tokens. Following this, LOQ performs additional computations to assign syntactic category to each grammatical clause, determine the main retrieval cues for phrases and clauses, explicitly annotate the obvious relations (e.g., an object is an argument of the verb in the same clause) and derive the inverse subcategorization and cohesive relations from the explicit annotations. These have been described in Section 7.2.2.

10.3 Processing in AWM′

The tokens are fed into the AWM′ component and processed according to the update rule: STORE the current token (stimulus) in an unoccupied location and then SEARCH for a match to the stimulus in the region delimited by the search radius. These steps have been covered in Chapters 8 and 9.

Currently, the effects of the stimulus on memory are not mapped to prosody during AWM′ processing. Instead, they are accumulated as a series of snapshots – at least one per time step and one per memory operation.[2] These snapshots capture the salient features of the contents of memory as seen from the pointer's current location. They record the spatial distributions of polarity, compressibility status and occupancy for each successive city block distance outward from the pointer. They also record the current stimulus and the memory operation that has occurred. This may be the

[2]Usually, these are synonymous, except when no comparison occurs at a particular distance.

recognition of a new token, the search for free space, the search for a match to the stimulus, incremental compression or the global compression and polarity resets that occur at the end of a paragraph or verse.

At the end of a simulation, the snapshots are used to reconstruct the sequence of memory operations and their consequences, and to map them to prosody as described in Chapter 9. The reason the mapping does not happen on the fly is because some parts of prosody depend on lookahead of a type that the model does not possess. For example, neither the duration of a word nor its phrase final contour can be assigned until all the related token have been collected, most especially the tokens that denote phrase and clause boundaries. This is in part an artifact of the bottom-up processing of the read speech. In spontaneous speech simulations, the sequence is presumably top down, from concept to lexical expression. Thus, in a LOQ-like simulation, the requisite phrase tokens would already be in memory by the time the word was encountered.

10.4 Mapping search and storage to pitch and timing

At the end of one simulation, LOQ has amassed a sequence of snapshots. The operations and outcomes they represent are mapped to pitch and timing on a word by word basis. The processing of phrase and clause tokens is always expressed in the prosody of a word such that its total duration is the sum of its own processing time plus that of all other associated tokens. The pitch accent of a word is calculated from the polarities before and after it is compared with the items in the search region. Its prominence is calculated from the search distance and polarities after the comparison. The processing of the tokens that follow the word, if any, determines the type and prominence of the phrase tones.

The pitch range is calculated from density – the ratio of the current occupancy to the maximum occupancy of a region (see Figure 6-7). It is currently calculated for *each word* and for *each boundary tone*. Usually, this produces the pitch range declination that is typical of an intonational phrase and allows the phrase final tone to reflect the most current polarity and occupancy information for the AWM′ space.

10.5 Interpreting the mapping for the synthesizer

The product of the mapping is a series of prosodic abstractions. The intonational types are symbols. Durations are expressed as ticks of the AWM′ clock. The work of the interpretation step is to translate the prosodic abstractions into synthesizer-specific instructions.

Fortunately, little change is needed for the intonational markings, since the TrueTalk synthesizer uses most of the same symbols. Only the H*+L accent must be translated for TrueTalk. It becomes a H* accent followed by a L phrase accent, and both are assigned the prominence of the original H*+L accent.

The quantitative information requires more interpretation. It must be turned into absolute numerical values and expressed in units that the synthesizer understands. Thus, pitch range, which starts as density in AWM′, is mapped onto a range of fundamental frequency values. The prominence of a pitch accent or phrase tone, which is a percent of a total in the LOQ calculations, is currently scaled by 1.5 for accents with H*, and by 1 otherwise. It remains an abstraction (a percentage of the pitch range) because TrueTalk also represents it in this manner. The word duration is originally counted as ticks of the AWM′ clock. These are mapped to an inverse percentage of the default TrueTalk speech rate. This percentage is then scaled by the search radius, as described in Section 9.3.1. The pause duration, also in clock ticks, is assigned the unit "centiseconds". Currently, it is not scaled for search radius and there is no maximum duration for a pause. Thus, the duration of a pause is a direct reflection of both the magnitude of the search radius and the distribution of items within the search region.

As described so far, the mapping from AWM′ to pitch and timing produces a pitch accent for every word and phrase final contours for even the smallest phrases. To constrain this, intonation is only expressible under two conditions: (1) if it exceeds the *prominence threshold* (currently set low, at .05); (2) if its type – pitch accent, phrase accent, boundary tone – adheres to the syntax of the intonational phrase. For example, a phrase accent with acceptable prominence will not be expressed unless a pitch accent is the most recent intonational event.

The prominence threshold represents the speaker's sensitivity to activity in AWM′. Prominence values at or below the threshold produce de-accenting, and values above produce accenting. Because LOQ annotates all the words of a text, the synthesizer tends to use citation pronunciation for each one. To make the articulation less abrupt, LOQ cliticizes de-accented function words.

The pre-set ranges for speech rate, prominence and pitch range represent the speaker's stylistic biases and physiological settings. These values are fixed for all simulations and represent influences beyond the current theory of the effect of limited attentional and working memory on prosody. As other influences are identified and integrated, the fixed parameters can be made to vary.

Table 10.3 shows an example of the output that is sent to the TrueTalk synthesizer. This particular example was synthesized with a 22x22 AWM′ memory, a search radius of 4 and a pointer step size of 1. Note that the boundary tone is preceded by pitch range specifications, and followed by prominence. The duration of cliticized words is calculated by LOQ but not specified in the instructions that are sent to the synthesizer, because any duration assigned by LOQ is shortened even more by the TrueTalk's

rate	top line	base line	accent	promi-nence	text	phrase accent	boundary tone			
.84	214	85	H*	1.40	*Remedios*	—	—	—	—	—
—	204	86	cliticized	—	*the*	—	—	—	—	—
.90	204	86	—	—	*Beauty*	L .83	199	86	L	.50
.89	195	87	—	0.27	*stayed*	—	—	—	—	—
1.02	199	86	H*	0.13	*there*	L 1.00	199	86	L	.60
.90	204	86	H*	0.36	*wandering*	L .80	195	87	L	.66
.79	195	87	H*	0.17	*through*	L .12	—	—	—	—
—	195	87	cliticized	—	*the*	—	—	—	—	
.92	195	87	L*	0.22	*desert*	L .80	195	87	L	1.00
—	187	88	cliticized	—	*of*	—	—	—	—	—
1.18	187	88	L*	1.00	*solitude*	L .80	204	86	H	.75
10cs	—	—	—	(pause)	,	—	—	—	—	—

Table 10.3: An example of the synthesizer instructions that LOQ produces. The words, phrase accents and boundary tones are followed by a a prominence value. Words and boundary tones are preceded by the topline and baseline specifications for the pitch range.

cliticization algorithm. Since the duration that LOQ calculates will be shorter than the default, allowing the TrueTalk to assign its own short duration is acceptable. The LOQ simulations use the TrueTalk's default male voice.

10.6 Discussion

Many of the necessary practical components of prosody are not covered by a model of limited attention and memory. As much as possible, LOQ represents such components by fixed parameters that have been set to conservative values. Most are explicitly stated, such as the ranges that restrict the prominence, speech rate and pitch range to affectively calm expressions. Others are implicit because they are implemented as inequalities. For instance, the H tone bias for pitch accents and the L tone bias for phrase tones are currently based on the ratio of 1/2 (see (39), (40) and (42)). As implemented in LOQ, all of the fixed-value parameters can be made to vary to incorporate a theoretical claim or to reflect the empirical data on human prosody.

The partial ordering of the match predicates is another source of implicit assumptions. The ordering is based on an estimation of mutual information, as explained in Section 8.1. However, it substitutes for an ordering based directly on a statistical analysis of empirical data.

One very practical consideration for speech synthesis is the synthesizer hardware. Fortunately, the TrueTalk imposes only a few tasks. One is to select the speaker

beforehand (the default male voice, in this case). Another is to select ranges of pitch and prominence values so that the speech is neither over-excited nor flat. A third concerns the effect of the commands that LOQ sends to the synthesizer. They are often longer and more complicated than is customary for TrueTalk. This produces choppy articulation because any word with prosodic specifications is more likely to be fully articulated. Therefore, LOQ explicitly cliticizes all function words whose prominence is below the prominence threshold. This makes for a smoother and more natural articulation.

10.7 Summary

In this section, I have described the components of the implementation of the LOQ system. As a practical matter, the implementation relies on a combination of manual and automatic natural language analysis. With the development of better natural language processing tools, the human role in the analysis will diminish.

Another practically motivated feature is the inclusion of fixed parameters and orderings. The match threshold is the main fixed parameter for the AWM′ computations. Its effect is dependent on the order of the match criteria, which is currently fixed as well, according to estimations of the relative amount of mutual information that a positive outcome contributes to the determination that there is a match. The other main use of fixed parameters is the pre-selection of the ranges onto which the LOQ prosodic abstractions are mapped.

Within the framework established by the fixed parameters, the variable control parameters test the hypothesis that attentional limits are reflected in prosody, and that variations in attentional capacities produce characteristic kinds of prosodic variation. Operations in AWM′ are affected by three control parameters: the search radius; the pointer step size; and the size of the AWM′ space. Their effects on prosody are described in the next section.

Part III

Results and Evaluation

Chapter 11

Results

In the previous chapters I discussed the fixed parameters and biases in the LOQ computations. In this chapter I discuss the effects of the control parameters, which are variable. As in Landauer's and Walker's work, the main control parameter is the search radius, which defines the focus of attention. Items within the radius are retrievable and therefore, salient. Items beyond the radius are forgotten.

The second control parameter is memory size. In both Landauer's and Walker's simulations, it is fixed. However, because storage in AWM′ is limited to one item per node, the memory size imposes a hard limit on global storage. It is a variable quantity for the simulations in order to explore the effects of small, medium and large memories.

The final control parameter affects the movement of the pointer. In the original AWM model, the combination of the random trajectory of the pointer and its step size of one city block are jointly responsible for its slow random walk across the memory space. This is the source of the correspondence between temporal proximity among stimuli in the world and their spatial proximity in the memory model. Increasing the step size so that more distance is covered per step is likely to weaken this correspondence. The motivation for doing so is to see whether this effect would expand the number of prosodic styles produced by LOQ. However, its main effect is to add more variation to the prosody within a speaking style.

Even with the addition of variable memory and step sizes, the search radius remains the primary influence on prosody. Different radii produce clusters of quantitative and categorical features that characterize three main styles. The smallest radii produce wide pitch ranges, large prominence values, steep phrase final falls and a predominance of H* accents, thereby mimicking the exaggerated prosody that a reader might use when reading to a young child, or the prosody that the young child herself might use. I will call this the "child-like" style.

Mid-range radii produce narrower pitch ranges, smaller pitch accent prominences, fewer accented words and fewer and less steep final falls. This provides enough prosodic variation to be both interesting and expressive but without producing the sing-song melodies of the child-like style. I will call this the "adult narrative" style.

The largest radii produce the fewest accented words, the most phrase final rises, and a wide pitch range. At times, the simulated speaker sounds bored, slightly annoyed or at least very familiar with the material. I will call this the "knowledgeable" style.

Because the storage algorithm is stochastic, simulations with identical parameter values produce prosody that differs on a per-word basis. This illustrates prosodic variation for a single speaker. The prosody that varies for different radii within a stylistic range illustrates inter-speaker variation. Finally, the existence of more than one style illustrates variation among groups of speakers and shows that more than one style is possible, even for the same text.

11.1 The text

As input, I chose texts from three genres: (1) fiction, represented by an excerpt[1] from *One Hundred Years of Solitude* by Gabriel Garcia-Marquez; (2) nonfiction, represented by a news report originally delivered on National Public Radio; (3) rhymed poetry, represented by *Jabberwocky*, by Lewis Carroll. Their word, sentence and segment counts have been reported in Table 10.1.

For the purposes of gathering statistics, I used smaller samples of comparable length and structure: the first paragraph of the fiction excerpt (63 words); the first paragraph of the new story (68 words); the first three verses of the poem (72 words). Their word, sentence and segment counts are shown in Table 11.1.

Text	Genre	Words	Sentences	Segments
100 Years of Solitude	fiction	63	1	1st paragraph
NPR news report	non-fiction	68	4	1st paragraph
Jabberwocky	rhymed poetry	72	8	title + 1st 3 verses

Table 11.1: Equivalent samples from each text.

Table 11.2 shows the some of the structural differences that are captured by the text mark-up. For example, the rhymed poetry is composed of the most tokens. This reflects its slightly higher word count but also its far greater number of punctuation and layout tokens. Especially, because line breaks in rhymed poetry are integral

[1]Translated from the original Spanish.

rather than accidental, it contains significantly more layout tokens – a total of 224 tokens in all, as compared with 180 for the fiction text and 175 for the news text.

Text	Words	Punc-tuation	Layout	Syntactic Phrases	Gram-matical Clauses	Full Clauses	Total Tokens
100 Years...	66*	7	1	54	40	12	180
NPR news report	68	8	1	53	34	11	175
Jabberwocky	72	16	13	53	53	17	224

Table 11.2: Number of LOQ tokens per text and per category. The word total for *100 Years of Solitude* includes three deletions.

In the LOQ mark-up, the presence of more syntactic phrases than grammatical clauses indicates recursion and embedding in the syntactic parse. For example, a complex noun phrase will be decomposed into a sequence that contains the main noun phrase and its anchoring and modifier phrases. Accordingly, Table 11.2 shows that the syntactic complexity is least for the rhymed poetry and greatest for the news story. This makes intuitive sense – the news story must pack a great deal of background information into a short text. Anchoring is one technique that allows the author to introduce new and relevant information without marking it as new. On the other hand, the rhymed poem occurs in a children's story. As such, it is not likely to be constructed from complex linguistic and attentional structures, and new information is likely to be introduced directly and explicitly rather than as background material.

11.2 Initial explorations

The results of initial simulations eliminated two kinds of control parameters. The first is dimensionality – the prosody produced by the two and three[2] dimensional memories for the fiction sample (see Table 11.3) showed no differences that could be attributed to dimensionality. Although this is not a typical result for operations in two and three dimensional spaces, it appears that for the AWM′ memory and mapping procedures, the number of nodes in the space is more important than the number of intersections at each node.

It is fortunate for practical reasons, because the two dimensional simulations run much more quickly while still retaining the important properties of AWM′. Like the three dimensional memories, they are regular and periodic, and they too support both

[2]A simulation series with a 4x4x4x4 dimensional memory also showed no differences due to dimensionality.

Memory Size	# Nodes in the Space	Maximum Distance	Step Size	Radii
8 x 8	64	8	1–6	6 total: 1 4 7 10 13 16
4 x 4 x 4	64	6	1–6	4 total: 1 4 7 9
22 x 22	484	22	1–6	8 total: 1 4 7 10 13 16 19 22
8 x 8 x 8	512	12	1–6	5 total: 1 4 7 10 12
50 x 50	2500	50	1–6	17 total: 1-49 by 3, and 50
14 x 14 x 14	2744	21	1–6	11 total: 1-28 by 3, and 30

Table 11.3: Spatial properties and simulation parameters for two and three dimensional memories. Because the spaces wrap around, all dimensions must be even. The two and three dimensional spaces have an equivalent number of nodes.

the pointer's random walk and search within a region defined by attentional limits. These are the required features for modeling the effects of attentional resources on cognitive processing.

Another equivalence occurred for the pointer step size – the city block distance (i.e., number of nodes) traversed for each move. The results for the even step sizes (two, four, six) were similar, as were the results for the odd step sizes greater than one (three, five). These equivalences eliminated the step sizes 4, 5 and 6 from the simulations.

In sum, the initial explorations removed dimensionality as a control parameter, and reduced the range of pointer steps for the simulations. Therefore, the results I report are for small, medium and large two dimensional memories, the pointer steps 1, 2 and 3, and radii from 1 to the maximum distance needed to cover all the nodes in the space.

11.2.1 Main results

The main control parameters of the simulations are the search radius, pointer step size and memory size. The search radius models retrieval capacities and affects the chance of a match. Therefore, it is the main control parameter for simulating the effects of limited attention. The pointer step size models how quickly the focus of attention changes. Increasing it produces a sparse storage pattern and thereby weakens the correspondence between temporal and spatial proximity. Its attentional consequence is to reduce the chance of a match for items that are temporally related and therefore to increase the number of items that are classified as new.

The memory size describes the amount of working memory that the speaker has made available, but makes no commitment as to whether this availability is innate or

situational. Its main effect is on the distribution of items in memory, and therefore affects any computational or mapping feature that relates to storage. For example, because storage is most dense for the smaller memories, attempts at storage are most likely to require incremental compression as a precursor to actually storing the stimulus. Therefore, storage in the smallest memories is most likely to take longer. In addition, because density is currently mapped inversely to pitch range, the smaller memories will produce the smallest pitch ranges.

All the data reported herein are the result of running five simulations for each combination of control parameter values. Each combination defines the attentional and storage capacities of one LOQ "speaker". The control parameters, their combinations and the total number of simulations for each memory size are shown in Table 11.4.

memory size		step size		radii		speakers		repetitions		total
Small	(10 x 10):	1,2,3	×	1–10	=	30	×	5	=	150
Medium	(22 x 22):	1,2,3	×	1–22	=	66	×	5	=	330
Large	(50 x 50):	1,2,3	×	1–50	=	150	×	5	=	750

Table 11.4: Control parameters and their combinations in the simulations.

The intonational tone data is reported only when the prominence exceeds the current threshold of .05. Pitch accents and phrase tones with lower prominence are realized only as the absence of intonational specifications, that is, as an unaccented word or the absence of a phrase break.

11.2.2 Search radius

The main attentional consequence of the search radius is the distinction between given and new information. Figure 11-1 shows that the pitch accent mapping preserves this distinction. Simulations with the smallest radii produce the fewest unaccented words and the greatest number of words with H* accents. Conversely, the largest radii produce the most unaccented words and a greater proportion of accents with L* main tones relative to the total number of accents.

The figure shows that the search radius is also an important determinant of the proportions among the accent types. While they are roughly the same for all radii, their absolute values are scaled by the ratio between the accented and unaccented words. Overall, the trends in the distributions coalesce into three regions, which are defined by a range of radii:

- The radii from 1 to 3, in which the H* accents are predominant but decreasing,

and de-accenting is minimal, but increasing.

- The radii from 4 to 14, in which de-accenting increases and the H* accents decrease slightly;
- The radii above 14, in which the H* accents decrease sharply and de-accented words predominate.

These divisions distinguish the child-like, adult narrative and (adult) knowledgeable styles, respectively.

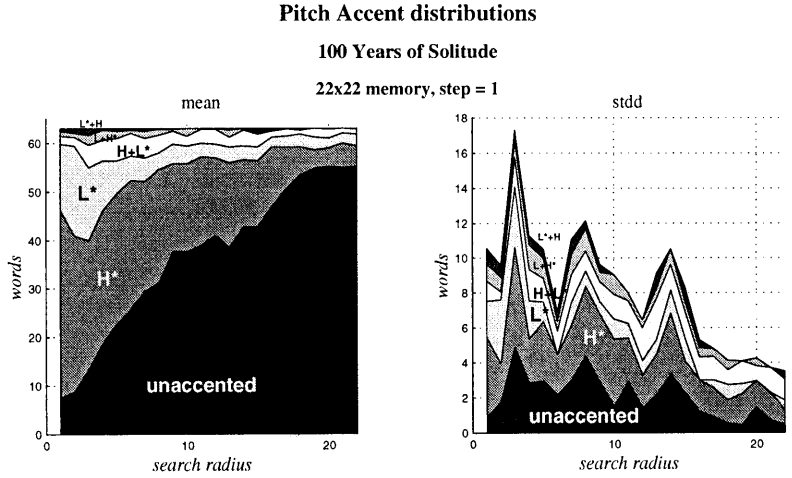

Figure 11-1: Mean and standard deviation for the distribution of pitch accent types, for a step size of 1, a medium memory capacity (22x22) and the fiction sample.

The effect of the search radii on phrase final contours is twofold. As shown in Figure 11-2, the smallest radii produce more intermediate phrases and hence, fewer full intonational phrases. In addition, as the search radii increase, the mean occurrence of falling contours (L L%) decreases, while rising contours (L H% and H H%) increase.

