A Computational Model for the Automatic Recognition of Affect in Speech

by

Raul Fernandez

B.S. in Electrical Engineering, University of Miami (1995)
S.M. in Computer Science and Electrical Engineering,
Massachusetts Institute of Technology (1998)

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 2004

Author .....
Program in Media Arts and Sciences,
School of Architecture and Planning
October 31, 2003

Certified by .....
Rosalind W. Picard
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Andrew B. Lippman
Chairman, Departmental Committee on Graduate Students

A Computational Model for the Automatic Recognition of Affect in Speech

by

Raul Fernandez

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on October 31, 2003, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Spoken language, in addition to serving as a primary vehicle for externalizing linguistic structures and meaning, acts as a carrier of various sources of information, including background, age, gender, membership in social structures, as well as physiological, pathological and emotional states. These sources of information are more than just ancillary to the main purpose of linguistic communication: Humans react to the various non-linguistic factors encoded in the speech signal, shaping and adjusting their interactions to satisfy interpersonal and social protocols.

Computer science, artificial intelligence and computational linguistics have devoted much active research to systems that aim to model the production and recovery of linguistic lexico-semantic structures from speech. However, less attention has been devoted to systems that model and understand the paralinguistic and extralinguistic information in the signal. As the breadth and nature of human-computer interaction escalates to levels previously reserved for human-to-human communication, there is a growing need to endow computational systems with human-like abilities which facilitate the interaction and make it more natural. Of paramount importance amongst these is the human ability to make inferences regarding the affective content of our exchanges.

This thesis proposes a framework for the recognition of affective qualifiers from prosodic-acoustic parameters extracted from spoken language. It is argued that modeling the affective prosodic variation of speech can be approached by integrating acoustic parameters from various prosodic time scales, summarizing information from more localized (e.g., syllable level) to more global prosodic phenomena (e.g., utterance level). In this framework speech is structurally modeled as a dynamically evolving hierarchical model in which levels of the hierarchy are determined by prosodic constituency and contain parameters that evolve according to dynamical systems. The acoustic parameters have been chosen to reflect four main components of speech thought to reflect paralinguistic and affect-specific information: intonation, loudness, rhythm and voice quality. The thesis addresses the contribution of each of these components separately, and evaluates the full model by testing it on datasets of acted and of spontaneous speech perceptually annotated with affective labels, and by comparing it against human performance benchmarks.

Thesis Supervisor: Rosalind W. Picard
Title: Associate Professor of Media Arts and Sciences

Thesis Advisor .....

Rosalind W. Picard

Associate Professor of Media Arts and Sciences

MIT Media Laboratory

Thesis Reader .....

Stefanie Shattuck-Hufnagel

Principal Research Scientist, Speech Communication Group

MIT Research Laboratory of Electronics

Thesis Reader .....

Julia Hirschberg

Professor of Computer Science

Columbia University

Acknowledgments

I would like to thank first of all the members of my thesis committee for their invaluable guidance and contribution to this work. My research advisor Roz Picard has fostered over many years a great intellectual climate to carry out this investigation, providing an ideal synthesis of guidance and independence that has not only strengthened the work, but also made it greatly enjoyable. Stefanie Shattuck-Hufnagel and Julia Hirschberg have provided most efficient, insightful and close readings of this document, elucidating many opaque issues and offering improving feedback. I also owe thanks to Stefanie for providing me with copies of prosodically annotated data for the development and testing of some of the work presented here. I feel fortunate to have benefited from the advice, supervision and contributions of this great committee; any remaining faults in this work remain, of course, my own.

Many undergraduate assistants have over many years lent their skills and dedication to this research. In somewhat chronological order I would like to thank: Daniel Roth, for crafting the original design of an interface for perceptual experiments and carrying out pilot studies with it; Margaret Whitman, for patiently poring over hours of recordings in search of effective and affective speech exemplars for constructing training databases; Jeffrey Bartelma, for helping write the code for vocal source modeling and for his entertaining code documentation style; Sie Hendrata Dharmawan, for further development of the experimental interface, and for very helpful and instrumental supporting code for graphical models; and finally Eugene Baik, for implementing the F0 stylization code.

Other colleagues and associates have contributed in many different ways to the work presented here. I would like to thank Anna Pandolfo for her help recruiting subjects and conducting perceptual experiments. Many present and former students from the MIT Media Lab, foremost in the Affective Computing and the former Vision and Modeling groups, have

facilitated countless interactions, exchanges and collaborations that have allowed me to learn and advance my research. Through what is doomed to be a very non-exhaustive list, I would like to thank Sumit Basu, Timothy Bickmore, Janet Cahn, Tanzeem Choudhury, Yuri Ivanov, Tony Jebara, Ashish Kapoor, Tom Minka, Selene Mota, Yuan Qi, Carson Reynolds, Tara Rosenberger Shankar, Jocelyn Scheirer, and Elias Vyzas.

I would like to thank my parents for having always encouraged my academic inclinations from a very early age, for having had the vision and the courage to take the steps that have allowed me to get here, and for their support through many years. Although my mother contemplates with some disbelief the fact that I will cease to be a student (for now!), she is nonetheless thrilled that I am finally graduating. I thank her for her patience through all these years, for which this conclusion constitutes but a small reward.

Many people have provided over the many years spent on this work the kind of backdrop, support, friendship and assistance, without which this would have been a far less enjoyable experience. I would like to thank Christine Donovan for her patience, understanding, words of encouragement, cooking, proofreading, and graphical design skills; for sharing the high and low points of this journey; for making sure that I did not go unfed and that I retained my sanity after long days at the lab; for making me smile and for being there for me. Liza Daly has watched this research slowly take shape probably longer than anyone directly involved in it (and, anticlimactically, managed to find herself half a continent away at the time of my defense to see it reach its symbolic conclusion). I would like to thank her for all the support she has lent over the years and for sharing the experience with me. Jocelyn Scheirer has played several roles during my stay at the lab: In addition to being colleague and office mate (and, in a perhaps less enviable capacity, having shared many demo sessions together), she has also proved to be a supportive friend, offering much-needed help at crucial moments.

I would finally like to extend my thanks to IBM and the multi-sponsor Digital Life consortium at the MIT Media Lab for their kind support of this research, and to British Telecom for kindly contributing their speech materials.

In memoriam R. F. M.,

who made this work at all possible and did not live to see me finish it.

Einzelnen Vokalen können Leben und Erfahrung einen Akzent verleihen, der sie ihrem alltäglichem Sinn völlig entfremdet und ihnen einen Schreckensnimbus verleiht, den niemand versteht, der sie nicht in ihrer füchterlichsten Bedeutung kennengelernt hat.

– Thomas Mann, *Doktor Faustus*

Das Verständliche an der Sprache ist nicht das Wort selber, sondern Ton, Stärke, Modulation, Tempo, mit denen eine Reihe von Worten gesprochen wird –kurz, die Musik hinter den Worten, die Leidenschaft hinter dieser Musik, die Person hinter dieser Leidenschaft: alles das also, was nicht geschrieben werden kann.

– Friedrich Nietzsche

Contents

List of Figures

List of Tables

Part I

Foundations

Chapter 1

Introduction and Thesis Outline

This thesis investigates the problem of building computational systems that can automatically recognize affective labels in speech. Much of the emphasis on machine processing of human language over the past few decades has been on the problem of automatic speech recognition; that is, the problem of decoding the series of words that a speaker produces. As we have moved toward systems capable of using voice for real-time human-machine interaction, the research has grown to encompass the area of natural language understanding, where the goal of the system is no longer just to produce a word transcription, but to extract from the interaction notions about the semantics and pragmatics of the dialogue. For many years, there have also been sporadic research efforts to try to build systems that infer the affective tone of a speaker. Progress in this area has been slower than in the speech recognition field, partially due to the difficulty of the problem and partially due to the absence of a common framework, an established research community, and a common problem definition. As we have continued to improve on systems that use the voice modality for human-machine interaction, however, it has become clear that automatic recognition of affect in speech is no longer a luxury that these systems can dispense with: It is a necessary ingredient for systems that adapt to the users, understand them and respond appropriately.

Automatic reliable recognition of affect in speech is a difficult problem with very large scope. Naturally, solving this problem goes beyond the work of a single doctoral thesis. The emphasis of the work proposed here is on building a computational model of the affective variation of prosodically defined acoustic parameters, using the framework of statistical learning for the automatic classification of affective labels from speech. There is a body of

work that has investigated various prosodic features to build automatic recognizers of vocal affective labels (Banse & Scherer, 1996; Dellaert et al., 1996; Roy & Pentland, 1996; Polzin, 2000; McGilloway et al., 2000; Batliner et al., 2000; Huber et al., 2000; Ang et al., 2002; Liscombe et al., 2003). A subset of this literature has focused on a series of states described as *stressed* and investigated the properties of speech produced under conditions that deviate from a neutral mode of production (Murray et al., 1996; Hansen & Womack, 1996; Sarikaya & Gowdy, 1997; Sarikaya & Gowdy, 1998; Zhou et al., 1998; Hansen et al., 1998; Steeneken & Hansen, 1999; Fernandez & Picard, 2003). Much of the work has looked at a series of parameters thought to be acoustical correlates of vocal affect, and used them as features in a vectorial space in conjunction with various statistical classification schemes to learn discriminant functions. The work proposed here still adheres to the principles of statistical classification but, motivated by theoretical notions from the field of prosodic phonology, goes beyond previous work by using the framework of graphical models or Bayesian networks to model hierarchical structure in speech.

1.1 Plan of the Thesis

This thesis is laid out in four parts. First, a Foundations section provides an introduction to the thesis (Chapter 1), as well as a background discussion on the major theoretical notions relevant to this thesis work in Chapter 2. The second part, Approach, covers the core of the thesis and the major contributions. In Chapter 3, we address the issue of producing automatic segmentations of speech in terms of structural components that, as this thesis argues, provide a suitable lens for the description and analysis of prosodic acoustic parameters. Chapters 4 through 7 stand together as a major component of this thesis, providing an account of the acoustic processing algorithms needed to support modeling of the four major components of spoken language that we have addressed: loudness, intonation, voice quality, and rhythm. The final chapter in this section (Chapter 8) then discusses the statistical learning models proposed for the automatic classification of affect from speech. This chapter provides the unifying ground for the material covered in chapter 3 with the material from the acoustic analysis chapters (4-7): As we will see, the structural description of speech provided by Chapter 3 will provide the acoustic processing modules described in Chapters 4-7 with suitable domains of analysis for the acoustic features therein proposed.

This description will also be used to generate the structured models covered in Chapter 8. An Experiments and Evaluation section follows, where we addresses the design of experiments for data annotation and the model evaluation. Chapter 9 represents a momentary departure from the analysis and modeling topic, as we address the design and execution of perceptual experiments needed for the annotation of databases of affective speech (Chapter 9). In Chapter 10, the different threads come together as we address the evaluation of the proposed approach on speech material. The thesis then concludes with some discussion of future work and some finishing remarks in Chapter 11.

1.2 Speech Materials

Throughout the discussion in this thesis, we will make reference to different data sets of speech material. We will first make use of a corpus of speech assembled from actors and annotated with affective annotations, plus a neutral category, to lend support to the analysis of the affect-relevant acoustic parameters that will be the subject of the thesis in Chapters 4-7. In addition we will make use of two data sets not explicitly annotated with affective content, but chosen for the variability of affective expression present. The first is a subset assembled from conversations of the CallHome English dataset. The second is a set of proprietary recordings from a customer service call center. We will develop experimental paradigms for gaining perceptually motivated ground truth labeling for these corpora, and then submit them, together with the actors data, to the evaluation of Chapter 10.

Chapter 2

Theoretical Background

2.1 Introduction

This chapter introduces a series of notions which provide the theoretical backdrop to the research in the thesis. In the next few sections we address the need for a descriptive framework of affect in order to approach the study of affective tone of voice. We next delineate a basic taxonomy of the means or mechanisms of speech communication in order to situate the study of affect within this whole. We conclude with an examination of the structure of spoken language and its relevance to this research.

2.2 Motivation and Applications

Why address the issue of automatic recognition of affect from speech? Until fairly recently, this endeavor may have seemed superfluous, or at least one that could be deferred until more fundamental and seemingly more pressing technical problems in Human-Computer Interaction or Artificial Intelligence had been solved more efficiently. A closer look reveals that this endeavor is neither.

Human affect is an essential component of communication between humans, and spoken language is one of the most effective channels humans have at their disposal to convey this kind of information. Humans emote, perceive the affect of their partners, and adjust their interactions accordingly. This effect is pervasive in human-to-human communication, and it is presumably more difficult to create a situation where it is altogether absent than one where it is an integral component. Much effort in human-computer interaction has been spent in

trying to endow the interaction with qualities that make it human-like. These efforts include giving computational systems such qualities as embodiment, speech recognition, language understanding and conversational abilities (Cassell et al., 2000; Breazeal & Aryananda, 2002; Bickmore, 2003). In so far as the naturalness of human-to-human communication remains the (lofty) goal of these computational systems, the affective component can not be dispensed with. Humans do not, nor should any system that attempts to mirror the nature of human interaction.

Not only is affective expression a factor that underlies human communication, and that therefore we would like to build into human-like artificial systems. It turns out that humans *already* direct their affect at technology, whether or not the technology does anything to acknowledge or react to this expression. The work of Reeves & Nass (1996) has shown that humans carry onto human-computer interaction the same attributes characterizing their interaction with other humans. In many cases technology elicits affect, often unwittingly and negatively: A recent study by Ceaparu et al. (2003) quantifying the frequency, cause, and degree of severity of frustrating experiences in computer use reported that users lost 31% to 46% of the time spent interacting with a computer due to frustration. More germane to the speech modality in particular, it has been reported that recognition rates of speech recognizers trained with neutral speech can drop by an average of 30% and as much as 60% in the presence of affective speech (Hansen et al., 1998; Steeneken & Hansen, 1999). These examples illustrate that affect *does* already permeate and qualify the interaction between humans and computers. Addressing it is therefore neither a superfluous luxury, nor one that can be dispensed with for too long.

Some of the problems highlighted in the last few paragraphs point to several straightforward applications of this kind of research. Systems can, for instance, benefit from detecting episodes of frustration, annoyance, and anger, particularly when those seem to be caused by the technology. We will later make this application more relevant in the domain of Customer Relation Management when we discuss a dataset containing phone calls with affective content. Although much work is still needed to make automatic speech recognition robust to the adverse effects of a factor like affective tone of voice, knowledge of the affective category (as well as a description of the ways in which a particular affective tone may alter the acoustic parameters used by a speech recognizer) can offer a starting point to start investigating adaptation or compensation techniques for these kinds of effect. Finally,

as conversational interfaces evolve to be more human-like, paying attention to the affective delivery of their users, and addressing them accordingly, will be a required and expected component of their performance.

2.3 Preliminaries and Definitions

Formulating a problem statement regarding the general goal of automatic recognition of emotions from speech naturally begs asking in turn for precise definitions of some of the ideas and notions that such research entails. Although the notion of emotion arises naturally and cross-culturally in the world of human interaction, the literature on the subject tends to be fragmented, and an agreed-upon definition of this term has remained elusive throughout various but differing theoretical accounts of emotion theory. This is illustrated in a review article by Kleinginna Jr. & Kleinginna (1981), who summarize 111 definitions of emotions from a literature spanning over one century.

Without appealing to a single definition, this chapter aims to introduce some relevant notions which might prove useful in describing and discussing emotion, and to briefly present some of the main models proposed for describing them. This review is not carried out with completeness in mind, but rather with the goal of abstracting a suitable framework that can be adopted when addressing the issues that arise in the development of computational systems.

The view has been advanced that building computational systems for the automatic recognition of emotion should not hinge on resolving a formal definition of the term (Picard, 1997; Cowie et al., 2001). Rather than trying to understand *emotion* in the classical conceptual framework whereby a concept is neatly tied to its extension, we can try to approach it as one would an *approximate concept*. John McCarthy has argued that most human common sense knowledge involves approximate concepts, that is, concepts that are imprecise, but on which one can base many statements having precise truth values. Offering such examples as the concept of *Mount Everest* (as well as the concept of *the social welfare of a chicken*) for discussion[1], McCarthy argues that "it is possible and even common to have a solid knowledge structure from which solid conclusions can be inferred based on

[1] "The exact pieces of rock and ice that constitute Mount Everest are unclear. For many rocks, there is no truth of the matter as to whether it is part of Mount Everest. Nevertheless, it is true without qualification that Edmund Hillary and Tenzing Norgay climbed Mount Everest in 1953 and that John McCarthy never set foot on it" (McCarthy, 2000).

a foundation built on the quicksand of approximate concepts without definite extensions" (McCarthy, 2000). Emotion may be a good candidate for an approximate concept: People uncontroversially claim to feel them and perceive them; they also, as we shall see below, engage in meaningful natural language discourse having emotion as its referent. One proposal to escape the impasse of the lack of definition is to arrive at descriptive tools for the study of emotion by first investigating how the space of emotional states is built and whether we can somehow characterize this space with a more parsimonious representation than a list of its members, perhaps by appealing to a few descriptive features, or perhaps by summarizing the space in terms of a few prototypes. A swift glance at how different authors have sought to undestand emotions can start to provide some background to the problem.

Historically, different theories have zoomed in on different aspects of emotion in order to provide a definition. Charles Darwin, for instance, viewed emotions within the context of evolutionary theory as particular action patterns beneficial to evolution. William James emphasized the physical component of emotion by asserting that emotions consist precisely of somatic (physiological) responses to particular stimuli. This classical view sets off a dichotomy between physical and cognitive components of emotion and finds its counterpart in the work of the cognitive appraisal theorists who view emotion primarily as a result of a sequence of mental appraisals, relegating the physical component to a peripheral role.

It is worth observing that in spite of the elusiveness of a formal definition of *emotion*, natural languages tend to generously allocate a variety of lexical items to denote different emotional phenomena. It is probably safe to venture that *all* languages contain *some* terms to refer to *some* emotions, although the way in which this happens is not uniform across languages: it is known that different languages focus differently on the emotional landscape and the emotional experience, different languages pointing to different emotional concepts with their lexical components. A study due to Whissel found over 60 "emotion words" to which subjects assigned distinctive emotional meaning (Cowie et al., 2001). Another classical study by Shaver et al. (2001) investigated human subjects' intuition of what an emotion is by having them rate a set of 213 candidate emotion words on a scale ranging from 1 (*I definitely would not call this an emotion*) to 4 (*I definitely would call this an emotion*). More than 150 words received an average rating of 2.5. This suggests that even if humans have difficulties defining what constitutes an emotion, they are however much more prolific at making references to specific examples and to exchange knowledge and information based

on mutual assumptions, which points to a tacit understanding of what emotions are.

There are a few other terms used in the literature on emotions which we will find useful to mention. *Affective state* is a term often used interchangeably, although it tends to designate a broader class of states than those encompassed by the term *emotion*. Some authors delineate the difference along an intensity dimension, reserving the term emotions for more "full-blown" and short-lived phenomenon (Cowie et al., 2001) and "affect" for a milder state (Fell, 2003). *Mood*, on the other hand, seems to refer to an emotional state that underlies the occurrence of emotional phenomena, an emotional baseline of sorts on which more transitory emotional phenomena take place. In the context of this work, we will also find it useful to make a key two-way distinction between, on one hand, the *internal affective state* of a subject and its *externalization*, (via whatever channels available to the subject), and on the other hand between this external display and the *perception* of this externalization by an observer. The first distinction is necessary since the externalization of an internal affective state can be modulated by a set of display rules that account for the influence of such factors as culture, social milieu, context, gender, etc. on the outward expression (Ekman, 1993). The latter distinction is needed to account for the mismatch that often occurs between intended and perceived emotional content, a fact that is empirically documented by various studies investigating, for instance, how subjects decode facial expressions or the affective content of spoken utterances. This is an aspect of emotional communication that is particularly relevant to designers of systems that aim to perceive and express emotion, and which deserves full understanding if we need to take into account the interaction with a user. This emotional mismatch can arise out of a complex interaction of factors, amongst which we can cite:

- discrepancies between the display rules mentioned above for observer and emoter (as might be the case when members of different cultures bring in different rules about facial expression display into the interaction),

- limitations of the medium encoding the affective expression, something which has been termed the *affective bandwidth* of a medium (this may be illustrated by the case when vocal linguistic messages are transcribed with a written medium and stripped of the oral component),

- polysemy or multifunctionality of the signs used to encode the affective component of

a message (relevant when disambiguating, for instance, whether lowered eyebrows are a sign of anger or concentration).

Having introduced some relevant terminology and raised some distinctions we feel are particularly relevant to this work, we now turn to take a look at models that we can use, particularly as computational tools, to describe and study affect.

2.4 Models of Affect

This section does not aim to be an overview of the vast existing literature on emotion theory. Its goal is to present a simplified account of the kinds of descriptive frameworks that have been proposed to study emotion. As suggested by the experiments of Whissel and Shaver mentioned earlier, humans are adept at using an extended emotion vocabulary to describe and communicate a range of emotional experience. Leaving aside the vexing question of what constitutes an emotion, we might ask instead what describes this set, hoping to perhaps arrive at a description in terms of a parsimonious set of features that can generate the set. Alternatively we might reduce the set to a few clusters where members share some similarity measures with each other, or perhaps just with a prototype that typifies the cluster ensemble. These two ways of approaching the problem parallel two basic kinds of models that have been proposed to describe emotion. The latter is a *prototype* or *categorical* model of affect; the first approach is a *dimensional* model.

Categorical models of emotion posit that emotions are discrete and belong to one of a few groups. Furthermore, some of these groups are more fundamental or "basic" in some sense. The number of such groups and the nature of what defines a fundamental set has been a contentious question for some time. Although proposals diverge, there are a few contenders that appear commonly on different authors' lists; among these we find *fear*, *anger*, *sadness*, and *happiness* (Picard, 1997). This kind of proposal suggests the question of what criterion defines an emotion as being basic. One answer to this question has been provided by the work of Ekman (1993), who has adduced anthropological arguments to favor his set of six basic emotions. (Ekman's set also includes *disgust* and *surprise* in addition to the four already mentioned.) Ekman has studied emotion in the context of facial expression and reported distinctive patterns associated with each of these that are observable in members of vastly different cultures. Further support may be given by the studies of Shaver et al.

(2001), who found that a hierarchical clustering of subjects' ratings of the similarity between 135 emotion words, yielded five of these six emotions as cluster prototypes.[2]

In contrast to discrete emotion models, dimensional models do not reduce emotion to a finite set. Instead, they attempt to find a finite set of underlying features or dimensions into which emotions can be decomposed. These models are generative in the sense that any combination of feature values can give rise to a different (conceptually, at least) affective state. If, furthermore, the dimensions considered are continuous, the model is able, in principle, to generate an infinite number of affective states.[3] One of the most discussed dimensional models of affect is originally due to Schlosberg (1953), who proposed a three-dimensional structure to emotion. Under this model, affective states may be described in terms of three components or axes: one describing the degree of pleasantness underlying the emotional experience, one describing the level of activation of the emotion, and one describing the level of attention or rejection.[4] A simplified version of this model using its first two components, also known as *valence* and *arousal*, has become quite popular in the study of emotion, particularly in computational studies, since it offers a very simple (though simplified!) view of affect, and it has a readily accessible semantic interpretation.

As a computational tool it has been used to characterize the eliciting affective content of images in the work of Lang (1995). This model has also been evaluated as a perceptual tool to gather subjects' appraisals of the affective content of speech (Pereira, 2000), and in Chapter 9 we discuss an application of the model to annotate corpora of affective speech.

Having brought up some general theoretical notions related to the study of affect, and described theoretical models that can be used as tools for the description of affective content, we now turn to the particular modality of interest in this thesis: speech.

2.5 Speech and Emotional States

There has been a significant amount of work produced on the topic of speech and emotional states. The focus of much of the traditional research in the linguistic literature has been on a class of states described as *attitudes*. Unfortunately, what distinguishes *attitudes*, and par-

[2] *Disgust* was the only emotion that did not match this result; the status of *disgust* as a basic emotion has also been challenged in other places in the literature (Banse & Scherer, 1996).

[3] This is an altogether different issue from whether a continuum of affective states is perceived.

[4] This last dimension, also called *stance* in the literature, describes the tendency to attend to a stimulus (such as when experiencing fear or surprise) or to withdraw from it (such as when experiencing contempt or disgust).

ticularly what sets them apart from other kinds of affective states, does not clearly emerge from a literature characterized by a proliferation of descriptive labels and inconsistent use of terminology. Informally speaking, they typically describe a position on the speaker's part toward a linguistic exchange. So while a label like *happy* can describe an emotion, the term *friendly* may be used to designate an attitude. In a review and critique of this body of work, Wichman (2000) proposes to separate these two notions by shifting *attitude* to the domain of pragmatics: attitudes are interpretations from the linguistic exchange that involve the complex interaction of contextual features (such as the relationship between hearer and speaker and commonly held beliefs between them as well as the text itself. She suggests, therefore, that the speech signal itself does not convey attitude, and allows that only *emotion* may be transmitted in the encoding of particular acoustic-phonetic patterns. We will not invoke a distinction in this work between attitudes and emotion. However, this distinction is worth mentioning since her position represents an attempt to provide an objective demarcation of these two commonly used labels, and because the research in this thesis –being purely based on acoustical analyses of the speech waveform– lies in line with her view of what emotions in speech are.

The study of the affective variation of speech naturally motivates more basic and fundamental questions regarding the vocal manifestations of language. What is the range and the nature of the vocal phenomena expressed by a speaker? How much of what is uttered by a speaker is *meaningful* in some sense? What aspects are used and controlled by the speaker with the intention of communicating, and how much of it are concomitant epiphenomena that exist as byproducts of linguistic communication? As we shall see in the next section, speech acts as a carrier of multiple sources of information, and it might be advisable to gain an understanding of the nature of this multiplicity to accord the study of the vocal affective variation its proper place within this complexity.

2.5.1 Linguistic, Paralinguistic, and Extralinguistic Aspects of Speech

We will address this issue by invoking a classical view that has tried to elucidate the multifunctionality of speech by way of a distinction between its *linguistic*, *paralinguistic* and *extralinguistic* aspects. Although (as seems to be the theme in this chapter) no consensus exists about the boundaries between these notions, the distinction is a fruitful one that can help us place the study of the affective variation of speech within its proper context.

Perhaps it is best to start with what is possibly the most agreed upon and delineated of these domains. Properly *linguistic* aspects of spoken language are those that use an *arbitrary* system of contrasts and can serve a communicative purpose. Following Laver (1994), we feel the need to stress the difference between communicative and informative aspects of a signal. *Communicative* aspects are those manipulated by the sender with the intention of making the receiver aware of a particular signification; *informative* aspects, on the other hand, are those that manage to signify something to the receiver irrespective of the sender's intentions, perhaps even beyond his control. One of the defining characteristics of linguistic systems is that its elements are quantal in nature, and that the encoding is therefore achieved by appealing to combinations of a finite set of units. Examples of these are the set of a language's phonemes used for phonological contrast. When discussing spoken language we will use the terms *linguistic* and *phonological* synonymously.

The term paralinguistic seems to have emerged in the 1950s to describe those aspects of vocal communication that were meaningful but resided outside the linguistic structured system just described (Ladd, 1996). For Ladd (1996) the difference between language and paralanguage resides strictly in the quantal or categorical nature of the former and the gradient or continuous nature of the latter. However, when he formulates this distinction he seems to have in mind some features of language which can be used both linguistically and paralinguistically (e.g., the manipulation of pitch to produce categorical intonational contour vs. manipulations to exact a higher pitch range).

Perhaps the most detailed and encyclopedic treatment of paralanguage has been given by Poyatos (1993), whose definition of the term is so precise and extensive that is worth quoting in full

> the nonverbal voice qualities, voice modifiers and independent utterances produced or conditioned in the areas covered by the supraglottal cavities (from the lips and the nares to the pharynx), the laryngeal cavity and the infraglottal cavities (lungs and esophagus), down to the abdominal muscles, as well as the intervening momentary silences, which we use consciously or unconsciously supporting or contradicting the verbal, kinesic, chemical, dermal and thermal or proxemic messages, either simultaneously or alternating with them, in both interaction and noninteraction.

For Poyatos, therefore, paralanguage involves an array of modalities (voice, gesture, touch, distance) as well as more than the communicative-interactive purpose.

We will use the term *extralinguistic* to designate any aspects of the speech signal which are informative but not communicative. They include information such as the age and gender of the speaker, or any aspects indicative of a voice pathology or a physiological state. However, it is important to stress that, as it has often been observed (Roach, 2000; Laver, 1994), any given speech feature can be multifunctional, and therefore it seems well advised to speak of *linguistic*, *paralinguistic* or *extralinguistic usages* of a feature rather than to describe the feature itself as belonging to one of these categories. For instance, according to the distinctions laid out above, a set of linguistic contrasts germane to a dialect associated with a particular social background can be used, additionally, in a paralinguistic fashion if implemented by a non-speaker of the dialect for a particular purpose during a social interaction. Likewise, the vocal patterns of a particularly young or old age can be used paralinguistically by speakers outside the age group.