The prominence of intonational tones is another feature that responds mainly to the search radius. This is by design, according to the current mapping, which adjusts prominence values for the radius to ensure that the prominence reaches its maximum for the maximum search radius. Consequently, the smaller radii produce the widest range of prominence values and but the least diversity among the actual values. This contributes to the sing-song melodies that characterize the child-like style.

Conversely, the larger radii produce a smaller range of values, but with a greater diversity of values. This contributes to the expressive but not exaggerated intonation of the narrative style and to the subdued intonation of the knowledgeable style (in which the same set of prominence values are realized for far fewer accents). Figure 11-3 shows that for all three memory sizes, the mean is largest for the smallest radii (1, 2, 3 and 4) and levels out after that. The standard deviation is largest for a radius of 1, and about equal for all other radii.

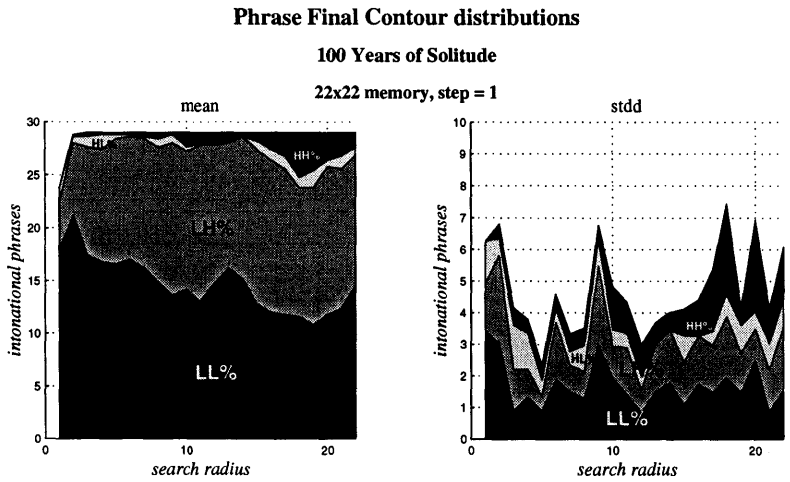

Figure 11-2: Mean and standard deviation for the distribution of phrase contour types, for a step size of 1, a medium memory capacity (22x22) and the fiction sample.

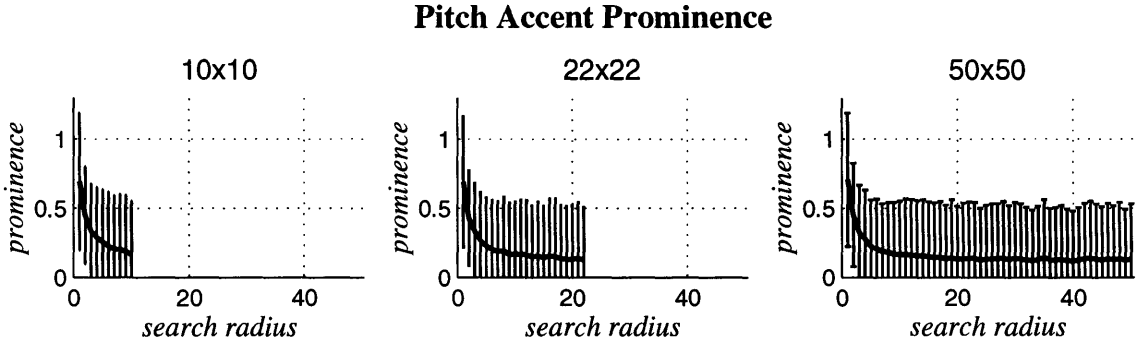

Figure 11-3: Mean and standard deviation for pitch accent prominence, for three memory sizes and a step size of 1 (fiction text sample).

Another feature that mainly responds to search radius is the duration of a pause. This is again by design, as a result of assigning the centisecond unit to the clock tick total. Thus, the smallest times are produced when matches are found close to the pointer, as illustrated by the mean times shown in Figure 11-4. The larger radii tend to produce larger search times, since the search may extend over the entire region. However, because matches may also be found close to the pointer, the variation among the pause duration values is greatest for the largest radii, as shown by the standard deviation plots in Figure 11-4.

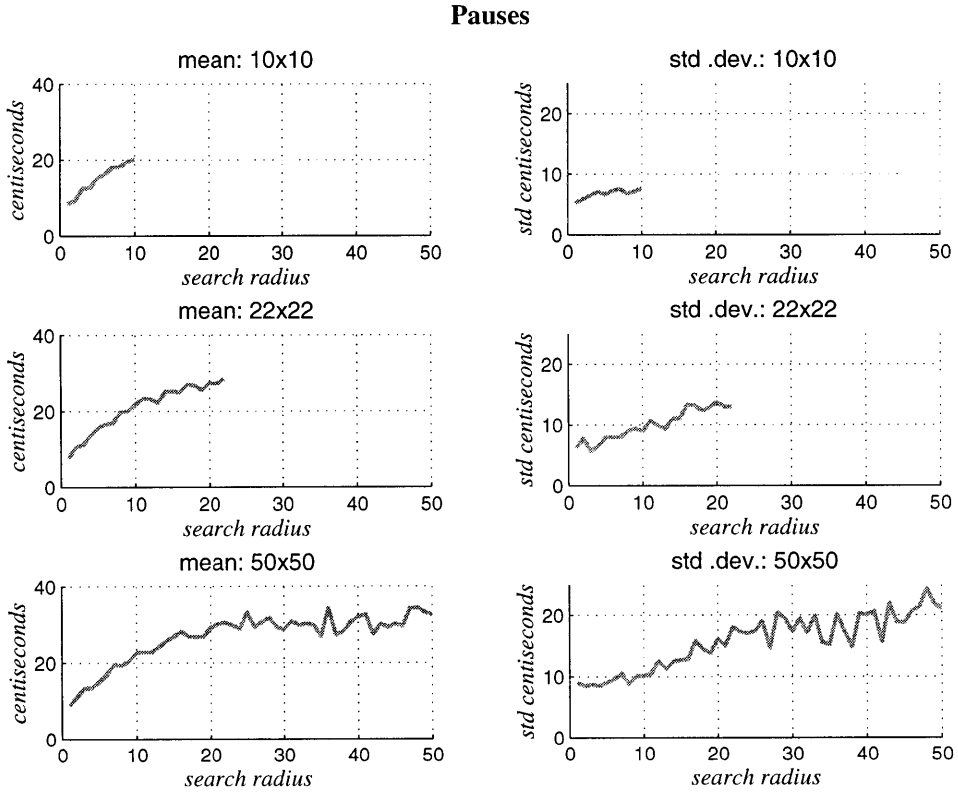

Figure 11-4: Mean and standard deviation for pause duration for three memory sizes and a step size of 1 (fiction text sample).

11.2.3 Pointer step size

In Landauer's and Walker's simulations, the pointer travels in a slow random walk, moving one city block per time step at most. This produces a distribution that is locally random but temporally coherent. The effect of increasing the step size is to increase the sparseness of the distribution of items in memory. This has the potential to disrupt the correspondence between the temporal proximity among the stimuli and their spatial proximity in the memory model. Yet, the main effect of the step size is to reduce the number of de-accented words for the larger radii and to increase the number of H* accents, as shown in Figure 11-5.

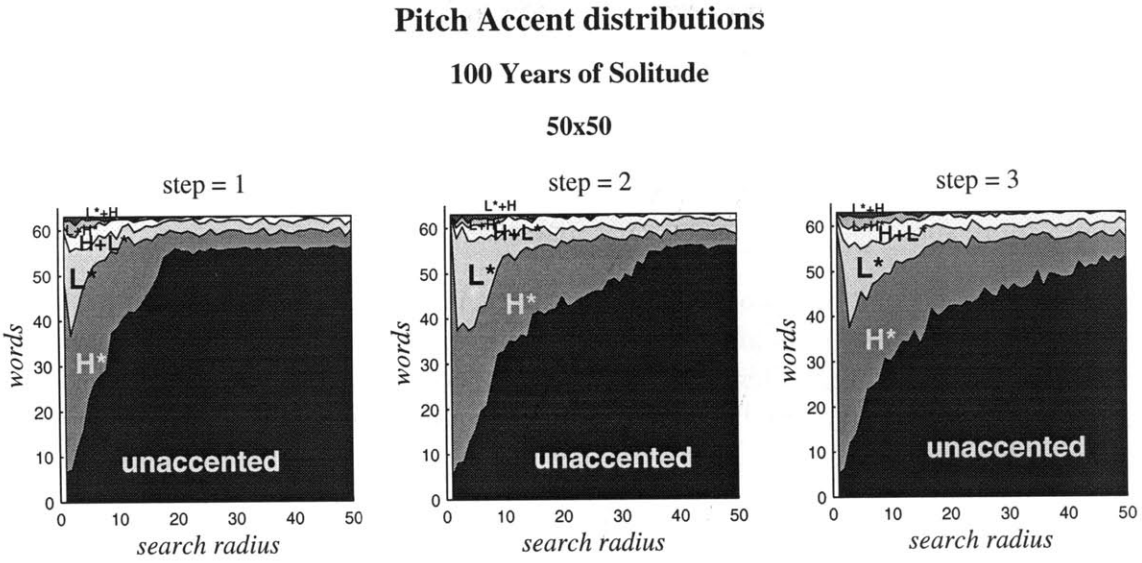

Figure 11-5: Mean distribution of pitch accent types, shown for the largest memory and for the pointer step sizes of 1, 2 and 3 (fiction text sample).

The general trend is that increasing the step size increases the sparsity of the distribution of items in memory. However, in a two dimensional Cartesian lattice, the even steps produce an additional and unintended affect. Because they only store items at locations whose coordinate sums are always even or always odd, (depending on the pointer's initial coordinates), they use only half the available space at most. Sometimes, this may create a distribution that is more sparse than those produced by larger odd step sizes. In addition, the checkerboard storage pattern it produces leads to aberrant outcomes for simulations with a search radius of one, because items stored every two steps cannot be accessed when the search radius is only one. Therefore, all pitch accents are new, as can be seen in Figure 11-5.

11.2.4 Memory size

Landauer found that a memory size of 20x20x20 produced output that correlated most clearly with the frequency and recency data for human experimental subjects. Walker found that a memory size of 16x16x16 was large enough to avoid having to choose either the overwrite or multiple storage strategy.[3] Because AWM′ stores only one item per node, disallows overwrite and does not include swap into a long term memory component, its memory size must be chosen beforehand to accommodate the maximum number of items likely to be in memory at any one time. The absolute lower bound on memory size is the number of incompressible tokens in the parsed input. Currently, this includes most full clauses, and the punctuation and layout

[3]Personal communication. October, 1998.

tokens that are not contained in a full clause.

In general, it is wise to set the memory size above the lower bound expectations. This reduces the time spent searching for free space, and avoids simulating memory pathologies such as thrashing or failure to store.[4] Because of these constraints, the different memory sizes mainly reveal the effects of a space that is populated sparsely (small), more sparsely (medium) or very sparsely (large).

In AWM′ calculations, any prosodic feature that is influenced by distribution patterns will be influenced by the memory size. For example, because a small memory will be more densely packed regardless of step size, and because LOQ calculates pitch range as the inverse of density, smaller memories will produce the smallest pitch ranges, as shown in Figure 11-6.

Memory size appears to interact with the prominence of phrase tones as well (Figures 11-7 and 11-8). The sparsity that is the typical consequence of the largest pointer step (of 3) is constrained by the smallest memory and only begins to have an effect as memory size increases. The low prominences for the pointer step of two occur because the storage pattern is always a checkerboard pattern whose sparsity is invariant, regardless of memory size.

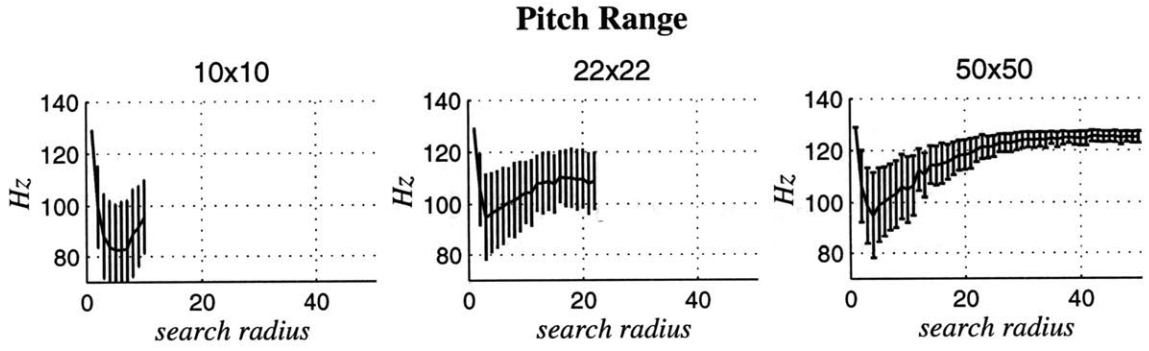

Figure 11-6: Pitch range mean and standard deviation for the fiction sample for three memory sizes and a step size of 1.

Word duration is also affected by memory size. The mean speech rate is slowest for the smallest memories, in part because more time is spent searching for free space and in part because word duration is adjusted for the search radius. The standard deviation (not shown) is about 33 words per minute for all radii and all memories.

[4]If AWM′ included a mechanism for moving fully compressed items into long term memory, its memory spaces could be smaller and less carefully chosen.

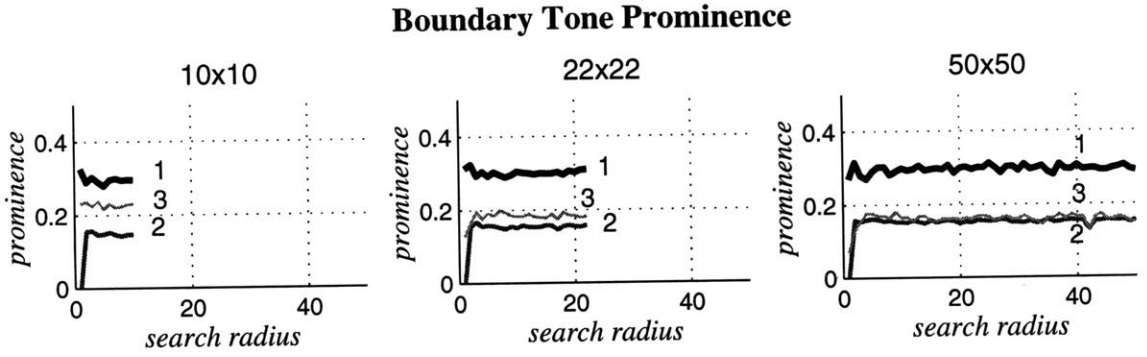

Figure 11-7: Mean boundary tone prominence for three memory sizes, and three step sizes (fiction text sample).

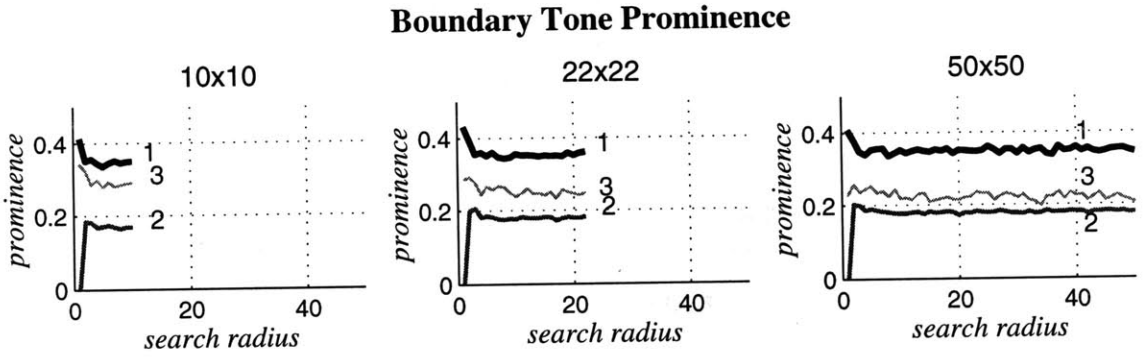

Figure 11-8: Standard deviation for boundary tone prominence, for three memory sizes and three step sizes (fiction text sample).

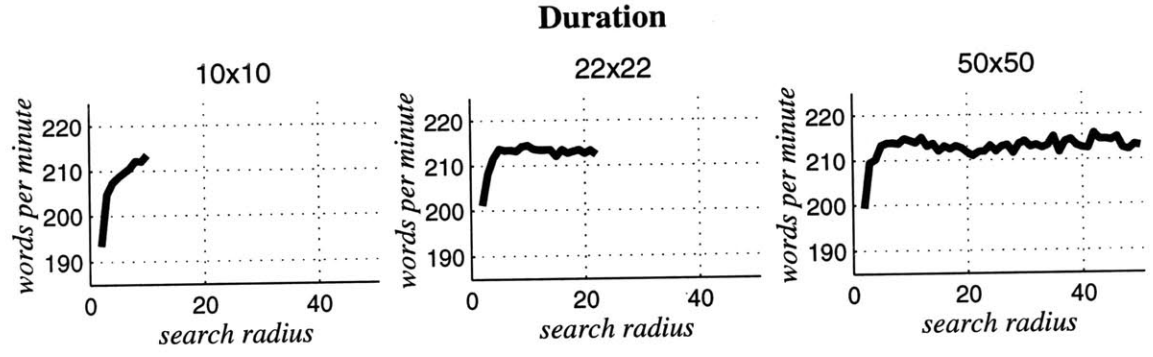

Figure 11-9: Mean durations for three memory sizes, for a step size of one (fiction text sample).

11.2.5 The text

Quantitative descriptions of authorship and genre focus on characteristic distributions of linguistic constituents. For example, simply from word frequencies of synonymous function words (such as *while* and *whilst*, *on* and *upon*), Mosteller and Wallace [MW64] claim to have determined which of three authors wrote each of the eighty-five Federalist papers.[5] Not only the unique voice of an author, but whole genres have characteristic distributions, often of syntactic categories as well as individual words or phrases [Bib93].

This is borne out in the distributions reported earlier in Table 11.2 and in Table 11.5, which show that the percentage of empty category tokens is greatest for the fiction sample. It is not explainable simply by differences in the number of attentional foci. For, although *Jabberwocky* is a rhymed narrative with three characters, the fiction text is a prose narrative with four, and the news story is a prose narrative with at least nine, it is the fiction text that contains the most empty categories. The interpretation I propose is that genres may be characterized in part by the assumptions they make about the listener's attentional and storage capacities, and that these assumptions are reflected in characteristic distributions of linguistic and textual features.

Text	Words	Grammatical Clauses	Total # Tokens
100 Years of Solitude	5%	33%	9%
NPR news report	—	24%	5%
Jabberwocky	—	18%	4%

Table 11.5: For the smaller samples, the percent of the tokens that represent empty categories.

For instance, *Jabberwocky* is part of a children's story. As such, it is written for listeners with limited attention and for whom many words and ideas are new. Therefore, so as not to unduly confuse, its references are lexicalized, its syntax is simple and its clauses (as identified in the mark-up) are short. Similarly, although addressed to an adult audience, the news report is new information for the listener, and in addition, is attended to in real time and in competition with other memory and attention tasks. Therefore, references are explicitly stated (lexicalized) for this audience as well. On the other hand, fiction in general is meant for an extended hearing or a silent read. An author of fiction may reasonably assume that there will be little competition for the reader's memory and attentional resources. Thus, the fiction sample contains the greatest number of empty categories. In addition, because the sample is taken from

[5]Papers written from 1787 to 1788 by John Jay, Alexander Hamilton and James Madison to convince Americans to ratify the U. S. Constitution.

the middle of the book, the author may also assume that by this point, the silent reader or (listener) has already populated his memory with the information needed to resolve the empty categories references.

The mean distributions of intonational types (Figure 11-10) reflect these structural differences. They are most similar for the prose texts and the most distinct for *Jabberwocky*, whose simulations produce the fewest unaccented words and a fairly large proportion of H* accents throughout. This is due in part to the content of the poem, which is composed of many nonsense words, but also to its shorter and more numerous phrases, each requiring its own nuclear accent.

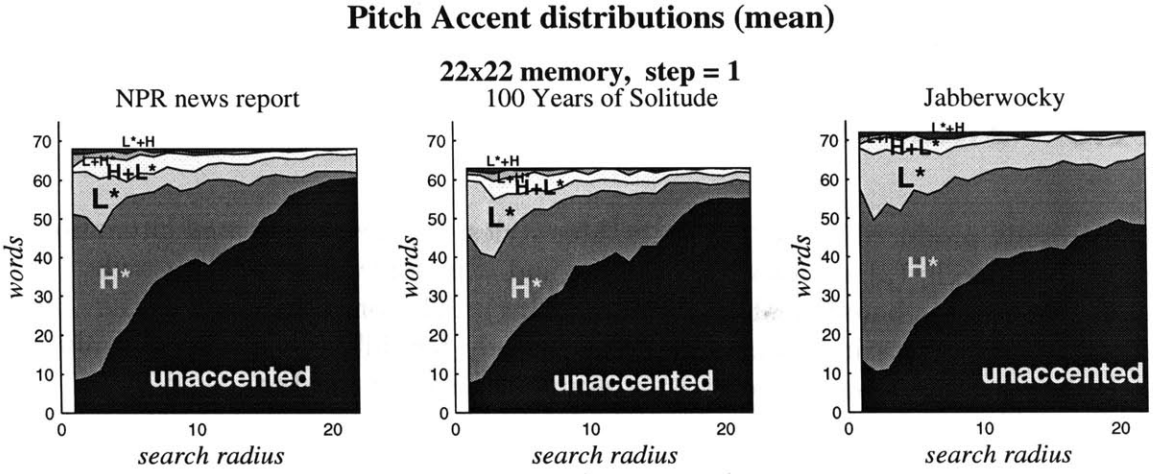

Figure 11-10: Area plots of the mean pitch accent distributions for all three texts, for medium-sized memory and a pointer step of 1.

The distributions of phrase final contours (Figure 11-11) show the influence of phrase and clause structure even more strongly. Again, the trends for *Jabberwocky* are the most distinct. They show a much smaller proportion of continuation rises (LH%), and a much larger proportion of high phrase final rises (HH%). This is likely due to the line by line structure that interrupts clauses, and therefore produces rising phrase final contours, which indicate continuation [Ran80].

In contrast, both the fiction and news texts show similar patterns. Both exhibit a preponderance of falling (LL%) contours, and proportionally more level (HL%) contours than the rhymed poetry. The fiction text shows more continuation contours, perhaps a reflection of a narrative with fewer main characters and therefore longer and more in-depth coverage for each one.

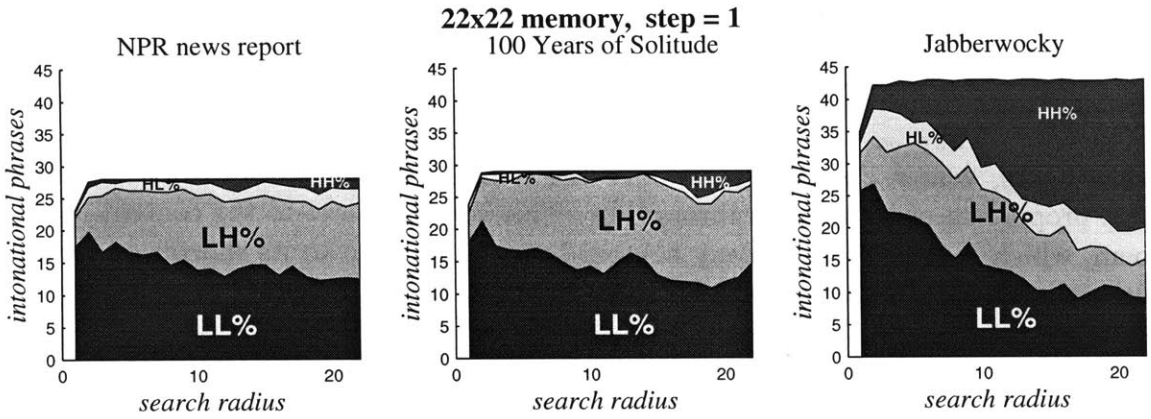

Figure 11-11: Area plots of the mean phrase final contour distributions for all three texts, for a medium-sized memory (22x22) and a step size of 1.

11.3 Summary

The prosody produced by LOQ exhibits three styles: (1) the exaggerated intonation of children's speech; (2) a somewhat intimate and expressive adult reading style; (3) a style that sounds bored (and sometimes boring). The prosody varies within these styles in two ways – it produces different prosody for different speakers within the style, but also produces varied prosody even for the same speaker.

Under the current mapping of AWM′ processing to prosody, the significant correlations between the control parameters and trends in the output are:

- The radius is the strongest determinant of the system's behavior for all the prosodic features, both quantitative and categorical. However, the pointer step size has an independent effect on the distribution of pitch accent types. Increasing it expands the radius range that produces the adult narrative style.

- Under the current mapping, memory size affects the distribution of items in the space. Therefore, the pitch range and duration show the effects of memory size.

- A completely unintended side effect of using Cartesian grids is that the combination of even pointer steps and a search radius of one (an outlier condition) exhibit the exact same intonation per text, regardless of the step size (2, 4, or 6). This comes from the checkerboard distribution of items for even pointer steps.

- The structure of the text imposes some form on the output. The pitch accent and phrase final contour distributions show the most distinct pattern for the rhymed poetry, whereas the two prose styles have more similar distributions.

Informal feedback indicates that while sections of the text sound natural, the prosody for readings of the whole text is not as coherent. In the next section, I report on a formal evaluation which compares the LOQ output with the natural speech readings of the same text.

Chapter 12

Evaluation

LOQ consistently produces three prosodic styles likely to be associated with attentional and memory differences: a child-like style for small search radii; an adult narrative style for the mid-range radii; and a knowledgeable style for the largest radii. Like human speech, the LOQ speech shows individual variation within these styles. In this section I report on a comparison that seeks to answer two questions: (1) Does the LOQ intonation exhibit patterns of variability that are the same or comparable to those of natural speech? (2) How closely does it match?

Ideally, a comparison would be possible for all three styles generated by LOQ. However, both the currently available intonationally annotated corpora – the Audix corpus of Associated Press news stories and the Boston University (BU) corpus National Public Radio news stories [OPSH95] – consist of news stories read by professional newscasters. Neither are likely to contain examples of child-like prosody, but rather, prosody that is closer to the adult narrative and knowledgeable styles. I elected to use the BU corpus, because, unlike the Audix data, it contains intonational data for more than one speaker – seven in all. This allows meaningful comparisons within and across groups of human and LOQ speakers. Moreover, because the corpus also includes duration, pause, F0 and phoneme data, it is likely to become a test suite for evaluating both recognition and synthesis algorithms (e.g., [RO96, Mag97]).

An alternative is to run listener tests. However, such tests tend to be costly in terms of experiment design (and re-design) and gathering a sufficient number of subjects for each stage of the experiment. In addition, asking subjects to rate naturalness may not yield the most useful data because naturalness is hard to define and its perceptual cues hard to disambiguate. Indeed, Ross and Ostendorf [RO96] report inconclusive results from their listener studies that compared the prosody of natural and synthetic speech. They attribute this to the difficulty of comparing a defect in one version against a different defect in another. The perception of naturalness may also be tied to the perception of speaker characteristics such as sex and age, as can be the case for the perception of emotions in synthetic speech [Cah90]. In all, this suggests

that we don't yet know which features (acoustical, social or otherwise) contribute most to a perception of naturalness and, in addition, whether they are dependent or independent.

I base the comparison on intonation because demonstrating a quantitative and processing effect on intonational category (pitch accents, in particular) is one of the main goals of this work.[1] I use the kappa statistic to measure the closeness of the intonation within and across the groups of natural and synthetic speakers. Typically, kappa is used to quantify the reliability of data classifications assigned by multiple human coders [Car96]. For my purposes, it is as much a measure of variability as reliability. The first application measures the intonational variability among the human speakers; the second measures variability within each group of LOQ speakers as defined by search radius, pointer step size and memory size; the third compares the LOQ and human intonation to find the LOQ simulation parameters most likely to duplicate the output of at least one of the natural speakers.

12.1 Measuring similarity using Cohen's kappa

The kappa statistic was developed by Cohen [Coh60] to assess inter-coder reliability among coders using nominal scales. It assumes qualified judges (coders) and determines how much of their agreement or disagreement is significant. In Cohen's example, the judges are clinical psychologists, the nominal categories are "schizophrenic", "neurotic" and "brain-damaged" and the coding units are psychological test protocols. In general, the kappa statistic is appropriate when the coded data meet these conditions (pp.38, Cohen(1960)):

1. The units are independent.
2. The categories of the nominal scale are independent, mutually exclusive, and exhaustive.
3. The judges operate independently.

These are also the preconditions for using Chi-squared. However, as Cohen points out, while Chi-squared measures how significantly classifications differ from chance, it does not report on whether the difference is due to agreement or disagreement. Kappa does both. A score of 1 indicates perfect agreement that is not due to chance (the score for any coder against him or herself is always 1). A score of 0 indicates that all agreement is due to chance. Negative scores indicate disagreement greater than chance.

An important aspect of the kappa statistic is that it compares agreement on a unit-by-unit basis. This distinguishes between two extremes that simple averaging will

[1] Although LOQ also generates quantities for prominence, pitch range and the duration of words and pauses, most of these have already been linked to individual processing capacity.

miss: (1) when judges A and B assign the same categories the same number of times but to none of the same units; (2) when judges A and B assign the same categories to the same units. When reliability of coding is key, the data for the first condition cannot be used because the judges disagree so completely. Its kappa statistic is negative, indicating agreement at less than chance (conversely, disagreement greater than chance). The kappa for the second case is 1, indicating that the two judges are in perfect agreement and most importantly, that the agreement is not due to chance. If they agree at a level significantly above chance, reliability can be assumed. Krippendorff [Kri80] maintains that kappa above .8 is highly significant for content analysis, while kappa between .67 and .8 allows only tentative conclusions to be drawn. However, the significance criterion is likely to be application dependent.

The prosody of a text when read multiple times, whether by one or many speakers, meets the conditions for application of the kappa statistic as long as the prosodic features can be expressed as categorical labels. This clearly holds for intonation, whose tonal inventory is composed of independent categories.[2] That is, the judges are the speakers themselves, the units are the words; and the categories are the pitch accents, phrase accents and boundary tones.[3]

Although other aspects of prosody may be expressed categorically, such categories are either not in widespread use (e.g., categorical prominence for words and phrases) or have ordinal as well as nominal attributes (categorical prominence and the ToBI break indices). Fortunately, the tonal inventory is standardized and its denotation, meaning and use well-studied. Moreover, intonation is a key communicative component of prosody and therefore a meaningful basis for comparing the LOQ and human data.

12.2 The natural speech data

The natural speech for this study comes from the Boston University corpus of acoustically and phonologically analyzed radio news broadcasts. It was collected and annotated with ToBI symbols under the direction of Ostendorf, Price and Shattuck-Hufnagel [OPSH95].[4] The entire corpus contains four news stories read by seven National Public Radio newscasters. It includes both original broadcasts and versions recorded off-air solely for the purpose of data collection. The off-air portion contains two versions of each new story. For the first version, the readers were instructed to use a conversational style. For the second, recorded 30 minutes later, they were instructed to use their normal newscasting style. Ostendorf *et al.* [OPSH95] report that the newscasters tended to slip into their radio style during the conversational sessions. Therefore, the most stylistically consistent data is found in the radio style

[2]Although some researchers such as [PS94] conflate H* with L+H*, and L* with de-accenting.

[3]I am indebted to Susan Brennan for this insight and for pointing me to Cohen's original paper.

[4]It is distributed by the Linguistic Data Consortium.

portion of the off-air recordings[5] and is the source of the comparison data for this study.

Using the kappa statistic is possible if one takes the perspective that speakers perform a categorical prosodic mark up of text as they deliver it. However, their actual output is an acoustical signal, whose features must be classified categorically. For example, the F0 component must be translated into intonational symbols. Currently, all such translations are done by trained human coders. This raises questions about the consistency of the coding as well as about how true it is to the original speaker's intent. Indeed, quantifying the accuracy of the human translations is precisely the case for which the kappa statistic was developed. In the case of the BU corpus, the bulk of the ToBI annotations were contributed by two coders who showed high agreement. Disagreements were resolved by the vote of a third expert.[6] For these reasons, I believe the coding to be sufficiently accurate and consistent.

One problem with the corpus is incomplete coverage – not all stories are accompanied by ToBI annotations, not all the speakers recorded all the stories and the ToBI annotations are sometimes incomplete for one speaker. My choice of text was dictated by the need to find the most complete and consistent data. I chose a six-paragraph, 442-word story about a retiring Massachusetts State Supreme Court chief justice.[7] Its ToBI data are available for six speakers – three female (F1A, F2B and F3A) and three male (M1B, M2B and M3B). The "B" speakers (F2B, M1B, M2B and M3B) typically wrote their own news stories and recorded them prior to broadcast; the "A" speakers (F1A and F3A) typically read their stories live and did not write the text.

Unfortunately, even this data set is incomplete. The second paragraph of the F3A data is missing its ToBI intonational mark up, and the first four paragraphs of the M1B data are annotated with pitch accent location but not type. Therefore, tests on pitch accent type must either exclude speaker M1B or use only the last two paragraphs of the story. Likewise, the second paragraph can not be used in analyses that include speaker F3A. For these reasons, I chose to exclude the data for M1B and to compare the prosody of the first paragraph only.

12.3 Test categories

I compared the natural and synthetic versions of the news text mainly for the conditions described in Ross and Ostendorf [RO96]: (i) a boolean test for pitch accent location; (ii) a multi-valued tests for pitch accent type; and (iii) a multi-valued test for phrase final contour types. The category divisions for each test are shown in

[5] In the LDC distribution, this data is located in the Radio subdirectory of the Labnews data.
[6] Personal communication, Mari Ostendorf, July, 1998.
[7] Ross and Ostendorf [RO96] also use this data for their comparisons.

Table 12.1. Because LOQ does not produce downstep as a categorical feature, but rather, does so by varying prominence (a continuous variable) I altered the pitch accent test *ii* to consider three categories instead of four. I also added a six-valued test for pitch accent category and a boolean test for intonational phrase boundary location. The five tests and their categories are shown in Table 12.2.

Test	Categories (Ross & Ostendorf, 1996)	
(i) Pitch accent location (2 categories)	Unaccented	
	Accented	
(ii) Pitch accent type (4 categories)	Unaccented	
	High:	H*, L+H*, H+!H*
	Downstepped:	!H*, L+!H*, X*?
	Low:	L*, L*+H, L*+!H
(iii) Boundary tones (3 categories)	Falling:	L-L%
	Level:	H-L%
	Rising:	L-H%, H-H%

Table 12.1: Tests used by Ross and Ostendorf to compare the intonational predictions of their synthesis algorithm to the natural intonation. (The X*? notation indicates the presence of an accent but uncertainty about its type.)

12.4 Trends and variability in the natural prosody

According to one of the main premises of this work, it would be unsurprising to find individual differences among the speakers while still finding some overall correlation corresponding to a National Public Radio intonational style. Table 12.3 gives the pitch accent and intonational phrase counts per speaker for the one and five paragraph samples. It shows that speakers are consistent about whether they accent a lot or a little and that phrase and pitch accent counts are somewhat independent. For example, in both texts, speaker F3A assign the most pitch accents and divide the text into the fewest intonational phrases.

Figure 12-1 shows the mean and standard deviations for accent and boundary tone occurrences, for both the one and five paragraph samples. Both show minimal deviation for the mean counts of phrase contour types, and the greatest deviation for the mean counts of unaccented, H* and L+H* accented word. The main difference is the minimal presence of L*+H in the longer samples.

	Test	Categories	
(i)	Pitch accent location (2 categories) *(same as Ross & Ostendorf, (i))*	Unaccented Accented	
(ii)	Pitch accent type (3 categories) *(equivalent to Ross & Ostendorf, (ii))*	Unaccented High: Low:	 H*, L+H* H+!H*, !H*, L+!H*, X*? L*, L*+H, L*+!H, H+L*
(iii)	Pitch accent type (6 categories)	Unaccented H*: L*: L+H*: L*+H: H+L*:	 H*, H+!H*, !H*, X*? L* L+H*, L+!H* L*+H, L*+!H H+L*
(iv)	Intonational phrase boundary location (2 categories)	None Marked	
(v)	Phrase final tones *(Ross & Ostendorf, (iii))* (4 categories)	None Falling: Level: Rising:	 L L% H L% L H%, H H%

Table 12.2: Tests on NPR and LOQ data.

	1st paragraph 68 words, 4 sentences		5 paragraphs 369 words, 20 sentences	
Speaker	**Accents**	**Phrases**	**Accents**	**Phrases**
F1A	31	12	187	61
F2B	35	12	197	70
F3A	38	10	208	57
M2B	32	12	183	68
M3B	30	12	163	64

Table 12.3: Pitch accent and intonational phrase counts for five BU corpus speakers, for the one and five-paragraph excerpts from chief justice news story.

12.4.1 Pitch accents

For both samples, the order of accenting occurrences is the same:

(55) $unaccented > H* > L+H* > L* > L*+H > H+L*.$

A little over half the words in both samples are unaccented and the use of accents with L* main tones is extremely low (L*, L*+H) or non-existent (H+L*). This may be due to coder bias but could also be a stylistic feature of NPR prosody or newscasts in general. Consider that very little background information can be assumed to be shared by newscaster and listeners beforehand, hence, accents with H* tones should predominate and the L* accent should be rare.

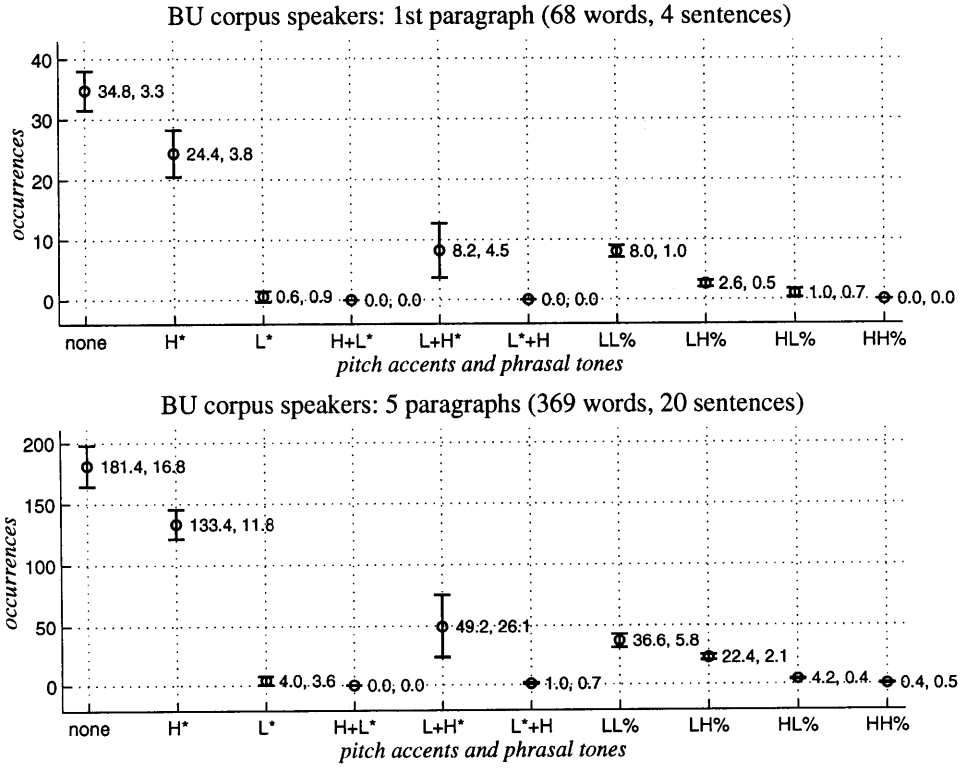

Figure 12-1: Accent and phrasal tone distributions for the natural data, for the one and five paragraph excerpts.

12.4.2 Intonational phrases

The ratios among phrase final contour occurrences are larger for the single paragraph. For example, the LL%:LH% ratio is approximately 3:1 for the first paragraph, and 1.5:1 for the five paragraphs. However, the order of occurrence is the same:

(56) $LL\% > LH\% > HL\% > HH\%$.

Falling contours (LL%) are the most frequent, followed by level contours (LH%). Rising contours are more rare – the HH% final contour is only used twice in the longer sample.