Where in this schema does the study of emotion fit in? The short answer is that it does not fit neatly into any component. Emotion is a factor that often appears classified both under paralinguistic and extralinguistic headings by different authors (Schötz, 2003). The view of affect as playing a role in paralanguage seems to be warranted by the linguistically non-contrastive nature of some of the resulting vocal variation (Poyatos, 1993) as well as by its communicative role (i.e. used by the speaker with the *intention* to communicate) (Laver, 1994). The view of emotion as an extralinguistic phenomenon, on the other hand, owes its status to the notion that affect is to some extent physiologically and biologically regulated. These differing classifications are, therefore, motivated by different aspects of emotion. Indeed, a research program that tries to delineate the differences between the vocal expression of affect (whether genuine, acted, elicited) on the one hand, and the congruence or incongruence between this expression and an underlying physiological-cognitive affective state, on the other, would be served well by this distinction. Such research may want to sift extralinguistic from paralinguistic effects and try to isolate distinctive perceptual and acoustic cues. This distinction, however, will not be invoked in this work, where the interest resides in the vocal expression of affect.[5]

[5] We will, nonetheless, be working with speech corpora, consisting of both acted and spontaneous speech, which might typify this distinction. This difference will be definitely be made explicit whenever relevant and discussed in more detail later on. However, the focus is not on characterizing or recognizing the differences

The research in this area leads us to believe that the perception of affect from speech is informed by all of these aspects of language. The connection between vocal emotion and paralanguage is clear since speakers often appeal to various sources of non-linguistic variation at their disposal for the purpose of encoding affective expression (e.g., raising the voice paralinguistically, speaking faster paralingustically). Less obvious, however, is the effect that employing categories that are thought to be part of phonological descriptions can have on the perception of affect, as we will see in later chapters (Scherer et al., 1984; Ladd et al., 1985; Mozziconacci, 2000). It also seems reasonable to entertain that extralinguistic aspects of the voice pointing to factors like age and gender may have a confounding effect on the perception of the affective content of an utterance. Attempting to delimit these notions, and determining where to place the study of affective vocal variation is more than a mere exercise in cataloging. Because of the multifunctionality already mentioned, it is good practice to keep in mind the range of factors that contribute to vocal variation along any dimension (intonation, loudness, voice quality, etc.), and then isolate the ones that are relevant to the study of affect. To approach the study of spoken language, however, we need to have an understanding or a framework for the description of how it is structured. To this question we turn next.

2.5.2 The Structure of Spoken Language

It is a commonplace observation that a description of a spoken utterance involves more than a transcription of its textual component. The latter, although in many cases sufficient to transmit essential semantic aspects, is stripped of much variability that adds to the transmission of other types of information, many of which are non-redundant and essential to communication. A characterization of a spoken utterance, therefore, calls for a description that goes beyond a specification of the linear arrangement of units over time; it calls for a *prosodic* description.

A survey of the extensive literature on *prosody* reveals that a common definition of the term is difficult to pin down. Several definitions link the concept of prosody to suprasegmentals. Suprasegmentals are "features whose arrangement in contrastive patterns in the time dimension is not restricted to single segments defined by their phonetic qualities (i.e., distribution of energy in the frequency dimension)" (Lehiste, 1970). In other words, *pa-*

between these two modes.

rameters of speech that are defined above the segment (phone) level and whose variation over time may exist independent of the arrangement of the segments. The classical prosodic parameters cited in the literature comprise pitch, intensity, and duration. This definition poses problems since these parameters are also known to correlate with distinctive features of phones even for non-tone languages (such as voicing and place of articulation), and therefore their function independent of the segmental level is suspect. In the classical generative phonological framework of Chomsky & Halle (1968), prosody also shared this status as a collection of suprasegmental features that did not clearly fit into a phonological system with the segment as its major organizing constituent.

Unlike this *classical* view, in which prosody is treated as a collection of parameters, more recent metrical theories have viewed prosody as the *structural organization* of speech into hierarchical constituents and the patterns of prominence between them. The classical view of prosody is restrictive in the sense that it limits itself to only three major articulatory/acoustic dimensions of speech (with the possible exclusion of other parameters which may be influenced by a speaker's organization of the segments into a particular surface form). In contrast, the *structural view* shifts the focus from a collection of speech parameters to a description of the organization of speech into meaningful units. This definition still allows one to describe changes in the classical parameters as prosodic phenomena, but as viewed through the structural constructs that influence these parameters in particular systematic ways. Beckman, working within the framework of the autosegmental and metrical theory, argues that "prosody is not just another word for 'suprasegmentals'; rather, it is a complex grammatical structure that must be parsed in its own right"; it "does not designate any set of distinctive features, but instead refers to the ... structural organization itself" (Beckman, 1996).

Shattuck-Hufnagel & Turk (1996) have suggested using a working definition that merges these two views of prosody by invoking into the definition (i) the patterns of variation in pitch, intensity, duration, and stress that can be best accounted for by high-level organizational structures and (ii) the high-level structures themselves that account for these patterns. This view of prosody is useful since it allows us to pay attention to measurable acoustical parameters from a waveform while considering the structure a defining characteristic. Much of the work on automatic recognition of affect from speech has emphasized different acoustic cues that might be salient predictors of affect, typically averaging out

values over some time span, without paying the conceptual construct its due attention. Unifying these under a single class of models will be one of the attributes of the approach presented in this thesis.

This notion of prosody as structure is a very general one that applies to several models. Most of these models share the view, however, that these structures are hierarchical in nature, similar in some respects to the way syntax is hierarchically organized in a sentence, but different also in many other respects. The main formulation of a hierarchy in prosody is given by the Strict Layer Hypothesis originally articulated by Selkirk (1984), and which may be rendered as follows. There exists a ranked series of prosodic categories or constituents forming a tree structure, such that constituents at a given level must exclusively parse or decompose into constituents at the immediately lower level (Hayes, 1989; Ladd, 1996). The inventory of constituents also vary in different versions of the Strict Layer Hypothesis but typical constructs include the notions of *utterance*, *intonational phrase*, *prosodic word*, *metrical foot*, and *syllable*. According to the definition of prosody reviewed above, describing the prosody of an utterance does not reduce to specifying a few acoustic parameters, but also includes a description of how these parameters are structurally associated with these constituents and how they are temporally arranged.

The Strict Layer Hypothesis is an attractive computational theory since it describes the prosodic structure of an utterance in terms of a fixed number of tiers. Furthermore, the relationships between constituents induce tree-like structures which can lend themselves to efficient tree-like computational structures. It is worth emphasizing that there are other existing theories of prosody that also invoke the notion of a structural organization of speech but use different representations to describe this structure. The work in metrical phonology and the use of metrical grids, for instance, is a noteworthy example. The Strict Layer Hypothesis, however, serves as a good point of departure to couple a theory about the structural organization of speech with a theory that makes use of these induced structures to motivate learning machines. This connection will be more fully explicated in Chapter 8 when we propose a hierarchical-dynamical network model for modeling the distribution of acoustic parameters along the time dimension and across different scales. The next few chapters, however, pave the way for this type of model, devoting attention to the modeling of acoustic parameters for the description of several components of speech (Chapters 4-7) and algorithms for parsing the speech waveform into some of the structural units which we

have discussed here (Chapter 3).

2.6 Chapter Summary

In this chapter we have discussed some fundamental theoretical issues relevant to the study of spoken language and the study of emotions. While avoiding the issue of defining what constitutes an emotion, a phenomenon not yet fully understood, we have tried to make the case that emotional phenomena may be approached armed with some theoretical descriptive framework which allows us to carve out the domain of study, reviewing in the process models that may be used as tools in the study of affective expression. We have introduced fundamental theoretical linguistic notions, including a broad description of the range and the nature of vocal phenomena as well as the structure of spoken language, which we believe informs the study of emotional expression in speech and will further motivate the line of research in this thesis.

Part II

Approach

Chapter 3

Prosodic Parsing

3.1 Introduction

It remains a challenge to obtain a fully automatic parsing of an acoustic waveform in terms of prosodic constituents, such as those proposed by the Strict Layer Hypothesis outlined in Chapter 2, especially if we aim to produce a full segmentation at every level of the hierarchy. Previous work in this area has been more successful when the constituents are intonationally characterized (i.e., when detecting minor or major intonational phrase boundaries) (Wightman & Ostendorf, 1991; Wightman & Ostendorf, 1992; Wightman et al., 1992; Wightman & Ostendorf, 1994; Kompe et al., 1994; Hirschberg & Nakatani, 1998). Such approaches have benefited from having access to a segmental alignment with the acoustic waveform (typically the output of a speech recognizer in the fully automatic case, or of hand-labeled transcriptions otherwise), from which segment-dependent properties (e.g., lengthening) can be assessed to diagnose the possible occurrence of an intonational boundary. Furthermore, the segmental representation enables, in principle, the alignment of the acoustic waveform with a syntactic parse, which in turn can shed light on the prosodic constituency of the waveform given the degree to which syntax and phonology are interrelated.

In this work, we wish to remove the assumption that a segmental alignment is available. Not only is the main goal of this thesis to propose a system that can make inference from prosodic cues alone, it is also expected that the output of a speech recognizer will be particularly unreliable in the presence of speech with affective variability, not to mention the effects that factors such as variety of dialects and noisy phone-quality speech will have on compounding this degradation.

In this chapter we propose a method which identifies salient structural units in speech from an acoustic waveform. The proposed approach works in a bottom-up fashion by successively building higher-level representations using information about structural constituency from lower-level units, and a set of features defined in terms of these. Beginning with a short-term spectral representation at the acoustic level in terms of fixed-sized intervals, the system first identifies acoustic segments using the blind segmentation algorithm described in Section 3.2. In Section 3.3, we describe how this first estimate of acoustic segment boundaries is then combined with information about syllable centers to locate the boundaries of syllable nuclei and refine the acoustic segmentation. At this stage, the system also identifies the presence of discrete acoustic events like pauses and breaths. The representation obtained in terms of nuclei, inter-nuclear material, pauses and breaths serves then as the basis for the algorithm presented in Section 3.4 for identifying major intonational boundaries.

3.2 Blind Acoustic Segmentation

For the present discussion we will informally consider an acoustic segment to be a region of the acoustic waveform during which some properties (e.g., typically spectral or cepstral) are somewhat stationary. In the simplest cases, an acoustic segment is thus related to a single phonetic unit when the properties under consideration do not vary greatly throughout the duration of the phone. Although this can be the case for certain classes, like fricatives, we will note that there are also classes, most notably stop consonants, vowels and glides, for which the production of a phone undergoes various distinct phases, resulting in noticeably non-stationary behavior in the acoustic waveform and its spectral content. For the present work we are not interested in determining whether an acoustic segment maps onto a single phone, or whether it is a constituent of a more complex one. The objective of a blind acoustic segmenter is to find the transitions between adjacent acoustic segments without having access to any linguistic information concerning the nature or total number of such segments. We will assume a spectral representation of speech throughout the rest of this discussion.

Finding the boundary between two segments is facilitated by the degree to which there is a sharp transition between them: We should expect to be able to better detect a transition between two segments whose acoustic properties are very different (e.g., a high-energy

fricative followed by a vowel) than a transition between segments of similar properties (e.g., a glide followed by a vowel). At any rate, thinking of acoustic segments as regions of stationary spectral content suggests that we can construct a function to monitor the variability between spectral frames and that we can expect to find this function peaking at the points where this variability is greatest, the boundary of interest. Let $X(k,m)$ denote the squared magnitude of the short-time Fourier transform of a speech signal $x(n)$ (spectrum), calculated every T_{step} seconds with a window of T_{win} seconds, and let $k = 1, \cdots, K$ and $m = 1, \cdots, M$ denote the frame index and the frequency bin index respectively. If we let $IFV(k)$ represent the inter-frame variation function in the vicinity of frame k, then the boundaries are the indices that (locally) maximize this function in a neighborhood (of width $2N$ centered around k):

$$\hat{k} = \underset{k \in (k-N, k+N)}{\operatorname{argmax}} \; IFV(k), \tag{3.1}$$

A good candidate function IFV should exhibit local maxima in the vicinity of a boundary with no spurious peaks. Spectral distortion measures involving some metric have been proposed as a first stage in segmentation algorithms (Eberman & Goldenthal, 1996; Alani & Deriche, 1999). However, because they tend to be sensitive to small changes between adjacent frames, further refinements are often required. For instance, in the segmentation algorithm proposed in Sharma & Mammone (1996), the Euclidean norm of the difference between adjacent frames (using cepstral coefficients instead) is used to help establish an upper bound on the number of segments since it tends to oversegment. This tendency is illustrated in panels (c) and (d) of Fig. 3-1 where the following spectral distortion using a 2-frame neighborhood around each frame has been evaluated using an Euclidean distance metric (c) and the Kullback-Leibler divergence (d).

$$IFV_d(k) = \sum_{r=k-1}^{k} \sum_{s=k+1}^{k+2} d\Big(X(r,m), X(s,m)\Big) \tag{3.2}$$

The purpose of this section is to introduce an algorithm that constructs a more robust measure of the temporal spectral variability of a signal in order to determine acoustic boundaries more reliably. The result can be previewed in Fig. 3-1 (e).

Since, as we have noted, a transition between acoustic segments brings about a concomitant change in the distribution of energy across some frequencies, we may concentrate on

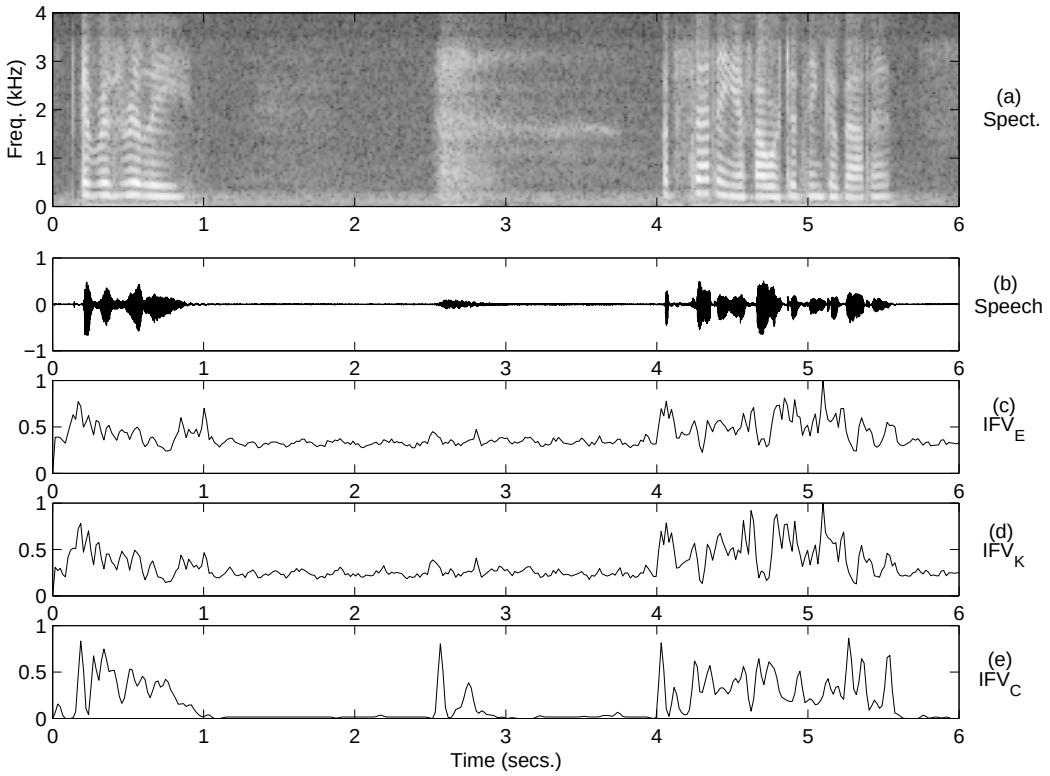

Figure 3-1: Spectral inter-frame variability. Panels (c)-(e) show different inter-frame variability functions for the spectrogram in (a) corresponding to the speech waveform in (b). Panel (c) and (d) show plots based on the Euclidean and Kullback-Leibler divergence metrics respectively. Panel (e) plots the spectral variation function built based on detecting strong edges on the spectrogram.
algorithm

the frequencies where this change is most sharply pronounced, and attribute less weight, or perhaps ignore altogether, the frequency bands containing smoother changes. In particular, for any pair of adjacent frames, we can locally examine small frequency regions and build a local indicator function that votes in favor or against a potential transition depending on whether or not it encounters a pronounced change in that region. By adding up all such votes across frequencies, we can build a function reflecting the total variability. Notice that this procedure effectively applies an early thresholding operation since it only gets to record potential boundaries if it considers them strong enough, ignoring them otherwise. Unlike the cumulative spectral distortion measures discussed earlier, this approach may build some resilience against the kind of noisy meandering that was observed through regions of small spectral change.

The second intuition we would like to invoke to motivate the algorithm is based on the

observation that these concomitant changes in energy often describe vertical edges on the spectrogram, resulting from the changes in spectral intensity over time.[1] Combining the last two remarks suggests a simple algorithm that cycles through the frequency spectrum, poses the problem of finding a strong enough transition as one of detecting local strong edges, and accumulates the findings to build an auxiliary function for detecting transitions.

Before fully specifying the details of the algorithm, let us address the issue of detecting the edges. As noted above, we wish to retain transitions that are sharply pronounced. By treating $X(k, m)$ as an array (image) of values where the (k, m) entry (pixel) contains a function of the spectrum (intensity levels), we are free to apply any edge detection algorithm that can pick out vertical edges (horizontal transitions) on the array. A suitable algorithm would be one that is able to sift out strong edges and perhaps disregard softer edges most of the time.

The algorithm originally described in Canny (1986) has become a powerful tool to detect edges in an image since it explicitly incorporates the notion of strong and weak edges. The algorithm uses two thresholds, rather than one, to define a hysteresis thresholding procedure whereby strong edges are unconditionally retained, and weak edges are rejected unless they are connected to a strong edge. The algorithm consists of three steps. First, gradients are estimated by convolving an image with two operators, first with a Gaussian and then with the first derivative of a Gaussian. In the second step, these gradients are linearly interpolated around each pixel to determine whether the pixel in question is maximal for some direction in the image. Finally, non-maximal pixels are set to zero, and the maximal pixels subjected to the hysteresis process mentioned. The details of the algorithm are fully given in Algorithm 3.2 (Fig. 3-3). This is the general algorithm to detect arbitrary edges in an image. Since we are interested in vertical edges (changes in intensity in the horizontal direction), we have restricted the search for maximal pixels in the directions specified by the last two panels in Fig. 3-4.

We are now ready to fully specify the segmentation procedure described by Algorithm 3.1 in Fig. 3-2. At every step of the algorithm, an "image" is constructed by taking a slice of width 3 (i.e., the frequency bin of interest plus the surrounding bins; larger neighborhoods could also be considered) from the spectrogram around a frequency of interest. This image

[1] We are quite adept, for instance, at finding major spectral transitions from a spectrogram image, a result that may presumably owe more to our visual ability to detect edges, than to any knowledge about spectrogram interpretation.

is submitted to Canny's algorithm to construct a binary edge map, and the row of the map associated with the frequency under investigation retained as an indicator function. This process is repeated by sweeping over the frequency dimension after every step. At the end of this pass, the indicator functions are averaged to construct the variability function IFV_C. A sample IFV_C is shown in panel (d) of Fig. 3-1. The resulting segmentation is shown in more detail for a speech fragment in Fig. 3-5. This algorithm has been tested on signals sampled at 8 and 16 kHz, and has been implemented starting with spectrograms computed from 20- msec. short segments of speech, in steps of 17 msecs. The edge detection step is carried out with a width parameter $\sigma = 2$, and only peaks at least 3 frames from each other are considered to be segment boundaries.

Algorithm 3.1: Let $s(n)$ be a speech signal

1. Find the short-time Fourier transform $S(k,m)$ from $s(n)$, using a window length of 20 msecs. every 17 msecs., zero-padded to the next integral power of 2.

2. Find the magnitude, and apply square-root compression to reduce the range of the spectrum: $X(k,m) = |S(k,m)|^{\frac{1}{2}}$

3. Let $k = 1, \cdots, k_{max}$ index the time frames, and $m = 0, \cdots, m_{max}$ index the frequency samples ($m = 0$ corresponds to DC). Then for $m = 2, \cdots, m_{max} - 1$:

   3a. Extract the mth local spectral band $I^{(m)}(k,j) = X(k,j)$ for $k = 1, \cdots, k_{max}$, $j = m-1, m, m+1$. This results in a $k_{max} \times 3$ array of local spectral values.

   3b. Apply Canny's algorithm (see Algorithm 3.2) to $I^{(m)}(k,j)$ and obtain the binary-valued edge map $E^{(m)}(k,j)$.

   3c. Let $A(k,m) = E^{(m)}(k,2)$ for $k = 1, \cdots, k_{max}$; that is, save the locations of estimated local vertical edges (sudden changes in spectrum over time) for the center frequency under consideration.

4. Compute the inter-frame spectral distortion function by averaging over local vertical edges:

$$IFV_C(k) = \sum_{m=2}^{m_{max}-1} A(k,m) \tag{3.3}$$

5. Find the local maxima of (3.3), and assign those to the set of frame boundaries.

Figure 3-2: Spectral acoustic segmentation algorithm.

Algorithm 3.2: Let $I(x,y)$ be the input image, and σ a width parameter. Let $E(x,y)$ be a binary-valued array representing the desired edge map.

- Gradient Estimation

  1. Smooth out the image by convolution with a 2D Gaussian $G(x,y)$

$$I_s(x,y) = I(x,y) * G(x,y) = I(x,y) * \exp\Big(-\frac{x^2+y^2}{2\sigma^2}\Big) \tag{3.4}$$

  2. Apply the directional Gaussian-derivative operators

$$\frac{\delta G}{\delta x} = G'(x,y) = -\frac{x}{\sigma^2}\exp\Big(-\frac{x^2+y^2}{2\sigma^2}\Big) \tag{3.5}$$

  (with $\frac{\delta G}{\delta y} = G'(y,x)$) to obtain the gradient estimators $I_g(x,y)$:

$$I_g(x,y) = G'(y,x) * \big(G'(x,y) * I_s(x,y)\big) \tag{3.6}$$

- Non-Maximum Suppression

  3. Let $E_g(x,y) = 0, \forall x,y$. For each center pixel (x_o,y_o) in the array $I_g(x,y)$ and each of the four gradient directions shown in Fig. 3-4, let (x_1,y_1) and (x_2,y_2) be the pixels flanking the p direction, and (x_3,y_3) and (x_4,y_4) the pixels flanking the q direction.
  4. Linearly interpolate the gradient

$$I_g^p(x_o,y_o) = \beta I_g(x_1,y_1) + (1-\beta) I_g(x_2,y_2) \tag{3.7}$$
$$I_g^q(x_o,y_o) = \beta I_g(x_3,y_3) + (1-\beta) I_g(x_4,y_4) \tag{3.8}$$

  5. If $I_g(x_o,y_o) > I_g^p(x_o,y_o)$ and $I_g(x_o,y_o) > I_g^q(x_o,y_o)$ for some p and q (i.e., (x_o,y_o) is a maximal pixel), then $E_g(x_o,y_o) = I_g(x_o,y_o)$.

- Hysteresis Thresholding

  6. Dynamic thresholds selection: Let $\tau_h = prctile_{70}\{I_g(x,y)\}$ and $\tau_l = 0.4\tau_h$. Let $E(x,y) = 0, \forall x,y$.
  7. Let $\mathcal{I}_l = \{x,y\}$ s.t. $E_g(x,y) \le \tau_l$ and $\mathcal{I}_h = \{x,y\}$ s.t. $E_g(x,y) \ge \tau_h$. Then for every pixel (x,y):
     7a. If $(x,y) \in \mathcal{I}_l, \Rightarrow E(x,y) = 0$.
     7b. If $(x,y) \in \mathcal{I}_h, \Rightarrow E(x,y) = 1$.
     7c. Otherwise, $E(x,y) = 1$ if there exists a path from (x,y) to a pixel in $\mathcal{I}_h$, and every intermediate pixel does not belong to $\mathcal{I}_l$. A path is any sequence of pixels, such that two pixels in the path are neighbors in $I(x,y)$ (using 4-pixel connectivity in this case).

Figure 3-3: Canny's algorithm for edge detection.

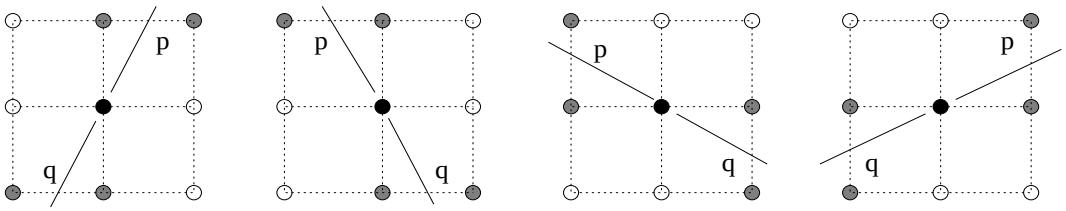

Figure 3-4: Gradient directions for the evaluation of step 3 of Canny's algorithm. Filled pixel represents the center pixel. Shaded pixels represent the pixels from which the gradient in directions p and q is calculated. For vertical edge detection, only the directions in the last two panels are considered.

3.3 Identification of Syllable Nuclei

In order to detect and establish the boundaries of syllable nuclei, we avail ourselves of the acoustic segmentation obtained through the algorithm described in the previous section, and of a function measuring the amount of energy in the region of the first few formants. The frequency band in the area around the first few formant is expected to have high energy for vowel segments, an acoustic cue that is often used as a basis for syllable detection (Mermelstein, 1975; Pfitzinger et al., 1996). Following Cummins & Port (1998) and references therein, we apply the following procedure to obtain an accurate estimate of an energy profile primarily influenced by the presence of a vowel, and use that to select the locations of nuclei.

- Band-pass filter the speech signal with a Butterworth filter $h(n)$ of low order ($N = 2$) with cutoff frequencies at $f_l = 500$ Hz. and $f_u = 1500$ Hz.:

$$s_{bp}(n) = s(n) \star h(n) \tag{3.9}$$

  The low-order filter has the effect of smoothly tapering the energy at frequencies outside the band of interest without introducing sharp discontinuities in the energy profile.

- Find the running short-term energy by summing a squared speech segment, windowed with a 100 msec. Hanning filter centered around each sample. This can be efficiently implemented via the convolution

$$e(n) = s_{bp}^2(n) \star w_h^2(n), \tag{3.10}$$

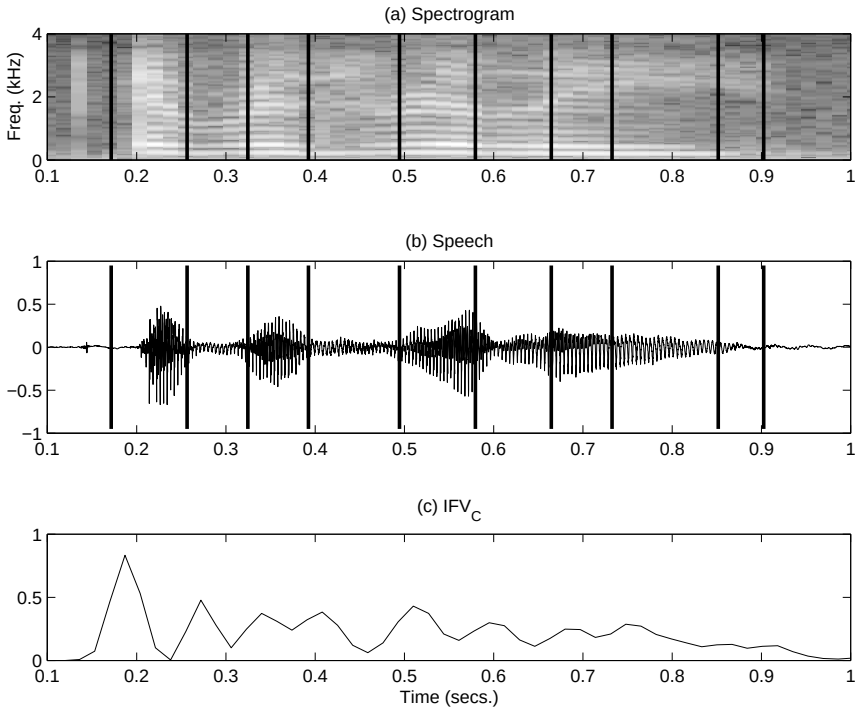

Figure 3-5: Result of running the acoustic segmentation on the spectrogram shown in (a) corresponding to the waveform in (b). The auxiliary spectral variability function in (c) peaks at the points to which acoustic boundaries are assigned.

where $w_h(n)$ contains the Hanning window coefficients.

- Find peaks in the resulting energy signal $e(n)$ separated by at least 80 msecs. Let p_k represent the location of these peaks. Subject to the same constraint, find the valleys between any two adjacent peaks, and let v_k be the location of a valley between p_k and p_{k+1}.

- Find the largest normalized jump associated with a peak:

$$\delta_k^l = e(p_k) - e(v_{k-1}) \tag{3.11}$$

$$\delta_k^r = e(p_k) - e(v_k) \tag{3.12}$$

$$\delta_k = \max\left(\frac{\delta_k^l}{e(p_k)}, \frac{\delta_k^r}{e(p_k)}\right) \tag{3.13}$$

- If δ_k is less than a threshold η ($\eta = 0.1$), then discard the location p_k from the set of peaks.

The final set of p_k contains the locations of syllable nuclei.

Having identified nuclei locations through this procedure, we use the acoustic segmentation previously obtained to identify segments containing these nuclei. Since the segmentation is exhaustive and non-overlapping, there is only one segment containing the kth nucleus given by p_k. The converse, however, is not guaranteed and, though it was not found to be the case with the output of the segmentation algorithm implemented here, it may be the case that a segment contains more than one nucleus as detected by the procedure above. More likely is the case (although still infrequent) that two adjacent segments contain two adjacent nuclei if the acoustic segmenter failed to detect a brief or smooth spectral change between the two nuclei. In order to correct this undersegmentation, we introduce the following refinement heuristic to ensure that there is always (at least) one segment of non-nuclear material between two nuclei: If two adjacent peaks located at p_k and p_{k+1} are assigned to the same or to two adjacent acoustic segments, we augment the boundary set with two additional boundaries located at p_k^* and p_{k+1}^*, where the locations are chosen such that the following holds:

$$\begin{aligned} e(p_k^*) &= e(p_k) - \xi\delta_k^r & p_k^* &\in (p_k, v_k) \\ e(p_k^{**}) &= e(v_k) + (1-\xi)\delta_{k+1}^l & p_k^{**} &\in (v_k, p_{k+1}), \end{aligned} \tag{3.14}$$

with δ_k^r and δ_k^l given as in (3.11) and (3.12) and $0 < \xi < 1$ ($\xi = 0.6$ in this implementation). This is graphically illustrated in Fig. 3-6. Essentially, starting at one peak, we slide down the slope of the energy profile toward its local minimum, and place a new boundary when we reach a point which equals ξ percent of the gap between the energy at the peak (assumed to roughly correspond to the center of a nucleus) and the energy at the minimum (assumed to roughly correspond to the point furthest away from the surrounding nuclei). One more step is added at this stage to eliminate any high-energy noise-like segments that may have caused a local maxima in the energy function, by eliminating candidate nuclei that do not contain more than 90% voiced frames in their pitch contour. By the end of this stage, we have obtained a possibly more refined segmentation where segments corresponding to syllable nuclei have been identified. It might also be desirable to identify other acoustic segments corresponding to discrete paralinguistic events occurring on the speech waveform. In particular, the presence of pauses and breaths are of importance to us, as they provide significant cues about the placement of major phrase boundaries. Furthermore, they can

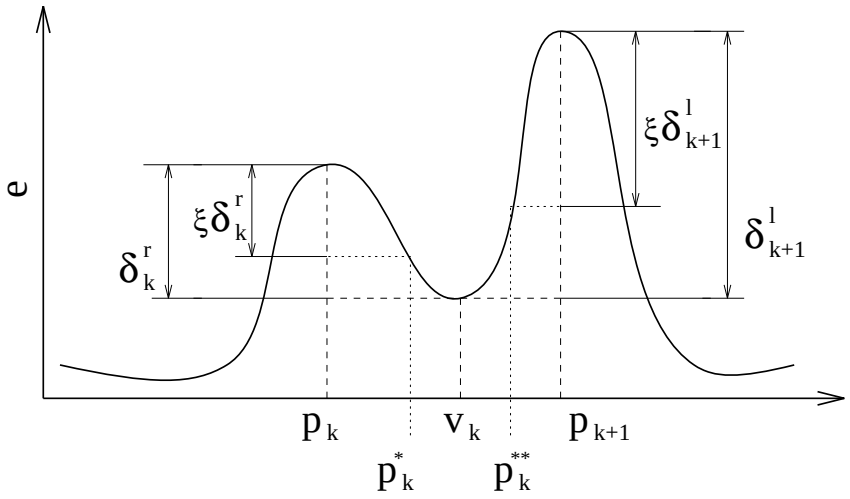

Figure 3-6: Schematic of inter-nuclear boundary insertion.

be taken into account when calculating other prosodic features, like speaking rate, or even more "behavioral" measures, like how talkative a speaker is or a speaker's proclivity to sigh under a particular affective state.

3.3.1 Detection of Pauses and Breaths

The detection of pauses and breaths is based on the same spectral representation used for the segmentation in Section 3.2 by submitting a feature vector derived from this representation to a classifier that learns the decision boundaries between the categories (i.e., pauses, breaths, or neither). We first transform each acoustic segment to a common fixed dimensionality by prototyping it with its median value taken along the time dimension. This modeling decision can be considered quite sensible given that each segment corresponds to a fairly stationary fragment of speech, and, as we shall see, some of the most relevant features to the three-way classification turn out to be global values measured on the segment as a whole (e.g, duration), and not features exploiting the dynamic evolution of the spectral values throughout the segment.

Let $X(m) = \underset{k}{\text{med}} X(m,k)$ for $m = 0, \cdots, M$ be the spectral slice associated with a segment $s(n)$ for $n = 1, \cdots, N$ sampled at F_s Hz. Let m_f denote the frequency bin containing frequency f. We consider the following features:

- The duration (in seconds) of the segment $s(n)$

$$l = \frac{N}{F_s} \tag{3.15}$$

- The zero-crossing rate (normalized by duration) of the segment $s(n)$

$$zcr = \frac{1}{2l} \sum_{n=2}^{N} |\text{sgn}(s(n)) - \text{sgn}(s(n-1))| \tag{3.16}$$

Next, we consider the spectral slice in the range 250-3500 Hz. and normalize it over the following frequency range (where m_f denotes the discrete bin index from the FFT associated with a frequency of f Hz.):

$$\tilde{X}(m) = X(m) / \sum_{m_{250}}^{m_{3500}} X(m) \quad m = m_{250}, \cdots, m_{3500} \tag{3.17}$$

so that the spectral values can be treated as a discrete distribution over the frequency range before extracting the following two measures:

- The normalized entropy

$$h = \frac{-\sum_{m_{250}}^{m_{3500}} \tilde{X}(m) \log_2 \tilde{X}(m)}{\log_2(m_{3500} - m_{250} + 1)}, \tag{3.18}$$

  where the denominator in 3.18 is the entropy of a uniform distribution defined on the same range. Since of all distributions defined on a given support, the uniform distribution attains maximum entropy, Eq. 3.18 constrains h to the range $0 < h \leq 1$.

- The interquartile range

$$iqr = \text{prctile}_{75}(s(n)) - \text{prctile}_{25}(s(n)) \tag{3.19}$$

  where $\text{prctile}_Q(s(n))$ is the first value greater than $Q\%$ of the values in $s(n)$. The interquartile range is a robust estimate of the spread of the data since it is not as sensitive to the presence of outliers.

Finally we divide the spectral slice into the critical bands defined by the following frequencies f_1 through f_{10} ($f = 203.1, 312.5, 437.5, 609.4, 812.5, 1109.4, 1484.4, 1968.8, 2625, 3484.4$ Hz.),

and, after proper normalization over the range (as above), define:[2]

- The distribution of spectral coefficients over critical bands

$$d_k = \sum_{m_{f_k}}^{m_{f_{k+1}}} \tilde{X}(m) \quad k = 1, \cdots, 9. \tag{3.20}$$

Eqs. 3.18-3.20 help summarize the shape of the spectral slice for the segment. The duration feature is particularly important in discriminating breaths and pauses from other vocal segments (since the latter tend to be considerably shorter), whereas the zero-crossing rate feature helps disambiguate pauses and breaths of similar lengths (since the latter tend to contain more high-frequency noise and hence a higher value for zcr). A feature vector is thus defined for each segment as

$$o_{pb} = [l, zcr, h, iqr, d_1, \cdots, d_9]^T. \tag{3.21}$$

The problem of detecting pauses and breaths is posed as a classification problem by training a classifier to discriminate between pauses, breaths, and other segments. At this stage, one could attempt to implement any of a number of classification schemes. However, this is a problem in which the *a priori* knowledge we have about the structure of the features, even if somewhat impressionistic, can help suggest some particular classifiers. For instance, very large duration values likely rule out speech-like acoustic segments, and shift the burden to establish a decision between a pause and a breath. Likewise, a high concentration of spectral coefficients in the region of the first few formants suggests a speech segment from which possibly only a breath needs to be ruled out as a candidate. These observations suggest that perhaps a classifier using a series of embedded local decisions can provide good discrimination between these categories. Based on this motivation we have decided to apply the classification decision tree framework proposed by Breiman et al. (1993), and known as CART, to solve this task.

In order to grow the tree, an initial node containing all the data is successively split into

[2]This particular critical band division has been proposed by (Greenwood, 1961) to explain auditory processing in the cochlea. In his model, each band corresponds to approximately 1 mm. along the basilar membrane. The Greenwood critical band spectral division has been previously applied to the task of syllable onset identification in (Shire, 1997).

further nodes by searching for a local split that reduces the impurity of the data defined as

$$\Delta i_{s,t} = i_t - \alpha^L i_t^L - (1 - \alpha^L) i_t^R, \tag{3.22}$$

where the form of impurity i_t implemented here is the Gini impurity at a node given by

$$i_t = \frac{1}{2}\Big(1 - \sum_j P_j^2\Big). \tag{3.23}$$

The terms i_t^L and i_t^R are the impurities associated with a split s that divides the data arriving at node t into a left and a right node, whereas α^L is the proportion of those samples sent to the left node. P_j and N_j denote, respectively, the empirical prior and class count associated with the jth class. A split s of the data at a given node consists of a partition of the input space along a given dimension.

Let the cost matrix C with $[c_{ij}]$ represent the cost associated with classifying class i as class j class respectively. In order to handle variable costs, the framework of altered priors is implemented here. The idea consists of transforming a problem with class priors P_j and unequal costs to a problem with equal costs (i.e., $c_{ij} = 1 - \delta_{ij}$) and a different set of priors P_j' such that the expected misclassification cost of the tree remains constant (Breiman et al., 1993).[3] The altered priors are given by

$$c_j = \sum_k c_{jk} \tag{3.24}$$

$$P_j' = \frac{c_j P_j}{\sum_j c_j P_j}. \tag{3.25}$$

If a partition s splits the data arriving at node t into a left and right set containing class counts N_j^L and N_j^R, then maximizing (3.22) is equivalent to maximizing the last two terms

$$s^* = \underset{s}{\text{argmax}} - \alpha^L\Big(1 - \sum_j P_j^{L^2}\Big) - (1 - \alpha^L)\Big(1 - \sum_j P_j^{R^2}\Big), \tag{3.26}$$

[3] Alternatively, one can fold the classification costs into the impurity measure. However, (Breiman et al., 1993) discuss how the Gini impurity index does not appropriately handle the case of highly non-symmetric costs and propose the altered priors alternative instead.

where the probabilities, scaled in terms of altered priors, are given by

$$\alpha^L = \frac{\sum_j P'_j N_j^L / Nj}{\sum_j P_j (N_j^L + N_j^R)/Nj} \tag{3.27}$$

$$P_j^L = \frac{P'_j N_j^L / N_j}{\sum_j P'_j N_j^L / N_j} \tag{3.28}$$

$$P_j^R = \frac{P'_j N_j^R / N_j}{\sum_j P'_j N_j^R / N_j}. \tag{3.29}$$

Each node is recursively split until it is pure or contains at least N_{min} observations ($N_{min} = 10$). The optimal pruning tree algorithm from CART (Breiman et al., 1993) is then applied to the grown tree to obtain a sequence of subtrees and their associated generalization errors (estimated through 10-fold cross-validation). This sequence of subtrees and their generalization performances can then be examined to retain an optimal tree, or one that exhibits suitable performance (e.g., one with a suitably diagonal confusion matrix).

Applying this procedure to the breath and pause classification task yields the decision tree shown in Fig. 3-7. The figure shows the percentage of points from each category at the root node and after the first split. As can be seen from the queries, the features of duration, zero crossing rate and amplitude spread provide a good initial split.

The generalization performance of the tree with optimal 10-fold cross-validation error was refined using leave-one-out cross validation. The distribution of the expected errors is shown in the confusion table in Table 3.1.

	Other	Pause	Breath
Other	87.49	1.15	0.60
Pause	0.76	2.46	0.65
Breath	0.60	0.33	5.96

Table 3.1: Confusion matrix showing the distribution of the generalization error estimated by leave-one-out cross-validation for the pause-breath classification task.

Acoustic segments of at least 100 msecs. in duration are considered candidates for breaths and pauses and submitted to the feature extraction procedure and the classification decision tree training procedure just outlined. Any adjacent segments that receive the same classification are merged at this stage into a single segment, and their boundaries adjusted. The result is an acoustic segmentation annotated with the locations of syllable nuclei, as well as those of any pauses and breaths found. This kind of representation provides the

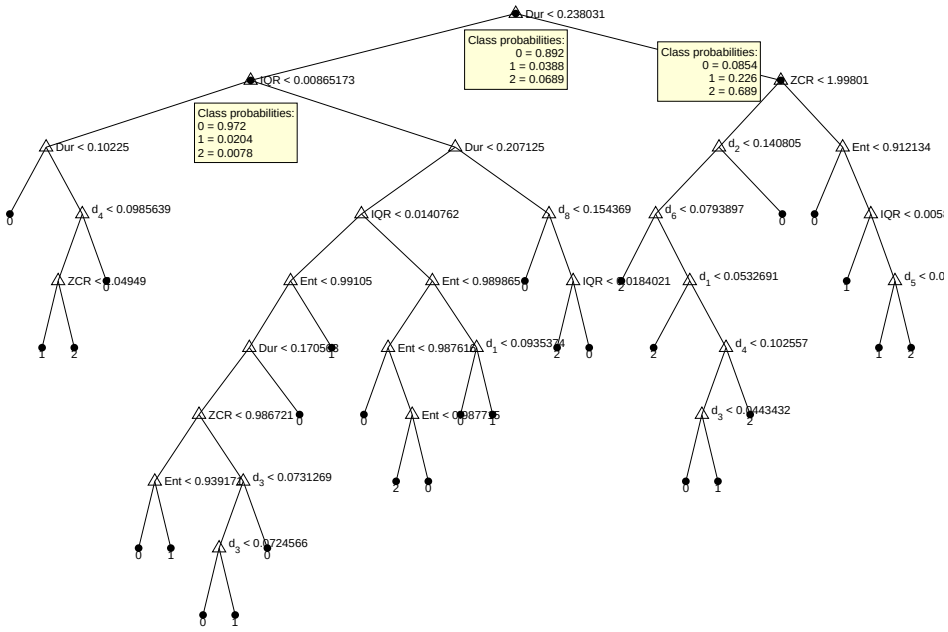

Figure 3-7: Classification tree for breath and pause detection. Insets show the class distribution after the top splits (1=pauses, 2=breaths, 0=rest). Left branches represent a positive answer to a node's query.

basis for the next step: the identification of major prosodic boundaries associated with intonational constituents.

3.4 Detection of Major Prosodic Boundaries

Determining the locations of major prosodic boundaries is of interest to us since they delimit the natural domain for the analysis of various phenomena of interest, particularly intonational phenomena. From the point of view of automatic recognition, it is worth noting that the percept of a "break" which seems to follow major intonation groups is accompanied in many cases by some distinct features that can be acoustically characterized. However, other factors (syntactic and semantic, for instance) play a non-negligible role in this assignment, as illustrated by the following quote

> The assignment of intonation-group boundaries is therefore something of a circular business; we establish some intonation-groups in cases where all the external criteria [i.e., phonetic cues] conspire to make the assignment of a boundary relatively certain; we note the sorts of internal intonational structure occurring

> in such cases and this enables us to make decisions in those cases where the external criteria are less unambiguous. And, in some difficult cases, we take grammatical or semantic criteria into account (Cruttenden, 1997).

Since our approach is purely acoustic, we shall first review some of the most prominent external criteria that have been cited in the literature as indicative of the location of a major intonational boundary.

Chief among these is the presence of pauses or breath groups after an intonation group, with the length of pauses correlating with the type of constituent. Major boundaries, therefore, are expected to correlate with the presence of longer pauses. Physiological constraints also enter into this, as speakers often wait for an intonational boundary to take a breath. Although this is particularly true of read and highly fluent speech, spontaneous speech tends to complicate the issue with the presence of disfluencies and hesitation pauses. Another cue is provided by a speaker's tendency to articulate at a faster rate following a major boundary, in particular over the span of unstressed syllables between the boundary and the first stress (the anacrusis) (Cruttenden, 1997). Segmental lengthening has also been proposed as a marker of phrase boundaries. In particular, it has been experimentally verified that the lengthening is restricted to the rhyme (i.e., the syllable nucleus and any material that follows within the syllable) preceding a boundary (Wightman et al., 1992), where lengthening in this context is understood to be with respect to the phonetic class of segments making up the rhyme. Changes in voice quality near a boundary have also been described in the literature, particularly in the form of vowel glottalization at the beginning of a new intonational phrase (Dilley et al., 1996) and at the end of intonational phrases and utterances (Redi & Shattuck-Hufnagel, 2001). It has also been observed that speakers tend to divide utterances into phrases of approximate equal length, sometimes favoring this constraint over the effects of syntactic constituency (Shattuck-Hufnagel & Turk (1996) citing Gee & Grosjean (1983)).

Not all these external cues are amenable to our working representation. Assessing final rhyme lengthening, for instance, requires a finer-precision parsing with knowledge about word boundaries, the assignment of inter-nuclear material to syllables (i.e., what are the onsets and codas of each syllable), and the segmental identity of the phones in the rhyme. Other criteria, like the tendency to isometry, are difficult to capture with simple features since knowledge of the locations of other potential boundaries is simultaneously needed.

Our approach is based on formulating major boundary detection as a problem of detecting whether there is a boundary between two adjacent syllable nuclei since prosodic boundaries are defined as occurring between syllables. By major boundaries we mean here those that receive a rating of 4, 5, or 6 on the annotated transcripts of the Boston University Radio News Corpus (Ostendorf et al., 1995). These values have been assigned through perceptual experiments where the degree of coupling between adjacent pairs of words was rated by several listeners (Wightman et al., 1992). We have not addressed the assignment of the inter-nuclear material to structures in the preceding and following syllable. In fact, we have stopped short of delimiting syllable boundaries, preferring instead to characterize the waveform, at the syllable level, in terms of nuclear and non-nuclear intervals. Clearly delimiting syllable boundaries is a more complex issue and, at the same time, one of minor practical importance if all we wish to have is an approximate parsing of speech into major phrases to which we can apply some higher-level processing. Based on this and previous remarks regarding the criteria for the presence of phrase markers, we have proposed the interval-based approach outlined below. The proposed features are intended to capture information about timing, pausing (and/or breathing), energy and $F0$ contour shape around the vicinity of a potential boundary. Similar feature sets have been proposed in the literature (Wang & Hirschberg, 1992; Kompe et al., 1994; Wightman & Ostendorf, 1994; Hirschberg & Nakatani, 1998; Nöth et al., 2000).

The following auxiliary intervals are first defined for the extraction of a feature set around the location of a potential boundary (where a potential boundary is understood to follow a nucleus):

- I_0: the interval spanning any segments contained within two adjacent nuclei
- I_1: the interval associated with the current nucleus
- I_2: the interval spanning the previous two syllables (i.e., the current and previous nuclei and intervening segments
- I_3: the interval associated with the next nucleus
- I_4: the interval spanning the next two syllables (i.e., the next and following nuclei and intervening segments

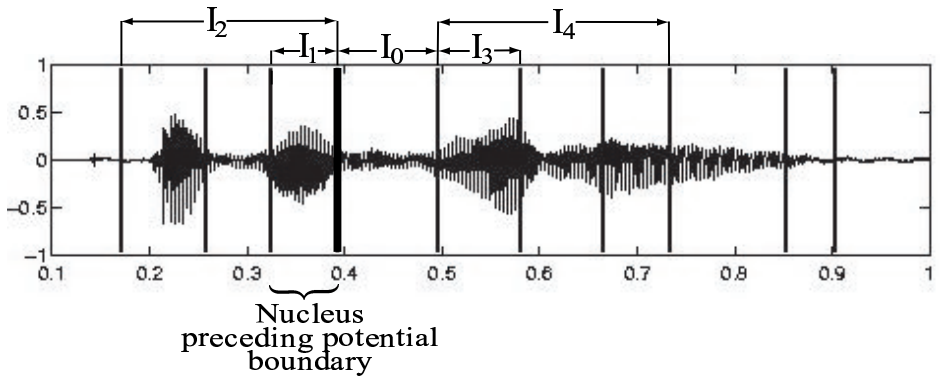

Figure 3-8: Illustration of the auxiliary analysis intervals, as defined in the text, used to define the domain of the features used in the intonational boundary detection task.

The features are defined as follows. It is assumed that $F0$-related measures are applied to the logarithm of voiced portions of $F0$ over a given interval:

1. the inter-nuclear distance (in secs.): the difference between the centroids of intervals I_3 and I_1.

2. a Boolean flag indicating whether I_0 contains a pause segment.

3. the duration of the pause (in secs.) in I_0 (0 if none).

4. a Boolean flag indicating whether I_0 contains a breath segment.

5. the duration of the breath segment (in secs.) in I_0 (0 if none).

6. the difference between the mean durations of the nuclei in I_4 and I_2: the duration of the nuclear segments in I_4 are averaged to obtain a coarse approximation to the number of syllables per second following the potential boundary; the same procedure is repeated on the interval I_2 and the results subtracted from those calculated on I_4 to roughly estimate the change in speaking rate across a boundary.

7. the inter-nuclear average energy on I_0: the average energy on an interval is defined as the Hanning-windowed sum of squared speech samples on the interval, divided by the interval length

8. the difference in average energy (as defined in 7) between the intervals I_3 and I_1.

9. the difference in average energy (as defined in 7) between the intervals I_4 and I_2.

10. the $F0$ jump across the boundary: the difference between the onset (first voiced value) of $F0$ on I_4 and the offset (last voiced value) on I_2.

11. the change in $F0$ baseline across the boundary: the difference between the means of $F0$ calculated over I_4 and I_2.

12. the inter-quartile range of $F0$ (as described in Section 3.3.1 and in Eq. 3.19) over I_4

13. the inter-quartile range of $F0$ over I_2

14. the value of the pre-boundary $F0$ slope: using linear regression, a straight line is fit to $F0$ to calculate its slope over I_2

15. the root mean square of the residual found by the regression in (14)

16. the change in $F0$ slope across the boundary: the difference between the linear-regression slope found over I_4 and the slope estimate obtained in (14).

17. the quadratic coefficient (a_2) of a logistic regression applied to the model $F0(n) = a_2 n^2 + a_1 n + a_0$ over the interval I_4

18. the linear coefficient (a_1) of the model in (17) over the interval I_4.

19. the quadratic coefficient (a_2) of the model in (17) over the interval I_2.

20. the linear coefficient (a_1) of the model in (17) over the interval I_2.

21. the normalized location $\tilde{t}_{min}$ of the $F0$ minimum on I_4; the normalization is carried on as follows to produce a value $0 \leq \tilde{t} \leq 1$ describing the bounded distance away from the offset: let T_{on} and T_{off} correspond to the time instants for the onset and offset of $F0$ over an interval, then $\tilde{t} = \frac{t - T_{off}}{T_{on} - T_{off}}$.

22. the normalized location $\tilde{t}_{max}$ (as described in 21) of the $F0$ maximum on I_4.

23. the normalized location $\tilde{t}_{min}$ (as described in 21) of the $F0$ minimum on I_2.

24. the normalized location $\tilde{t}_{max}$ (as described in 21) of the $F0$ maximum on I_2.

These features are extracted from a speech waveform after applying the segmentation, syllabification, and pause-breath detection algorithms described in the preceding sections.

Using a subset of the Boston University Radio News Corpus (Ostendorf et al., 1995), we built a training set of features for one speaker containing 314 major boundaries (final utterance boundaries were eliminated since most of the features would be undefined given the lack of corresponding intervals I_3 and I_4). This number represents 11.76% of the total number of existing inter-nuclear segments. The remaining 88.24% are therefore associated with minor prosodic breaks, or no break at all. Using these figures as priors on the classes, we proceeded to design a classifier that could be used for the detection of major intonational breaks. Once again we appealed to decision trees as a classification scheme for reasons similar to those outlined in Section 3.3.1: different features can play more or less decisive roles in the classifier's decision, with some features only being called in to elucidate ambiguities unresolved at a higher level (i.e., the presence of a breath group, for instance, is rarely associated with the presence of a minor break). In addition, the feature set contains a mixture of categorical and continuous variables which decision trees can handle with no special provisos.

This binary classification problem can be posed as a detection problem for the major boundary class. In order to assess the performance of the detection scheme, we would like to quantify not just the overall estimated generalization error of the classifier on the two categories considered (we will call this the accuracy of detection), but also the trade-off between proper detection of the boundaries of interest (i.e., the hit rate) and the misslabeling of non-major boundaries as such (i.e., the rate of false alarm). We can investigate the trade-off that the classifiers are able to provide by incorporating a variable parameter γ (quantifying the cost associated with a false alarm) into the cost matrix

$$C_\gamma = \begin{pmatrix} 0 & \gamma \\ 1-\gamma & 0 \end{pmatrix}. \tag{3.30}$$

The non-zero entries in Eq. 3.30 regulate the trade-off between the so-called type-1 and type-2 errors. As γ is varied, therefore, we obtain classifiers with different receiver operating characteristics.

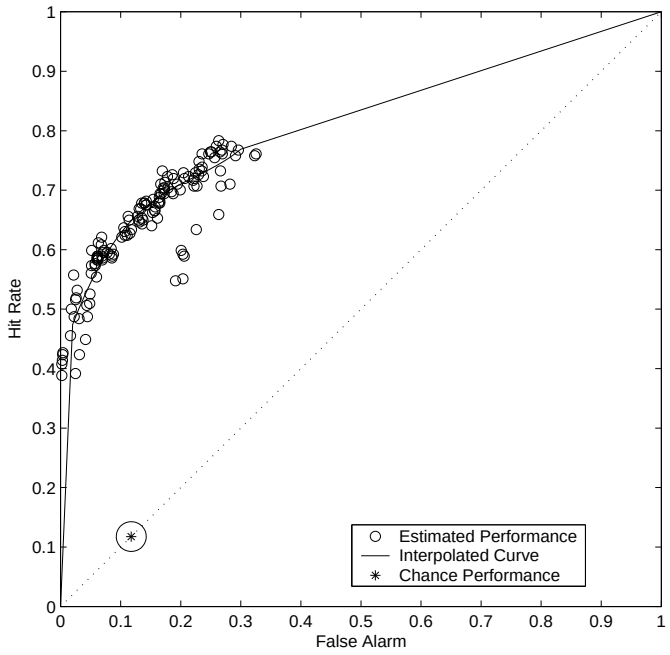

Figure 3-9: Receiver Operating Characteristics summarizing the generalization errors (estimated through 10-fold cross-validation) for the sequence of classification trees obtained by varying γ in Eq. 3.30. The vertical axis represents the probability of correctly detecting a major boundary (Hit Rate), whereas the horizontal axis represents the probability of incorrectly classifying a non-major boundary as major (False Alarm). Each data point on the curve corresponds to a value of γ. The trade-off for a classifier producing a decision drawn randomly according to the class priors is also shown.

3.5 Evaluation

This framework was implemented by fitting classification trees to the feature set previously described with different cost matrices C_γ. For each value of γ in the range $\gamma \in (0.0005, 0.975)$, the performance of each subtree (in the sequence of optimally pruned subtrees found by the algorithm in CART) was assessed using 10-fold cross validation. The subtree minimizing the sum of type-1 and type-2 errors was chosen as the "optimal" tree for a given γ value, and performance statistics were collected at this stage (these include the confusion matrix, the type-1 and type-2 errors, and the accuracy of the tree). These results are partially summarized by the Receiver Operating Curve shown in Fig. 3-9.

We also consider the performance of a straw-man classifier which randomly produces a label according to the class priors (with the priors on major prosodic boundaries estimated from the class proportions on the training set, as listed above). This results in a detection scheme operating at the point $(11.76, 11.76)$ on the false-alarm vs. hit-rate diagram (also

$\gamma = 0.425$

	Minor	Major
Minor	82.21	6.03
Major	4.46	7.30

T_1	$1 - T_2$	Acc.
6.83	62.10	89.51

$\gamma = 0.0805$

	Minor	Major
Minor	75.66	12.58
Major	3.74	8.02

T_1	$1 - T_2$	Acc.
14.26	68.15	83.67

Table 3.2: Confusion matrices, Type-1 error (or False Alarm), complement of Type-2 error (or Hit Rate), and Accuracy describing the generalization performance of two classifiers trained with different values of γ. (Values are expressed in %.)

shown in Fig. 3-9), and an accuracy $Accuracy = 100 \times \big(1 - Pr(error)\big) = \sum_j P_j^2 = 79.25\%$. The receiver-operating point for this classifier is also shown in Fig. 3-9. The performance results for two different points in the ROC curve are more explicitly tabulated in Fig. 3.2, which shows the confusion matrices, false alarm, hit rate, and accuracy of the decision trees.

As highlighted at the onset of this chapter, the goal of this component of the thesis was to create a system that would produce a segmentation of an acoustic waveform into structural prosodic components without the use of any lexical knowledge (hand-labeled, or output by a speech recognizer). This represents a different set of constraints from that underlying most of the research literature on this topic, and the novel contribution of this work lies in the methodology proposed and tested here to accommodate this set of constraints. It should be clear, however, that this constraint was motivated by a desire to build a system unaffected by the performance of automatically recognized lexical information, which, as we argued in the introduction, should be expected to vary greatly in the presence of speaker-dependent affective speech. It should also be noted that having access to lexical cues contributes greatly to the identification of prosodic boundaries, and that, therefore, some of the approaches already described in the literature may be preferred whenever such cues can be reliably obtained. We would like to motivate this work not as an alternative to such approaches, but rather as an approximation under a different set of constraints. As a point of comparison, table 3.3 lists the performance of other approaches where lexical information (leading to the use of a language model or of lexical features, such as duration) and/or hand-labeled annotations were available. The work in Wightman & Ostendorf (1991) and Wightman & Ostendorf (1994), furthermore, makes use of a subset of utterances from the

Reference	Hit Rate	False Alarm Rate	Accuracy
Wightman & Ostendorf (1991)†	69	3	90
Wang & Hirschberg (1992)†,‡	79	7	91
Wightman & Ostendorf (1994)†	78	7	90
Kompe et al. (1994)†,*	89	10	–
This work	62	7	90
	69	14	84

Table 3.3: Comparison of prosodic parser performance with other approaches. Values are expressed in %. († = includes lexical cues; ‡ = includes hand-labeled annotations; * = includes stochastic language model; blank entries indicate that the figure was not readily computable from the reference.)