The strong agreement on phrase contour type is indicated by the small standard deviations (Figure 12-1). It also shows up in the kappa scores for each speaker in pairwise comparison with the other four. As shown in Figure 12-2, the greatest agreement among the speakers is for phrase boundary location and type. Even so, none of the kappas for these tests reach Krippendorff's significance criterion of .8. At best, the mean scores for the phrase boundary location test (*iv*) enter the "tentative conclusions only" zone, and for only three of the five speakers (F1A, F2B and M3B). Thus, even within these samples, there is sufficient variability such that no one speaker's intonation matches that of any other at or above significance.

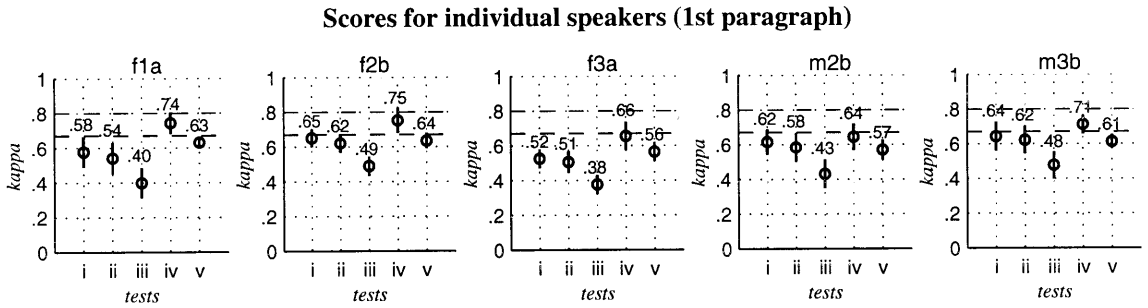

Figure 12-2: Mean and standard deviation for pairwise kappa comparisons for all speakers, for the first paragraph of the news story. As indicated by the dotted lines, significant kappa is above .8; possibly significant values are above .67.

12.4.3 Discussion

The data for the first paragraph show two strong trends. One is inter-speaker variability (also noted in [RO96]) – only rarely does one speaker's intonation match another's at significance. As shown by the standard deviations, only speakers F1A and F2B achieve significant kappas (above .8) with other speakers and only for the boundary location test. The other trend is the seeming independence of pitch accent and boundary location strategies. As shown in Table 12.3, a speaker who applies many pitch accents is not necessarily likely to mark many phrase boundaries. This shows up in the kappa scores. The ordering per speaker of mean scores for the accent tests is:

(57) $F2B > M3B > M2B > F1A > F3A$

while for phrase tests, it is:

(58) $F2B > F1A > M3B > F3A > M2B.$

Only speaker F2B (the author and original newscaster for the text) has the highest mean kappa scores for both pitch accent and phrase final phenomena.

Averaging over all individual pairwise comparisons shows that the location tests (*i* and *iv*) produce the highest kappa scores. This is not surprising, since they test only for presence or absence. However, it does suggest that the NPR prosodic style may not be distinct with regard to tone type. The mean scores for all tests are shown in Figure 12-3. Their order by decreasing kappa is:

(59) *iv* (.71) > *i* (.64) > *ii* (.62) > *v* (.61) >> *iii* (.48).

I will use these values to determine the minimum scores that a comparison between the LOQ and the natural output must attain to be counted as success.

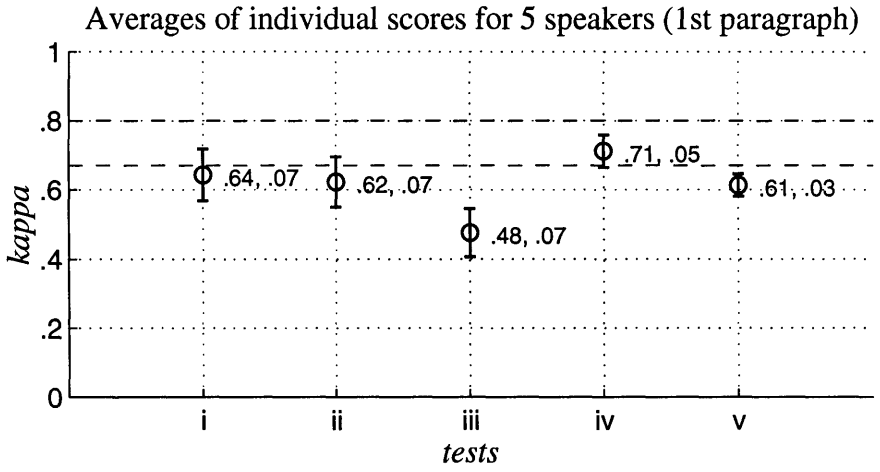

Figure 12-3: Mean and standard deviation for averaged pairwise kappa comparisons for all speakers, for the first paragraph of the news story.

12.5 Trends and variability in the LOQ prosody

The LOQ data consists of simulations for three memory sizes, three pointer step sizes and all the radii up to the maximum that each memory can accommodate. Five simulations were run for each parameter configuration yielding the totals shown in Table 12.4. Of the three parameters, memory size has the least (and negligible) influence on the kappa values. Therefore, data are reported only for the 50x50 memories, which most comprehensively show the behavior of the system as the radius increases.

Memory size		Total simulations (radii x 5 speakers x 3 step sizes)
small:	10x10	10 x 5 x 3 = 150
medium:	22x22	22 x 5 x 3 = 330
large:	50x50	50 x 5 x 3 = 750

Table 12.4: LOQ simulation parameters and the total number of runs.

12.5.1 Prominence

Initially, LOQ produces prominence and tone for every word and phrase boundary, thereby expressing the observation that every linguistic item requires some processing. However, in most speech genres, not all words are accented and not all possible phrase breaks are realized. Therefore, to generate intonation within a more natural range, LOQ includes methods for reducing the prominence of a pitch accent by prior occurrence probabilities for the word (Section 9.1.2). LOQ also imposes a minimum prominence, set conservatively low at .05 (Sections 9.1.2 and 9.2.2). With prominence below the threshold, words are not accented and phrase tones are not realized.

12.5.2 Pitch accents

The mean distributions[8] of accenting phenomena as a function of search radius and step size are shown in Figure 12-4. Judging by the area spanning all radii and step sizes, the most common order by frequency of occurrence is:

$$(60)\ \textit{unaccented} > H* > L* > H + L* > L + H* > L* + H.$$

As with the natural data, the unaccented and H* accented words occur most frequently. With the exception of the lower radii, the unaccented count is always greatest. The most noticeable difference between the natural and LOQ distributions are that the L+H* accent in the natural data shares the same position and ratio as the L* accent in the LOQ data. This, along with the different ordering among the rare accents suggests that the LOQ pitch accent mapping needs some revision to better accommodate the NPR style.

Figure 12-5 shows the mean and standard deviations of the kappa scores for the pitch accent tests. Since the standard deviations are small, discussions of trends are meaningful. The main trend for the pitch accent tests (*i*, *ii* and *iii*) is that the

[8]Mean distributions only are shown since the standard deviations are small.

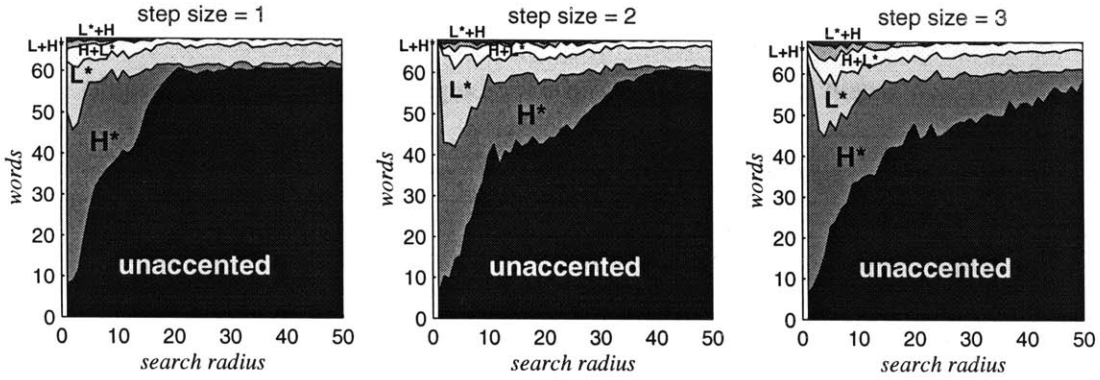

Figure 12-4: Mean distributions for accenting phenomena as a function of search radius and step size.

mean kappa scores are lower for the lowest radii and their standard deviation wider. This reflects the wider variation of accent location and type that typifies the prosody generated from the smaller radii.

As with the other prosodic features generated by the LOQ mapping, the effect of larger step sizes on intonation is to expand the range of lower radii in which the greatest individual variation occurs. Greater variation maps to lower kappa scores, thus raising the radius at which the kappa values reach significance. For the pitch accent tests, the first significant kappa occurs at a radius of 19 for a step size of one, at 35 for a step size of two and at 50 for a step size of three. As discussed earlier, the outlier behavior of the model for a search radius of one and a step size of two produces identical intonation and prominence for every run, and therefore a kappa score of 1 for every test.

12.5.3 Intonational phrases

Figure 12-6 shows that the phrase boundary type distributions exhibit the same order as in the natural speech:

$$\texttt{(61)}\ \ LL\% > LH\% > HL\% \ > HH\%.$$

However, the ratio of falling to level contour occurrences is smaller (1:1 versus 3:1 in the natural data). In addition, far more intonational phrase boundaries occur in the LOQ data – twenty-eight occurrences for all radii above 1, versus a mode of twelve in the natural data. This figure is slightly less than the number of full grammatical clauses (thirty-four) into which the first paragraph is divided (see Table 11.2). This difference alone will lower the kappa scores for the phrase tests in comparisons between the LOQ and natural data.

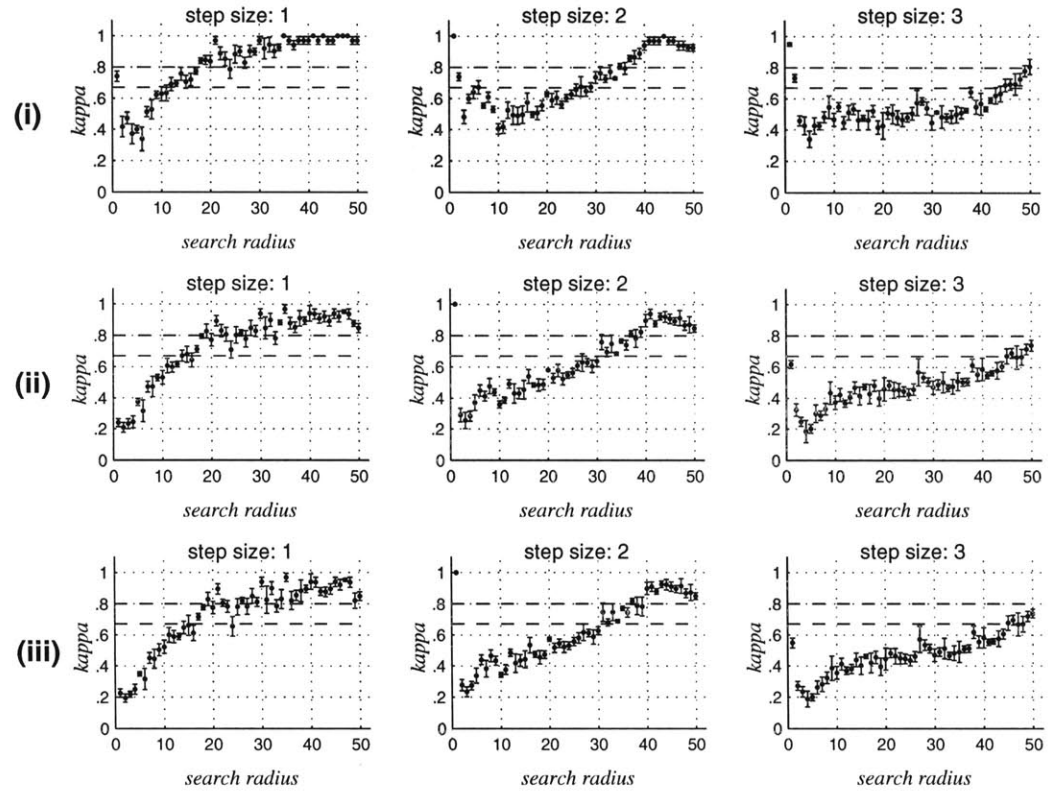

Figure 12-5: Mean and standard deviation of the kappa scores for the pitch accent tests for the LOQ simulations of the NPR news story (first paragraph). Five simulations were run per data point.

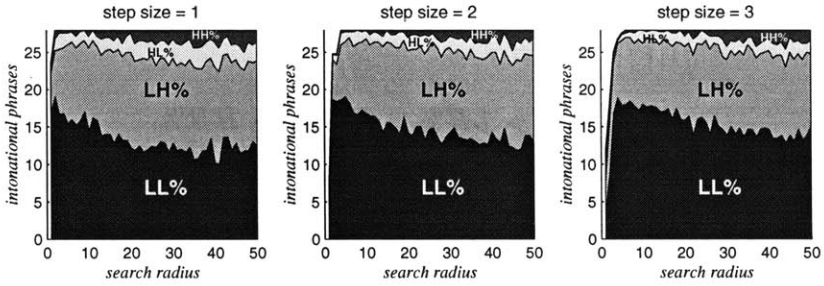

Figure 12-6: Mean distributions of phrase contour types in LOQ simulations as a function of search radius and step size.

As with the pitch accent tests, though to a much lesser extent, larger step sizes expand the radius range in which kappa is not significant, again reflecting greatest within-style variation for the lower radii. As shown in Figure 12-7, most of the kappa values for the LOQ boundary phenomena are significant – almost all the boundary location scores are 1, indicating perfect agreement. In contrast, the scores of the natural data approach significance for the location of the boundary tone, but not for its type. Thus, the natural data shows more variability of phrase boundary location and type.

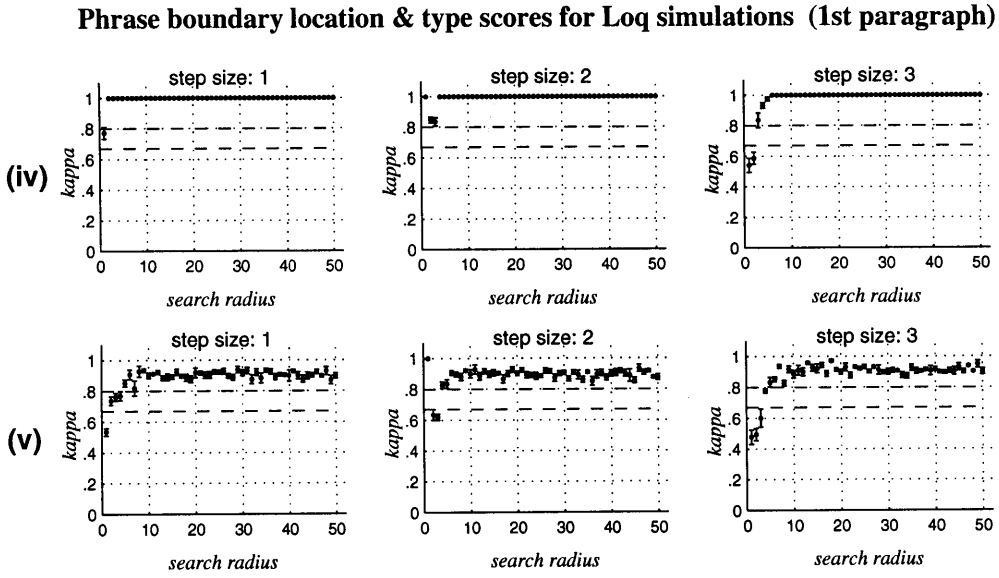

Figure 12-7: Mean and standard deviation of kappa scores for the phrase location and boundary tests on the LOQ versions of the NPR news story (first paragraph).

12.5.4 Discussion

A few observations about the differences between the natural and synthesized versions of the same text have been made in this section. The first concerns the ordering of accenting phenomena. Both the LOQ and natural versions show mostly unaccented words, followed by H* as the next most numerous accenting phenomenon. However, the natural speech has fewer accents with L* tones, and the LOQ speech has few L+H* accents. The second observation is that, although the order of occurrence for phrase final contour types is the same for the natural and simulated speakers, the LOQ simulations produce many more intonational phrases. This alone will lower the phrase test scores in the comparisons of LOQ and natural intonation.

12.6 Kappa comparison between LOQ and natural intonation

In this section I describe the results of using the kappa statistic to compare the LOQ and natural intonation. The *a priori* prediction is that the NPR intonational style will be a combination of knowledgeable (because the speakers are familiar with the text) and expressive (because they are communicating with an audience that is not). This predicts the best matches at the boundary between the two styles, approximately at a radius of ten. On this count, the pitch accent tests succeed. Their scores above the match criterion are not random but rather, fall mainly within the radius range from seven to fourteen, with the most and best matches for a radius of nine. Fewer comparisons achieve a matching score on the boundary tests, mainly because of the greater number of phrase boundaries in the LOQ output. Therefore, it is not possible to claim that LOQ has successfully imitated the full intonation of any one NPR speaker. It may be that none of the LOQ styles are adequate and that other genres besides NPR newscasts are more appropriate for comparison.

12.6.1 Method

The comparisons are pairwise. Every LOQ intonational mark up is compared to that of each of the five human speakers for a total of 3750 comparisons in all. The aim is to locate the LOQ parameters that consistently give the best kappa scores. Neither the extreme of all significant scores or no significant scores is desirable. The first outcome would show that LOQ only produces one prosodic style regardless (especially) of search radius; the second would show that it produces no matches at all. Therefore, the search is instead for parameter clusters that produce the best kappa scores.

I apply the kappa statistic in four passes such that each raises the threshold that classifies a score as success. The first and second criteria use the performance of the natural speakers (Figure 12-3) as the lower bound for success. The first is the most lenient and counts as success any scores at or above the mean score less one standard deviation, per test. The second counts as success any score at or above the mean score (per test) in the averaged pairwise comparisons for the natural data. The third and fourth criteria use Krippendorff's standards and count as success any test scores above possible significance (.67) and significance (.8). The lower bounds per criterion and per test are shown in Table 12.5.

Figure 12-8 shows that no comparison achieved a kappa above .8 and very few (nine in all) achieved kappa above .67. Therefore, I mainly discuss the results for criteria #1 (mean− standard deviation), #2 (mean) and #3 (possibly significant). In addition, For all four thresholds, test *v*, on phrase final contour type, produced no scores above the lowest threshold, at .56. Therefore, I mainly discuss the results for tests *i* through

Criterion	Accent test			Phrase test	
	(*i*)	(*ii*)	(*iii*)	(*iv*)	(*v*)
#1 (mean − standard deviation)	.57	.53	.41	.66	.56
#2 (mean)	.64	.62	.48	.71	.61
#3 (possibly significant)	.67	.67	.67	.67	.67
#4 (significant)	.8	.8	.8	.8	.8

Table 12.5: Minimum kappa scores for the four threshold criteria.

iv – all the pitch accent tests, and the test on the location of intonational phrase boundaries.

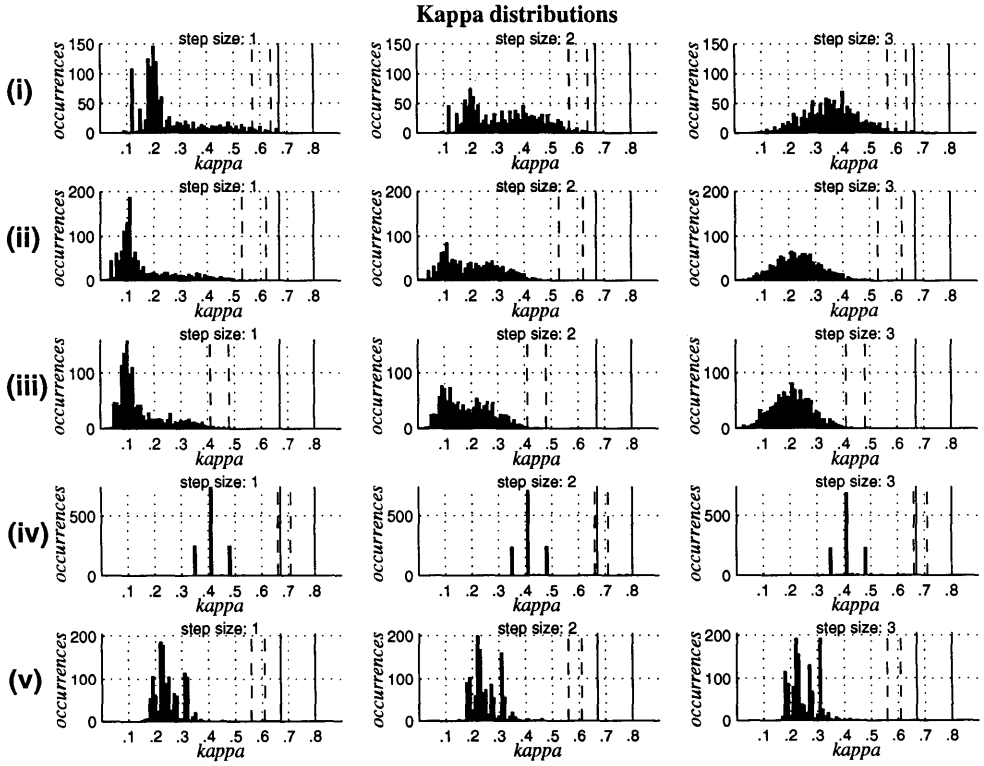

Figure 12-8: Counts of kappa values per radius. Dotted lines indicate the minimum values for criteria #1 and #2. The solid lines mark Krippendorff's tentative and actual significance criteria of .67 and .8, respectively.

12.6.2 General characteristics of the data

As with other prosodic features, the general effect of larger pointer step sizes on intonation is to increase the range of radii that produce interesting variation. This effect is complicated by the interaction of the Cartesian topology and even step sizes,

which produces a checkerboard distribution of items in the AWM′ space. In general, trends that change monotonically with increasing step size show the influence of increased sparseness, trends that change nonmonotonically with increasing step size show the effect of decreased global randomness for the even step sizes.

Figure 12-8 illustrates trends that are specific to the kappa comparison. One is that the mean and mode score increase with step size. However, the overall range decreases slightly with increased step size. Another trend is that the phrase boundary test scores exhibit smaller ranges and less variation within the range, mainly because there is very little variation in the LOQ phrase boundary features to begin with. Finally, the figure shows that few of the scores exceed even the minimum match thresholds.

12.6.3 Results

What kind of LOQ output is the NPR newscasting style(s) most likely to resemble? Speaker F2B wrote and delivered the original newscast and so was the most familiar with the text. However, by the time of the off-air radio recordings, all the speakers had already recorded the story once in their non-radio style and were also familiar with the text. A reasonable prediction is that the natural intonation would most resemble the LOQ intonation of a knowledgeable speaker. However, because the intent is to communicate with listeners unfamiliar with the material, one would also expect some adaptation toward the more expressive styles, as generated via the smaller radii.

Pitch accents

As predicted, the highest mean and standard deviations occur between the radii of seven and fourteen. Figure 12-9 shows that this effect is strongest for a step size of one, indicating that the temporal proximity it preserves is best suited to the NPR styles. Figure 12-10 shows that pitch accent location scores that pass criterion #1 are for simulations with radii that cluster around ten and with peaks at nine and thirteen (the larger step sizes contribute matches at the largest radii). It also reveals that simulations with a radius of nine are the most successful, with admissible scores for the pitch accent type tests and criteria #2 and #3 as well. The matches are not evenly distributed for the natural speakers. As Figure 12-11 shows, the number of matches tends to be largest for speakers M2B, F3A and occasionally, F1A.

Table 12.6 examines these results in more detail. Most importantly, it shows that these patterns occur jointly for the pitch accent location tests. Simulations for a step size of one and radii of nine to thirteen most frequently match the pitch accent location patterns of M2B and F3A, and often do so twice (out of five possible).

It is encouraging that the predicted patterns exist. However, at best they do so for

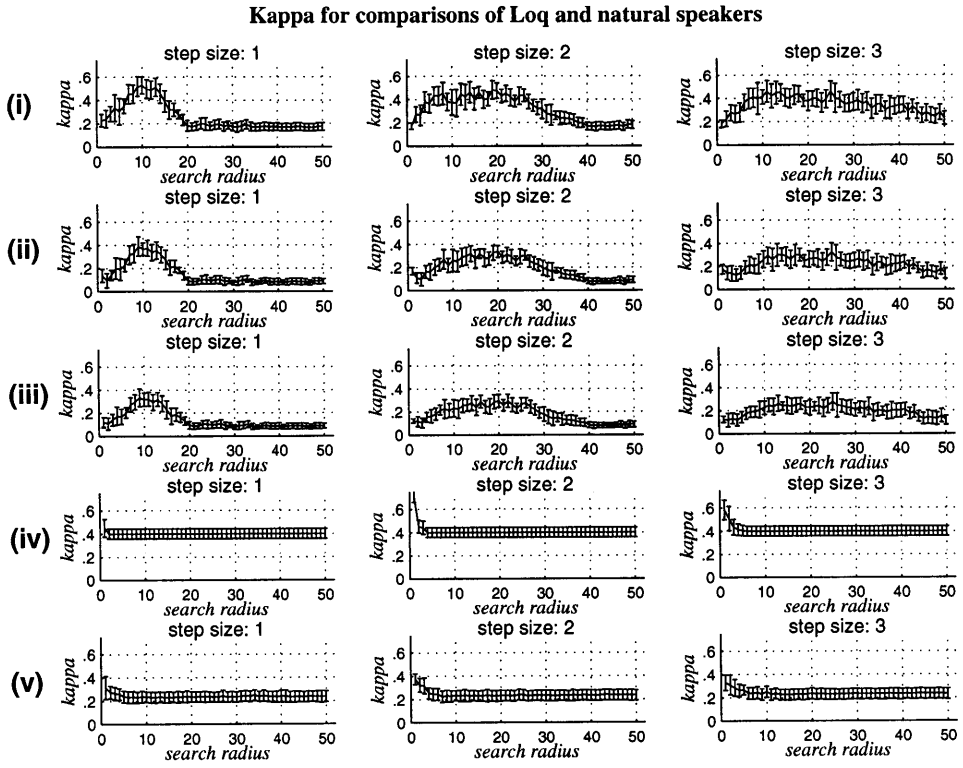

Figure 12-9: Mean and standard deviation kappa values, per radius, for pairwise comparisons between LOQ and natural speakers.

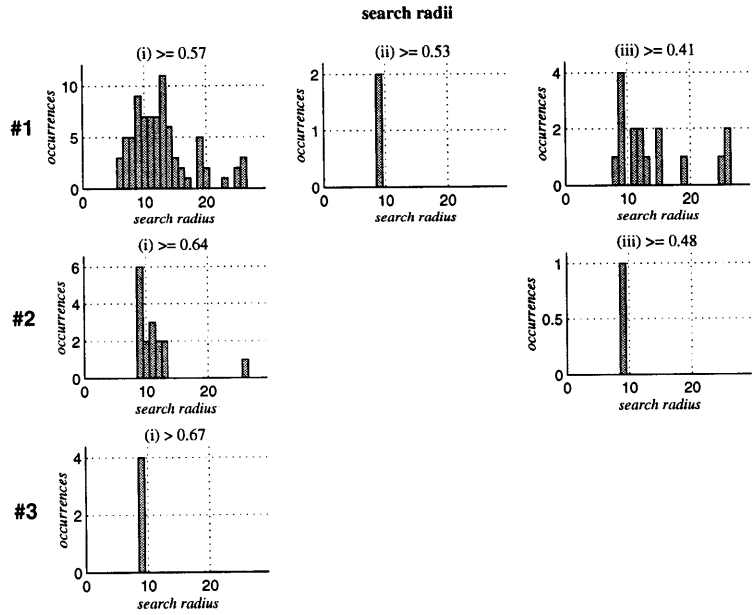

Figure 12-10: Pitch accent test results by radius (for all step sizes) and for the first three match criterion.

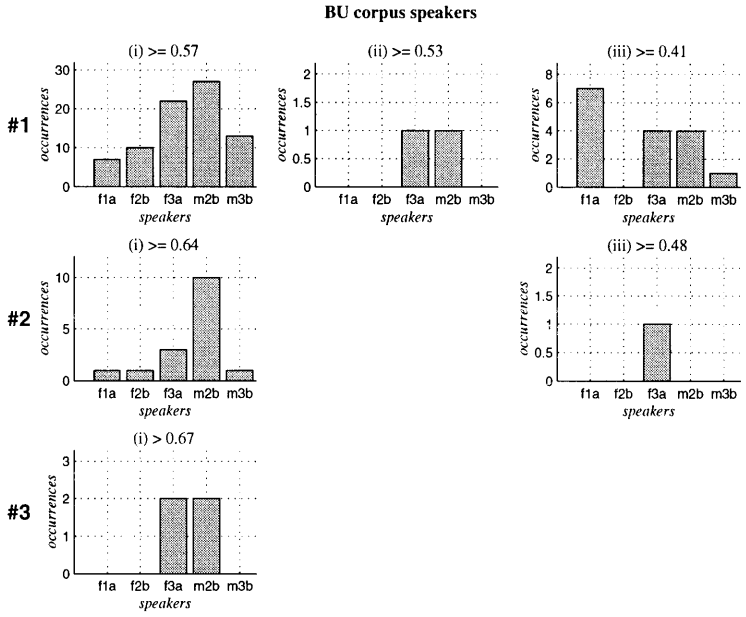

Figure 12-11: Pitch accent test results according to the speaker they best match, usually speakers M2B and F3A, for all radii and all step sizes.

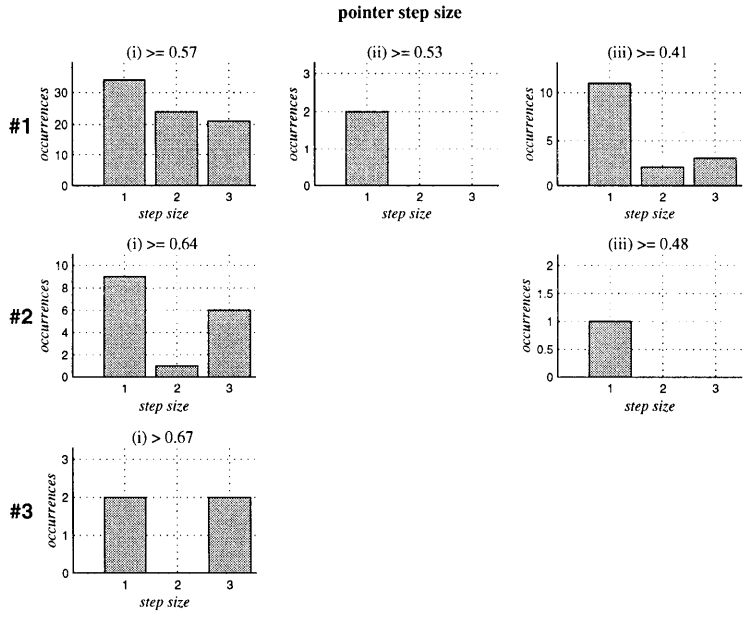

Figure 12-12: Pitch accent tests by step size for the first three match criterion. The step size of one consistently produces the greatest number of matches.

Speaker	Occurrences	#1 $(i) \geq .57$ step size 1	2	3	#2 $(i) \geq .64$ step size 1	2	3	#3 $(i) > .67$ step size 1	2	3
F1A	1	11 12	8 19 23	25 26	—	—	26	—	—	—
F2B	1	7 10 11 13	—	9 10 13 14	—	—	9	—	—	—
	2	—	13	—	—	—	—	—	—	—
F3A	1	7 11 12 13 14	7 8 15 19	7 9 12 13	9 13	—	9	9	—	9
	2	9 10	6	14 17 26	—	—	—	—	—	—
M2B	1	7 8 11	6 8 12 13 15 16	9 11 12 13 16 26	11 12 13	12	9	9 11	—	9
	2	9 10 12 14	19	—	9 10	—	—	—	—	—
M3B	1	8 10 13 15	11 14 19	11 25	—	—	11	—	—	—
	2	9	20	—	—	—	—	—	—	—

Table 12.6: Matches on pitch accent location (test *(ii)*) for each of the three step sizes times each of the three match criteria, and by count (1 or 2) for the number of matches per radius.

less than half of simulations for which matches are predicted. This is in part due to the emphasis on individual variation within a style – mainly from the pointer's random walk. In all the totals are low and are greatest for the lowest threshold. They are reported in Table 12.7.

Criterion	**Accent test**			**Phrase test**	**Totals**
	(*i*)	(*ii*)	(*iii*)	(*iv*)	
#1 (mean − standard deviation)	79	2	16	5	102
#2 (mean)	16	0	1	2	19
#3 (possibly significant)	4	0	0	(5)	9

Table 12.7: Simulations that match according to the lowest three match criteria.

Intonational phrases

The phrase boundary type test (test *v*) yields no hits and as shown in Figure 12-13, the boundary location test yields very few, and only for the lowest radii. This is due entirely to the slightly fewer intonational phrase boundaries produced by LOQ for a radius of one (twenty-five versus twenty-eight).

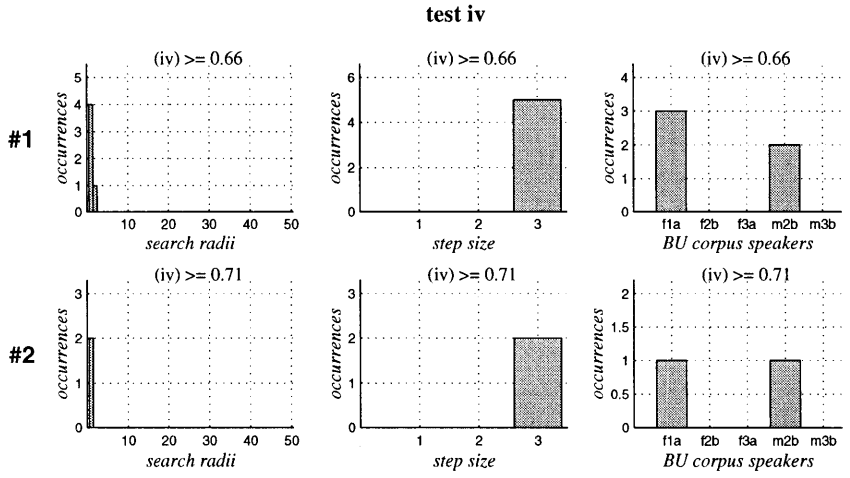

Figure 12-13: Phrase boundary location test results.

Although useful as a diagnostic for the phrase boundary algorithms, the relative success for the radius of one is not meaningful. To see how lowering the number of intonational phrase boundaries would affect the scores, I removed all phrasal tones whose prominence was below .1, .2 and .5 in three independent passes, while preserving the syntax of phrase final contours (boundary tones whose prominence was at or above the threshold were only counted if they were also preceded by a phrase accent with prominence also at or above). This accounts for the possibility that the

coders applied some kind of threshold below which F0 fluctuations were not treated as meaningful but as noise.

The first two passes provided very little reduction in the number of phrase boundaries. The third may have removed too many but produced more matching scores. As shown in Figure 12-14 simulations with all radii produce matching scores. This appears to correlate with the total number of boundaries. Recall that the mode for the natural speakers is twelve. Most of the simulations for a step size of one produce twenty-three phrases and therefore, the fewest matching scores. The simulations for step sizes two and three, typically produce eight and eleven intonational phrases, respectively, and therefore have the greater number of matching scores.

Figure 12-15, shows that some matches were produced for the contour type test (v) and moreover, that they were produced mainly for the lowest radii. As with the pitch accent tests, speakers M2B, F3A and F1A had the best scores for both phrase tests. This, along with the significant boundary location test scores for radii between one and twelve, suggests that with a better mapping for the phrasal tones, it may be possible to more closely match the intonational output of one speaker. One difference between the accent and phrase tests is that the best results for the latter are for step sizes two and three, rather than for one. This correlates with the number of boundaries. Clearly raising the minimum prominence for phrasal tones is only partly successful for the NPR style.

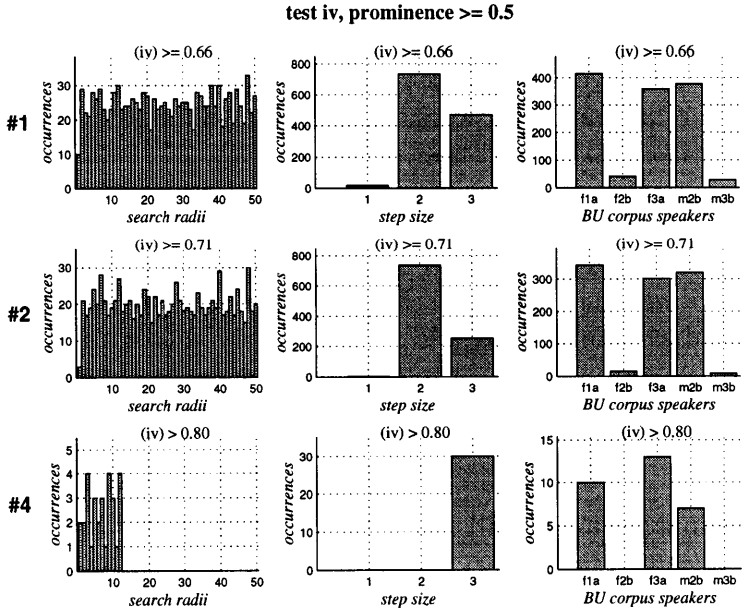

Figure 12-14: Results for test *iv*.

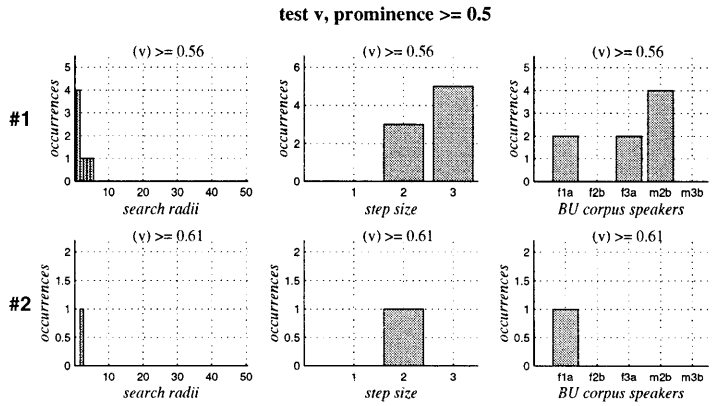

Figure 12-15: Results for test *iv*.

12.7 Discussion

The comparisons produce some of the predicted patterns, but in low numbers. On the positive side, the simulation parameters that produce the best pitch accent scores are, as predicted, at the boundary between the expressive and knowledgeable styles. However, because at best, only two simulations match out of a possible five for any one LOQ parameter configuration, this is only suggestive rather than conclusive. One argument against expecting more is that it is a priority for LOQ to produce individual variation within styles.

The kappa statistic is a strict criterion because agreement is offset by the number of disagreements. Thus, very few of the best scores approach significance. On this count, the LOQ simulations do no worse than the lowest of the scores of the natural speakers in comparison with each other.

Restricting the admissible prominence values produces better scores on the phrase tests. This suggests that one way to bring the LOQ style closer to the NPR styles is to impose different minimum prominence thresholds for pitch accents and phrasal tones. That accenting and phrasing strategies might be independent or at least different is also suggested by the natural data (Table 12.3). However, the prominence algorithms may need revision independently of whether they apply to pitch accents or phrasal tones.

According to the means (and small standard deviations), the LOQ prominence algorithms for pitch accents are more successful, and except for the lowest radii, tend to duplicate the trends in the natural data for unaccented versus accented words. The tone mapping algorithms are the most successful for the phrasal tones, showing the same order of distribution as in the natural data. They also appear to produce the same overall distribution ratios for the H* accent (excepting those of the lowest radii). However, the order of occurrence for the more rare accents differs.

12.8 Conclusion

Intonation is the most closely related to text content and is the main categorical feature of prosody. Therefore, it is an appropriate basis for a comparison. Using mainly the kappa statistic, the results are encouraging but not conclusive. However, the perception of naturalness (and hence its evaluation) may respond as well to quantitative aspects of prosody, such as prominence, duration and pitch range. For example, the phrasal tones removed by raising the minimum prominence may have a perceptible influence on naturalness but not on phrasing. In all, the correct test for evaluating naturalness or its components has not yet been designed.

Part IV

Conclusion

Chapter 13

Contributions, Future Directions, Conclusion

This dissertation investigates the claim that prosodic variation is correlated with attentional capacity, thus linking two phenomena: (1) the ubiquitous prosodic variation that occurs within and among speakers; (2) the established correlation of intonation and timing with salience.