Boston University Radio News Corpus, so we should expect that speaking styles are similar in the development data used in the evaluation of their approaches and that described in this work. This table further illustrates the contribution of lexical information to the problem of intonational boundary detection. (There are differences as well between the automatic learning schemes employed in each case; it is not unreasonable to assume, however, that the performance is at least partially improved by the extended sources of knowledge.) The approach proposed here allows a decrease in performance while operating under a different set of constraints.

3.6 Chapter Summary

In this chapter we have described a series of algorithms to build a representation of speech in terms of salient structural units which can be of assistance to further prosodic processing (e.g., by demarcating natural domains to extract syllable- and phrase-level prosodic parameters). We proposed a method to quantify spectral distortion based on finding strong transitions on a spectral function (or edges on a spectrogram), and used this to segment an acoustic waveform into homogeneous segments of nearly stationary spectral content. As we suggested in section 3.2 this method shows some robustness effects compared to other spectral metric-based segmentation approaches. Using this as the starting point of a bottom-up approach, we then built higher-level segmentations at the syllable and phrase levels, where syllables have been treated, for simplification, as sequences of nuclei with ambiguously assigned inter-nuclear material. The phrases identified by these algorithms are those delimited by major prosodic boundaries

We have argued that we wish to constrain our approach to identifying prosodic units such that we exclude any kind of hand-labeled annotations (since we are interested in a fully automatic algorithm), as well as any kind of automatically-obtained lexical information (since we would like to abstract the performance of the algorithm from the performance of a speech recognition system that is likely to degrade in the presence of affective speech). The contribution of this chapter has been to motivate a particular architecture that addresses these constraints and recursively builds a structural decomposition from a low-level acoustic representation to higher-level units represented by intonational phrases.

In the next few chapters (4-7) we will shift attention momentarily to a different component of this thesis: the modeling of different acoustic parameters related to several aspects of spoken language which are investigated for their contribution to affect prediction. The structural parsing of speech that has been the object of this chapter will become relevant once again in Chapter 7 when we discuss how to extract and model features related to rhythm. The concepts of this chapter will also return fully fleshed in Chapters 8 and 10, when we discuss the structurally motivated learning models for integrating acoustic information at various prosodic levels and evaluate this model on speech datasets.

Chapter 4

Analysis of Loudness

4.1 Introduction

In this chapter we turn to the analysis of the loudness of spoken utterances. Loudness, as understood here, refers primarily to the perceptual ability whereby listeners are able to compare and rank the degree of audibility of different sounds (and of the corresponding abilities in production to manipulate sounds to appeal to such percepts in a listener). Although this ability has been the focus of previous investigations that look at how variabilities in this domain correspond with affective variations, the term has been used somewhat differently across studies. In particular, the object of investigation is often the *intensity* of a signal, by which simple properties derived from temporal or spectral representations of the acoustic waveform (e.g., the amplitude of a sinusoid, the root-mean-square value of a time-varying signal, etc.) are understood, regardless of any perceptual implications. Although such properties are an integral part of the perception of loudness, they fail to reflect certain aspects of human auditory processing. Chief among these is the notion of *critical-band* processing, the notion that listening devotes unequal emphasis and specificity to different areas of the audible spectrum. Also fundamental is the notion of *masking*, the process whereby the loudness contributed by a sound of a certain frequency in a sound complex is influenced by adjacent artifacts both in time and in frequency.

Recently, Scherer (2003) has suggested addressing processes of the hearing mechanism in affect modeling as a fruitful research direction. With few exceptions –notably the work of Quast (2001), who has investigated loudness features based on a simplified version of the perceptual loudness model described here, this has not been a trend in the growing

literature on this subject. We wish for a model that can go beyond merely quantifying the properties of the physical stimulus by addressing these known characteristics of human auditory processing. It is also particularly important that we look for a model that quantifies the variation of loudness across different affective states while diminishing the effects of such practical issues as the gain of a channel, or the recording sound level. In addition to judging which of two sounds is louder, a human listener may also have the ability to judge that a speaker, standing at a distance considerable enough to weaken the acoustic signal, is speaking louder than a second speaker in close proximity, and whose emissions reach the listener with higher intensity. In this sense the perception of loudness may be related to the perception of the degree of vocal effort exerted by a speaker in producing sounds, aside from the degree of attenuation or enhancement that the sound my undergo after production. In what follows we summarize a model that attempts to address these issues.

4.2 A Model of the Time-varying Loudness of Utterances

4.2.1 Some Fundamental Properties of Auditory Processing

Human hearing is capable, in principle, of processing sounds ranging from approximately 20 Hz to 16 kHz, with the area of highest sensitivity localized around 4 kHz. The ear is likewise able to process sound pressure levels ranging from about 0 dB to 130 dB, these two bounds defining what are known as the *threshold in quiet* and *threshold of pain* respectively. It is noteworthy, however, that these thresholds are not constant throughout the 20 Hz - 16 kHz spectrum, but are rather frequency dependent. The threshold in quiet, in particular, shows a rather non-linear behavior throughout this bandwidth, increasing as we approach lower frequencies to signal the ear's need to be excited with a higher sound pressure level at very low frequencies in order to be able to perceive it (e.g., a tone of 4 kHz can be perceived around a 0 dB level, whereas a 20 Hz tone needs to reach approximately 60 dB to be perceived), (Zwicker & Fastl, 1999).

A second important property of auditory processing involves the concept of masking. Intuitively, masking refers to the disturbance that the perception of the loudness of a particular sound undergoes when pitted against a different sound of different intensity. We encounter abundant examples of this, such as when a normal-level conversation is momentarily suppressed (partially masked) or drowned (fully masked) by a passing siren. We can

see from this example that loudness is not necessarily an additive phenomenon: the loudness of the siren and the drowned conversation do not add up to the total perceived loudness; rather, one overcomes the other. A tone may not only be fully masked (i.e., rendered inaudible) by a masking tone of higher intensity, it can also be partially masked, such that its loudness is diminished. It may, of course, also be the case that no masking occurs and that the perceived loudness, indeed, equals the separate contributions of each tone.

To understand when and what kind of masking takes place leads us to the notion of *critical band*. This concept has been discussed in the literature to explain the fact that our auditory system analyzes a broad spectrum into a discrete set of bands. Various scales have been proposed to model this quantization. The model of loudness described here makes use of the Bark scale, a division of the audible frequency range into 24 regions of variable bandwidth as shown in Table 4.1, a mapping for which Zwicker & Fastl (1999) have proposed the following approximate analytical expression (with f in kHz and B in Bark):

$$B = 13 \arctan(0.76f) + 3.5 \arctan(f/7.5)^2. \tag{4.1}$$

The Bark scale assumes that each unit of increment in the scale corresponds to a constant length increment of 1.3mm along the basilar membrane.[1] The first five bands exhibit a constant bandwidth of 100 Hz, which increases logarithmically from 500 Hz.

B [Bark]	1	2	3	4	5	6	7	8	9	10	11	12
f [kHz]	.1	.2	.3	.4	.5	.63	.77	.92	1.08	1.27	1.48	1.72

B [Bark]	13	14	15	16	17	18	19	20	21	22	23	24
f [kHz]	2	2.32	2.7	3.15	3.7	4.4	5.3	6.4	7.7	9.5	12	15.5

Table 4.1: Critical-band Bark scale and corresponding frequencies.

Masking can be modeled by appealing to critical bands. It has been observed through psychoacoustical experiments that as the frequency of a test tone approaches the frequency of a masker tone, the loudness of the test tone diminishes until, once this difference lies below a critical bandwidth, the loudness of each tone does not contribute to the percept, and only the loudness of the masker predominates. Likewise, it has been observed that if the frequency of the test and masker are fixed, then by steadily increasing the amplitude of

[1] The closely related ERB scale makes a similar assumption in terms of approximately 0.86mm increments per ERB unit.

the masker, the test tone continuously transitions from a fully audible (unmasked) tone to a fully masked tone. This suggests a process whereby a masker tone "extends" its full loudness to the neighboring tones within its critical band, and partially "radiates" its loudness to tones outside the band. This process is illustrated in Fig. 4-1, where the loudness of the masker tone interacts with the loudness of the test tone through its upper band slope by creating a masking area under which no test tones are perceived distinctly. It is known that a similar masking process applies to test tones lying to the left of the masker. However, this effect is significantly less prominent since the lower band slope that would define a masking area to the left of the critical band has a very sharp fall, and hence, will not be considered here.

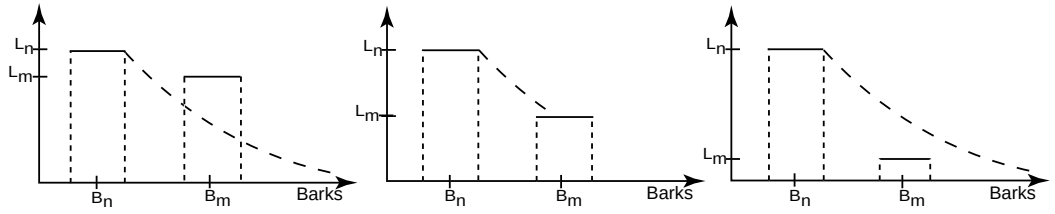

Figure 4-1: Masking: Interaction of loudness levels between two arbitrary neighboring critical bands B_n and B_m (in Bark) with respective levels L_n and L_m. The level at B_n creates a "masking area" (delimited by the dashed downward slope in the figure) which may occlude, partially (middle panel) or completely (third panel), the levels at critical bands to the right of it. Leftward masking effects are ignored.

4.2.2 Zwicker's Model of Absolute Loudness

In this section, we will summarize some of the main aspects of a model of absolute loudness proposed by Zwicker (Zwicker & Fastl, 1999). This model is the basis of an International Standard (ISO 532B), the software code for which has been published in FORTRAN (Paulus & Zwicker, 1972) and BASIC (Zwicker et al., 1984; Zwicker et al., 1991), and which has more recently been made available in a Matlab version (Hastings, 2002).

Zwicker's model addresses four stages of loudness processing:

1. Transmitting (attenuating) the sound through the outer and middle ear.

2. Filtering by a bank of auditory filters to produce excitation levels.

3. Transforming an excitation pattern into specific loudness by a power-law relationship.

4. Integrating specific loudness across critical bands taking into account any masking which results from the excitation levels of adjacent bands.

The filtering of the sound prior to reaching the cochlea (inner ear) consists of the frequency-selective attenuation a_0. This step is subsumed directly into step (3) by adjusting the excitation levels by this factor prior to transforming them to a specific loudness pattern. Step (2) implements a critical-band filter bank which attempts to model the cochlear transformations that the sound undergoes at different points along the basilar membrane. Since the shape of auditory filters is not always available, Zwicker suggested approximating the ear's selectivity with 1/3 octave band-pass filters (i.e., filters whose bandwidth is one third of the center frequency) centered at the frequencies F_c tabulated on the first row of Table 4.2. This approximation is acceptable for frequencies above 300 Hz. For smaller frequencies, however, 1/3 bandwidths are too small compared to critical bands. To circumvent this problem, Zwicker proposes combining the outputs of more than one 1/3 filter to fill an approximate critical band (ACB); the combination of 1/3 octave filter outputs to approximate critical bands is shown by the relation between rows 1 and 2 of Table 4.2 (for instance, the outputs of the first 5 1/3 octave filters contribute to the first approximate critical band). Another consequence of the proposed approximation is a division of the spectrum into critical-band rate differences that are no longer 1 Bark. The usual 24-point Bark scale is effectively resampled in terms of 20 uneven bands, the upper edge of which is shown by the third row (BUE) in Table 4.2.

After filtering a sound through the filter bank just specified, the approximate-critical-band dependent *excitation levels* are found by summing over all frequencies the filtered outputs (and combining the output of more than one filter for frequencies below 300 Hz as described). The result of this stage is an excitation pattern, a series of critical-band-dependent excitation levels L_k that are next transformed into a *specific loudness* pattern via the relationship

$$L_{a_k} = L_k - a_{0_k} - L_{DCB_k} \tag{4.2}$$

$$N'_k = \begin{cases} 0.635 \cdot 10^{\frac{L_{ETQ_k}}{4}} \left(\left(\frac{3}{4} - \frac{1}{4} 10^{0.1(L_{a_k} - L_{ETQ_k})} \right)^{0.25} - 1 \right) & \text{if } L_{a_k} > L_{ETQ_k} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

The quantity a_{0_k} is the critical-band dependent attenuation factor due to the filtering of

F_c	.025	.0315	.04	.05	.063	.08	.1	.125	.16	.2	.25	.315	.4	.5
ACB	1						2			3		4	5	6
BUE	.9						1.8			2.8		3.5	4.4	5.4
a_0	0						0			0		0	0	0
L_{DCB}	-.25						.-6			-.8		-.8	-.5	0
L_{ETQ}	30						18			12		8	7	6

F_c	.63	.8	1	1.25	1.6	2	2.5	3.15	4	5	6.3	8	10	12.5
ACB	7	8	9	10	11	12	13	14	15	16	17	18	19	20
BUE	6.6	7.9	9.2	10.6	12.3	13.8	15.2	16.7	18.1	19.3	20.6	21.8	22.7	23.6
a_0	0	0	0	0	-.5	-1.6	-3.2	-5.4	-5.6	-4	-1.5	2	5	12
$LDCB$	.5	1.1	1.5	1.7	1.8	1.8	1.7	1.6	1.4	1.2	.8	.5	0	-.5
$LETQ$	5	4	3	3	3	3	3	3	3	3	3	3	3	3

Table 4.2: Summary of the main variables involved in the critical band-rate analysis of Zwicker's loudness model, and their values for the 1/3 octave band approximation scheme: F_c (center frequencies of 1/3 octave bandpass filters), ACB (resulting approximate critical bands), BUE (upper edge, in Bark, of each approximate critical band), a_0 (outer and middle-ear attenuation factor), L_{DCB} (correction factor to compensate for the approximation), L_{ETQ} (threshold-in-quiet attenuation). See text for details.

the sound through the outer and middle ear, the proposed first stage that is explicitly modeled at this point, whereas L_{DCB_k} is a correction factor introduced by Zwicker to further compensate for artifacts introduced by the approximation of critical band auditory filters with 1/3 octave filters. The values of a_0 and DCB are shown in rows 4 and 5 of Table. 4.2 for the approximate critical band divisions. These two values are used to obtain L_{a_k}, the adjusted level that gets transformed into specific loudness through Eq. 4.3 Finally, L_{ETQ_k} is the threshold-in-quiet level mentioned earlier, the minimum level a sound of a certain frequency must reach before becoming audible. These figures are also tabulated in Table 4.2 for each of the critical bands.

The output of this stage is a specific-loudness pattern, a distribution over critical bands of specific loudness values. The absolute loudness is then obtained by integrating the specific loudness over the Bark scale, that is, by finding the area under the curve traced by the specific loudness levels after the masking effects associated with the downward masking slope of each level have been incorporated. Fig. 4-2 illustrates this process graphically. The curve begins with a rise to the first specific loudness level and remains at this level throughout the band. When the next band is reached, the curve rises to the next level if this level is higher. Otherwise, the curve follows the downward masking slope until it intercepts a new band level. At this point, the curve remains leveled throughout the remaining band, and the process continues as described. The result of this procedure is a trace that may

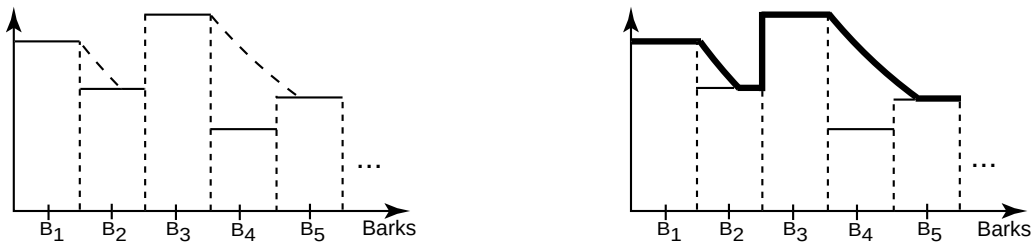

Figure 4-2: Integration over critical bands, taking into account masking effects, to determine absolute loudness.

fully mask the specific loudness on certain bands (B_4 on the example), and partially mask it on others (B_2 and B_5). The value of the slopes are band and level dependent, and have been tabulated in Zwicker et al. (1991). Finding the area under the curve so defined yields the absolute loudness of the sound. Zwicker's algorithm is summarized in Fig. 4-3.

In the implementation discussed here, 3rd-order Butterworth filters have been used to realize the 1/3 octave filters, and the power spectral density of a speech segment has been estimated using the method of averaged periodograms (Oppenheim & Schafer, 1989).

4.2.3 Variability of Instantaneous Loudness for Time-Varying Sounds

The loudness model just described is particularly suited for modeling the loudness of a stationary sound. To obtain a time-varying representation of loudness, we can extend this procedure by applying a short-time (frame) analysis to a windowed segment of speech sound and building a profile of instantaneous loudness over time. It was mentioned at the beginning of this chapter that it was desirable, for the purposes of this research, to implement a loudness algorithm which could reflect the vocal effort exerted in producing a sound. It would be particularly suitable to model this difference in production if global gain properties of the stimulus were altered due to channel gain, or if some root-mean-square normalization were to be applied to the signal. As we shall see when discussing some specific datasets in later chapters, such normalizations are often required when analyzing sounds recorded under conditions that yield very different loudness (e.g., a phone conversation recorded at one end typically has a more prominently audible channel; *in-situ* recordings can show fluctuations due to distance from the microphone, etc.). To address this issue, the sounds processed here undergo first an RMS-normalization as an attempt to compare

Algorithm 4.1: Zwicker's Loudness Model

1. Design a set of 1/3 octave bandpass filters with center frequencies as given by F_c in Table 4.2. Let $|H_m(\omega)|^2$ be the power spectrum of the mth filter.
2. Find the power spectral density $S|(\omega)|^2$ of the speech signal.
3. Filter the sound with each filter in the filter bank: $X_m = \sum_\omega |S(\omega)|^2 |H_m(\omega)|^2$ for $m = 1, \cdots, 28$.
4. Transform these values to approximate critical band levels (in dBs) using the correspondence from rows 1 and 2 from Table 4.2:

$$L_k = \begin{cases} 10 \log_{10} \sum_{m=1}^{6} X_m & k = 1 \\ 10 \log_{10} \sum_{m=7}^{9} X_m & k = 2 \\ 10 \log_{10} \sum_{m=10}^{11} X_m & k = 3 \\ 10 \log_{10} X_{k+8} & k = 4, \cdots, 20 \end{cases}$$

5. Using Eqs. 4.2 and 4.3, and the values tabulated in Table 4.2, transform the excitation level pattern into a specific loudness pattern N'_k for $k = 1, \cdots, 20$.
6. Integrate the area under the specific loudness curve taking into account accessory loudness: Let $N = 0$. At each increment ΔB on the Bark scale, determine whether the main loudness, or accessory loudness from previous critical bands, contributes to the total loudness; add the contributing area under the curve to the value in N.

Figure 4-3: Zwicker's loudness model.

sounds of approximately equal gain:

$$s_n(n) = A \frac{s(n)}{\sqrt{\sum_n s(n)^2}}. \tag{4.4}$$

To compare the effects of the normalization with a value of $A = 0.25$ (empirically chosen) in Eq. 4.4, two test sounds of the utterance *Hello*, one soft and one loud, recorded under similar conditions, are shown in Figs. 4-4 and 4-5. The center panels show the fluctuation of the instantaneous loudness calculated for 40 msec. windowed segments every 10 msecs. The lower panels show the fluctuation of the instantaneous RMS values calculated with the same parameters. As expected, the loud sound shows a considerable increase in average loudness and average RMS (shown boxed in each respective panel) when the test signals are processed without any normalization (Fig. 4-4): the average loudness and average RMS values are greater by a factor of approximately 5 and 2 respectively.

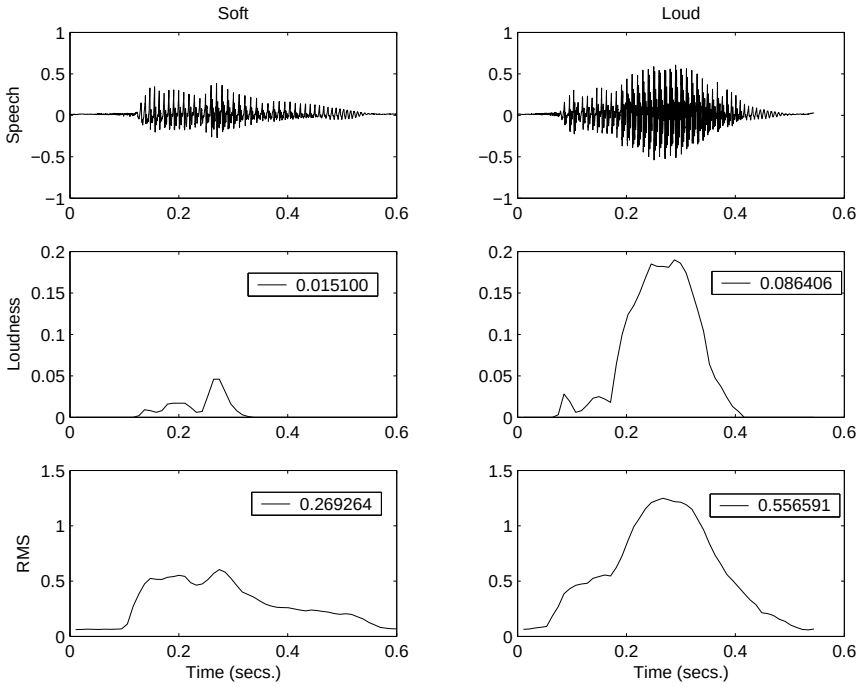

Figure 4-4: Loudness and RMS profiles (no normalization). Average loudness and average RMS values are shown boxed in each panel.

When the sounds are normalized (Fig. 4-5), the value of the RMS average is not significantly different. However, the average loudness for the normalized loud sound is still twice as large as it is for the normalized soft sound. Since the model of loudness presented here takes into account the distribution of excitation levels over critical bands, rather than global properties of the auditory stimulus, it is expected that such a model will be better at capturing the vocal effort exerted by the speaker in producing a louder sound, even when the gain has been adjusted. This suggests that RMS normalization followed by the loudness calculation discussed in the preceding section is a suitable procedure to apply if we wish to model the loudness of a sound while reducing artifacts like channel attenuation.

4.2.4 Loudness Features

Based on the preceding discussion, we propose to use features derived from the discussed loudness model to test whether affective variations can be partially captured with such a feature set.

Let $s_n(n)$ be the RMS-normalized speech signal as in Eq. 4.4 with $A = 0.25$. Let K be the number of short-time frames obtained from the short-term analysis of the speech

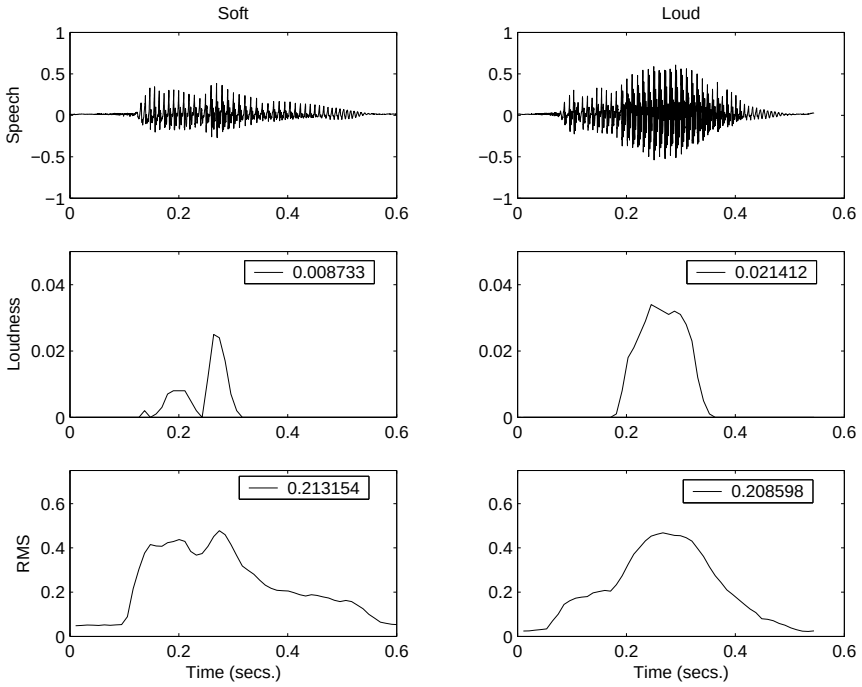

Figure 4-5: Loudness and RMS profiles (with normalization).

segment. Let $N(k)$ and $R(k)$ be the perceptual loudness and RMS value of the kth frame, for $k = 1, \cdots, K$. Let $N'(k, m)$ be the specific loudness pattern for the kth frame, where m is chosen to cover the Bark scale from $z_1 = 4.25$ to $z_{14} = 17.25$ in increments of 1 Bark. As evidenced by Figs. 4-4 and 4-5, there may be some frames for which the perceived loudness is zero (possibly due to the frame intensity not exceeding the threshold-in-quiet levels). In the following we restrict the analysis to the set of frames $K_{nz} = \{k : N(k) \neq 0\}$ for a which a non-zero perceived loudness is obtained. Let the superscript nz denote a signal evaluated only at frames in K_{nz} and $|K_{nz}|$ be the size of this set.

Let the mean perceived loudness be

$$\mu_N = \frac{1}{|K_{nz}|} \sum_{k \in K_{nz}} N_k, \tag{4.5}$$

Let p be the p-th percentile value of the signal $N^{nz}(k)$, and define the p^+-percentile mean $\mu_L^{(p)}$ as the mean of the K_p values exceeding p

$$\mu_N^{(p)} = \frac{1}{K_p} \sum_{k:N^{(nz)}(k)>p} N^{(nz)}(k). \tag{4.6}$$

Define, analogously, the p^+-percentile mean $\mu_R^{(p)}$ quantity for $R^{(nz)}(k)$ as

$$\mu_R^{(p)} = \frac{1}{K_p} \sum_{k:R^{(nz)}(k)>p} R^{(nz)}(k). \tag{4.7}$$

Finally, let the mean specific loudness pattern for the m-th band, delimited by z_m and z_{m+1}, be given by

$$\mu_{N'_m} = \sum_{k \in K_{nz}} N'(k,m). \tag{4.8}$$

The vector of loudness-related features is defined as follows, using Eqs. 4.5, 4.6, 4.7, and 4.8, and letting p be the percentile values 25, 50, and 75:

$$FS_{loud} \doteq [\mu_N, \mu_N^{(25)}, \mu_N^{(50)}, \mu_N^{(75)}, \mu_R^{(25)}, \mu_R^{(50)}, \mu_R^{(75)}, \mu_{N'_1}, \cdots, \mu_{N'_{13}}]^T. \tag{4.9}$$

The 20-feature vector in Eq. 4.9 is intended to capture information about the distribution of the loudness and RMS values over the course of a sound, as well as information about the specific loudness on different critical bands. Notice that μ_R, the average of RMS values, is not included, as, due to the RMS normalization, this value is expected (and intended) to be less critical. The p^+-percentile features are included, however, since they reflect the distribution of RMS values over time, a feature which may still prove useful (and does, in fact, turn out to be, as discussed in the next section).

4.3 Evaluation and Discussion

The feature set just defined was used to build binary classifiers of emotion pairs to assess its feasibility in discriminating between categories of affect for different subjects. The objective of training classifiers at this stage is not to aim for a state-of-the-art recognition performance. Rather, it is to establish, in at least a weak sense, the potential of the feature set under investigation to *partially* model affect from the cues under investigation. All other features not related to the current feature class (loudness) are purposely withheld and examined separately. Although other classification schemes and combinations of feature subsets can improve the recognition performance, this preliminary analysis is carried out with partial information concerning a homogeneous class of features in order to test the particular contribution of this set. We will also apply this methodology to test the acoustic

parameters discussed in the next few chapters.

	Ang	Hap	Ntr	Sad
Afr	**0.28**†	**0.24**†	0.38‡	0.48
Ang		**0.32**†	0.41	**0.32**†
Hap			0.43	**0.34**†
Ntr				0.56

Subject 1

	Ang	Hap	Ntr	Sad
Afr	**0.26**†	**0.21**†	**0.23**†	0.45
Ang		**0.21**†	**0.24**†	**0.33**†
Hap			0.38‡	**0.25**†
Ntr				0.35‡

Subject 2

	Ang	Hap	Ntr	Sad
Afr	**0.25**†	**0.28**†	**0.27**†	0.45
Ang		**0.22**†	0.42	**0.28**†
Hap			**0.28**†	**0.24**†
Ntr				**0.26**†

Subject 3

	Ang	Hap	Ntr	Sad
Afr	**0.18**†	**0.11**†	**0.24**†	**0.18**†
Ang		**0.25**†	**0.29**†	**0.25**†
Hap			**0.29**†	**0.19**†
Ntr				0.41

Subject 4

	Ang	Hap	Ntr	Sad
Afr	**0.25**†	**0.31**†	**0.21**†	**0.25**†
Ang		0.35‡	0.37‡	**0.31**†
Hap			**0.31**†	**0.30**†
Ntr				**0.34**†

Subject 5

	Ang	Hap	Ntr	Sad
Afr	**0.33**†	**0.33**†	0.39	0.53
Ang		**0.26**†	**0.26**†	0.41
Hap			**0.30**†	**0.34**†
Ntr				0.43

Subject 6

Table 4.3: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent linear classifiers trained on emotions pairs using loudness feature set for subjects 1 through 6 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

The results are summarized in Tables 4.3 and 4.4. These tables show the generalization error of a simple Bayesian classifier, estimated through leave-one-out cross-validation. The features have been extracted from the set of actors data discussed earlier. There are 50 samples per subject of each category. Equal covariance matrices have been used to model each pair of classes for each classifier, leading to a linear boundary in feature space. The result tables indicate the classifiers achieving a generalization error which differs significantly from a random classifier (with expected generalization error of 0.5) at the $p < 0.05$ and $p < 0.01$ levels (the latter cases have also been typeset in bold for easy inspection). Consult Appendix D for details on the significance test applied to classifier outputs.

These results show that the loudness feature set described provides, by itself, better-than-chance discrimination for most emotion pairs, for most subjects. The inter-speaker variation is noteworthy: significantly small error rates are obtained for almost all classifiers

	Ang	Hap	Ntr	Sad
Afr	**0.16**†	**0.28**†	**0.28**†	0.39
Ang		**0.16**†	**0.23**†	**0.23**†
Hap			**0.25**†	**0.33**†
Ntr				0.37‡

Subject 7

	Ang	Hap	Ntr	Sad
Afr	0.37‡	**0.30**†	0.52	0.46
Ang		**0.27**†	0.37‡	0.45
Hap			**0.32**†	0.42
Ntr				0.56

Subject 8

	Ang	Hap	Ntr	Sad
Afr	0.34‡	**0.30**†	0.42	0.56
Ang		0.42	0.41	0.36‡
Hap			0.41	**0.36**†
Ntr				0.46

Subject 9

	Ang	Hap	Ntr	Sad
Afr	**0.20**†	**0.21**†	**0.30**†	**0.34**†
Ang		0.45	**0.17**†	**0.27**†
Hap			**0.20**†	**0.26**†
Ntr				0.54

Subject 10

	Ang	Hap	Ntr	Sad
Afr	**0.34**†	**0.28**†	0.37‡	0.35‡
Ang		**0.34**†	**0.26**†	**0.23**†
Hap			**0.22**†	**0.21**†
Ntr				0.43

Subject 11

Table 4.4: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent linear classifiers trained on emotions pairs using loudness feature set for subjects 7 through 11 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

trained on Subjects 2, 3, 4, 5, 7, 10 and 11. However, only two classifiers achieve similarly significant recognition for Subject 9. It is also interesting to see the subject-dependent variation on the performance on emotion categories. For Subject 4, the features provide uniform better-than-chance discrimination between the *Afraid* categories and all others. For Subject 5, this is the case between the *Sad* categories and the rest. The emotion pair classifier with worst generalization error is obtained for the *Sad-Afraid* pair, a result which seems counterintuitive, as we would expect loudness to provide reasonable separation between a potentially high-arousal affective state (fear) and a typically low-arousal state (sadness). Likewise, the better discrimination between the pair *Afraid-Angry* (for all subjects, except 8 and 9) may seem surprising if we think of these as high-arousal categories.

Table 4.5 shows the generalization error of the same classifier structures used in the pairwise discrimination task when trained instead to discriminate between each emotion category and the remaining ones. The sets have been randomly subsampled to retain equal

Subject	Afr	Ang	Hap	Ntr	Sad
1	**0.33**†	**0.33**†	**0.36**†	0.60	0.42
2	**0.33**†	**0.34**†	**0.24**†	0.46	0.40
3	**0.34**†	**0.31**†	**0.32**†	0.45	0.43
4	**0.20**†	0.37‡	**0.20**†	0.50	0.42
5	**0.29**†	**0.34**†	0.45	**0.36**†	0.40
6	0.42	**0.27**†	**0.28**†	0.40	0.55
7	**0.29**†	**0.20**†	**0.33**†	**0.33**†	0.41
8	0.40	0.38‡	**0.26**†	0.50	0.61
9	0.50	**0.37**†	**0.37**†	0.56	0.51
10	**0.28**†	**0.26**†	**0.34**†	0.43	0.42
11	**0.37**†	**0.29**†	**0.25**†	**0.30**†	**0.29**†

Table 4.5: Generalization error (estimated by leave-one-out cross validation) for subject-dependent linear classifiers trained to discriminate between each emotion category and the remaining ones pooled together, and using feature set containing loudness features only ($\dagger$and $\ddagger$denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

priors on all categories prior to training (chance level discrimination is, therefore, 50%). As evidenced by the leave-one-out generalization error estimate, the categories *Happy* and *Angry*, followed closely by the category *Afraid*, are significantly discriminated from the rest by the loudness features for most subjects. The loudness features proposed, however, do not offer consistent significant separation for the *Neutral* and *Sad* categories, suggesting that the way the speakers employ these features when affecting neutral or sad speech overlaps considerably with their usage to convey the remaining categories.

4.4 Chapter Summary

In this chapter we have reviewed Zwicker's model of absolute loudness and used it to model the time-varying loudness of speech signals. The model was shown to capture differences between loud and soft utterances even when a root-mean-square normalization is imposed *a priori*, a step which we argue is needed when dealing with speech samples obtained under different gain conditions. In such cases, global root-mean-square values of the signals become unreliable indicators of the loudness produced by a speaker, and are more reflective of a property of the channel than of the utterance itself. The model described here more closely reflects the vocal effort associated with producing louder utterances by taking into account critical-band dependent specific loudness measures, and integrating these contributions over

the spectrum.

The novel contribution of the chapter resides in the proposal of an extensive set of features based on Zwicker's perceptual model of loudness and the demonstration of their predictive abilities to model differences between affective categories using automatic learning schemes. With the exception of the work of Quast (2001) (who implemented a simpler formulation of Zwicker's model and investigated the correlation between a single feature (specific loudness) and conveyed affect), this perceptual loudness model has not received much attention in the literature on affect modeling. In this work, however, we have contributed an approach that uses the International Standard (ISO 532B) based on Zwicker's model to create loudness profiles which serve as the basis for various novel measures that yield significant discrimination between affective categories.

Chapter 5

Analysis of Fundamental Frequency

5.1 Introduction

In this chapter we turn to the analysis of fundamental frequency contours for modeling intonational aspects of speech that may be relevant to the detection of affect from spoken language. Compared to the paralinguistic use of other aspects of speech, the study of paralinguistic uses of F0 has received prominent attention. In particular and until recently, the effect of affective variations on F0 has been studied more prominently than the effect of affective variation on other aspects of speech. This focus may be purely circumstantial (e.g, the accessibility of algorithms for recovering the fundamental frequency of phonatory vibration), but it is not unreasonable to suppose that the experimental inquiry in the laboratory is also matched by a very strong intuition, likely shared even by the layman, that some aspects of pitch is often at the mercy, sometimes even beyond our control, of our affective state.

We will first review a series of algorithms that have been adapted from published work and which are fundamental to the acoustic modeling of F0 realized in this research. We will then propose a series of measures for discriminating between affect categories and will test this assumption on a set of utterances labeled according to affective content. Throughout this chapter, we will invoke three major views of F0 contours for the analysis of affect: the view of F0 as a sequence of values representing the fundamental frequency of vibration of

the focal folds, smoothed out over a short-segment of speech containing a few cycles; F0 as a sequence of stylized segment types defined between landmark (or turning) points (a view more closely in line with a F0-as-targets description of a phonological nature); and finally F0 as an unfolded sequence over time of elements belonging to frequency clusters defining pitch intervals. Each of these views will, in turn, suggest a different set of acoustic features which we will examine for their relevance to affect modeling.

5.2 F0 Determination Algorithm

In this section we report the algorithm used to extract the F0 contour that serves as the basis for the remaining analysis in this chapter. The algorithm is to a large extent an implementation of the one described by Boersma (1993), and available in the Praat software package. Since some specifics of implementation differ, the algorithm is reviewed next.

F0 extraction is performed using an autocorrelation analysis of short segments of speech. The novelty of the algorithm consists of a normalization of the autocorrelation function of the speech segment by the window autocorrelation, a step which Boersma (1993) reports had not been noted in the literature prior to this formulation. It is known that choosing a locally optimal F0 estimate, such as the one indicated by the position of the global maximum of the autocorrelation function, typically leads to errors in the F0 contour (e.g., frequent transitions between voiced and unvoiced segments, octave doubling, etc.). To compensate for this, it is common to apply a post-filtering step, such as median filtering (Rabiner & Schafer, 1978). The approach proposed here, however, accrues a series of candidates from every analysis frame, and defers selecting the optimal candidate until all frames have been processed. An optimal sequence is then defined to be one that minimizes a cost on the entire F0 contour, and dynamic programming is used to find the globally optimal sequence.

For every frame m, up to N_{max} voiced candidates are found from local maxima of the normalized autocorrelation function in a desired range of interest $(f0_{min}, f0_{max})$ (to exclude periodicities which are known to lie outside normal F0 ranges), and each candidate receives a local "strength" of

$$r_{m,k} = R_s(\tau^k_{max}) - oct_{cost} \log 2(f0_{min}\tau^k_{max}), \tag{5.1}$$

where $R_s(\tau)$ is the normalized autocorrelation function of the segment, and τ^k_{max} is the

k-th lag value at which R_s reaches a maximum. The value oct_{cost} is a free parameter in the algorithm whose function is to favor high vs. low fundamental frequencies to model the tradeoff between *perceived* F0 (pitch) and *acoustic* vocal fold periodicity (fundamental frequency). The parameter also helps reduce the number of local downward octave jumps caused by noisy signals (Boersma, 1993).

In addition to the voiced candidates, an unvoiced candidate is always considered at every frame with a local strength of

$$r_{m,0} = v_{th} + \max\Big\{0, 2 - \frac{(lap)(1+v_{th})}{(gap)(sil_{th})}\Big\}. \tag{5.2}$$

The values v_{th} and sil_{th} denote voicing and silence thresholds, and are free parameters of the algorithm: a frame is likely to be classified as voiceless if no autocorrelation peaks above v_{th} are found, or if the local absolute peak (lap) is less than sil_{th} percent of the global absolute peak (gap). The values of gap and lap are found globally at the start of the algorithm and then for every frame, respectively.

The steps just outlined describe the basis of the algorithm to build a sequence of $F0$ candidates. A further refinement of the algorithm calls for an interpolation around every local maxima to identify with better precision the position of each maxima (since the autocorrelation function of the sampled speech is, after all, a sampled version of a continuous autocorrelation function, and the maxima may not correspond with sampled points). This step is carried out by using cubic spline interpolation. The windowing step involved in the short-time analysis has been carried out using the following Gaussian-like window, following the robustness results which Boersma (1993) reports for the extraction of F0 under various noise conditions using this window:

$$w(n) = \frac{\exp\big(-12((n/N - 0.5)^2\big) - \exp(-12)}{1 - \exp(-12)}. \tag{5.3}$$

This procedure is summarized in Fig. 5-1. The result of running algorithm 5.1 is a sequence of M sets of frequency-strength pairs $\big\{\{f_{k,m}, r_{k,m}\}_{k=1}^{K_m}\big\}_{m=1}^{M}$ corresponding to each short-time analysis frame. This sequence defines a trellis which summarizes every possible path corresponding to every possible pairwise transitions between F0 candidates in adjacent frames. This view allows us to associate the following cost to each path $\{m, l_m\}$, where m is a frame index, and l_m is a number between 1 and the maximum number of candidates

Algorithm 5.1: Let $s(n)$ be a speech waveform with sampling rate F_s. Let $f0_{min}$ and $f0_{max}$ denote the minimum and maximum values allowed on the F0 contour.

- Preliminaries

  1. Low-pass filter the signal with a 10th-order Butterworth filter with cutoff frequency of 4kHz. Let $s_{lp}(n)$ denote the low-pass filtered signal.
  2. Find the global absolute peak $gap = \max|s_{lp}(n)|$.
  3. Choose a window length L to accommodate 3 pitch periods: $N = \frac{3F_s}{f0_{min}}$.
  4. Using Eq. 5.3, evaluate the window function; append half a window length of zeros to $w(n)$, and, to the resulting signal, append zeros until its length is an integer power of 2. Let N_{zp} denote the final length of the zero-padded signal.
  5. Evaluate the normalized autocorrelation function of $w(n)$, $R_w(\tau) = \frac{\sum_n w(n)w(n+\tau)}{\sum_n w(n)^2}$.

- Short-term Analysis. For every segment $s_{lp}^{seg}(n)$ of length N_l, in steps of N_s:

  6. Subtract the local average to obtain $s_z(n) = s_{lp}^{seg}(n) - \frac{1}{N}\sum s_{lp}^{seg}(n)$.
  7. Find the local absolute peak $lap = \max|s_z(n)|$.
  8. Append zeros to $s_z(n)$ until its length is N_{zp}.
  9. Multiply it by the window to obtain $a(n) = s_z(n)w(n)$.
  10. Evaluate $R_a(\tau) = \frac{\sum_n a(n)a(n+\tau)}{a(n)^2}$, the segment's normalized autocorrelation.
  11. Divide the segment autocorrelation by the window autocorrelation: $R_s(\tau) = \frac{R_a(\tau)}{R_w(\tau)}$.
  12. Let $f_{m,0} = 0$ and $r_{m,0}$ as in Eq. 5.2 be the value and strength of the unvoiced candidate for the m-th frame being analyzed.
  13. Select up to N_{max} local maxima from $R_s(\tau)$ in the lag range determined by the interval $(f0_{min}, f0_{max})$.
  14. Using cubic splines, interpolate around each peak and its left- and rightmost neighbor to refine the location and height of each extremum. Let τ_{max}^k and $R_s(\tau_{max}^k)$ be the locations and values of the interpolated extrema for k up to N_{max}.
  15. Let $f_{m,k} = \frac{F_s}{\tau_{max}^k}$ be the voiced candidates with strength as in Eq. 5.1.

Figure 5-1: Algorithm to obtain sequence of F0 candidates.

- Dynamic Programming to Obtain Optimal F0 Sequence. Let M be the maximum number of frames obtained in Algorithm 5.1, and let K_m be the number of candidates obtained from the m-th frame.

  16. Initialization. Let $\delta(1,k) = -r_{1,k}$ and $\psi(1,k) = 0$ for $k = 1, \cdots, K_1$

  17. Recursion. For $2 \leq m \leq M_{max}$, $1 \leq k \leq K_m$, and using Eq. 5.4, calculate

     a. $\delta(m,k) = -r_{m,k} + \min\limits_{1 \leq l \leq K_{m-1}} \Big(\delta(m-1,l) a_m(l,k) \Big)$

     b. $\psi(m,k) = \operatorname{argmin}\limits_{1 \leq l \leq K_{m-1}} \Big(\delta(m-1,l) a(l,k) \Big)$

  18. Backtracking. Let $l_M^* = \operatorname{argmin}\limits_{1 \leq l \leq K_M} \delta(M,l)$.

     a. For $m = M-1, \cdots, 1 : l_m^* = \psi(m+1, l_{m+1}^*)$

     b. $F0(m) \leftarrow l_m^*$

Figure 5-2: Dynamic programming algorithm to obtain optimal F0 sequence.

K_m found for the m-th frame:

$$C(\{m, l_m\}) = \sum_{m=2}^{M} a(f_{m-1,l_{m-1}}, f_{m,l_m}) - \sum_{m=1}^{M} r_{n,l_n}, \tag{5.4}$$

and

$$a(f_{m-1,l_{m-1}}, f_{m,l_m}) = \begin{cases} 0 & \text{if } f_{m-1,l_{m-1}} = 0 \text{ and } f_{m,l_m} = 0, \\ OctJump_{cost} \left| \log_2 \frac{f_{m-1,l_{m-1}}}{f_{m,l_m}} \right| & \text{if } f_{m-1,l_{m-1}} \neq 0 \text{ and } f_{m,l_m} \neq 0, \\ VcUnv_{cost} & \text{otherwise} \end{cases} \tag{5.5}$$

defines a local cost function for transitions between adjacent frequency values. The parameters $VcUnv_{cost}$ and $OctJump_{cost}$, also free parameters of the algorithm, penalize the cost between voiced and unvoiced frames and the cost associated with octave jumps, respectively. A dynamic programming algorithm to find the path that minimizes Eq. 5.4 is presented in Fig. 5-2. The optimal $F0$ contour is then taken to be the sequence of frequency candidates from the optimal path.

5.3 Stylization of F0

The previous section described a series of algorithms for arriving at a representation of the frequency of oscillation of the vocal folds in terms of a smooth curve. In this representation, smaller scale period-to-period variability is ignored in favor of a representation that is stable over a short time scale of usually a few dozen milliseconds. (We will return, however, to a characterization of this period-to-period fluctuation as a component of voice quality in Chapter 6.) The smooth representation of F0 allows us, for instance, to describe some general tendencies of the trajectory in terms of descriptive statistics. Through this approach, several authors have investigated and discovered correlates between affective categories and general statistics of the F0 contour (Williams & Stevens, 1972; Banse & Scherer, 1996). It has also been proposed, however, that affect may be encoded not just by adjusting some general continuous parameters of the F0 contour (such as pitch baseline, pitch range, etc.), but also by particular choices of more abstract categories or pitch types, such as those that could form the inventory of a phonological system of contrasts for intonation (Banse & Scherer, 1996; Mozziconacci, 2000; Mozziconacci, 2002). The purpose of this section is to review algorithms that yield a more parsimonious representation of the F0 curve in terms of primitive contour types. No claim about the phonological plausibility of this representation is made at this point; rather, the stylization is carried out only as a first-order approximation to a categorical description of intonational phenomena. The next section then proposes some features that can be extracted from this representation to augment a feature set derived from population statistics.

The stylization algorithm implemented here is adapted from that described by Bagshaw (1994) (which in turn is based on the work of Scheffers (1988)). The algorithm takes as its input an F0 contour as described in the previous sections, as well as a parsing of the speech waveform with syllable nuclei assignment as obtained with the algorithms introduced in Chapter 3. It returns a schematization of F0 consisting of a series of piecewise linear segments, joined at turning points, as well as information concerning whether each stylized piece constitutes a pitch accent. The objective of the stylization algorithm is to reduce the phonetic variation of production in order to arrive at a simplified representation in terms of types from a finite inventory. Turning points are landmark points of the F0 curve at which two (possibly distinct) types are joined, potentially creating a pitch accent.

The algorithm proceeds in three major sections. First, turning points are found from contiguous sections of voiced F0 contours. After interpolating the estimated turning points, each linear piece is schematized as a *rise*, a *fall*, or *level*. Finally, the schematized sequence is input into a rule-based filter that detects potential pitch accents by combining the schema with syllable nuclei information.

Let us first describe the turning point detection algorithm. We will assume that we have an F0 contour in semitones (where $F0_{semitones} = 12\log_2(F0_{Hz}/55)$), and that the contour is separated into maximal sections of contiguous voiced values. That is, we consider the largest segments of $F0$ with no voiced-to-unvoiced transitions. Short segments, corresponding to less than 40 msecs. in duration, are not submitted to the algorithm. To the rest of the segments, the algorithm shown in Fig. 5-3 is applied to detect significant turning points within the spans of voiced $F0$ sections. The turning-point detection proceeds as follows: Starting at the beginning of an $F0$ section, an initial window size of 5 points is selected (Step 1), and a straight line is fit using logistic regression (Steps 2 and 3). The fitted straight line is evaluated on the next point. If a good fit is found (i.e., the value predicted by the line does not deviate by more than 1 semitone from the actual $F0$ value), then it is assumed that the next $F0$ value does not constitute a turning point. The analysis window length is then incremented by one, and the process repeated (Step 4.1.) until this condition is violated, or the end of the voiced section is found. If, however, the next $F0$ value deviates from the prediction by more than the allowed amount, this value is considered to be either a turning point or a disturbance. In order to test which is the case (Step 5), the following sub-procedure is implemented.

A subinterval is setup containing all $F0$ frames starting with the candidate turning point and spanning up to 100 msecs. or up to the end of the voiced section (whichever occurs first). If *every* frame within the subinterval consistently deviates from the straight line prediction, then the candidate is considered to be a genuine turning point. The value of the candidate's location is recorded, and the positions of the analysis window are re-initialized to start at the newly found turning point, and to span 5 frames. The process already described above is then applied to detect new turning points. If, however, we encounter a frame within this subinterval that is well predicted by the fitted line, then the candidate turning point is considered to be a disturbance. The analysis interval is then readjusted to contain this new frame, but not any of the intermediate frames that deviated from the

Algorithm 5.2:

1. Let $N_i \leftarrow 1$ $N_f \leftarrow N_i + 4$, let $I = [N_i, Nf]$ and $T_p = N_i$.
2. If $N_f = N_{max}$ (i.e., end of voiced segment), $T_p \leftarrow T_p \cup N_f$. Go to step 6.
3. Apply logistic regression to the voiced segment $f_0(n)$ on the interval I to obtain the slope m_I and intercept b_I of a straight line fit over the interval.
4. Let $\epsilon(m_I, b_I) \doteq \sum_{n \in I} |f_0(n) - m_I n - b_I|$
   - 4.1 If $\epsilon < \epsilon_{th}$ ($\epsilon_{th} = 1$ semitone) (i.e., N_f is *not* a turning point): $N_f \leftarrow N_f + 1$; $I = [N_i, N_f]$; go to step 2.
   - 4.2 Otherwise, N_f is a *candidate* turning point. Go to step 5 and test.
5. Test candidate turning point:
   - 5.1. $k \leftarrow 1$, $N_k = N_f + k$
   - 5.2. Apply logistic regression on the interval $I = [N_i, N_k]$ to obtain the straight line parameters m_I and b_I.
   - 5.3 If $\epsilon(m_I, b_I) \geq \epsilon_{th}$, $\forall I = [N_i, N_k]$, such that $k \leq k_{max}$ (where $N_{k_{max}}$ is the smallest of the last frame N_{max} or a frame 100 msecs. away from N_f): Then N_f is a turning point. $T_p \leftarrow T_p \cup N_f$, $N_i \leftarrow N_f$, $N_f \leftarrow N_i + 4$; go to step 2.
   - 5.4 Otherwise, let N_r be the first frame encountered for which $\epsilon(m_I, b_I) < \epsilon_{th}$. Augment the analysis interval I: $I \leftarrow [N_i, N_f] \cup N_r$; update the interval endpoints: $N_f \leftarrow N_r$; and go to step 2.
6. Done. Return T_p

Figure 5-3: Algorithm for finding turning points on the F0 contour.

prediction by the allowed amount (that is to say, we augment the analysis window with the first point beyond the original candidate turning point that is well-predicted by the line, but eliminate any of the intermediate outliers that caused the disturbance). With the new analysis interval thus defined, the logistic regression analysis proceeds as described above until the end of the voiced section is encountered. The initial and final points of a maximal voiced section are always included in the set of turning points.

Since most of the turning points returned by the algorithm define a transition between two piecewise sections, a linear interpolation is next applied in order to assign each turning point t_p to a unique $F0$ value without causing a discontinuity (discontinuities are only allowed across turning points that belong to different maximal $F0$ sections). If n_{tp}, m_{tp} and

b_{tp} denote the location of the *pth* turning point, and the slope and intercept value of the piecewise section preceding this turning point, then

$$\hat{F0}(n_{tp}) = \begin{cases} \frac{1}{2}(m_{tp}n_{tp} + b_{tp} + \\ \qquad m_{tp+1}n_{tp} + b_{tp+1}) & \text{if frames } n_{tp}-1 \text{ and } n_{tp}+1 \text{ are voiced} \\ m_{tp+1}n_{tp} + b_{tp+1} & \text{if frame } n_{tp}-1 \text{ is unvoiced and } n_{tp}+1 \text{ is voiced} \\ m_{tp}n_{tp} + b_{tp} & \text{if frame } n_{tp}-1 \text{ is voiced and } n_{tp}+1 \text{ is unvoiced.} \end{cases} \tag{5.6}$$

A stylized contour can then be obtained by converting the interpolated $\hat{F0}$ values corresponding to the found turning points back to the Hertz scale, and leaving every unvoiced frame unchanged.

In order to detect the locations of potential pitch accents, a further level of schematization is invoked at this point to turn every linear piecewise section found through the stylization algorithm into one of three types of $F0$ segments: a *fall*, a *rise*, or a *level*. This basic kind of schematization can provide a reasonable account of the $F0$ behavior, such that fundamental perceptual properties of intonation are retained in the schematical categorization after removing $F0$ fluctuations. This forms the basis of the perceptual IPO approach to modeling intonation proposed by 't Hart and colleagues (t'Hart et al., 1990; Mozziconacci, 2000).

The set of turning points $\{T_p\}$ are first normalized to z-scores by taking into account their deviation from a straight line describing the overall tendency of the entire contour (typically a declination). A logistic regression is once again applied to the final set of turning points to obtain a straight-line fit described by the parameters m_{decl} and b_{decl}, from which the following utterance-dependent normalization is obtained:

$$\sigma^2_{log} = \frac{1}{|T_p| - 1} \sum_{n_{tp} \in T_p} \left(\hat{F0}(n_{tp}) - m_{decl} \times n_{tp} - b_{decl} \right)^2 \tag{5.7}$$

$$z(n_{tp}) = \frac{\hat{F0}(n_{tp}) - m_{decl} \times n_{tp} - b_{decl}}{\sigma_{log}}. \tag{5.8}$$

If $z(n_{tp})$ and $z(n_{tp+1})$ denote the beginning and end of a normalized piecewise section, then the trajectories are categorized by taking into account the height of the jump in any stylized

piece as a function of the spread of the normalized $F0$ values:

$$F0_{trajectory}(n_{tp}, n_{tp+1}) = \begin{cases} Fall & \text{if } z(n_{tp}) - z(n_{tp+1}) > 0.2 \times iqr\{z(n_{tp})\} \\ Rise & \text{if } z(n_{tp}) - z(n_{tp+1}) < -0.2 \times iqr\{z(n_{tp})\} \\ Level & \text{otherwise.} \end{cases} \tag{5.9}$$

Piecewise sections are considered part of a potential pitch accent if either at least 50% of the section overlaps a syllable nucleus or at least 50% of a syllable nucleus is overlapped by the piecewise section. In order to carry out this evaluation, we make use of the automatic segmentations and syllable nuclei classification already discussed in Chapter 3. When more than one piece is assigned to a nucleus, pieces are iteratively combined in a left-to-right fashion until there is a two-type combination involving rises, falls, and levels, or all pieces have been merged into a single type. Merging two pieces (i.e., three consecutive turning points $n_{tp}, n_{tp+1}, n_{tp+2}$) involves redefining a new piecewise segment on (n_{tp}, n_{tp+2}) and re-evaluating Eq. 5.9. It is therefore possible that segments change schema types throughout this merging process (e.g., two level types can become a rise or a fall). Once this process is complete, the pitch-accent classification filter is applied to each qualifying section.

Any nucleus that ends up receiving any combination of rise and fall types is automatically assigned a pitch accent. All other remaining types are classified according to the rule-based decision filter shown in Fig. 5-4 (following the pitch-accent detection filter described in Bagshaw (1994) and developed by Hieronymus (1989)). The filter operates by considering the type of the nucleus to be classified, as well as its left- and right-contexts (i.e., the scheme types associated with the preceding leftmost and rightmost vowels for which a known type is available). Fig. 5-4 summarizes the "grammar" of allowed combinations which receive pitch accents. The rules implement a combination of type checks followed in some cases by a condition requiring a step-up (s_u) or a step-down (s_d) within certain allowed ranges. Any combination of conditions specified through a connected path on the graph defines a legitimate sequence to receive a pitch accent. For instance, a nucleus classified as level, preceded by a combination of a rise with a step-up, and followed by step-down of at least Δ_f Hz to a rise type (or, alternatively, followed by a step-down to a fall type) receives a pitch accent.