13.1 Main contribution

Using a simulation system, I demonstrate the correlation of attentional and storage capacities in working memory to prosody. An important part of the implementation is to devise a reasonable mapping from search and storage in working memory to pitch and timing in speech. The one I develop is designed to reproduce the empirical findings whenever possible. Since most of the research that correlates attention and prosody is not phrased in terms of limited attentional capacity, the strongest claim is that the mapping is plausible.

Nonetheless, it is well-founded enough to produce prosody that varies consistently and plausibly with attentional capacity. Simulations of speakers with minimal attention produce a child-like prosody, while simulations of speakers with maximal attention produce prosody that sounds adult and knowledgeable, if not bored. Thus, the main contribution of this work is to demonstrate a correlation between attentional capacity and prosody, and by doing so, to expand the expressive range of synthetic speech.

13.2 Additional contributions

The work produces additional contributions as a consequence of designing and building the simulation system: a cognitively-based architecture for speech applications and an extended model of attention and working memory.

13.2.1 A cognitively-based architecture for speech applications

The LOQ design specifies a cognitively-based architecture for speech generation that applies to both text-to-speech and concept-to-speech synthesis. Its main conceptual contribution is to include both working and long term memory components in one system. The working memory component models the influence of attentional and storage capacities. The long term memory components model the speaker's background knowledge about the structure, function and use of language.

The more specific design contributions are by example. LOQ makes extensive use of online linguistic databases as a means of incorporating domain knowledge into its operations. It identifies the match criteria as the (domain-specific) key to distinguishing between given and new information and uses a match threshold as a simple binary detector of a match.

The system is implemented as a text-to-speech system. However, because it relies on annotated text, the output of a text generation system could easily replace the pre-selected and pre-analyzed texts used by the current implementation.

13.2.2 An extended model of attention and working memory

The appeal of Landauer's spatial model of attention and working memory is that it reproduces key features of learning and interaction from simple and locally applied rules. Its main problem is that its storage capacity is underspecified. Addressing this, AWM′ restricts storage to one item per node. The result is that compression and expansion[1] become part of the update rule for the model.

As Landauer notes, his model stops short of transferring items into long term memory. The addition of the compression step motivates such transfers more precisely – items

[1] As previously noted, the expansion operation is not explicitly implemented in AWM′ because the associated cognitive claims are unclear. However, retrieval from a compressed memory item is still possible via tree descent.

that are fully compressed are probably the least salient. Therefore, they may be retired from working memory.

A second extension is the attribution of intrinsic state to the items in memory. In the original model, the state of a memory item is its memory address. In AWM', each item also has an intrinsic state that reflects its participation in memory computations. The component of intrinsic state consists of two boolean variables that determine the compressibility status of the item. The value of the first variable indicates whether the item is part of a structure that is still under construction. If so, it is treated as salient by the model and consequently cannot be compressed. The second variable indicates whether one compression attempt has already been made. It is included to delay the compression of recently stored items.

Another state variable for a memory item depicts the result of the most recent comparison of that item with a stimulus. It too is binary-valued. Its value is L if the memory item has primed for a stimulus, and H if it has not. Because its value determines the tone type, this schema provides a way for previous stimuli to influence current processing. It adds to the model a kind of cause and effect for memory processing that is not present in prior implementations.

13.3 Future Directions

Both the implementation and the subject matter are rich in directions for further exploration. They include: (1) refining the current implementation; (2) extending the architecture to include other processing components; (3) extending the model of attention and working memory.

13.3.1 Refining the current system

The refinements to the current implementation take three forms: (1) varying the fixed parameters; (2) improving the mapping algorithms; (3) enriching the linguistic data and text representations. The parameter testing consists of:

- Varying the fixed parameters to determine their full effect and to delineate a reasonable range of values. The main candidates for this phase are the match threshold, the prominence threshold and the ranges for the acoustical values into which prosodic specifications are finally mapped.

- Testing the effects of changing the biases in the inequalities that derive tone type.

Improvements to the mapping algorithms have been discussed in detail in Chapter 9. Aside from parameter tuning, they mainly consist of exploring alternative algorithms for calculating the phrase accents, and of a better means for deriving pauses from empty categories, especially the deletions that follow conjunctives (see Section 9.4).

Both the linguistic data and the textual representations can be enriched, often by replacement with components that are more empirically accurate and more theoretically sound. The enrichment steps include:

- Adding more online databases as they become available.
- Replacing the manual analyses with natural language processing tools as they become available. What is especially needed are tools that can produce a shallow grammatical parse and identify subcategorization and cohesive links within and among clauses.
- Replacing the current compression strategy (chunking and deletion) with true compression, using techniques based on classical information theory. This step carries with it the pre-requisite that the text structure and content are represented as bit strings. Such representations are more amenable to the precise calculation of mutual information, and have the additional advantage of being universal. Their use generalizes the model so that it can apply to complex data from realms other than language and speech.
- Determining a data-driven set of match criteria and their relative importance to identifying a match. Machine learning techniques are probably ideal for this step.

13.3.2 Extending the architecture for speech processes and production

Currently, LOQ operates on the assumption that both reading and speaking are trouble-free and that the only limited resources are attention and storage in working memory. However, a limited resource approach extends quite readily to include the influence of variable reading spans and the limited availability of the speech articulators. Integrating both as dynamic components of a simulation system incorporates the influence of perception and physiology.

Operationally, the addition of the reading and articulation modules adds feedback links among the perceptual, memory and articulation component. This adds more dynamics to the system and may perhaps help to increase the number of speaking styles that LOQ is able to generate. Certainly, it is an extension consistent with spontaneous speech processes, which often show the effect of race conditions between articulation, perceptual and cognitive functions [Lev89].

13.3.3 Alterations and extensions of the working memory model

The alterations I suggest are motivated by the limitations of the current model as revealed by the simulation results. The extensions are motivated by the need for greater coverage of speech and memory processes.

Because the current AWM′ spaces are both two dimensional and Cartesian, even pointer step sizes produce an invariant checkerboard pattern of storage. This works against the more varied distribution pattern that would normally occur from the pointer's random walk. In addition, its effect is that no items in memory are accessible when the step size is larger than the search radius. Two alterations are suggested. The first is to use three dimensional models, as Landauer and Walker do. This allows further exploration of the curious fact that dimensional differences do not produce prosodic differences.[2] The second alteration retains the two dimensional space but replaces the Cartesian grid with other regular topologies, such those are that triangular (three intersection per node) or octagonal (four intersections).

In hopes of expanding the stylistic range, the processing within the AWM′ component can be made more dynamic by turning the static parameters into adaptive ones. The main candidates for this alteration are the search radius, the pointer step size and the match threshold.

The explicit inclusion of lookahead is an extension that better reflects speech processing. Currently, LOQ instantiates lookahead in two ways: (1) by manual annotation to clauses for the function of upcoming punctuation; (2) by taking into account the associational expectations that are typical for linguistic constituents. An explicit mechanism for lookahead may be also approximated with a delay between the time that a stimulus is processed and the time that its processing is mapped to prosody. This allows some of the subsequent processing to affect its prosody. However, the extent to which subsequent stimuli influence the processing of earlier stimuli is an open question.

The most far-reaching functional extension is the inclusion of criteria and mechanisms for transferring items between working memory and long term memory. The compression operation in AWM′ provides a partial criterion – a fully compressed item is a more likely candidate for transfer. However, it is further addressed with the addition of parallel processing such that items in memory interact with each other as well as with the stimulus, perhaps to produce combinations that are appropriate for storage in long term memory.

[2]The mapping may obscure the differences between multi-dimensional spaces.

13.4 Conclusion

In this thesis I have developed and implemented a model for producing prosodic variation as the consequence of the processing capacities of the simulated speaker. Its dynamic memory component makes it a production model. Its focus on individually varying attentional and working memory capacities makes it a performance model. It demonstrates that prosody is not determined solely by the text but also by the particular speaker who utters it.

The model produces three recognizable styles that correlate with attentional capacity: a child-like style for smallest capacities, an adult narrative style for medium capacities, and a knowledgeable style for the largest capacities. It also produces prosodic variation within each style for multiple readings by the same simulated speaker. Although its prosody has not matched natural prosody at significance, the closest matches show clear trends based on attentional differences. Future work will reveal whether the model mainly needs refinement or whether it should also be augmented to include other processing components. In conclusion, the current work advances both the explanation of prosodic variation in human speech and the methods for reproducing it in synthetic speech.

Appendix A

Match Criteria and Scores

IDENTITY

major category	minor category	predicate
Co-reference (9/45)	Grammatical role (1/6)	grammatical-role(x) = grammatical-role(Y).
	Co-reference forms (1/2)	referent(X) = referent(Y). referent($X_{unlexicalized}$) = referent($Y_{unlexicalized}$). referent($X_{unlexicalized}$) = referent(Y). referent(X) = referent(Y) AND form(X) $\neq$ form(Y). referent(X) = referent($Y_{unlexicalized}$).
	Coreference form changes (1/3)	form(X) $\leq$ form(Y). form(X) = form(Y). form(X) $\geq$ form(Y).
Form (8/45)	Orthography (2/3)	orthography(X) = orthography(Y). orthography(stem(X)) = orthography(stem(Y)).
	Acoustic (1/3)	pronunciation(X) = pronunciation(Y). coda(X) = coda(Y). $coda_{compressed}$(X) = $coda_{compressed}$(Y). phonemes(X) = phonemes(Y). stress(X) = stress(Y).
Semantic Identity (6/45)	Sibling (2/3)	X = synonym(Y). X = antonym(Y). antonym(X) = antonym(Y). parent(X) = parent(Y). child(X) = child(Y).
	Cross-generational (1/3)	X = child(Y). X = parent(Y). category(X) = category(Y). category(X) is related to category(Y).

Table A.1: Predicates in order of application and their contribution to the total match score. X is the stimulus; Y is an item in memory;

ASSOCIATION/COMPOUND CUE

major category	minor category	predicate
Collocation 7/45	Predicted by prior usage in the text (1)	Y → X. Y → stem(X). stem(Y) → X. stem(Y) → stem(X). Y → syntactic-category(X). stem(Y) → syntactic-category(X). syntactic-category(Y) → X. syntactic-category(Y) → stem(X). syntactic-category(Y) → syntactic-category(X).
Semantic Association (5/45)	Part of (4/10)	X = part-of(Y). X = part-of(Y). part-of(X) = part-of(Y). X = part-of(Z) AND Y = part-of(Z).
	Causes (3/10)	X causes Y. Y causes X. X causes Z AND Y causes Z.
	Entails (2/10)	X entails Y. Y entails X Y. X entails Z AND Y entails Z.
	Attribute (1/10)	X = attribute(Y) OR Y = attribute(X).

Table A.2: Predicates in order of application and their contribution to the total match score. X is the stimulus; Y is in memory; Z is either in memory or in one of the online databases.

ASSOCIATION/COMPOUND CUE

major category	minor category	predicate
Collocation (weak) (4/45)	herald (1)	X → Y. X → parent(Y).
Structural Association (3/45)	Argument Relations (10/30)	*argument-relation(X,Y). X = argument(Y). Y = argument(X). X = argument(Z) AND Y = argument(Z). argument(X) = argument(Y).
	Modifier Relations (8/30)	*modifier-relation(X,Y). X = modifier(Y). Y = modifier(X). X = modifier(Z) AND Y = modifier(Z). mod(X) = mod(Y).
	Anchoring (6/30)	*anchor-relation(X,Y). X = anchor(Y). Y = anchor(X). X =anchor(Z) AND Y = anchor(Z). anchor(X) = anchor(Y).
	Related to same item (4/30)	*cohesion-relation(X,Y). relation1(X,Z) ≠ relation2(Y,Z). and relation1 and relation2 are: argument *of*, modifier *of* or anchor *of*. relation1(X,Z) ≠ relation2(Y,Z) and they are : *has* argument, *has* modifier, or *has* anchor. relation1(X,Z) ≠ relation2(Y,Z) and none of the above conditions hold.
	Same relation, different item (2/30)	*cohesion-analog-relation(X,Y). relation1(X,Z)≠ relation2(Y,ZZ) and the relations are both either *of* or *has* (as in the cohesion-relation above). relation1(X,Z) ≠ relation2(Y,ZZ) and none of the above conditions hold.

Table A.3: Predicates in order of application and their contribution to the total match score. X is the stimulus; Y is in memory; Z and ZZ are either in memory or in one of the online databases. "*" indicates an additive score. For example, if X is an argument of Y, both the *argument-relation* and the X = *argument*(Y) predicates contribute to the total match score.

INTENSIFIER (WEAK IDENTITY)

major category	minor category	predicate
morphology and syntax (2/45)	inflection (4/10)	inflection(X) = inflection(Y).
	wn-verb (3/10)	case-frame(X) = case-frame(Y).
	Word Role (2/10)	X and Y head same phrase type and at the same depth. X and Y head the same type of syntactic phrase. X = phrase-head AND Y = phrase-head.
	Syntactic (1/10)	syntactic category(X) = syntactic category(Y) . X and Y are phrases, OR X and Y are orthographic tokens.
parallel structure (1/45)	tree-struct (2/3)	tree-structure(X) = tree-structure(Y). tree-structure(X) is similar to tree-structure(Y).
	Word Position (1/3)	position(X) = position(Y) in their respective phrases. X and Y hold equivalent positions in their respective phrases.

Table A.4: Predicates in order of application and their contribution to the total match score. X is the stimulus; Y is in memory.

Appendix B

The texts

B.1 Fiction

Excerpt from *One Hundred Years of Solitude* by Gabriel Garcia-Marquez, (pages 242–243).

Remedios the Beauty stayed there wandering through the desert of solitude, bearing no cross on her back, maturing in her dreams without nightmares, her interminable baths, her unscheduled meals, her deep and prolonged silences that had no memory until one afternoon in March, when Fernanda wanted to fold her brabant sheets in the garden and asked the women in the house for help.

She had just begun when Amaranta noticed that Remedios the Beauty was covered all over by an intense paleness. "Don't you feel well?" she asked her. Remedios the Beauty, who was clutching the sheet by the other end, gave a pitying smile. "Quite the opposite." she said, "I never felt better."

She had just finished saying it when Fernanda felt a delicate wind of light pull the sheets out of her hands and open them up wide. Amaranta felt a mysterious trembling in the lace on her petticoats and she tried to grasp the sheet so that she would not fall down at the instant in which Remedios the Beauty began to rise.

Ursula, almost blind at the time, was the only person who was sufficiently calm to identify the nature of that determined wind and she left the sheets to the mercy of the light as she watched Remedios the Beauty waving good-bye in the midst of the flapping sheets that rose up with her, abandoning with her the environment of beetles and dahlias and passing through the air with her as four o'clock in the afternoon came to an end, and they were lost forever with her in the upper atmosphere where not even the highest flying birds of memory could reach her.

B.2 Nonfiction: News story

From National Public Radio, broadcast on WBUR in Boston. First paragraph is read by another speaker; remainder is read by Margo Melnicove).

Wanted: Chief Justice of the Massachusetts Supreme Court. In April, the S.J.C.'s current leader Edward Hennessey reaches the mandatory retirement age of seventy, and a successor is expected to be named in March. It may be the most important appointment Governor Michael Dukakis makes during the remainder of his administration and one of the toughest. As WBUR's Margo Melnicove reports, Hennessey will be a hard act to follow.

In nineteen seventy-six, Democratic Governor Michael Dukakis fulfilled a campaign promise to de-politicize judicial appointments. He named Republican Edward Hennessey to head the State Supreme Judicial Court. For Hennessey, it was another step along a distinguished career that began as a trial lawyer and led to an appointment as associate Supreme Court Justice in nineteen seventy-one. That year Thomas Maffy, now president of the Massachusetts Bar Association, was Hennessey's law clerk.

The author of more than eight hundred State Supreme Court opinions, Hennessey is widely respected for his legal scholarship and his administrative abilities. Admirers give Hennessey much of the credit for sweeping court reform that began a decade ago, and for last year's legislative approval of thirty-five new judgeships and three hundred million dollars to restore crumbling court houses. Despite the state's massive budget deficit, Hennessey recently urged colleagues in the bar association not to retreat from these hard won gains.

Hennessey is the S.J.C.'s thirty-second chief justice. Holding the court system on the course he has set and plotting its future agenda won't be an easy job for his successor.

Attorney Haskell Kassler chairs the Judicial Nominating Council, eighteen attorneys and laypeople charged with screening applicants for vacancies on the bench. Usually the J.N.C. refers three nominees to the Governor. His top choice is rated by bar associations and grilled by the Governor's executive council. Kassler says, unlike the Federal Supreme Court, there's no litmus test on particular issues that Massachusetts high court nominees must pass.

All but one of the Chief Justices since eighteen ninety-nine, when Oliver Wendell Holmes was appointed, came from the ranks of S.J.C. associate justices. If he sticks with tradition, Dukakis is likely to elevate one of his appointees to chief. That means Paul Leocos or Ruth Abrams, the only woman on the court. Another possible choice is Herbert Wilkins, a Governor Sargent appointee, and next to Hennessey, the court's most senior member. The other three associate justices were put on the bench by Governor Edward King. And many lawyers say that despite Dukakis' promise to keep the judiciary above the political fray, it's unlikely Dukakis will choose a King appointee to run the state's highest court. For WBUR, I'm Margo Melnicove.

B.3 Rhymed poetry

Jabberwocky, from *Alice in Wonderland*, by Lewis Carroll.

Jabberwocky

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal sword in hand:
Long time the manxome foe he sought –
So rested he by the Tumtum tree,
And stood awhile in thought.

And, as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

One, two! One, two! And through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went galumphing back.

"And hast thou slain the Jabberwock?
Come to my arms, my beamish boy!
O frabjous day! Callooh! Callay"
He chortled in his joy.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

Appendix C

Analyzed text example

C.1 The mark-up tokens

annotation	description
:cl	Clause.
:s, :v, :o, :io, :pre, :post, :br	Grammatical role clauses: subject, verb, object, indirect object, pre, post, bridging constituent.
:np, :vp, :adjp :advp :pp :gen-np :relp	Syntactic phrase categories: noun phrase, verb phrase, adjective phrase, adverb phrase, prepositional phrase, genitive noun phrase, relativizer phrase.
:punct	Punctuation.
:ortho	Layout token (for formatted text).
:=probe	Retrieval cue
:=co	Co-refers with.
:=arg-of	Required argument of a verb.
:=mods	Modifier for a noun phrase (vs. specifier) or a verb (vs. argument).
:=anchors	Anchors a noun phrase.
=pro=	PRO subject.
=0=	Deletion.
=tr=	Trace.
:=lookahead	For a clause, denotes the "sense" of the upcoming punctuation.
:=interp	Punctuation "sense".

Table C.1: Text mark-up symbols.