The series of algorithms describe in this section yield a schematized F0 contour in

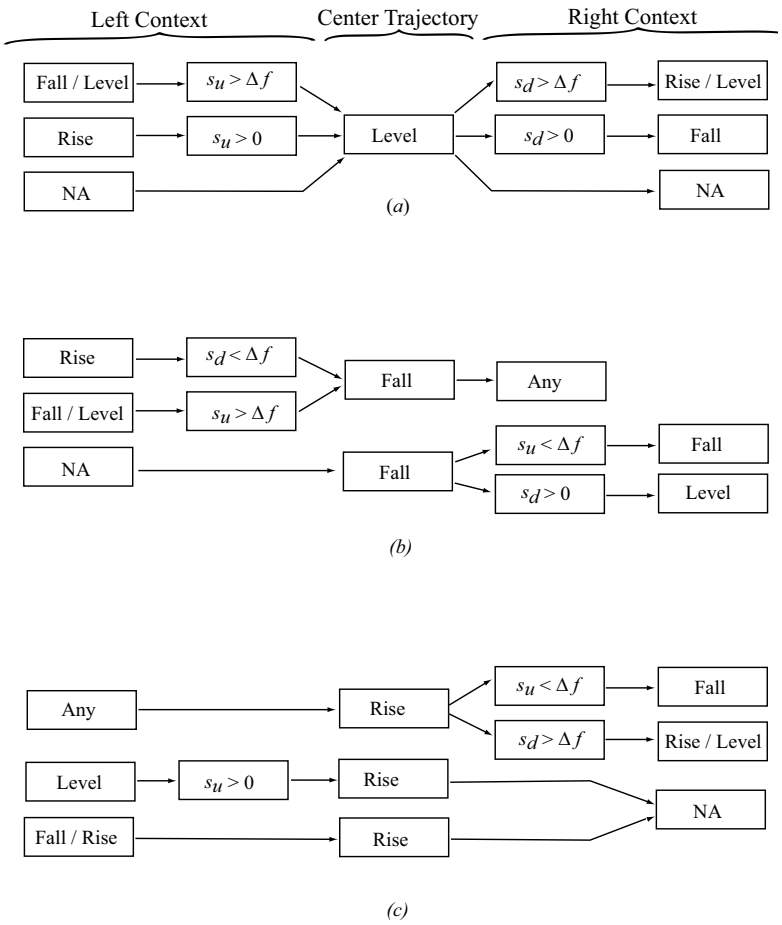

Figure 5-4: Rule-based pitch accent filter.

terms of a considerably downsampled set of turning points which define piecewise sections associated with three categorical types, and marked pitch accents. This representation can now be used, in conjunction with the raw, unschematized $F0$ contour obtained through the algorithm in Section 5.2 to define a set of features for further analysis.

5.4 Analysis of Features from Raw and Stylized F0 Contours

Fundamental frequency has been one of the components of spoken language most studied in the literature on speech and emotion (Williams & Stevens, 1972; Scherer et al., 1984; Ladd et al., 1985; Banse & Scherer, 1996; Paeschke et al., 1999; Mozziconacci, 2000; Paeschke & Sedlmeier, 2000; Mozziconacci, 2002). Since fundamental frequency serves a properly linguistic function, this work has placed emphasis on trying to sift out what aspects of F0 are made available for encoding linguistic meaning (such as the contrast between statements

and questions) and what aspects of it serve a paralinguistic function. Several studies have looked at this question differently, typically seeking to understand the linguistic function of F0 by appealing to a description (or decomposition) of the fundamental frequency curve in terms of basic categorical elements (where the number and nature of these elements varies according to theoretical accounts, such as the tone group account of the British school, or Pierrehumbert's finite-state grammar of intonational sequences (Cruttenden, 1997)). The paralinguistic function of F0, however, has been assumed to reside in the continuous variation of parameters (or "gradient features," to use the terminology of Scherer et al. (1984) and Ladd (1996)), such as range and F0 baseline, which can be superimposed on categorical configurations dictated by phonological constraints. Despite differences in theoretical assumptions regarding intonational models, independent studies have been able to confirm that the perception of emotion from F0 contours does not merely arise from paralinguistic variation, but rather that the intonational categories under consideration contribute to the perception of affect. For instance, the studies of Mozziconacci (2000), where F0 contours were analyzed using the IPO framework, reveal that although no F0 contour type is strictly necessary to convey any of the affect conditions under study, the phonological choice of contour type does contribute to the perception of emotion.

Motivated by these observations, this work incorporates two basic kinds of analyses of fundamental frequency: a series of statistics summarizing the behavior of the F0 curve (and conceptually closer to the paralinguistic gradient features discussed by Scherer and Ladd), and a set of features from the stylized F0 contour (as obtained through the algorithm described in the previous section). No claim is made that the stylization is close to any phonological description of F0 contours. Rather, it is just an attempt to reduce some of the surface phonetic realization to arrive at more abstract descriptions that might be useful for affect modeling.

Let, therefore, $F0(n)$ be a fundamental frequency contour in semitones, and $F0_v(n)$ be the subset of N_v voiced values in $F0$, and consider the following statistics:

$$\mu_v = \frac{1}{N_v} \sum_{n=1}^{N_v} F0_v(n) \tag{5.10}$$

$$\sigma_v^2 = \frac{1}{N_v} \sum_{n=1}^{N_v} \Big(F0_v(n) - \mu_v \Big)^2 \tag{5.11}$$

$$skew_v = \frac{1}{N_v} \frac{\sum_{n=1}^{N_v} \Big(F0_v(n) - \mu_v \Big)^3}{\sigma_v^3} \tag{5.12}$$

$$F0_{>\mu_v} = \frac{1}{N_v} \sum_{n=1}^{N_v} 1 - \delta(F0_v(n) > \mu_v) \tag{5.13}$$

$$iqr_v = \text{prctile}_{75}(F0_v(n)) - \text{prctile}_{25}(F0_v(n)) \tag{5.14}$$

$$range_v^+ = \text{prctile}_{95}(F0_v(n)) - \mu_v \tag{5.15}$$

$$range_v^- = \mu_v - \text{prctile}_5(F0_v(n)). \tag{5.16}$$

Eq. 5.12 is the sample skewness of $F0_v(n)$ and Eq. 5.13 the proportion of voiced values that lie above the mean. These provide measures of how asymmetrical the F0 curve is with respect to its sample mean. Eq. 5.14 has been used to model more robustly the spread, or deviation, of F0 around its mean, and to reduce the sensitivity to outliers that may not have been eliminated from the F0 estimate by the dynamic programming smoothing discussed earlier. Eqs. 5.15 and 5.16 capture the range of the F0 signal, above and below its sample mean (where the 95% and 5% percentile values have been used instead of the maximum and minimum values, again, to be robust to outliers). From these values we define the feature vector

$$FS_{F0-raw} = [skew_v, F0_{>\mu_v}, iqr_v, range_v^+, range_v^-]^T. \tag{5.17}$$

Let us next assume that by applying the stylization algorithm discussed in the preceding section we have arrived at a description of the F0 curve in terms of a set of turning points $\{T_p\} = n_1, \cdots, n_P$ (defining linear piecewise sections), and that every syllable s_k, $k = 1, \cdots, K$, associated with a section of F0, has been classified using the method described as receiving a pitch accent or being unstressed. Let PA_k be an indicator function equal to 1 if the kth syllable is classified as accented, and 0 otherwise. Since several turning points may be associated with the span of a syllable nucleus, let $B_p^k \subset T_p$ be the subset of turning points occurring within the kth syllable. Finally, let us assume that $\{Sl_{flr}\} = sl_1, \cdots, sl_{P-1}$ is a set of indicators specifying the slope type between two adjacent turning points as a fall (-1), level (0) or a rise $(+1)$, and recall the z-scores defined by Eq. 5.8 and the slope m_{decl} of the linear regression fit through the turning points. With these specified quantities, we

define the following measures:

$$PA_{perc} = \frac{1}{K}\sum_{k=1}^{K} PA_k \tag{5.18}$$

$$\Delta Sl_{perc} = \sum_{p=1}^{P-1} \frac{|sl_{p+1} - sl_p|}{2} \tag{5.19}$$

$$k^* = \text{argmax}_k \Big(\max_{n_p \in B_p^k} z(n_p) \Big) \tag{5.20}$$

$$PA^*_{max} = \max_{n_p \in B_p^{k^*}} z(n_p) \tag{5.21}$$

$$PA^*_{range} = \max_{n_p \in B_p^{k^*}} z(n_p) - \min_{n_p \in B_p^{k^*}} z(n_p) \tag{5.22}$$

$$m_{last} = \frac{z(n_P) - z(n_{P-1})}{n_P - n_{P-1}}. \tag{5.23}$$

Eq. 5.18 defines the proportion of syllables to receive a pitch accent and Eq. 5.19 the proportion of stylized slope changes throughout the F0 contour. The value of k^* (5.20) signals the syllable position with the strongest pitch accent in the contour, from which we derive the height (in terms of the normalized z-scores) and range of this pitch accent (Eqs. 5.21 and 5.22 respectively). Finally, m_{last} represents the slope of the last stylized section. The feature vector of stylized F0 features is defined as

$$FS_{F0-styl} = [m_{decl}, PA_{perc}, \Delta Sl_{perc}, PA^*_{max}, PA^*_{range}, m_{last}]^T. \tag{5.24}$$

Features like m_{decl} provide a description of the overall declination (or inclination) of the F0 contour whereas ΔSl_{perc} provides information about the amount of peaks in the stylized F0 contour . The feature PA_{perc} is proposed to capture information about the degree to which syllables are accented, a feature which intuitively should reflect much about speaking style and which has been exploited before in affective speech synthesis (Cahn, 1990). PA^*_{max} and PA^*_{range} provide some descriptive statistics about the strongest pitch accent in the contour. Information regarding boundary tones and phrase accents (following Pierrehumbert's context-free grammar of intonation) has been considered before (typically in the form of hand labeled annotations) in the automatic analysis of affect from intonation (Liscombe et al., 2003). The emphasis in this work is on having all subcomponents, feature extraction algorithms as well as the learning system, work automatically. The feature m_{last} is included

as an attempt to facilitate an automatic measure that might correlate with boundary tone information.

A model of the F0 contour is therefore obtained by combining the features derived from the continuous F0 representation and its stylization:

$$FS_{F0} = [FS^T_{F0-raw}, \quad FS^T_{F0-styl}]^T. \tag{5.25}$$

This is the final representation that we submit to the analysis in the next section.

5.4.1 Evaluation and Discussion

Tables 5.1 and 5.2 show the results obtained from running subject-dependent Bayesian classifiers (with pooled covariance matrices) on a set containing the features defined by Eq. 5.25. for the 11 subjects of the actors database considered. Classifiers have been trained to discriminate between emotion pairs, and their generalization performance estimated by using leave-one-out cross-validation. Significant results at the $p < 0.05$ and $p < 0.01$ level are indicated on the tables, with the latter cases typeset in bold for ease of inspection. (Details of the significance test are provided in Appendix D.)

As the results show, the features considered provide significant discrimination for several emotion pairs for various subjects. It is also the case that the level of discrimination varies with the subject, with only one case of significant discrimination obtained for subjects 6 and 11 while significant recognition is attained for all but two emotion pairs for subject 3. A fairly stable result throughout the set is that the F0 features tends to provide significant discrimination between the *Happy* category and the remaining affect categories considered, attaining in some cases (e.g., subject 7) very good generalization. It is also noteworthy to examine how different subjects appeal to F0 when encoding affect: Whereas the results for subject 7 suggest that she makes good use of the F0 cues considered to set apart her *Happy* vocalizations, the performance on subject 5 suggests that she uses the cues to distinguish *Neutral* speech from the rest. This once again brings to the foreground the difficulty of building subject-independent affect recognizers from speech: even if reliable acoustic features can be identified, the subject-dependent usage of those cues can be highly idiosyncratic and may serve very different communicative purposes. Finally, it should be pointed out that removing either the FS_{F0-raw} or the $FS_{F0-styl}$ entries from the analysis of

	Ang	Hap	Ntr	Sad
Afr	**0.31**†	**0.33**†	0.40	**0.24**†
Ang		**0.26**†	0.38‡	0.44
Hap			0.44	**0.20**†
Ntr				**0.31**†

Subject 1

	Ang	Hap	Ntr	Sad
Afr	**0.34**†	**0.23**†	**0.30**†	**0.32**†
Ang		**0.33**†	0.35‡	**0.32**†
Hap			**0.31**†	0.43
Ntr				0.49

Subject 2

	Ang	Hap	Ntr	Sad
Afr	**0.24**†	**0.27**†	**0.25**†	0.50
Ang		**0.25**†	0.48	**0.21**†
Hap			**0.20**†	**0.30**†
Ntr				**0.22**†

Subject 3

	Ang	Hap	Ntr	Sad
Afr	**0.32**†	**0.21**†	0.48	**0.33**†
Ang		**0.15**†	0.38‡	0.38‡
Hap			**0.19**†	**0.15**†
Ntr				0.37‡

Subject 4

	Ang	Hap	Ntr	Sad
Afr	0.39	0.43	**0.26**†	0.41
Ang		0.42	**0.27**†	0.49
Hap			**0.17**†	0.41
Ntr				**0.28**†

Subject 5

	Ang	Hap	Ntr	Sad
Afr	0.56	0.40	0.51	0.45
Ang		0.50	0.39	0.37‡
Hap			**0.30**†	0.39
Ntr				0.46

Subject 6

Table 5.1: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Bayesian classifiers (with pooled covariance matrices) trained on emotions pairs using feature set derived from raw F0 contour and its stylization for subjects 1 through 6 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

the feature vector in Eq. 5.25 increased the overall generalization error rate, suggesting that both components (global statistics and stylized features) are contributing independently to the performance of the classifiers.

Table 5.3 shows the generalization error exhibited by the same Bayesian classifier when trained to discriminate between each emotion category and the remaining labels in the set. The data has been subsampled to retain equal priors, and chance-level discrimination is 50%. The columns of table 5.3 show that the best discrimination across subjects for the features considered in this section is obtained for the speech tokens labeled as *Happy*. The F0 features also provide an improvement for the *Neutral* and *Sad* categories over the performance obtained with the loudness features in Chapter 4, demonstrating the need for independent sources to capture differences between the affective labels considered.

	Ang	Hap	Ntr	Sad
Afr	0.40	**0.11**†	0.38	0.40
Ang		**0.08**†	0.42	0.41
Hap			**0.08**†	**0.07**†
Ntr				0.36‡

Subject 7

	Ang	Hap	Ntr	Sad
Afr	**0.30**†	0.42	**0.33**†	0.40
Ang		**0.31**†	**0.34**†	0.45
Hap			**0.28**†	0.35‡
Ntr				0.42

Subject 8

	Ang	Hap	Ntr	Sad
Afr	**0.32**†	0.44	**0.31**†	0.46
Ang		**0.28**†	0.51	0.49
Hap			**0.21**†	**0.30**†
Ntr				0.38‡

Subject 9

	Ang	Hap	Ntr	Sad
Afr	0.42	**0.26**†	0.35‡	0.55
Ang		**0.26**†	**0.33**†	**0.34**†
Hap			**0.19**†	**0.31**†
Ntr				0.46

Subject 10

	Ang	Hap	Ntr	Sad
Afr	0.53	**0.28**†	0.37	0.46
Ang		0.44	0.46	0.49
Hap			0.40	0.48
Ntr				0.44

Subject 11

Table 5.2: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Bayesian classifiers (with pooled covariance matrices) trained on emotions pairs using feature set derived from raw F0 contour and its stylization for subjects 7 through 11 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

5.5 "Chordal" Analysis of Intonational Contours

The analysis of the previous section relied on a description of the fundamental frequency contour in terms of global statistics derived from the voiced sections (Eq. 5.17), as well as a stylization that identified certain landmarks and produced descriptive statistics as a function of this stylization (Eq. 5.24). The analysis introduced in this section takes an alternative look by considering how speakers structure their intonational contours. This approach focuses on how speakers make quantal use of their pitch range by establishing certain F0 modes –values of F0 which are statistically recurrent throughout a phrase– and the perceptual consonance properties of the intervals that result from this division.

It has recently been argued by Cook (2002) that basic statistics derived from the F0 contour (e.g., mean, variance, range, etc.) fail to show a structure to which speakers may

Subject	Afr	Ang	Hap	Ntr	Sad
1	0.40	0.40	**0.33**†	0.49	**0.33**†
2	**0.32**†	0.41	**0.32**†	0.45	0.38‡
3	**0.37**†	**0.36**†	**0.36**†	**0.35**†	**0.36**†
4	**0.34**†	**0.35**†	**0.23**†	0.45	**0.25**†
5	0.45	0.55	**0.33**†	**0.25**†	0.51
6	0.59	**0.38**†	0.44	**0.34**†	0.51
7	**0.33**†	0.49	**0.10**†	0.42	**0.36**†
8	0.49	0.42	**0.29**†	0.39	0.40
9	0.40	0.53	**0.34**†	**0.37**†	0.55
10	0.40	**0.35**†	**0.22**†	**0.31**†	0.52
11	**0.36**†	0.56	0.40	0.46	0.58

Table 5.3: Generalization error (estimated by leave-one-out cross validation) for subject-dependent Bayesian classifiers trained to discriminate between each emotion category and the remaining ones pooled together, and using feature set derived from raw F0 contour and its stylization (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

appeal when signaling different affective states. Cook (2002) and colleagues set out to investigate whether F0 contours exhibit a chordal structure, and, by borrowing concepts from traditional harmonic theory, whether there is any triadic substructure in the resulting chords. In particular, the authors were interested in investigating whether the structure of major and minor chords, with their long-held impressionistic associations with happy and sad moods, was also reflected in speech that conveyed those qualities. Although their study did not confirm such a link, and the generality of their results remains to be tested, it did suggest that speakers may organize the F0 contours of emotional speech so as to exploit harmonic properties of the resulting intervalic structure. In the following analysis, we focus on how to quantify this structure, and how to describe its effects by using psychoacoustic models about the perception of frequency intervals.

The first noteworthy property of F0 contours relevant to this analysis is the multimodal distribution, as opposed to unimodal or uniform, that F0 contours tend to exhibit. In fact, Cook (2002) found that most of the emotional speech data in his analysis of 24 Japanese speakers (males and females) showed bimodal or trimodal F0 distributions To quantify this phenomenon, we can estimate the distribution of voiced F0 values by fitting a multimodal distribution using a mixture of Gaussian components, and then use the resulting parameters (i.e., the means, variances and mixture weights) to estimate the location of the modes and

therefore to quantify the intervalic structure (in terms of the means), the degree to which each "tone" is well-defined or spread out (in terms of the mixture variances), and the percent of time a mode is invoked throughout the phrase (in terms of the weights). Let the distribution of F0, therefore, be modeled by a trimodal distribution

$$p(F0) = \sum_{k=1}^{3} \pi_k \mathcal{N}(F0; \mu_k, \sigma_k), \tag{5.26}$$

where we will assume, without losing any generality, that $\mu_1 < \mu_2 < \mu_3$.

One of the fundamental psychoacoustic properties of frequency intervals, known since the work of Helmholtz in the 19th century, is that they carry an intrinsic dissonance (or, alternatively, consonance) level. Although dissonance is a concept that is hard to define (and one which has received differing definitions throughout the literature; consult Sethares (1998) for a review), we will concern ourselves with *sensory dissonance*, the notion of dissonance which describes the gentle or rough quality perceived when two pure tones of different frequencies interact with each other. There are a few well-known robust results concerning this phenomenon: listeners perceive the interval between the tones as most consonant when they are in unison; then their perception of consonance sharply falls as the interval increases as a 'rough beating' is perceived and, after reaching maximum dissonance, slowly increases toward consonance again as the two tones separate perceptually. This perceptual pattern follows for the interaction of two pure tones. When higher harmonics (upper partials) are taken into account, the "consonance curve" can adopt various shapes depending on the location and strength of the interacting harmonics. There have been models proposed to quantify this perceptual phenomenon. For this work, we will adopt the model proposed by Sethares (1998), and based on the earlier work of Plomp & Levelt (1965). This model is separately reviewed in Appendix B. Let us assume for the rest of the discussion, then, that we have a dissonance model D available, which maps a pattern of frequencies $f_1, \cdots, f_k$ and respective amplitudes $a_1, \cdots, a_k$ to a single-valued variable.

A second psychoacoustical claim relevant to the present analysis (although perhaps one less universally accepted) posits that the perception of chords is distinct from the perception of the constituting intervals. That is, chordal consonance is not just the summation of interval consonance; rather, a certain gestalt process seems to be at work in this perceptual process where some other properties relevant to the identity of the chord distinguish it from

other chords of similarly constituting intervals. Cook (2002) refers to this aspect as the chord tension, to distinguish it from the interval dissonance described earlier, and proposes to model it with the following expression:[1]

$$T(f_1, f_2, f_3) = \frac{\Delta_1 + \Delta_2}{2} \cdot \exp\big(-(|\Delta_1 - \Delta_2|/0.6)^2\big), \tag{5.27}$$

where Δ_1 and Δ_2 are the two smallest intervals (in semitones). The semitone scale adopted here is the 12-tone equitempered scale with $f_0 = 55Hz$, and octave values defined as in $f_n = r^n f_0, n = 0, \cdots, 11$ with $r = \sqrt[12]{2}$. Continuous semitone values k are then obtained from F0 via the following equation:

$$k = 12 \log_2\Big(\frac{F0}{55}\Big), \tag{5.28}$$

and values outside the range $[0, 12)$ are folded modulo-12 to keep the range within an equivalent octave.

Cook (2002) proposes an additive model of intervalic dissonance and chordal tension to capture a global property of a chord which he terms *instability*

$$I = D_{interval} + T(f_1, f_2, f_3) \tag{5.29}$$

To apply this model to the analysis of F0 contours, we can use the estimated parameters of the Gaussian mixture distributions. The intervalic dissonance component is calculated by creating an augmented frequency pattern containing the modes of the distribution (with relative amplitudes of 1), and second and third "partials" with relative amplitudes of 1/2 and 1/3. The intrinsic dissonance of the resulting frequency pattern

$$F = \{\mu_1, 2\mu_1, 3\mu_1, \cdots, \mu_3, 2\mu_3, 3\mu_3\} \tag{5.30}$$

with amplitude

$$A = \Big\{1\frac{1}{2}, \frac{1}{3}, \cdots, 1, \frac{1}{2}, \frac{1}{3}\Big\} \tag{5.31}$$

is then evaluated using the dissonance model of Appendix B (Eq. B.9).

[1] The original equation in Cook (2002) includes an additional scaling factor that is expressly introduced to capture the differences between augmented and diminished chords, and which has been omitted from this analysis.

The rest of the analysis directly involves the parameters obtained from the Gaussian fit. Letting the mode corresponding to the highest mixture component represent the "tonic" mode (i.e., the maximum of the distribution), and be indexed by k_{ton}, and the minimum and median mixture components be indexed by k_{min} and k_{med}, we define the following features:

$$\pi^* = \pi_{k_{ton}} \tag{5.32}$$

$$\pi^n_{min} = \frac{\pi_{k_{min}}}{\pi_{k_{ton}}} \tag{5.33}$$

$$\pi^n_{med} = \frac{\pi_{k_{med}}}{\pi_{k_{ton}}} \tag{5.34}$$

$$\sigma^n_{min} = \frac{\log \sigma_{k_{min}}}{\log \sigma_{k_{ton}}} \tag{5.35}$$

$$\sigma^n_{med} = \frac{\log \sigma_{k_{med}}}{\log \sigma_{k_{ton}}} \tag{5.36}$$

$$FR_k = \frac{\mu_k}{\mu_{k-1}}, \quad k = 2, 3 \tag{5.37}$$

Eqs. 5.33 and5.34 represent the strength of the weakest and median mode with respect to the tonic mode. Similarly, Eqs. 5.35 and 5.36 measure how concentrated or spread out each constituent of the chord is, with respect to the spread of the tonic mode. Finally, Eq. 5.37 measures the intervals in terms of ratios. These features are chosen to parsimoniously represent the trimodal structure of an F0 contour.[2] Together with the intrinsic dissonance feature discussed above, they define the feature vector

$$FS_{ch} \doteq [I, \pi^*, \pi^n_{min}, \pi^n_{med}, \sigma^n_{min}, \sigma^n_{med}, FR_1, FR_2]^T, \tag{5.38}$$

which is submitted to the statistical analysis described in the next section.

5.5.1 Evaluation and Discussion

In order to assess the potential of the features defined by Eq. 5.38 in discriminating between emotion categories, we trained subject-dependent classifiers on this feature set alone, withholding any remaining features under study in this work. Cook (2002) investigated the chordal structure of F0 contours and proposed further investigating the use of intervalic

[2]Suitable values of these parameters subsume the case when the contour may not exhibit three modes (e.g., $FR_k \approx 1$), so it is not fundamental that the assumption hold as long as the features capture the structure.

structure to denote affect. Although their study looked at tendencies in the structure of F0 contours for various examples of affective utterances, the current investigation further examines the *predictive* power of these features in discriminating between affect categories. The results of the generalization error of the trained classifiers (Bayesian classifiers with arbitrary class-dependent covariance matrices), estimated through leave-one-out cross-validation, are summarized in Tables 5.4 and 5.5 for the data set of 11 speakers from the actors data. Significant results at the $p < 0.05$ and $p < 0.01$ level are indicated on the tables, with the latter cases typeset in bold for ease of inspection. (See Appendix D for details on the significance test.)

	Ang	Hap	Ntr	Sad
Afr	0.49	0.39	**0.29**†	0.44
Ang		**0.26**†	**0.27**†	0.45
Hap			0.42	**0.34**†
Ntr				**0.28**†

Subject 1

	Ang	Hap	Ntr	Sad
Afr	0.44	0.41	0.37‡	0.44
Ang		**0.28**†	0.47	0.47
Hap			0.47	0.44
Ntr				0.41

Subject 2

	Ang	Hap	Ntr	Sad
Afr	**0.34**†	0.45	**0.29**†	0.57
Ang		**0.25**†	0.49	0.40
Hap			**0.24**†	0.40
Ntr				**0.30**†

Subject 3

	Ang	Hap	Ntr	Sad
Afr	**0.31**†	**0.26**†	0.36‡	0.36‡
Ang		**0.34**†	0.41	0.51
Hap			**0.26**†	0.46
Ntr				0.49

Subject 4

	Ang	Hap	Ntr	Sad
Afr	0.39	0.53	**0.30**†	0.38‡
Ang		**0.30**†	0.38‡	**0.28**†
Hap			**0.25**†	0.41
Ntr				0.46

Subject 5

	Ang	Hap	Ntr	Sad
Afr	0.47	0.40	0.50	0.43
Ang		0.39	0.44	0.38
Hap			**0.33**†	**0.29**†
Ntr				0.48

Subject 6

Table 5.4: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent quadratic classifiers trained on emotions pairs using feature set derived from chordal representation of F0 for subjects 1 through 6 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

With the exception of subject 8, the feature set proved to discriminate significantly better than chance for some emotion categories for all subjects. In particular, there seems to be a tendency across subjects for this feature set to provide significant discrimination between the *Happy* categories and other classes. This behavior is most notably clear for

subject 7, where significant discrimination is achieved between all pairwise combinations involving this category. This result is consistent with what was obtained through the analysis of the F0 features using global F0 statistics and stylization features (summarized in Tables 5.1 - 5.3). However, the analysis of chordal consonance proposed in this section is not just merely replicating the results obtained earlier; in many cases it is also providing complementary information. A comparison of the results obtained for subjects 8 and 11 provides a good illustration (Tables 5.2 and 5.5). It should be noted that the particular distinction *Happy* vs. *Angry*, a distinction primarily of valence, was significantly predicted for 7 of 11 subjects. The difference between *Happy* and *Neutral* utterances was also predicted for 8 of these subjects.

	Ang	Hap	Ntr	Sad
Afr	0.41	**0.27**†	0.50	0.40
Ang		**0.21**†	0.41	0.41
Hap			**0.21**†	**0.25**†
Ntr				0.35‡

Subject 7

	Ang	Hap	Ntr	Sad
Afr	0.55	0.44	0.43	0.43
Ang		0.41	0.39	0.46
Hap			0.47	0.50
Ntr				0.48

Subject 8

	Ang	Hap	Ntr	Sad
Afr	0.44	0.47	0.40	0.40
Ang		0.43	0.52	0.53
Hap			**0.32**†	0.48
Ntr				0.48

Subject 9

	Ang	Hap	Ntr	Sad
Afr	0.51	0.35‡	0.49	0.50
Ang		0.45	0.55	0.44
Hap			**0.29**†	**0.33**†
Ntr				0.45

Subject 10

	Ang	Hap	Ntr	Sad
Afr	**0.34**†	0.39	0.39	0.35‡
Ang		**0.28**†	0.52	0.42
Hap			**0.25**†	**0.23**†
Ntr				0.44

Subject 11

Table 5.5: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent quadratic classifiers trained on emotions pairs using feature set derived from chordal representation of F0 for subjects 7 through 11 ($\dagger$and $\ddagger$denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

Table 5.6 shows the generalization error of the same classifiers when trained to predict the difference between each affect label and the rest. The data has been sampled to reflect equal priors before training, and chance level is 50%. Only the speech tokens labeled as

Happy are significantly discriminated based on these features for about half of the subjects. As noted, the results seem to complement those shown in table 5.3 for the discriminants based on features from F0 (raw and stylized): Features derived from F0 and its stylization provided significant separation for the *Happy* category for all subjects except 6 and 11 whereas features derived from the chordal representation of F0 attain significance for those two subjects. These suggests that the two feature sets, and the two ways to treat the F0 contour, may be offering complementary information, and that this may be employed by different subjects to encode the affective differences associated with this affective category.

Subject	Afr	Ang	Hap	Ntr	Sad
1	0.43	0.42	**0.38**†	0.41	0.48
2	0.52	**0.33**†	0.40	0.41	0.48
3	0.45	**0.37**‡	0.36‡	0.40	0.56
4	0.38‡	0.43	**0.30**†	0.46	0.44
5	0.42	0.40	0.39	0.42	0.38‡
6	0.58	0.54	**0.33**†	0.46	0.47
7	0.41	0.44	**0.26**†	0.37‡	0.40
8	0.44	0.44	0.49	0.46	0.53
9	0.50	0.53	0.37†	0.52	0.43
10	0.42	0.51	**0.34**†	0.49	0.42
11	0.53	0.45	**0.29**†	0.48	0.39

Table 5.6: Generalization error (estimated by leave-one-out cross validation) for subject-dependent quadratic classifiers trained to discriminate between each emotion category and the remaining ones pooled together, and using feature set derived from chordal representation of F0 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

5.6 Chapter Summary

This chapter has looked at the analysis of F0 contours for modeling affective categories from speech. The fundamental frequency of voice has received attention in the literature as one of the principal aspects of speech to carry affective content. In addition to re-examining some of the classical features containing global statistics of F0 and considered elsewhere in the literature, we have applied a stylization algorithm to distill the F0 contour to a set of landmark points and primitive contour types that attempts to obtain a more parsimonious representation of F0 as a first approximation to a categorical description of intonational phenomena. We have proposed a feature set obtained from the more global and the more

parsimonious representations of F0 and shown that they provide significant discrimination between emotion pairs for several subjects. An alternative approach has treated F0 contours as the unfolded representation of the components of a three-chord structure and appealed to properties of these frequency clusters, including measures summarizing perceptual properties of the intervalic distance, to define a set of features for automatic discrimination of affect categories. Although such view of F0 contours has been proposed elsewhere, the inclusion of consonance features for automatic affect discrimination is, to the best of our knowledge, a novel approach, as is the inclusion of an extended feature set derived from this analysis.

Chapter 6

The Analysis of Voice Quality

6.1 Introduction

In this chapter we turn to the analysis of voice quality. By voice quality, we intend to compound here a host of perceptual effects associated with variations in the voice source, or glottal excitation. In broad terms, the production of many eggressive pulmonic speech sounds (i.e., sounds driven by an air stream from the lungs being expelled outwardly) can be described as the interaction of a source, originated at the glottis, with the vocal tract, set to a particular configuration to achieve a desired effect (e.g., segmental targets). One particular model, known as the source-filter model of speech production, models the interaction of these two components through a convolution between the excitation and a linear filter, assumed also to be time-invariant during a portion of nearly stationary speech. It is thought that all languages appeal to a very basic distinction between somewhat regular pulse-like sources and noise-like sources to encode the phonological differences known as voicing and voicelessness. Furthermore, some languages employ other aspects of vocal source variations –involving the shape and regularity of the pulses, amount of airflow, etc.– for the purpose of linguistic contrast (Laver, 1994). The contribution of voice quality to signaling paralinguistic and extralinguistic information such as attitude, emotion, and vocal pathological states has been often remarked upon in the literature. In particular, the work of Gobl & Ní Chasaide (2003) has explored the perceived affective correlates associated with variations in the glottal source, and Scherer et al. (1984) (through experiments that attempted to mask different aspects of speech such as intonation, verbal content, and voice quality) have proposed that voice quality may play a fundamental role in the perception of

affect. In what follows, we introduce a series of supporting algorithms for modeling aspects of the glottal excitation. From this analysis, we propose three different feature sets related to voice quality and evaluate their performance in predicting affect categories.

6.2 Estimation of the Glottal Flow Derivative

Following the linear source-filter theory, an estimate of the glottal flow derivative may be obtained by inverse filtering a stationary segment of the speech signal with a vocal tract transfer function estimate obtained from the segment. Letting $S(z)$, $U(z)$, $H(z)$, and $R(z)$ be respectively the z-transforms of the acoustic speech signal $s(n)$, the glottal pulse $u(n)$, the impulse response of the vocal tract $h(n)$ and the lip radiation effects $r(n)$, the source-filter model of speech production describes this process by the relations

$$\begin{aligned} S(z) &= U(z)H(z)R(z) = G(z)H(z) && (6.1) \\ \updownarrow & \\ s(n) &= u(n) * h(n) * r(n) = g(n) * h(n), && (6.2) \end{aligned}$$

where the effect of the lip radiation, typically modeled by a differentiator (i.e., with a single-pole transfer function with a root on, or near, $|z| = 1$) has been combined with the glottal pulse to define the glottal flow derivative term $g(n)$. Inverse filtering then consists of forming a suitable estimate of the vocal tract transfer function $\hat{H}(z)$ to obtain an estimate of the glottal flow derivative. If, furthermore, the vocal tract transfer function is assumed to be modeled by the all-pole pth-order filter $H(z) = 1/A(z) = 1/(1 - \sum_{i=1}^{p} a_i z^{-i})$, then inverse filtering consists of applying the following recursion

$$\begin{aligned} \hat{G}(z) &= \frac{S(z)}{\hat{H}(z)} = S(z)A(z) && (6.3) \\ \updownarrow & \\ \hat{g}(n) &= s(n) * a(n) = 1 - \sum_{i=1}^{p} a_i s(n-i). && (6.4) \end{aligned}$$

We then switch attention to examine the conditions under which we can form a suitable estimate of the effects of the vocal tract, and methods to estimate the set of coefficients $\{a_i\}$.

During modal phonation, the behavior of the vocal folds, and thus the shapes of the glottal pulse $u(n)$ and its derivative $g(n)$, can be qualitatively described in terms of three disjoint phases: an *open phase* starting when the vocal folds abduct, and air is let through until the excitation is highest (and the glottal pulse derivative attains its minimum), a *return phase* describing the portion of a glottal cycle starting from this point onwards until the vocal folds close; and a *closed phase* during which the vocal folds remain closed.

As Eq. 6.1 shows, the output speech is the result of the interaction (via a convolution operation) of two quantities, neither of which are known directly. In order to arrive at an estimate of these two unknowns, different algorithms proceed differently. A class of algorithms based on blind deconvolution attempt to separate the two sources simultaneously by, for instance, transforming the speech signal to a domain where the two sources are more easily separated (e.g., cepstral domain), and then applying an operation that filters out the relevant regions of this domain (e.g., lifting). Inverse filtering algorithms, on the contrary, proceed in a two-step approach where the effort is first shifted to explicitly identifying a parameterized model of the vocal tract. Once this estimate is available, the problem reduces to a simple deconvolution with only one unknown (Eq. 6.4). The basis of most inverse filtering algorithms lies in recognizing that during the closed phase of a glottal cycle, the input to the vocal tract is zero. Following the closure of the glottis, therefore, the vocal tract becomes uncoupled from the glottal source, and the speech waveform becomes a freely decaying oscillation. Formally, assuming the all-pole model introduced earlier, Eq. 6.2 becomes

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + g(n), \tag{6.5}$$

which, in the presence of zero input ($g(n) = 0$), corresponds to a p-th order autoregressive model of the speech waveform

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) \qquad N_c < n < N_o, \tag{6.6}$$

where N_c and N_o correspond to the instants of closing and opening of the glottis and the interval (N_c, N_o) describes the closed phase. These observations outline the following approach to estimating the vocal tract transfer function. For each glottal cycle,

1. identify the closed-phase region;

2. apply a parameter estimation algorithm to the speech segmented from the closed phase (i.e., such that Eq. 6.6 holds) to obtain the coefficients $\{a_i\}$.

6.2.1 Closed-Phase Identification

There exist a variety of algorithms for identification of the closed phase of a glottal cycle. One straightforward approach is derived from the fact that Eq. 6.6 only holds during the closed phase, and we should expect that the prediction error (i.e., the difference between the value predicted by the left-hand of Eq. 6.6 and the actual speech sample at time n) is small and relatively constant during the closed phase and increases outside of it. Therefore, monitoring the prediction error can help identify a closed-phase region (Wong et al., 1979). This approach is known to fail when the speech deviates from the idealized assumptions under which this equation is derived (e.g.; when there's air leakage through the glottis). As an alternative, Plumpe et al. (1999) have proposed a more robust approach where a closed-phase region is identified by monitoring the modulation of the first formant, which is expected to remain stationary during the closed-phase region. The algorithm takes as input a first formant track, finds the point at which a functional measuring formant change attains its minimum, and grows a region around this point. In this work we have followed the outline of this approach, summarized by the algorithm in Fig. 6-1. Throughout this and the next section, we will assume that we are dealing with only voiced segments of speech, and that we know the instances of maximum excitation within a glottal cycle. In Section 6.2.3 we will introduce the details of the algorithm we have implemented to obtain a robust identification of such instances.

6.2.2 Discrete All-Pole Modeling and Inverse Filtering

Once a closed phase region has been identified for each glottal cycle, we can use the speech samples from this region to try to derive an all-pole model that best fits these data according to some criterion. One of the most commonly used error-minimization criteria is the linear prediction error, which can be shown to be equivalent to the quantity (Makhoul, 1975)

$$\mathcal{E}_{\mathcal{LP}} = \frac{1}{N}\sum_{k=1}^{N}\frac{P(\omega_k)}{\hat{P}(\omega_k)}, \tag{6.7}$$

Algorithm 6.1: Let $s_{seg}(n)$ be the speech segment between two consecutive excitation instants, N its length (i.e., the pitch period in samples), and F_s the sampling frequency.

- Formant Tracking

  1. Perform LPC analysis with the covariance method on $s_{seg}(n)$ with a one-sample shift, a window length $N_w = N/4$ and an order $p_{seg} = min\{16, N_w - 3\}$. (The order must be at least $N_w - 3$ to ensure that the Cholesky decomposition attempted while performing the LPC analysis is successful; the order is defaulted to 16 as long as this condition is satisfied).
  2. For each set of AR coefficients
     2a. Let $P(z)$ be the polynomial of coefficients with roots $\{z_k\}$. For each complex-conjugate pair, find the formant candidates $fn = \frac{F_s}{2\pi}\text{angle}(z_k)$ and their associated bandwidths $bn = -\log(|z_k|)/\pi$
     2b. Let $f1$ be the smallest fn for which $bn \geq 5fn$.
  3. Let $F1$ be the set of $f1$, the track of first formant candidates. Let $F1_{med}$ be a median filtered version of $F1$ with a 4-point window.
  4. For every value m of $F1$ exceeding an allowed threshold of 1050 Hz, let m^* be the closest time index not exceeding this threshold, and let $F1(m) = F1_{med}(m^*)$.

- Initial Identification of Stationary Region

  5. Given the formant track $F1(m)$ found in the previous step, define the formant modulation function $D(n_0) = \sum_{m=n_0}^{n_0+4} |F(m) - F(m-1)| \quad 1 \leq n_0 < N - N_w - 5$ (a cumulative first difference over a 5-point interval), and let $n_0^* = \text{argmin}_{n_0} D(n_0)$.
  6. Let $[N_i, N_f] = [n_0^* - 1, \cdots, n_0^* + 4]$ be an initial stationary region. Let μ_F and σ_F be the sample mean and variance of the first formant over the interval $[N_i, N_f]$

- Growing the Stationary Region to the Right

  7. While $|F1(N_f + 1) - \mu_F| < 2\sigma_F$
     7a. $N_f \leftarrow Nf + 1$.
     7b. Update μ_F and σ_F.

- Growing the Stationary Region to the Left

  8. While $|F1(N_i - 1) - \mu_F| < 2\sigma_F$, let $N_i \leftarrow N_i - 1$

Figure 6-1: Algorithm for identifying the closed-phase region of a glottal cycle.

where $P(\omega)$ is the spectrum of the given signal and $\hat{P}(\omega)$ is the parametric spectrum of the all-pole model

$$\hat{P}(\omega) = |H(\omega)|^2 = \frac{1}{|\sum_{k=0}^{p} a_k e^{-j\omega k}|^2}. \tag{6.8}$$

It is known that minimizing Eq. 6.8 corresponds to matching the autocorrelation functions of the respective spectra defined above for lag values up to p

$$\hat{R}(\tau) = R(\tau), \qquad 0 \leq \tau \leq p, \qquad \text{where} \quad \hat{R}(\tau) \overset{DTFT}{\longleftrightarrow} \hat{P}(\omega) \qquad R(\tau) \overset{DFT}{\longleftrightarrow} P(\omega). \tag{6.9}$$

El-Jaroudi & Makhoul (1991) have shown that sampling the spectrum $P(\omega)$ induces aliasing in the autocorrelation function $R(\tau)$, and that as a result, the error minimization criterion in Eq. 6.7 leads to estimates $\{a_k\}$ corresponding to a spectrum which does not match the desired spectral envelope or the sampled spectral points. To compensate for this, they have proposed the following error criterion based on the Itakura-Saito spectral distance

$$\mathcal{E}_{\mathcal{IS}} = \frac{1}{K} \sum_{k=1}^{K} \frac{P(\omega_k)}{\hat{P}(\omega_k)} - \ln \frac{P(\omega_k)}{\hat{P}(\omega_k)} - 1, \tag{6.10}$$

and shown an equivalency between minimizing this error criterion and the following matching conditions

$$\hat{R}(\tau) = R(\tau), \qquad 0 \leq \tau \leq p, \qquad \text{where} \quad \hat{R}(\tau) \overset{DFT}{\longleftrightarrow} \hat{P}(\omega) \qquad R(\tau) \overset{DFT}{\longleftrightarrow} P(\omega), \tag{6.11}$$

where $\hat{R}(\tau)$ is now the inverse Fourier transform of a sampled version of the all-pole spectrum. By taking into account the autocorrelation aliasing and matching two functions that have been aliased the same way, this approach produces better spectral matches for signals whose spectra contain strong harmonic (i.e., discrete) components, such as is the case with voiced speech. El-Jaroudi & Makhoul (1991) have presented an iterative algorithm to solve for the coefficients $\{a_i\}$. Fig. 6-2 summarizes one version of that algorithm where an adaptive step has been added to speed up convergence.

Once an all-pole model has been identified from the speech data corresponding to the closed-phase region, an estimate of the glottal pulse may be obtained by implementing the filtering in Eq. 6.4. There are, however, some practical issues that should be addressed before carrying out the convolution with the output of the DAP algorithm. It is often the

Algorithm 6.2: Let $R(\tau)$ (the autocorrelation of the signal spectrum) and $\hat{h}(-\tau)$ (the time-reversed impulse response of the discrete-frequency sampled all-poled model) be given by

$$R(\tau) = \frac{1}{K}\sum_{k=1}^{K} P(\omega_k)e^{j\omega_k\tau} \tag{6.12}$$

$$\hat{h}(-\tau) = \frac{1}{K}\sum_{k=1}^{K} \frac{e^{-j\omega_k\tau}}{A(\omega_k)} = \frac{1}{K}\sum_{k=1}^{K} \frac{e^{-j\omega_k\tau}}{\sum_{m=0}^{p} a_m e^{-j\omega_k m}} \tag{6.13}$$

Also, let $\mathbf{a} = [a_0, \cdots, a_p]^T$, $\mathbf{r} = [R(1), \cdots, R(p)]^T$, $\hat{\mathbf{h}} = [\hat{h}(0), \cdots, \hat{h}(-p)]^T$, and R_p be the Toeplitz matrix with $[R(0), \cdots, R(p-1)]^T$ in its first column. Select $0 < \epsilon << \alpha < 1$.

1. Pick K peaks from the spectrum $P(\omega)$ of the speech signal. Let ω_k and $P(\omega_k)$ be the locations and magnitudes of the peaks.
2. Compute $R(\tau)$ from Eq. 6.12 given $\{\omega_k\}$, $1 \le k \le K$
3. Using standard LP methods (e.g.; Levinson), find an initial estimate of $\mathbf{a}$ by solving the normal equations $R_p\mathbf{a}_0 = \mathbf{r}$.
4. Given $\mathbf{a}_m$, compute $A_m(\omega_k)$ for $1 \le k \le K$ and $\hat{h}_m(\tau)$ for $0 \le \tau \le p$ using Eq. 6.13.
5. Let $\mathbf{a}_m = (1-\alpha)\mathbf{a}_{m-1} + \alpha R_p^{-1}\hat{\mathbf{h}}_{m-1}$
6. Evaluate the error function $\mathcal{E}_{IS_m}$ (Eq. 6.10) given the current estimate $\mathbf{a}_m$
7. Let $\Delta\mathcal{E}_m = \frac{\mathcal{E}_{IS_m} - \mathcal{E}_{IS_{m-1}}}{\mathcal{E}_{IS_m}}$
   - 7a. If $\Delta\mathcal{E}_m > 0$: $\alpha \leftarrow \alpha/2$; go to step 5.
   - 7b. If $\Delta\mathcal{E}_m < -\epsilon$: $m \leftarrow m+1$; go to step 4.
   - 7c. If $-\epsilon \le \Delta\mathcal{E}_m \le 0$: go to step 8.
8. Normalize the coefficients in $\mathbf{a}$ such that $\frac{1}{K}\sum_{k=1}^{K}\frac{P(\omega_k)}{\hat{P}(\omega_k)} = 1$ holds.

Figure 6-2: Discrete All-Pole (DAP) estimation algorithm.

case, for instance, that the solution obtained by the DAP algorithm needs minor adjustments in order to ensure that the estimate of the coefficients corresponds to formants of the vocal tract. Since formants are defined from complex conjugate pole pairs, any real roots that appear in the estimated polynomial do not reflect the vocal tract formants that we want to filter out. Since the estimates in $A(z)$ are guaranteed to be real, the roots of the polynomial will occur either in complex-conjugate pairs, or on the real axis, corresponding to $f = 0$ Hz (on the positive real axis) or to the Nyquist frequency (on the negative real axis). Wong et al. (1979) suggest that a narrow bandwidth pole at the Nyquist frequency may correspond to a formant location nearby, and suggest leaving the pole as part of the inverse polynomial since it may be modeling the spectral shape. In our experience, leaving out negative real roots did indeed worsen the qualitative results of the inverse filtering, so we have opted to eliminate only real roots corresponding to 0 Hz. In addition, any complex conjugate pairs with center frequency below 50 Hz. are also eliminated from the estimated polynomial. After factoring the polynomial $A(z)$ and eliminating unwanted roots, a new polynomial $\hat{A}(z)$ of possibly lower order is constructed by retaining the remaining roots and normalizing the coefficients such that the square norm of the coefficients in $A(z)$ is preserved.

Another practical consideration pertains to which segment of speech $s(n)$ to use in the convolution defined by Eq. 6.4 once a suitable set of coefficients has been constructed. Our approach is to use the coefficients derived from the closed-phase portion to inverse filter the entire glottal cycle containing that closed phase. A glottal cycle is thus identified (in terms of the glottal landmarks we have at our disposal) as the portion of speech containing only one instant of maximum excitation N_e, beginning (from the left) one sample after the closed-phase region associated with the previous excitation, and extending (to the right) to the last sample of the closed-phase region associated with the current excitation.

One last modification is motivated by the difficulty in extracting the glottal pulse derivative from speech signals with high fundamental frequency. Even if a closed phase actually occurs and can be reliably identified for such signals, the length of the closed phase is too small and often leads to less reliable estimates of the vocal tract (when not to a numerically unstable problem). Since the shape of the vocal tract does not radically change over a few glottal cycles, using closed-phase information from adjacent periods could be a practical way to circumvent the lack of data since the extraneous cycles don't *smear out* the behavior

of the present glottal cycle significantly. In this work we have considered a neighborhood of three glottal cycles, centered at the *analysis* glottal cycle, when deriving the inverse filter coefficients for each period (cycles that occur at the boundaries of a voiced region are accordingly analyzed with a two-cycle neighborhood). The neighborhood is advanced by one cycle at a time, and then the coefficients re-estimated.

6.2.3 Identification of Instances of Maximum Excitations

We have assumed in the preceding sections that we are working with voiced regions of speech where the instances of maximum excitation have been identified. We next turn to specifying the algorithm used to determine the instants of excitation. Although particular details of the implementation differ, the architecture of the algorithm is that proposed by Smits & Yegnanarayana (1995) and Yegnanarayana & Smits (1995), where properties of the average group delay of minimum-phase signals are exploited in order to reliably locate the maximum excitations. Speech signals can be suitably modeled as the impulse response of a minimum-phase system. It is a characteristic of such systems that the average slope of the unwrapped phase response is zero, or, if the impulse response is shifted in time, proportional to the time shift. This property is retained when the signal is windowed, as long as the signal contains the instant of maximum excitation and the windowing does not introduce discontinuities: If the analysis window is centered around the excitation, the average slope of the phase response of the short-time signal should be close to zero; otherwise, it exhibits a slope proportional to the offset of the excitation with respect to the center of the analysis window. This observation suggests a very intuitive algorithm which, by using a time window small enough to capture primarily one impulse, tracks the short-time frequency response and examines the average slope of the unwrapped phase response (since the negative instantaneous slope is defined to be the delay of the system, the global slope of the unwrapped phase response can be thought of as the overall group delay for the short-time segment). Before committing the signal to this processing, however, the speech is whitened to emphasize the excitation and reduce the correlation between the remaining adjacent samples. Note that simpler algorithms for determining the excitations proceed by picking local peaks from this whitened signal (Plumpe et al., 1999). These algorithms depend on a threshold value, however, and can often miss minor or weaker excitations. The full details of the algorithm are specified below in Fig. 6-3.

Algorithm 6.3 Use a pitch tracking algorithm (e.g.;(Boersma, 1993)) to segment the speech into M disjoint voiced segments. Let $f_0(m)$ denote the mean fundamental frequency of the mth segment, and let $s(n)^m$ be the mth voiced segment. For $m = 1, \cdots, M$:

1. Calculate the 10th-order LP residual of $s(n)^m$ using the autocorrelation method, a Hanning analysis window of 25 msecs, and a frame rate of 100 frames/sec.
2. Find the short-time fast Fourier transform (STFT) of the residual from [1] using a Hanning window of length $1.5/f_0(m)$ secs., zero-padded to the next integer-power of 2. Advance the analysis window by one sample.
3. For each frame n of the STFT in [2]:
   - 3.1 Unwrap the phase.
   - 3.2 Using linear regression, find the best linear fit to the unwrapped phase. Let $\phi(n)$ denote the slope of the fit line found for the nth frame.
4. Smooth out the phase slope function $\phi(n)$ with a Hanning window of approximately 4 msecs, and remove the mean from the smoothed phase slope function.
5. Assign the zero-crossing instants from the zero-mean smoothed phase slope function to the instants of maximum excitation.

Figure 6-3: Algorithm for finding the instants of maximum excitation.

6.3 Parametrization of the Flow Derivative

It is often convenient to be able to represent an estimate of the glottal flow derivative (such as one obtained by inverse filtering) with a parametric model. A parametric representation offers the advantage of producing a smoother waveform than what is typically obtained by inverse filtering means and also allows modeling the waveform in terms of a few time-varying parameters (i.e.; varying with each glottal cycle). This approach also lends itself to applications of statistical modeling techniques that can view these parameters as feature vectors in a vectorial space for such applications as speaker identification or voice quality classification.

6.3.1 The Liljencrants-Fant Model

One of the most commonly used parametric representations of glottal sources is provided by the Liljencrants-Fant (LF) model of the glottal pulse derivative, a discrete version of

which is described by the following piecewise equations

$$g(n) = \begin{cases} -\frac{E_e}{\exp(\alpha N_e)\sin\left(\frac{\pi N_e}{T_p}\right)} \exp(\alpha n)\sin\left(\frac{\pi n}{T_p}\right) & 0 \le n \le N_e \\ -\frac{E_e}{1-\exp(-\beta(N_0-N_e))}\Big(\exp\big(-\beta(n-N_e)\big) - \exp\big(-\beta(N_0-N_e)\big)\Big) & N_e < n \le N_c \\ 0 & N_c < n < N_0. \end{cases} \tag{6.14}$$

The LF model fully specifies the shape of a glottal cycle in terms of the parameter vector $\zeta = [E_e, T_p, \alpha, \beta, N_e, N_c, N_0]^T$ with the intervals $(0, N_e)$, (N_e, N_c), and (N_c, N_0) defining the open, return, and closed phases respectively. N_e is the instant of maximum glottal excitation when $g(n)$ attains its minimum value $-E_e$. N_c is the instant of glottal closure, and N_0 is the discrete-time counterpart of the fundamental period T_0. The parameter α and T_p control the shape of $g(n)$ during the open phase, with T_p defining the zero-crossing point, whereas β controls the shape during the return phase.

6.3.2 Optimizing the LF Model Parameters

Fitting an LF model to a glottal flow derivative estimate $\hat{g}(n)$ (such as one obtained through the inverse filtering approach described earlier) consists of finding a value of ζ (as defined in the previous section) that minimizes some error criterion between the estimate and the model. Before formulating such an optimization criterion, however, it is important to observe that for Eq. 6.14 to be meaningful or defined, the parameters have to be constrained. We can constrain the parameters by invoking several sources of knowledge. Clearly, for the model to be properly defined, the following relation between the landmark points on the time axis has to be obeyed: $T_p < N_e < N_c < N_0$. The model can be further constrained if we incorporate external knowledge about the speech production mechanism and observe that the closed phase of a glottal cycle is generally much shorter than the open phase (a property that is quantified by the open quotient N_e/N_0). Furthermore, if the frequency of oscillation $\omega_p = \frac{1}{T_p}$ of the damped sinusoid in the open phase is not carefully constrained, yielding more than one oscillation in $(0, N_e)$, the model is not physically meaningful even if the expression $g(n)$ in Eq. 6.14 is well defined. Lastly, we can safely impose sensible bounds on the value of E_e (even if a bound doesn't exist in the mathematical model) by noting that the maximum negative amplitude is constrained by the dynamic range of the

sampled speech signals.

Some of these constraints induce dependencies between the parameters (e.g.; T_p is lower bounded by a fraction of N_e, itself a parameter) adding complexity and difficulty to the optimization. In this section, we shall make a series of assumptions to simplify the LF model introduced previously and be able to formulate the optimization as a minimization problem subject only to constant bounded constraints (i.e.; boxed constraints). We first note that some of the landmark points for each glottal cycle may be omitted from the optimization, as they can be reliably estimated through –and in fact, are available as the output of, some of the algorithms introduced earlier for inverse filtering the speech waveform. We will assume that we have a segmentation of a voiced region into adjacent glottal cycles containing the locations of the N_e values (the result of running the algorithm in Section 6.2.3), as well as the instants of glottal opening (we will assume that glottal opening immediately follows the last sample of the closed-phase region identified through the algorithm in Fig. 6-1). These assumptions allow us to extract one glottal cycle as the neighborhood around a value of N_e, and delimited by two excitations obtained as in in Section 6.2.3 The last change that we introduce consists of omitting the explicit zero-value closed phase in Eq. 6.14 and let $N_c = N_0$. This change does not simplify the problem further, as $g(n)$ is constant throughout this phase and need not be optimized. However, allowing a non-zero phase prior to the opening instant may be a more realistic model for many speakers and varieties of speech. If a closed or nearly closed phase in fact exists, the value of β can always model this accurately with a sharp return to negligible values.[1]

All the simplifications introduced thus far allow us to reconsider the optimization for each cycle separately with a reduced parameter vector $\theta = [E_e, T_p, \alpha, \beta]^T$ and a two-component piecewise function $g(n, \theta)$ (where the dependency on θ has now been made explicit). If we consider a square-error criterion, then finding an optimal value of θ consists of solving the following problem

$$\theta^* = \underset{\theta_l < \theta < \theta_u}{\operatorname{argmin}} \frac{1}{2} \sum_{n=1}^{N_0} \big(g(n, \theta) - \hat{g}(n)\big)^2, \tag{6.15}$$

where θ_l and θ_u are vectors containing lower and upper bounds respectively for each of the parameters in θ.

Some approaches to trying to fit an LF model to a glottal flow derivative estimate

[1] Note, also, that although the existence of a closed phase was a necessary assumption to derive the inverse-filtering results, no such assumption is required at this stage.

have considered general optimization techniques for unconstrained problems and ensured a feasible solution by monitoring of the estimates to ensure they remain within feasible ranges (Riegelsberger & Krishnamurthy, 1995). We will instead apply an iterative algorithm which generates strictly feasible estimates of the parameters (i.e.; each estimate lies in the feasible set defined by the lower and upper bounds of Eq. 6.15), and which can handle lower and/or upper bounds on some or all of the variables. The algorithm, discussed in greater detail in Appendix A, is an algorithm for nonlinear optimization subject to bounds. It is one of a class of trust-region algorithms which proceeds by replacing the objective function in Eq. 6.15 with a simpler model function resembling the true objective within a neighborhood (the trust region), and solving the simplified optimization. After each iteration, the trust region is enlarged if the step in parameter space proposed by the iteration leads to a successful decrease in the objective function; otherwise, the step is discarded and the region decreased to reflect the increased uncertainty.