C.2 Example text mark-up

From "100 Years of Solitude", by Gabriel Garcia Marquez (p242-243)

```
(:cl (:s    (:np "REMEDIOS")
            (:np (:=anchors "Remedios") "THE" "BEAUTY"))
     (:v    (:vp  "STAYED"))
     (:post (:=arg-of "stayed")(:advp "THERE")))

(:cl (:=lookahead series)
     (:s    =0= (:=co "Remedios"))
     (:v    (:vp  "WANDERING"))
     (:post :pp (:=mods "wandering")
            (:pp  "THROUGH"(:np  "THE" "DESERT"))
            (:pp (:=anchors "desert") "OF" (:np "SOLITUDE"))))

(:punct @comma (:=interp series))

(:cl (:=lookahead series)
     (:s     =0=   (:=co "Remedios"))
     (:v    (:vp "BEARING"))
     (:o    (:np "NO" "CROSS"))
     (:post :pp (:=mods "bearing")
            (:pp "ON" (:np (:gen-np (:=co "Remedios") "HER") "BACK"))))

(:punct @comma (:=interp series))

(:cl (:=lookahead series)
     (:s    =0= (:=co "Remedios"))
     (:v    (:vp "MATURING"))
     (:post :pp (:=mods "maturing")
            (:pp "IN"
                 (:np (:gen-np (:=co "Remedios") "HER") "DREAMS"))
            (:pp (:=mods "dreams") "WITHOUT" (:np "NIGHTMARES"))))

(:punct @comma (:=interp series))

(:cl  (:=lookahead series)
      (:s =0= (:=co "Remedios"))
      (:v =0= (:=co "maturing"))
      (:post  (:=mods "maturing")
              (:pp (p =0= (:=co  "in"))
                   (:np  (:gen-np (:=co "Remedios") "HER")
```

```
                        (:adjp "INTERMINABLE")
                        "BATHS"))))

(:punct @comma (:=interp series))

(:cl (:=lookahead series)
     (:s =0= (:=co "Remedios"))
     (:v =0= (:=co "MATURING"))
     (:post  (:=mods "maturing")
             (:pp (p =0= (:=co "in"))
                  (:np (:gen-np (:=co "Remedios") "HER")
                       (:adjp  "UNSCHEDULED")
                       "MEALS"))))

(:punct @comma (:=interp series))

(:cl (:s =0= (:=co "Remedios"))
     (:v =0= (:=co "maturing"))
     (:post  (:=mods "maturing")
             (:pp (p =0= (:=co "in"))
                  (:np (:gen-np (:=co "Remedios") "HER")
                       (:adjp  "DEEP" "AND" "PROLONGED")
                       "SILENCES"))))

(:cl (:=mods "silences")
     (:s (:=co "silences") (:np "THAT"))
     (:v "HAD")
     (:o (:np "NO" "MEMORY")))

(:cl (:=lookahead more)
     (:s =0= (:=co "Remedios"))
     (:v =0= (:=co "stayed"))
     (:post  :advp(:=mods "stayed")
             (:pp "UNTIL"(:np "ONE" "AFTERNOON"))
             (:pp (:=mods "afternoon") "IN" (:np "MARCH"))))

(:punct @comma (:=interp more))

(:cl (:=mods "afternoon")
     (:br :relp  (:=co "afternoon") (:relp "WHEN"))
     (:s  (:np "FERNANDA"))
     (:v  (:vp "WANTED")))

(:cl (:=arg-of "wanted")
     (:s     =pro= (:=co "Fernanda"))
```

```
     (:v    (:vp "TO"  "FOLD"))
     (:o    (:np (:gen-np (:=co "Fernanda") "HER")
                 (:adjp "BRABANT") "SHEETS"))
     (:post :pp (:=mods "fold")(:pp "IN" (:np "THE" "GARDEN"))))

(:cl (:=lookahead end)
     (:br   (:conj "AND"))
     (:s    =0= (:=co "Fernanda"))
     (:v    (:vp "ASKED"))
     (:o    (:np "THE" "WOMEN")
            (:pp (:=anchors "women")  "IN" (:np "THE" "HOUSE")))
     (:post :pp  (:=arg-of "asked")  (:pp "FOR" (:np "HELP"))))


(:punct @period    (:=interp end))
(:ortho @paragraph (:=interp start))
```

Bibliography

[Abe96] Masanobu Abe. Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System. In Jan P. H. van Santen and Richard W. Sproat and Joseph P. Olive and Julia Hirschberg, editor, *Progress in Speech Synthesis*, chapter 39, pages 495–510. Springer-Verlag, 1996.

[AHG98] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. Evaluating a Focus-Based Approach to Anaphora Resolution. In *Proceedings of the 17th International Conference on Computational Linguistics.* COLING-ACL, August 1998.

[AK96] Jean E. Andruski and Partricia K. Kuhl. The Acoustic Structure of Vowels in Mothers' Speech to Infants and Adults. In *Proceedings.* Fourth International Conference on Spoken Language Processing, October 1996.

[And83] J.R. Anderson. *The Architecture of Cognition.* Harvard University Press, Cambridge, Massachusetts, 1983.

[APL84] Mark Anderson, Janet Pierrehumbert, and Mark Liberman. Synthesis by Rule of English Intonation Patterns. In *Proceedings of the Conference on Acoustics, Speech, and Signal Processing*, page 2.8.1 to 2.8.4, 1984.

[AR86] Eneko Agirre and German Rigau. Word Sense Disambiguation using Conceptual Density. In *Proceedings.* COLING, 1986.

[Ata76] Bishnu A. Atal. Automatic Recognition of Speakers from Their Voices. *Proceedings of IEEE*, 64(4):460–473, April 1976.

[Aus62] J. L. Austin. *How to do Things with Words.* Harvard University Press, Cambridge, MA, 1962.

[Aye94] Gayle M. Ayers. Discourse functions of pitch range in read and spontaneous speech. In *OSU Working Papers, volume 44.* Ohio State University, Spring 1994.

[BD87] Paula M. Brown and Gary S. Dell. Adapting Production to Comprehension: The Explicit Mention of Instruments. *Cognitive Psychology*, 19:441–472, 1987.

[BDSP93] John Bear, John Dowding, Elizabeth Shriberg, and Patti Price. A System for Labeling Self–Repairs in Speech. Technical Note 522, Stanford Research International, February 1993.

[BE91] Mary E. Beckman and Jan Edwards. Lengthenings and shortenings and the nature of prosodic constituency. In *Laboratory Phonology*, chapter 9, pages 152–178. Cambridge University Press, 1991.

[BF90] J. Bachenko and E. Fitzpatrick. A Computational Grammar of Discourse–Neutral Prosodic Phrasing in English. *Computational Linguistics*, 16(3):155–170, 1990.

[BFP87] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A Centering Approach to Pronouns. In *Proceedings of the 25th conference*, pages 155–162. Association for Computational Linguistics, 1987.

[Bib93] Douglas Biber. The Multi–Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities*, 26:331–345, 1993.

[Bin98] Kim Binsted. Character Design for Soccer Commentary. In *Proceedings*, pages 22–35. The RoboCup Federation, 1998.

[Bir91] Steven Bird. Focus and Phrasing in Unification Categorial Grammar. In Steven Bird, editor, *Declarative Perspectives on Phonology*, volume 7 of *Working Papers in Cognitive Science*, pages 139–166. Centre for Cognitive Science, University of Edinburgh, 1991.

[Bol58] Dwight Bolinger. A Theory of Pitch Accent in English. *Word*, 14(2–3):109–149, 1958.

[BP86] Mary E. Beckman and Janet B. Pierrehumbert. Intonational structure in Japanese and English. In Colin Ewen and John Anderson, editors, *Phonology Yearbook 3*, pages 255–309. Cambridge University Press, 1986.

[Bre82] Joan Bresnan, editor. *The Mental Representation of Grammatical Relations*. M.I.T. Press, Cambridge, MA, 1982.

[Bre90] Susan E. Brennan. *Seeking and Providing Evidence for Mutual Understanding*. PhD thesis, Stanford, 1990.

[Bre95] S. E. Brennan. Centering Attention in Discourse. *Language and Cognitive Processes*, 10(2):137–167, 1995.

[Bro83] Gillian Brown. Prosodic Structure and the Given/New Distinction. In A. Cutler and D. R. Ladd, editors, *Prosody: Models and Measurements*, chapter 6, pages 67–77. Springer Verlag, 1983.

[BT94] Alan W. Black and Paul Taylor. Assigning Intonation Elements and Prosodic Phrasing for English Speech Synthesis from High Level Linguistic Input. In *Proceedings*, pages 715–718, Yokohama, Japan, 1994. Third International Conference on Spoken Language Processing.

[Cah] Janet Cahn. Synthetic Emotional Speech, (interactive installation). *AI–Based Art Exhibit*, AAAI-92, July, 1992. *Smart Art* (permanent exhibit), Boston Computer Museum, 1993–present.

[Cah88] Janet Cahn. From Sad to Glad: Emotional Computer Voices. In *Proceedings of Speech Tech '88, Voice Input/Ouput Applications Conference and Exhibition*, pages 35–37, April 1988.

[Cah89] Janet E. Cahn. Generating Expression in Synthesized Speech. Master's thesis, Massachusetts Institute of Technology, May 1989. Unpublished.

[Cah90] Janet E. Cahn. The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society*, 8:1–19, July 1990.

[Cah92] Janet Cahn. An Investigation into the Correlation of Cue Phrases, Unfilled Pauses and the Structuring of Spoken Discourse. In *Proceedings of the IRCS Workshop on Prosody in Natural Speech, Technical Report IRCS-92-37*, pages 19–30. University of Pennsylvania, Institute for Research in Cognitive Science, Philadelphia, PA., 1992.

[Cah95] Janet Cahn. The Effect of Pitch-Accenting on Pronoun Referent Resolution. In *Proceedings of the Assocation for Computational Linguistics, 33rd Conference, Student Session*, pages 290–292, June 1995.

[Cam95] W. N. Campbell. From read speech to real speech. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 2, pages 20–27, 1995.

[Car96] Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[CB91] Herbert H. Clark and Susan E. Brennan. Grounding in Communication. In L.B. Resnick, J. Levine, and S.D. Teasley, editors, *Perspectives on Socially Shared Cognition*, chapter 7, pages 127–149. APA, 1991.

[CGN92] R. Carlson, B. Granström, and L. Nord. Experiments with emotive speech – Acted utterances and synthesized replicas. In *Proceedings*, pages 671–674, Banff, Alberta, Canada, October 1992. Second International Conference on Spoken Language Processing.

[CH68] Noam Chomsky and Morris Halle. *The Sound Pattern of English.* Harper and Row, 1968. Reprinted by M.I.T. Press, 1991.

[Cha94] Anil S. Chakravarthy. Toward semantic retrieval of pictures and video. In *Proceedings of RIAO94*, 1994.

[Cho65] Noam Chomsky. *Aspects of the Theory of Syntax.* M.I.T. Press, 1965.

[Cho82] Noam Chomsky. *Some Concepts and Consequences of the Theory of Government and Binding.* M.I.T. Press, 1982.

[CHW98] Sin-Horng Chen, Shaw Hwa Hwang, and Yih-Ru Wang. An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech. *IEE Transactions on Speech and Audio Processing*, 6(3):226–239, May 1998.

[CKPS91] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A Practical Part–of–Speech Tagger. In *Proceedings*, pages 133–140. Third International Conference on Applied Natural Language Processing, 1991.

[CM81] Herbert H. Clark and Catherine R. Marshall. Definite reference and mutual knowledge. In Webber Joshi and Sag, editors, *Elements of Discourse Understanding*, chapter 1, pages 10–63. Cambridge University Press, 1981.

[Coh60] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, XX(1):37–46, 1960.

[Col85] Gary Collier. *Emotional Expression.* Lawrence Erlbaum Associates, 1985.

[Coo88] V. J. Cook. *Chomsky's Universal Grammar: An Introduction.* Basil Blackwell, Ltd., 1988.

[Cor96] Jean-Pierre Corriveau. *Time-Constrained Memory: A Reader-Based Approach to Text Comprehension.* Lawrence Erlbaum Associates, Inc., 1996.

[Cru86] Alan Cruttenden. *Intonation.* Cambridge University Press, 1986.

[Cru93] Alan Cruttenden. The De-Accenting and Re-Accenting of Repeated Lexical Items. In *Proceedings, Workshop on Prosody*, pages 16–19. European Speech Communication Association, 1993.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1991.

[Cut84] Anne Cutler. Stress and Accent in Language Production and Understanding. In Dafydd Gibbon and Helmut Richter, editors, *Intonation, Accent and Rhythm: Studies in Discourse Phonology*, volume 8 of *Research in Text Theory*, pages 76–90. Walter de Gruyter, Berlin/New York, 1984.

[Cut96] Anne Cutler. Prosody and the Structure of the Message. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody*, chapter 5.1, pages 63–66. Springer-Verlag, New York, 1996.

[Dav64] Joel Davitz. *The Communication of Emotional Meaning.* McGraw-Hill, 1964.

[Dav89] James R. Davis. *Back Seat Driver: Voice Assisted Automobile Navigation.* PhD thesis, Massachusetts Institute of Technology, September 1989.

[DCC95] E. Douglas-Cowie and Roddy Cowie. The forms and functions of intonation in the phone voice. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 4, pages 240–243, 1995.

[Del85] Gary S. Dell. Positive Feedback in Hierarchical Connectionist Models: Applications to Language Production. *Cognitive Science*, 9(1):3–25, January–March 1985.

[Del88] Gary S. Dell. The Retrieval of Phonological Forms in Production: Tests of Prediction from a Connectionist Model. *Journal of Memory and Language*, 27:124–142, 1988.

[DH88] James R. Davis and Julia Hirschberg. Assigning intonational features in synthesized spoken directions. In *Proceedings of the Association for Computational Linguistics*, pages 187–193, 1988.

[DT87] James R. Davis and Thomas F. Trobaugh. Direction assistance. Technical Report 1, MIT Media Laboratory Speech Group, December 1987.

[Dun72] Starkey Duncan, Jr. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[EC86] Stephen J. Eady and William E. Cooper. Speech intonation and focus location in matched statements and questions. *Journal of the Acoustic Society of America*, 80(2), August 1986.

[Fai40] G. Fairbanks. Recent experimental investigations of vocal pitch in speech. *Journal of the Acoustic Society of America*, (11):457–466, 1940.

[FC77] David Fay and Anne Cutler. Malapropisms and the Structure of the Mental Lexicon. *Linguistic Inquiry*, 8(3):505–520, Summer 1977.

[FH87] C. Fowler and J. Housum. Talkers' Signalling of "New" and "Old" Words in Speech and Listeners; Perception and Use of the Distinction. *Journal of Memory and Language*, 26:489–504, 1987.

[FK92] Susan R. Fussell and Robert M. Krauss. Coordination of Knowledge in Communication: Effects of Speakers' Assumptions About What Others Know. *Journal of Personality and Social Psychology*, 62(3):378–391, 1992.

[FLB97] Carol A. Fowler, Elena T. Levy, and Julie M. Brown. Reductions of Spoken Words in Certain Discourse Contexts. *Journal of Memory and Language*, 37:24–40, 1997.

[FP39] G. Fairbanks and W. Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotions. *Speech Monographs*, 6:87–104, 1939.

[Fro71] Victoria A. Fromkin. The non-anomalous nature of anomalous utterances. *Language*, 47:27–52, 1971.

[FS84] A. Fernald and T. Simon. Expanded Intonation Contours in Mother's Speech to Newborns. *Developmental Psychology*, 20:104–113, 1984.

[GE58] Frieda Goldman-Eisler. Speech Analysis and Mental Processes. *Language and Speech*, 1:59–75, 1958.

[GE61] Frieda Goldman-Eisler. A Comparative Study of Two Hesitation Phenomena. *Language and Speech*, 4:18–26, 1961.

[GG83] James Paul Gee and Francois Grosjean. Performance Structures: A Psycholinguistic and Linguistic Appraisal. *Cognitive Psychology*, 15:411–458, 1983.

[GG87] Francois Grosjean and James Paul Gee. Prosodic structure and spoken word recognition. *Cognition*, 25:135–155, 1987.

[GH92] Barbara Grosz and Julia Hirschberg. Some Intonational Characteristics of Discourse Structure. In *Proceedings*, Banff, Alberta, Canada, October 1992. International Conference on Spoken Language Processing.

[Gib84] Dafydd Gibbon. Intonation as an Adaptive Process. In Dafydd Gibbon and Helmut Richter, editors, *Intonation, Accent and Rhythm: Studies in Discourse Phonology*, volume 8 of *Research in Text Theory*, pages 165–192. Walter de Gruyter, Berlin/New York, 1984.

[Gib87] Dafydd Gibbon. The Role of Discourse in Intonation Theory. In Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Pfeiffer, and John R. Rennison, editors, *Phonologica 1984*, pages 49–57. Cambridge University Press, 1987. Proceedings of the Fifth International Meeting.

[GJW95] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 2(21):203–225, June 1995.

[GL83] Gina Geffen and Mary A. Luszcz. Are the spoken durations of rare words longer than those of common words? *Memory & Cognition*, 11(1):13–15, 1983.

[GL90] Michael Gasser and Chan-Do Lee. Networks that Learn about Phonological Feature Persistence. *Connection Science*, 2(4):265–278, 1990.

[GNH92] Beverly Gable, Helen Nemeth, and Martin Haran. Speech Errors and Task Demand. In *Proceedings*, volume II, pages 1103–1105. Second International Conference on Spoken Language Processing, October 1992.

[Goo96] Marjorie Harness Goodwin. Informings and announcements in their environment: prosody within a multi-activity setting. In Elizabeth Couper-Kuhlen and Margaret Selting, editors, *Prosody in conversation*, pages 436–461. Cambridge University Press, 1996.

[GS86] Barbara J. Grosz and Candace L. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[GT91] Prahlad Gupta and David S. Touretzky. Connectionist models and linguistic theory: Investigations of stress systems in Language. *Cognitive Science*, 18:1–50, December 1991.

[GvH95] Charlotte Gooskens and Vincent J. van Heuven. Declination in Dutch and Danish: Global versus local pitch movements in the perceptual characteristics of sentence types. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 2, pages 374–377, 1995.

[Hal67] Michael A. K. Halliday. Notes on transitivity and theme in English. *Journal of Linguistics*, 3:177–274, 1967.

[Hal87] Morris Halle. Grids and Trees in Metrical Phonology. In Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Pfeiffer, and John R. Rennison, editors, *Phonologica 1984*, pages 79–93. Cambridge University Press, 1987. Proceedings of the Fifth International Meeting.

[Hay89] Bruce Hayes. The Prosodic Hierarchy in Meter. In Paul Kiparsky, editor, *Rhythm and Meter*, Phonetics and Phonology. Academic Press, San Diego, 1989.

[HE95] Caroline Henton and Bradly Edelman. Generating and Manipulating Emotional Synthetic Speech on a Personal Computer. *Multimedia Tools and Applications*, 3(2):105–126, September 1995.

[Hee91] Peter Anthony Heeman. A Computational Model of Collaboration on Referring Expressions. Technical Report CSRI–251, Computer Systems Research Institute, University of Toronto, Canada, 1991.

[Hen95] Henrietta J. Cedergren and Hélène Perreault. On the analysis of syllable timing in everyday speech. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 4, pages 232–235, 1995.

[HFLL93] Merle Horne, Marcus Filipsson, Mats Ljungqvist, and Anders Lindström. Referent Tracking in Restricted Texts Using a Lemmatized Lexicon: Implications for Generation of Intonation. In *Proceedings*, pages 2011–2014. Eurospeech, 1993.

[HG92] Julia Hirschberg and Barbara Grosz. Intonational Features of Local and Global Discourse Structure. In *Working Notes of DARPA Workshop on Speech and Natural Language*, pages 441–446, Arden House, Harriman, New York, February 1992.

[HH76] Michael A. K. Halliday and Ruqaiya Hasan. *Cohesion in English.* Longman Press, 1976.

[HHS96] Norio Higuchi, Toshio Hirai, and Yoshinori Sagisaka. Effect of Speaking Style on Parameters of Fundamental Frequency Contour. In Jan P. H. van Santen and Richard W. Sproat and Joseph P. Olive and Julia Hirschberg, editor, *Progress in Speech Synthesis*, chapter 33, pages 417–428. Springer-Verlag, 1996.

[Hir90] Julia Hirschberg. Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech. In *Proceedings.* American Assocation for Artificial Intelligence, 1990.

[Hir93] Julia Hirschberg. Pitch Accent in Context: Predicting Intonational Prominence from Text. *Journal of Artificial Intelligence*, 63:305–340, 1993.

[Hir95] Julia Hirschberg. Prosodic and Other Acoustic Cues to Speaking Style in Spontaneous and Read Speech. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 2, pages 36–43, 1995.

[HL87] Julia Hirschberg and Diane Litman. Now let's talk about now: Identifying cues phrases intonationally. In *Proceedings of the 25th Conference of the Association for Computational Linguistics*, pages 163–171, July 1987.

[HLPW87] Julia Hirschberg, Diane Litman, Janet Pierrehumbert, and Gregory Ward. Intonation and the Intentional Structure of Discourse. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 1987.

[HP86] Julia Hirschberg and Janet Pierrehumbert. The intonational structuring of discourse. In *Proceedings of the 24th annual meeting of the Association for Computational Linguistics*, pages 136–144, July 1986.

[HP94] Julia Hirschberg and Pilar Prieto. Training Intonational Phrasing Rules Automatically for English and Spanish Text-to-Speech. In *Proceedings of the ESCA/IEEE Workshop on Speech Synthesis*, 1994.

[HPC97] Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. Semantic and Discourse Information for Text-to-Speech Intonation. In *Proceedings of the ACL Workshop on Concept to Speech Generation Systems*, pages 47–56, 1997.

[JC92] Marcel Adam Just and Patricia A. Carpenter. A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychological Review*, 99(1):122–149, 1992.

[Joh96] M. E. Johnson. Synthesis of English Intonation Using Explicit Models of Reading and Spontaneous Speech. In *Proceedings*. Fourth International Conference on Spoken Language Processing, 1996.

[Kam94] Megumi Kameyama. Stressed and unstressed pronouns: Complementary preferences. In *Proceedings*. The FOCUS and NLP Conference, 1994.

[KF67] Henry Kučera and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, Rhode Island, 1967.

[Kla75] Dennis H. Klatt. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3:129–140, 1975.

[Kla87] Dennis H. Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustic Society of America*, 82:737–783, September 1987.

[Koh95] K. J. Kohler. Articulatory reduction in different speaking styles. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 12, pages 12–19, 1995.

[Kri80] Klaus Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, California, 1980.

[Lad83] D. Robert Ladd. Peak features and overall slope. In A. Cutler and D. R. Ladd, editors, *Prosody: Models and Measurements*, chapter 4, pages 39–52. Springer-Verlag, 1983.

[Lad86] D. Robert Ladd. Intonational phrasing: the case for recursive prosodic structure. In Colin Ewen and John Anderson, editors, *Phonology Yearbook 3*, pages 311–340. Cambridge University Press, 1986.

[Lad88] D. Robert Ladd. Declination "reset" and the hierarchical organization of utterances. *Journal of the Acoustic Society of America*, 84(2):530–544, August 1988.

[Lad93] D. Robert Ladd. Notes on the Phonology of Prominence. In *Working Papers 41, (ESCA Workshop on Prosody 1993)*, pages 10–15. Department of Linguistics and Phonetics, Lund, Sweden, 1993.

[Lan75] Thomas K. Landauer. Memory Without Organization: Properties of a Model with Random Storage and Undirected Retrieval. *Cognitive Psychology*, 7:495–531, 1975.

[Lev83] Willem J. M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[Lev89] Willem J. M. Levelt. *Speaking.* M.I.T. Press, Cambridge, MA, 1989.

[LFP83] Paul A. Luce, Timothy C. Feustel, and David B. Pisoni. Capacity Demands in Short-Term Memory for Synthetic and Natural Speech. *Human Factors*, 25(1):17–32, 1983.

[LP77] Mark Liberman and Alan Prince. On Stress and Linguistic Rhythm. *Linguistic Inquiry*, 8(2):249–336, 1977.

[LP84] Mark Y. Liberman and J. Pierrehumbert. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle, editors, *Language Sound Structure*, chapter 10. M.I.T. Press, 1984.

[LS74] Mark Y. Liberman and Ivan Sag. Prosodic form and discourse function. In *Proceedings of the 10th Regional Meeting*, pages 416–427. Chicago Linguistic Society, 1974.

[Lyo77] John Lyons. *Semantics 2.* Cambridge University Press, 1977.

[MA93] Iain R. Murray and John L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America*, 93(2):1097–1108, February 1993.

[Mac87] James N. MacGregor. Short–Term Memory Capacity: Limitation or Optimization. *Pyschological Review*, 94(1):107–108, 1987.

[Mag97] Arman Maghbouleh. Prosody phrasing without syntax via part-of-speech amalgamation. In *Proceedings, ESCA Workshop on Intonation*, Athens, Greece, 1997. European Speech Communication Association.

[MAN88] I. R. Murray, J. L. Arnott, and A. F. Newell. HAMLET – Simulating Emotion in Synthetic Speech. In *Speech '88; Proceedings of the 7th FASE Symposium.* Institute of Acoustics, Edinburgh, 1988.

[Mar80] Mitchell Marcus. *A Theory of Syntactic Recognition for Natural Language.* M.I.T. Press, 1980.

[Mar87] Boguslaw Marek. The Prosodic Structure of Intonational Contours. In Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Pfeiffer, and John R. Rennison, editors, *Phonologica 1984*, pages 187–193. Cambridge University Press, 1987. Proceedings of the Fifth International Meeting.

[MB82] Lisa Menn and Suzanne Boyce. Fundamental frequency and discourse structure. *Language and Speech*, 25(4):341–381, 1982.

[MBF+90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.

[McL91] Cynthia Ann McLemore. *The Pragmatic Interpretation of English Intonation: Sorority Speech.* PhD thesis, University of Texas at Austin, 1991.

[MF91] George A. Miller and Christiane Fellbaum. Semantics networks of English. *Cognition*, 41:197–229, 1991.

[Mil56] G. A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63:81–97, 1956.

[MO59] Howard Maclay and Charles E. Osgood. Hesitation Phenomena in Spontaneous English Speech. *Word*, 15:19–44, 1959.

[Mon90] A. I. C. Monaghan. Rhythm and stress–shift in speech synthesis. *Computer Speech and Language*, 4:71–78, 1990.

[MTKI96] Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech Synthesis using HMMS with Dynamic Features. In *Proceedings*, pages 389–392, Atlanta, 1996. International Conference on Acoustics, Speech, and Signal Processing. (IEEE signal processing society).

[MW64] Frederick Mosteller and David L. Wallace. *Inference and disputed authorship: The Federalist.* Addison-Wesley, Reading, Mass., 1964.

[Nak95] Christine Nakatani. Discourse Structural Constraints on Accent in Narrative. In Jan P. H. van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, pages 139–156. Springer-Verlag, 1995.

[NT82] S. G. Nooteboom and J. M. B. Terken. What Makes Speakers Omit Pitch Accents? An Experiment. *Phonetica*, 39:317–336, 1982.

[NV83] M. Nespor and .I Vogel. Prosodic structure above the word. In Anne Cutler and D. R. Ladd, editors, *Prosody: Models and Measurements*, pages 123–140. Springer–Verlag, New York, 1983.

[NV89] M. Nespor and .I Vogel. On clashes and lapses. In *Phonology 6.* Cambridge University Press, 1989.

[OC74] R. B. Oschman and A. Chapanis. The effects of ten communication modes on the behavior of teams during co-operative problem solving. *International Journal of Man/Machine Studies*, 6:579–619, 1974.

[O'C88] Daniel C. O'Connell. *Critical Essays on Language Use and Psychology.* Springer-Verlag, 1988.

[Oeh91] Richard T. Oehrle. Prosodic Constraints on Dynamic Grammatical Analysis. In Steven Bird, editor, *Declarative Perspectives on Phonology*, volume 7 of *Working Papers in Cognitive Science*, pages 167–195. Centre for Cognitive Science, University of Edinburgh, 1991.

[OPSH95] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical report, Boston University, February 1995.

[O'S92] Douglas O'Shaughnessy. Analysis of False Starts in Spontaneous Speech. In *Proceedings*, volume II, pages 931–934. Second International Conference on Spoken Language Processing, October 1992.

[OV94] M. Ostendorf and N. Veilleux. A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location. Draft: 1992. To appear in *Computational Linguistics*, 1994.

[PH90] Janet B. Pierrehumbert and Julia Hirschberg. The Meaning of Intonation Contours in the Interpretation of Discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT Press, 1990.

[Pie80] Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, 1980.

[Pie81] Janet B. Pierrehumbert. Synthesizing intonation. *Journal of the Acoustic Society of America*, pages 985–995, October 1981.

[Pin94] Steven Pinker. *The Language Instinct.* William E. Morrow and Company, 1994.

[Pla94] Tony A. Plate. *Distributed Representations and Nested Compositional Structure.* PhD thesis, University of Toronto, Canada, 1994.

[PM97] Shimei Pan and Kathleen R. McKeown. Integrating Language Generation with Speech Synthesis in a Concept to Speech System. In *Proceedings of the ACL Workshop on Concept to Speech Generation Systems.* ACL/EACL, 1997.

[PM98] Shimei Pan and Kathleen R. McKeown. Integrating Language Generation with Speech Synthesis in a Concept to Speech System. In *Proceedings of the 17th International Conference on Computational Linguistics.* COLING-ACL, August 1998.

[Pol90a] Jordan B. Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1-2):77–105, November 1990.

[Pol90b] Martha E. Pollack. Plans as complex mental attitudes. In Phillip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication.* MIT Press, 1990.

[Pre96] Scott Prevost. Modeling Contrast in the Generation and Synthesis of Spoken Language. In *Proceedings.* Fourth International Conference on Spoken Language Processing, October 1996.

[Pri81] Ellen F. Prince. Toward a Taxonomy of Given–New Information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, 1981.

[PS94] Scott Prevost and Mark Steedman. Specifying Intonation from Context for Speech Synthesis. *Speech Communication*, 15:139–152, 1994.

[Qui92] Philip T. Quinlan. *The Oxford Psycholinguistic Database.* Oxford University Press, 1992.

[RA95] German Rigau and Eneko Agirre. Disambiguating bilingual nominal entries against WordNet. In *Proceedings of the Workshop on The Computational Lexicon.* ESSLLI, 1995.

[Ram94] T. V. Raman. *Audio System for Technical Readings.* PhD thesis, Cornell University, 1994.

[Ran80] Emily Rando. Intonation in discourse. In Linda R. Waugh and C. H. van Schooneveld, editors, *The Melody of Language*, pages 243–278. University Park Press, 1980.

[Res95] Philip Resnik. Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proceedings of the 3rd Workshop on Very Large Corpora*, June 1995.

[Ris87] Jorgen Rischel. Is There Just One Hierarchy of Prosodic Categories? In Wolfgang U. Dressler, Hans C. Luschützky, Oskar E. Pfeiffer, and John R. Rennison, editors, *Phonologica 1984*, pages 253–259. Cambridge University Press, 1987. Proceedings of the Fifth International Meeting.

[RM88] Roger Ratcliff and Gail McKoon. A Retrieval Theory of Priming in Memory. *Psychological Review*, 95(3):385–408, 1988.

[RM94] Roger Ratcliff and Gail McKoon. Retrieving Information From Memory: Spreading-Activation Theories Versus Compund-Cue Theories. *Psychological Review*, 101(1):177–184, 1994.

[RO96] K. Ross and M. Ostendorf. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, 1996.

[RPL+91] James C. Ralston, David B. Pisoni, Scott E. Liveley, Beth G. Greene, and John W. Mullenix. Comprehension of Synthetic Speech Produced by Rule: Word Monitoring and Sentence-by-Sentence Listening Times. *Human Factors*, 33(4):471–491, 1991.

[Sac67] J. D. Sachs. Recognition Memory for Syntactic and Semantic Aspects of Connected Discourse. *Perception and Psychophysics*, 2:437–442, 1967.

[SBD92] Elizabeth Shriberg, John Ber, and John Dowding. Automatic Detection and Correction of Repairs in Human-Computer dialog. In M. Marcus, editor, *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 491–494, 1992.

[SBP+92] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. ToBI: A standard for labeling English prosody. In *Proceedings*. Second International Conference on Spoken Language Processing, October 1992.

[SBSP92] Kim Silverman, Eleonora Blaauw, Judith Spitz, and John F. Pitrelli. Towards Using Prosody in Speech Recognition/Understanding Systems: Differences Between Read and Spontaneous Speech. In *Proceedings of the Fifth DARPA Workshop on Speech and Natural Language, Arden House Conference Center, New York.*, February 1992.

[SC80] John M. Sorensen and William E. Cooper. Syntactic coding of fundamental frequency in speech production. In Ronald A. Cole, editor, *Perception and Production of Fluent Speech*, chapter 13, pages 399–437. Lawrence Erlbaum, 1980.

[Sch82] Emanual A. Schegloff. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In D. Tannen, editor, *Analyzing Discourse: Text and Talk*. Georgetown University Roundtable on Language and Linguistics, 1982.

[SDvTC92] P. Spyns, F. Deprez, L. van Tichelen, and B. Van Coile. Message-to-Speech: A Practical Compromise between Concept-to-Speech and Text-to-Speech Strategies for Dialogue Systems. In *Workshop on Spoken Dialogue Systems*. ACL/EACL, July 1992.

[Sea69] John Searle. *Speech Acts*. Cambridge University Press, 1969.

[Sel84] Elizabeth Selkirk. (1) The relation between syntax and phonology, (2) Rhythmic patterns in language. In *Phonology and Syntax*, chapter 1 and 2, pages 1–35. M.I.T. Press, 1984.

[Sel86] Elizabeth Selkirk. On derived domains in sentence phonology. In Colin Ewen and John Anderson, editors, *Phonology Yearbook 3*, pages 371–405. Cambridge University Press, 1986.

[SG94] M. G. J. Swerts and R. Geluykens. Prosody as a Marker of Information Flow in Spoken Discourse. *Language and Speech*, 37(1):21–43, 1994.

[SH86] Stefanie Shattuck-Hufnagel. The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech. In Colin Ewen and John Anderson, editors, *Phonology Yearbook 3*, pages 117–149. Cambridge University Press, 1986.

[SHOR93] S. Shattuck-Hufnagel, M. Ostendorf, and K. Ross. Pitch Accent Placement within Lexical Items in American English. *Journal of Phonetics*, October 1993. To appear.

[SHY92] Richard Sproat, Julia Hirschberg, and David Yarowsky. A Corpus-Based Synthesizer. In *Proceedings*, volume I, pages 563–566. Second International Conference on Spoken Language Processing, October 1992.

[Sid79] Candace Sidner. *Towards a computational theory of definite anaphora comprehension in English discourse.* PhD thesis, Massachusetts Institute of Technology, 1979.

[Sid86] Candace L. Sidner. Focusing in the Comprehension of Definite Anaphora. In Barbara J. Grosz, Karen Sparck-Jones, and Bonnie Lynn Webber, editors, *Readings in Natural Language Processing*, pages 363–394. Morgan Kaufman Publishers, Inc., 1986.

[SOWH96] David L. Swofford, Gary J. Olsen, Peter J. Waddell, and David M. Hillis. Phylogenetic Inference. In David M. Hillis, Craig Moritz, and Barbara K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer Associates, Inc., 1996. Second Edition.

[SP95] S. Sclaroff and A. Pentland. Modal Matching for Correspondence and Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(6):545–561, 1995.

[Spi85] Murray F. Spiegel. Pronouncing surnames automatically. In *Proceedings of 1985 Conference.* American Voice I/O Society, September 1985.

[Spr98] Richard Sproat, editor. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer Academic Publisher, 1998.

[SR87] Terrence J. Sejnowski and Charles R. Rosenberg. Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, 1:145–168, 1987.

[SSH96] Marc Swerts, Evan Strangert, and Mattias Heldner. F0 Declination in Read-Aloud and Spontaneous Speech. In *Proceedings*, volume 3, pages 1501–1504. Fourth International Conference on Spoken Language Processing, October 1996.

[SSJ74] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A Simple Systematics for the Organization of Turn-taking for Conversation. *Language*, 50(4):696–735, 1974. Reprinted in *Studies in the Organization of Conversational Interaction*, J. Schenken, ed., Academic Press 1978.

[Ste90] Mark J. Steedman. Syntax and Intonational Structure in a Combinatory Grammar. In Gerry T. M. Altmann, editor, *Cognitive Models of Speech Processing*, chapter 21, pages 457–482. M.I.T. Press, 1990.

[SWB96] Marc Swerts, Anne Wichmann, and Robert-Jan Beun. Filled pauses as markers of discourse structure. In *Proceedings*, volume 2, pages 1033–1036. Fourth International Conference on Spoken Language Processing, October 1996.

[Ter84] J. M. B. Terken. The distribution of pitch accents in instructions as a function of discourse structure. *Language and Speech*, 27(Part 3):53–73, 1984.

[TH94] Jacques Terken and Julia Hirschberg. Deaccentuation of Words Representing 'Given' Information: Effects of Persistence of Grammatical Function and Surface Position. *Language and Speech*, 37(2):125–145, 1994.

[The97] Mariët Theune. Constrastive accent in a data-to-speech system. In *Proceedings of the 35th annual meeting of the ACL / the 8th Conference of the EACL*, pages 519–521, 1997.

[TOKdP97] Mariët Theune, Jan Odjik, Esther Klabbers, and Jan Roelof de Pijper. From Data to Speech: A Generic Approach. Manuscript 1202, Instituut voor Perceptie Onderzoek, 1997.

[Tou95] Paul Touati. Pitch range and register in French politcal speech. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 4, pages 244–247, 1995.

[Ume82] Noriko Umeda. F0 declination is situation dependent. In *Journal of Phonetics*, volume 10, pages 279–290, 1982.

[vDvB96] Monique E. van Donzel and Florien J. Koopmans van Beinum. Pausing Strategies in Discourse in Dutch. In *Proceedings*. Fourth International Conference on Spoken Language Processing, October 1996.

[VHA88] Atro Voutilainen, Juha Heikkilä, and Arto Anttila. Constraint Grammar of English: A Performance-Oriented Introduction. Publication 21, University of Helsinki, 1988.

[VJ95] Atro Voutilainen and Timo Järvinen. Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, Dublin.* Ohio State, 1995.

[Vou92] Atro Voutilainen. *NPtool*, a detector of English noun phrases. In *Proceedings of the Workshop on Very Large Corpora.* Ohio State, June 1992.

[Wal92] Marilyn A. Walker. When Given Information is Accented: Repetition, Paraphrase. In *Proceedings of the IRCS Workshop on Prosody in Natural Speech, IRCS Report No, 92–37.* Institute for Research in Cognitive Science, University of Pennsylvania, 1992.

[Wal93] Marilyn A. Walker. *Informational Redundancy and Resource Bounds in Dialogue.* PhD thesis, University of Pennsylvania, December 1993. (Institute for Research in Cognitive Science report IRCS-93-45).

[Wal96a] Marilyn A. Walker. Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264, June 1996.

[Wal96b] Marilyn A. Walker. The Effect of Resource Limits and Task Complexity on Collaborative Planning in Dialogue. *Artificial Intelligence Journal*, 85(1–2):181–243, 1996.

[Was97] Thomas Wasow. End-Weight from the Speaker's Perspective. *Journal of Psycholinguistic Research*, 26(3):347–361, 1997.

[WCW97] Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. Improvising Linguistic Style: Social and Affective Bases for Agent Personality. In *Proceedings of the Agents '97 Conference.* Association for Computing Machinery, 1997. To appear.

[Web84] Bonnie Lynn Webber. So what can we talk about now. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse*, chapter 6. MIT Press, 1984.

[WH85] Gregory Ward and Julia Hirschberg. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61(4):747–776, 1985.

[WH92] Michelle Q. Wang and Julia Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196, 1992.

[Whi95] Sandra P. Whiteside. Temporal-based speaker sex differences in read speech: A sociophonetic approach. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 3, pages 516–519, 1995.

[WO92] Colin W. Wightman and Mari Ostendorf. Automatic recognition of intonational features. In *Proceedings*, pages 221–224. International Conference on Acoustics, Speech, and Signal Processing, March 1992.

[WS72] Carl E. Williams and Kenneth N. Stevens. Emotions and Speech: Some Acoustical Correlates. *Journal of the Acoustic Society of America*, 52(4 (Part 2)):1238–1250, 1972.

[WSHOP92] Colin W. Wightman, Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti J. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustic Society of America*, 91(3):1707–1711, March 1992.

[WT85] John A. Waterworth and Cathy M. Thomas. Why is Synthetic Speech Harder to Remember than Natural Speech? In *Proceedings, 1985 Conference on Human Factors in Computing Systems*, pages 201–206. ACM SIGCHI, 1985.

[WW90] Steve Whittaker and Marilyn A. Walker. Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. In *Proceedings*, pages 1–9. Association for Computational Linguistics, 1990.

[Yar94] David Yarowsky. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd Conference of the Association for Computational Linguistics*, page 8 Pages, 1994.

[Yar96] David Yarowsky. Homograph Disambiguation in Text-to-Speech Synthesis. In Jan P.H. Van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*, chapter 12, pages 157–172. Springer, 1996.

[YH92] Nicholas Youd and Jill House. Generating Intonation in a Voice Dialogue System. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 1287–1290, 1992.

[Yng70] Victor H. Yngve. On Getting a Word in Edgewise. In *Papers from the Sixth Regional Meeting*, pages 567–578. Chicago Linguistic Society, 1970.

[You92] Nicholas John Youd. *The Production of Prosodic Focus and Contour in Dialogue.* PhD thesis, The Open University, Great Britain, 1992.

[ZDG+89] Victor Zue, Nancy Daly, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, Stephanie Seneff, and Michal Soclof. The collection and preliminary analysis of a spontaneous speech database. In *2nd DARPA Speech and NL Workshop*, October 1989.