Given the simplifications introduced in Section 6.3.2, optimizing the LF model consists of finding a vector θ of four parameters for a two-segment piecewise nonlinear function defined over the intervals $(0, N_e]$ and $(N_e, N_0]$. Since the optimization detailed in Appendix A makes use of first and second order information in the form of the gradient and Hessian of the objective function, it is worth examining these quantities in detail. The objective function is given by the square error criterion $f(\theta) = \frac{1}{2}\sum_{n=1}^{N_0}\big(g(n,\theta) - \hat{g}(n)\big)^2$ with gradient

$$\nabla_\theta f(\theta) = \sum_{n=1}^{N_0}\big(g(n,\theta) - \hat{g}(n)\big)\nabla_\theta g(n,\theta) = \sum_{n=1}^{N_0} e(n)\Psi(n,\theta), \tag{6.16}$$

where $e(n) \doteq g(n,\theta) - \hat{g}(n)$, $\Psi(n,\theta)$ is the Jacobian matrix containing the partial derivatives $\left[\frac{\partial g}{\partial E_e}, \frac{\partial g}{\partial T_p}, \frac{\partial g}{\partial \alpha}, \frac{\partial g}{\partial \beta}\right]^T$, and

$$\frac{\partial g}{\partial E_e} = \begin{cases} -\dfrac{\exp(\alpha n)\sin\left(\frac{\pi n}{T_p}\right)}{\exp(\alpha N_e)\sin\left(\frac{\pi N_e}{T_p}\right)} & 0 \le n \le N_e \\ \dfrac{\exp(-\beta(N_0 - N_e)) - \exp(-\beta(n - N_e))}{1 - \exp(-\beta(N_0 - N_e))} & N_e < n \le N_0 \end{cases} \tag{6.17}$$

$$\frac{\partial g}{\partial T_p} = \begin{cases} \frac{E_e \exp(\alpha n)}{\exp(\alpha N_e)} \left[\frac{\pi n \cos\left(\frac{\pi n}{T_p}\right) \sin\left(\frac{\pi N_e}{T_p}\right) - \pi N_e \sin\left(\frac{\pi n}{T_p}\right) \cos\left(\frac{\pi N_e}{T_p}\right)}{T_p^2 \sin^2\left(\frac{\pi N_e}{T_p}\right)} \right] & 0 \le n \le N_e \\ 0 & N_e < n \le N_0 \end{cases} \tag{6.18}$$

$$\frac{\partial g}{\partial \alpha} = \begin{cases} -E_e(n - N_e) \exp\big(\alpha(n - N_e)\big) \frac{\sin\left(\frac{\pi n}{T_p}\right)}{\sin\left(\frac{\pi N_e}{T_p}\right)} & 0 \le n \le N_e \\ 0 & N_e < n \le N_0 \end{cases} \tag{6.19}$$

$$\frac{\partial g}{\partial \beta} = \begin{cases} 0 & 0 \le n \le N_e \\ \frac{E_e(n-N_e) \exp\left(-\beta(n-N_e)\right)}{\left(1-\exp\left(-\beta(N_0-N_e)\right)\right)^2} - & \\ \qquad \frac{nE_e \exp\left(-\beta(n-2N_e+N_0)\right) - (N_0-N_e) \exp\left(-\beta(N_0-N_e)\right)}{\left(1-\exp\left(-\beta(N_0-N_e)\right)\right)^2} & N_e < n \le N_0. \end{cases} \tag{6.20}$$

The Hessian may then be approximated by applying finite differences to the gradient in Eq. 6.16. The feasible set for the optimizations in this work has been defined as the set of values constrained by the following bounds:

$$0 < E_e < \infty, \quad .51N_e < T_p < .95N_e, \quad 0 < \alpha < 1, \quad 0 < \beta < 1. \tag{6.21}$$

Finally, it is worth remarking that speech does not always obey the assumptions underlying the closed-phase inverse filtering algorithm. Although the techniques discussed earlier are bound to find a region of minimal first formant modulation, there are no guarantees that the glottis fully closes during this period, and the degree to which this assumption holds will affect the results obtained. Once the effects of the vocal tract have been estimated and inverse filtered, there may remain oscillations throughout the assumed closed phase which cannot be removed by linear filtering. Depending on the application, the obtained estimate of the glottal flow derivative may then be post-processed to de-emphasize these oscillations while preserving the behavior throughout the open phase. One such operation, for instance, could be a simple pitch-synchronous window multiplication with a window $w_p(n)$ defined

as

$$w_p(n) = \begin{cases} \frac{1}{1+\exp\left(a(n-0.1N_e)\right)} & 0 \le n \le Ne \\ \\ w_e \exp\left(\frac{\log(w_0/w_e)}{N_0-N_e}(n-N_e)\right) & Ne < n < N_0, \end{cases} \tag{6.22}$$

where $a = \frac{1}{0.1N_e}\log(\frac{1-w_0}{w_0})$. The first segment of this piecewise function is a sharply rising logistic function (it reaches its half point at 10% of the open phase duration) which is nearly constant throughout most of the open phase, whereas the second segment is a decaying exponential that tapers down the effect of any residual oscillations. The parameters $0 < w_0 < w_e \le 1$ ($w_p(N_e)$ and $w_p(N_0)$ respectively) regulate how much to taper these oscillations. We explore the results of the algorithms presented so far in the next section.

6.3.3 Evaluation of Glottal Flow Parametrization Algorithm

Figs. 6-4 through 6-6 show the results of applying the algorithms described in Secs. 6.2 and 6.3 to inverse filter the speech to obtain an estimate of the derivative of the glottal flow, and then a parametrization in terms of the LF model. The figures also illustrate some of the variability that the algorithms need to support in order to produce significant results. Fig. 6-4 shows the results obtained from the speech of a male speaker uttering a sustained vowel. This example typifies a best-case scenario. As one can see from panel (b), there is an identifiable nearly-closed phase during phonation (or at any rate, the air flow during the expected closed phase is smaller than the strength of the excitation). Also, the pitch of the speaker (approximately 110 Hz) allows a good interval of data between excitations to approximate the vocal tract, and since the vowel is sustained, the assumption that the vocal tract characteristics do not change drastically over the span of the three cycles used to derive the vocal tract coefficients holds particularly well.

A male speaker sustaining a vowel with nearly modal phonation represents a favorable point in a continuum of voicing for the purpose of the algorithms described so far. Figs. 6-5 and 6-6, in contrast, show the results of applying the algorithms to a section of continuous speech from two female speakers. Fig. 6-5 shows a fragment of speech, where each cycle resembles a damped sinusoidal. In this case, we expect to find some glottal closure since, by inspection of the speech waveform, we see a primarily decaying signal where the interaction with the excitation may be minimal. By comparing panels (a) and (b), we can see that

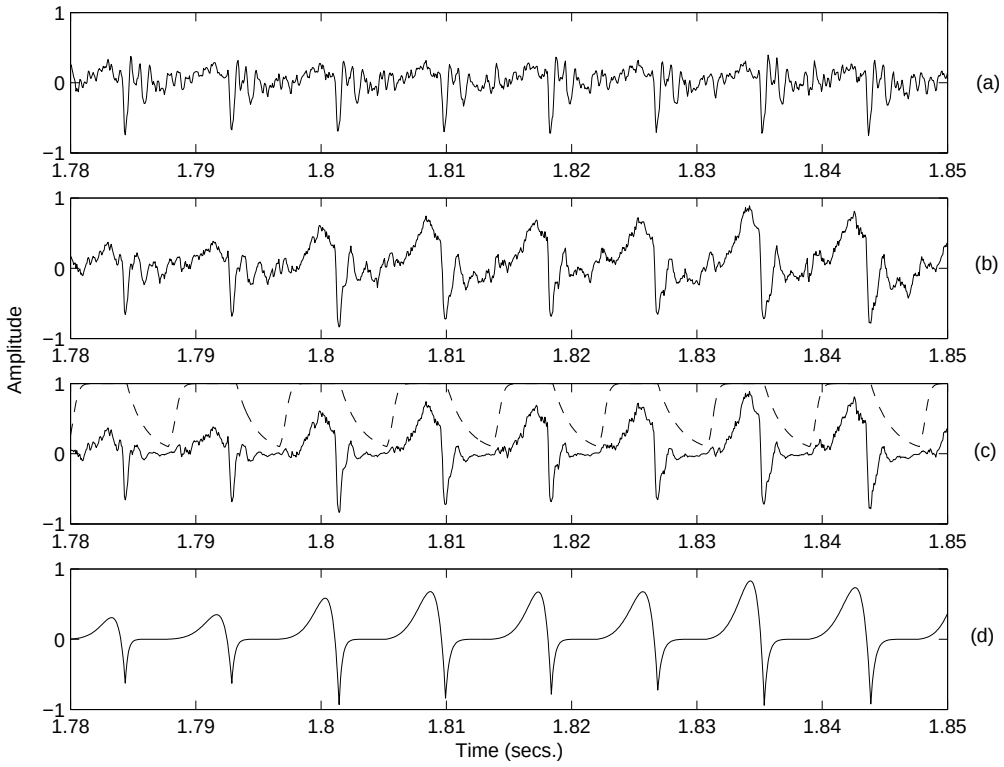

Figure 6-4: (a) Speech signal from male speaker sampled at 16 kHz, $f_0 \approx 110$ Hz; (b) Estimate of the glottal flow derivative by inverse filtering; (c) Post-multiplied flow derivative (solid), and window (dashed) with $w_0 = 0.1$ and $w_e = 1$; (d) LF model fit.

inverse filtering is able to emphasize the excitation (particularly during the region (T_p, N_e) where the flow derivative goes negative), but significant *ringing* remains throughout the assumed closed phase. Noticed that the average fundamental frequency is about $250Hz$, which imposes constraints on the length of the closed-phase.

Finally, Fig. 6-6 shows a "harder" case where the speech waveform closely resembles a sinusoidal with minimal or no damping. The effect of the vocal tract is discernible on the modulations superimposed on this sinusoidal. In spite of the absence of a strong peak associated with the excitation, the algorithm described in Section 6.2.3 is able to reasonably predict the excitation instant. However, the shape of the waveform and the lack of damping during the "closed phase" suggests that the uncoupling assumed during the formulation of the inverse filtering algorithm may hold less well, an observation which seems to be confirmed by the remaining oscillations in the inverse filtered signal in panel (b). Panel (c) shows the effect of applying the post-processing windowing to the inverse filtered signal to reduce the remaining oscillations throughout the closed phase. Although this transformation is

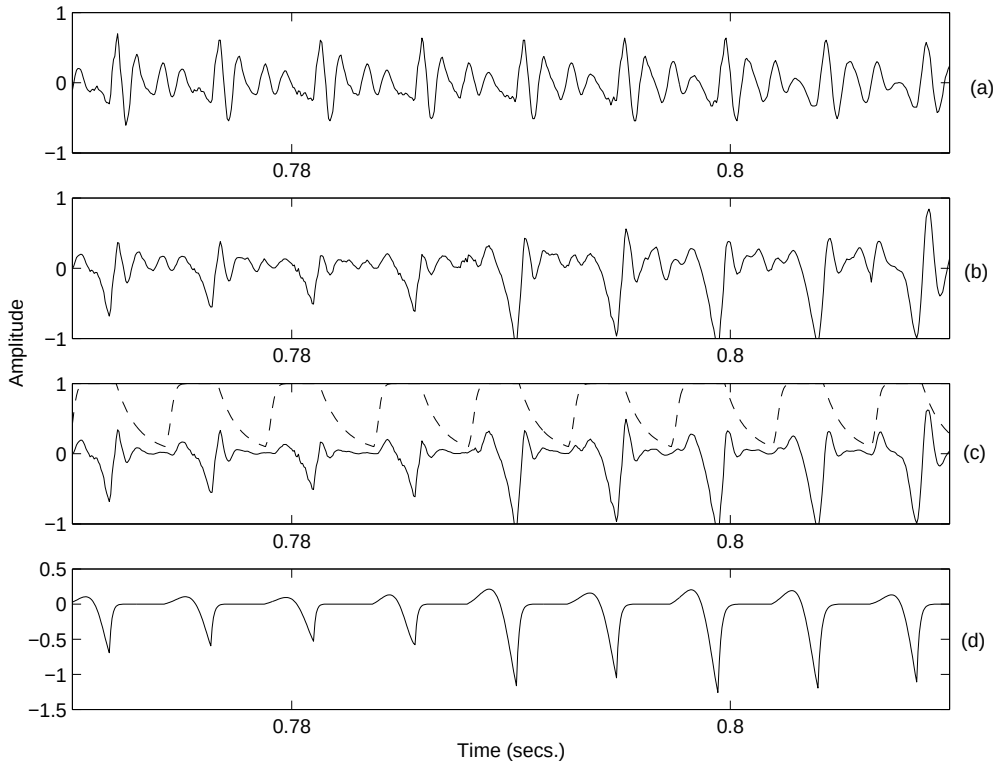

Figure 6-5: Speech signal from female speaker sampled at 16 kHz, $f_0 \approx 250$ Hz; (b) Estimate of the glottal flow derivative by inverse filtering; (c) Post-multiplied flow derivative (solid), and window (dashed) with $w_0 = 0.1$ and $w_e = 1$; (d) LF model fit.

somewhat artificial, the resulting waveform approximates more closely a modal-phonation glottal flow derivative. Panels (d) in Figs. 6-4 through 6-6 show the parametric LF model fits found for each of the cases. Since the LF model assumes modal phonation, and therefore a clearly defined closed phase, we find the behavior outside the open phase interval not to be greatly affected by the shape of the inverse-filtered signal once the return phase has been taken into account. This is more explicitly shown in Fig. 6-6, where LF fits have been obtained for the inverse-filtered signals with and without post-windowing. The results are plotted in panel (d) in solid and dashed lines, which coincide throughout most of the speech segment. Throughout the open phase, however, the LF fit will be influenced by the shape of the data, as can be seen by comparing Figs. 6-4 (d) and 6-5 (d).

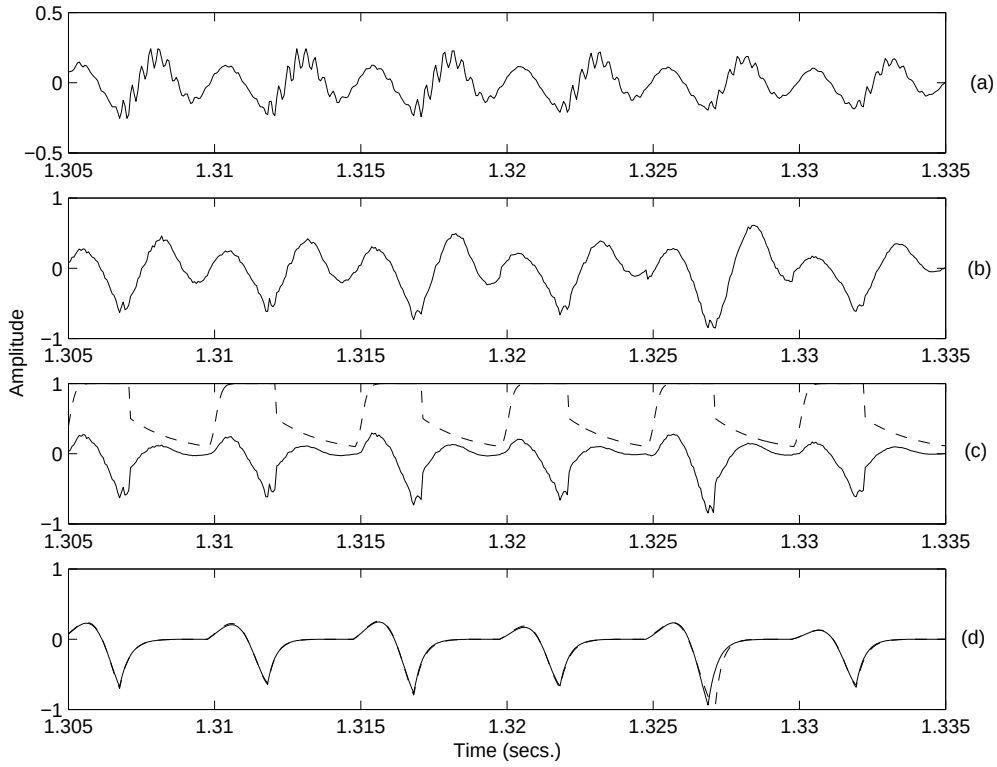

Figure 6-6: Speech signal from female speaker sampled at 16 kHz, $f_0 \approx 200$ Hz; (b) Estimate of the glottal flow derivative by inverse filtering; (c) Post-multiplied flow derivative (solid), and window (dashed) with $w_0 = 0.1$ and $w_e = 0.5$; (d) LF model fits of the signal in (b) –solid– and of the signal in (c) –dashed.

6.4 Discrimination Based on Features Derived from the LF Parametrization

The algorithms and optimization procedures just described may be used to extract information about the landmarks defining each phonatory cycle (i.e., instant of maximum excitation, instant of glottis closing, etc.), to inverse-filter the speech waveform defined over one cycle, and then to fit a parametrized model to the result. The latter procedure manages to summarize the glottal waveform over one cycle in terms of a few sparse parameters which can be used, either directly or as inputs to a feature extraction procedure, to model the glottal source in terms of a vectorial quantity and to allow automatic learning techniques to learn regions associated with different categories. We next turn to applying this procedure to modeling affect.

Let us assume that $\hat{g}_k(n)$ represents the estimated inverse-filtered glottal volume velocity

flow waveform over the *kth* glottal cycle, and let $g_k(n;\theta_k^*)$ be the optimized LF fit, with

$$\theta_k^* = [E_{e_k}^*, T_{p_k}^*, \alpha_k^*, \beta_k^*]^T \tag{6.23}$$

representing the optimized parameters for the *kth* cycle. Let the error signal between the signals be decomposed as in

$$\epsilon_k^{LF} = \sum_{n=1}^{N_{0_k}} (g(n;\theta^*) - \hat{g}(n))^2 \tag{6.24}$$

$$= \sum_{n=1}^{N_{e_k}} (g(n;\theta^*) - \hat{g}(n))^2 + \sum_{n=N_{e_k}}^{N_{0_k}} (g(n;\theta) - \hat{g}(n))^2 \tag{6.25}$$

$$= \epsilon_{o_k}^{LF} + \epsilon_{c_k}^{LF} \tag{6.26}$$

where N_e^k and N_0^k are the instant of maximum excitation and the fundamental period of the *kth* cycle and where, without losing any generality, we have assumed that the cycle begins at sample 1. The error in Eq. 6.26, therefore, consists of the contributions over the open ($\epsilon_{o_k}^{LF}$) and closed phases respectively ($\epsilon_{c_k}^{LF}$).

Since the parameter T_p (the point on the open phase at which the glottal pulse reaches its maximum, or alternatively, when the glottal volume velocity reaches zero) can take a wide range depending on the value of the fundamental period or the value of the open phase, it seems reasonable to introduce a normalization that expresses it as a fraction of one of these constants. We have opted to define

$$\gamma_k^* = \frac{T_{p_k}^*}{N_{e_k}}, \tag{6.27}$$

a measure which reflects the asymmetry of the pulse during the open phase.[2] We also define the open quotient OQ_k as

$$OQ_k = \frac{N_{e_k}}{N_{0_k}}, \tag{6.28}$$

the proportion of the open phase duration to the fundamental period. We now define the

[2]This parameter is related to what some authors (Childers & Lee, 1991) have referred to as the speed quotient (SQ), the ratio of rising time (to the pulse maximum) to the falling time (to zero flow). However, the literature is not consistent with this terminology; Cummings & Clements (1995b), for instance, refer to the speed quotient as the ratio of open duration to pitch period. To avoid confusion, we will just label this parameter γ, and point out the link to similarly proposed parameters in the literature.

following set of four statistics on an arbitrary sequence x_k:

$$S(\{x\}) = \Big[prctile_{25}(x_k), \quad median(x_k), \quad prctile_{75}(x_k), \quad iqr\Big(\frac{x_{k+1} - x_k}{x_{k+1}}\Big) \Big]^T. \tag{6.29}$$

The first three components of the vector $S(x)$ are the three quartile points and give a sparse representation of the distribution of the sequence; the last element is the range of the normalized first difference and gives a measure of the average range of the sample-to-sample variation in the sequence. We next define the feature vector as

$$\begin{aligned} FS^{LF} = \Big[& S(E^*_{e_k})^T, \quad S(\gamma^*_k)^T, \quad S(\alpha^*_k)^T, \quad S(\beta^*_k)^T, \cdots \\ & S(OQ_k)^T, \quad S(\epsilon^{LF}_{o_k})^T, \quad S(\epsilon^{LF}_{c_k}) \Big]^T. \end{aligned} \tag{6.30}$$

Studies of glottal excitations have been able to show that the shape of the glottal pulse can change in systematic ways under different affective and stressed conditions (Cummings & Clements, 1995a). Studies have also been able to show a link between different phonation types, manipulated through synthesis, and perceived affect (Gobl & Ní Chasaide, 2000). Some of these phonation types may not be adequately modeled by the LF parametrization. However, in such cases, we might expect a larger deviation between the inverse-filtered waveform and the parametrized model, particularly for styles that exhibit a significant flow throughout what the LF models as basically a near-zero flow phase. Following these observations, the 28 components of the feature vector FS^{LF} have been defined to summarize statistics about the LF parametrization of the glottal waveform over various cycles of speech, providing a concise description of the distribution of the glottal shape over the speech segment. The error parameters have been included to provide a description of the accuracy of the LF fit, and to quantify to what extent the glottal cycles deviate from the idealized LF model. In particular, ϵ^{LF}_o has been included to target phonation types (e.g., breathy or whispery) during which there is some or considerable flow throughout the open phase (Laver, 1994; Gordon & Ladefoged, 2001; Hanson et al., 2001). We next evaluate the performance of this feature vector in discriminating between affect categories.

6.4.1 Evaluation of Proposed LF Features and Discussion

Tables 6.1 and 6.2 summarize the results of building subject-dependent Bayesian classifiers (with pooled covariance matrices) of emotion pairs based on the features defined by Eq. 6.30. The clear overall trend is that the features provide significant discrimination between most emotion pairs for all subjects. For subjects 3 and 5 they provide significant discrimination between all combinations of emotions. In spite of the difficulty one may encounter in reliably estimating the glottal volume velocity flow from the waveform, especially under affectively stylized speech, the stylization using a parametric model, together with the statistics of the optimized parameters that have been used to define the feature vector, manage to capture enough style-dependent information as to prove useful in categorical prediction of affect.

	Ang	Hap	Ntr	Sad
Afr	**0.13**†	**0.21**†	**0.27**†	**0.29**†
Ang		**0.16**†	**0.30**†	**0.18**†
Hap			0.37‡	**0.16**†
Ntr				**0.23**†

Subject 1

	Ang	Hap	Ntr	Sad
Afr	**0.17**†	**0.23**†	**0.20**†	**0.17**†
Ang		**0.25**†	**0.30**†	**0.30**†
Hap			0.41	**0.33**‡
Ntr				0.47

Subject 2

	Ang	Hap	Ntr	Sad
Afr	**0.04**†	**0.20**†	**0.11**†	**0.32**†
Ang		**0.15**†	**0.30**†	**0.10**†
Hap			**0.26**†	**0.27**†
Ntr				**0.20**†

Subject 3

	Ang	Hap	Ntr	Sad
Afr	**0.08**†	**0.06**†	**0.06**†	**0.07**†
Ang		**0.13**†	**0.30**†	**0.17**†
Hap			**0.18**†	**0.16**†
Ntr				0.44

Subject 4

	Ang	Hap	Ntr	Sad
Afr	**0.11**†	**0.21**†	**0.05**†	**0.14**†
Ang		**0.26**†	**0.30**†	**0.21**†
Hap			**0.20**†	**0.26**†
Ntr				**0.23**

Subject 5

	Ang	Hap	Ntr	Sad
Afr	**0.21**†	0.41	**0.25**†	**0.30**†
Ang		**0.23**†	**0.29**†	0.36‡
Hap			**0.33**†	**0.32**†
Ntr				0.46

Subject 6

Table 6.1: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Bayesian classifiers (with pooled covariance matrix) trained on emotions pairs using feature set derived from the LF parametrization of the glottal volume velocity waveform for subjects 7 through 11 ($\dagger$and $\ddagger$denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

The results also show some other notable trends at the $p < 0.01$ level. The category *Afraid*, for instance, is consistently separated from the *Anger* category. The separation

between these two distinctly valenced categories is observable for all subjects with the exception of 8. For subject 4, the discrimination between the *Afraid* category and all others is particularly striking. The *Happy* category is also significantly discriminated from the remaining in the set for subjects 3, 4, 5, 7, 9 and 11 whereas the *Neutral* category is significantly discriminated from the rest for 4 of 11 subjects.

	Ang	Hap	Ntr	Sad
Afr	**0.06**†	**0.27**†	**0.16**†	**0.18**†
Ang		**0.09**†	**0.22**†	**0.09**†
Hap			**0.15**†	**0.23**†
Ntr				**0.28**†

Subject 7

	Ang	Hap	Ntr	Sad
Afr	0.35‡	**0.33**†	0.41	0.41
Ang		**0.30**†	0.41	0.35‡
Hap			**0.32**†	0.38‡
Ntr				0.49

Subject 8

	Ang	Hap	Ntr	Sad
Afr	**0.14**†	**0.32**†	**0.23**†	0.32‡
Ang		**0.17**†	0.42	0.35‡
Hap			**0.25**†	**0.29**†
Ntr				0.44

Subject 9

	Ang	Hap	Ntr	Sad
Afr	**0.06**†	**0.14**†	**0.13**†	**0.21**†
Ang		**0.25**†	**0.21**†	**0.24**†
Hap			**0.27**†	0.43
Ntr				0.43

Subject 10

	Ang	Hap	Ntr	Sad
Afr	**0.07**†	**0.30**†	**0.16**†	0.40
Ang		**0.07**†	**0.26**†	**0.10**†
Hap			**0.20**†	**0.24**†
Ntr				**0.22**†

Subject 11

Table 6.2: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Bayesian classifiers (with pooled covariance matrix) trained on emotions pairs using feature set derived from the LF parametrization of the glottal volume velocity waveform for subjects 7 through 11 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

Table 6.3 shows the generalization errors when the classifiers are trained to discriminate between each affect category and the rest. The training data are sampled to obtain equal priors, so chance-level discrimination remains at 50% for this task. The categories *Afraid* and *Angry* are consistently predicted at a significance level for most subjects, followed by the *Happy* category. The labels *Neutral* and *Sad* are still modeled less accurately, a pattern that has been observed for the loudness and intonational features as well. Table 6.3, nonetheless, confirms that the LF parametrization offers a good representation for obtaining features that

can aid in the automatic disambiguation of affective categories.

Subject	Afr	Ang	Hap	Ntr	Sad
1	**0.27**†	**0.30**†	0.38‡	0.43	**0.34**†
2	**0.28**†	**0.33**†	0.45	0.43	0.40
3	**0.29**†	**0.14**†	0.38‡	**0.33**†	0.45
4	**0.10**†	**0.32**†	**0.17**†	0.41	**0.28**†
5	**0.16**†	0.38‡	0.41	**0.27**†	**0.28**†
6	**0.36**†	0.40	**0.32**†	0.45	0.46
7	**0.23**†	**0.18**†	**0.27**†	**0.36**†	**0.27**†
8	0.50	0.35‡	**0.30**†	0.42	0.51
9	**0.31**†	**0.26**†	**0.19**†	0.42	0.46
10	**0.16**†	**0.19**†	0.44	0.35†	0.50
11	**0.27**†	**0.15**†	**0.29**†	**0.27**†	0.38‡

Table 6.3: Generalization error (estimated by leave-one-out cross validation) for subject-dependent Bayesian classifiers (with pooled covariance matrix) trained to discriminate between each emotion category and the remaining ones pooled together, and using only feature set derived from the LF parametrization of the glottal volume velocity waveform (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

The results presented here confirm a point repeatedly raised in the literature by several authors stressing the contribution of voice quality features to the communication of affect (Scherer et al., 1984; Ladd et al., 1985; Gobl & Ní Chasaide, 2003). We further explore this contribution in the next sections.

6.5 Other Source Features

The previous section undertook modeling voice quality by extracting the glottal volume velocity flow through a pitch-synchronous inverse-filtering of the speech waveform, and by directly parameterizing each phonatory cycle in terms of the Liljencrants-Fant model for the derivative of the glottal pulse. As a result of running Algorithm 6.3, we gained access to a parsing of the voiced portions of speech in terms of the instances of maximum excitation, from which we directly gain a local estimate of the fundamental frequency of voicing. We have already analyzed *intonational* fluctuations in the fundamental frequency in Chapter 5, that is, longer-term changes of F0 associated with larger-scale phenomena affected by various factors like the type of speech act, syllable accent placement, etc. There is another source of variation of the fundamental frequency, this one correlating perceptually with voice quality,

that is of interest here: the short-term fluctuations, typically between consecutive periods, of the fundamental frequency of voicing known as *jitter*. Jitter is a fluctuation observable even when a speaker is attempting to sustain a vowel, and which may have underlying physiological underpinnings (Klatt & Klatt, 1990). The level of jitter in voiced speech may be an indication of psychologically conditioned, or stressed speech as well as a sign, when exceeding certain levels, of a speech pathology. Analogous to this random period-to-period durational variation is a quantity known as *shimmer*, the random short-term changes in the glottal pulse amplitude. Shimmer measures have also been considered in the context of voice quality assessment as a parameter for describing the kinds of irregularities associated with vocal pathology (Michaelis et al., 1998). Voice pathologies might represent one end of the continuum which these parameters can describe; it is conjectured that between normal and pathological ranges, these parameters may also take on different sets of values corresponding to other vocal qualifications.

Although, impressionistically, jitter and shimmer are relatively straightforward to qualify, in discussing the literature, Michaelis et al. (1998) point out that many definitions and ways of quantifying these parameters have evolved. Following Kasuya et al. (1993), they make use of the following two auxiliary measures, known as the perturbation factor (PF) and perturbation quotient (PQ_K), defined for an arbitrary sequence x_n of length N as

$$PF(\{x_n\}) = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{x_{n+1}-x_n}{x_{n+1}}\right| \tag{6.31}$$

$$PQ_K(\{x_n\}) = \frac{1}{N-K}\sum_{n=(K-1)/2}^{N-((K-1)/2)-1}\left|\frac{x_n-\mu_K(x_n)}{\mu_K(x_n)}\right| \tag{6.32}$$

$$\mu_K(x_n) = \frac{1}{K}\sum_{k=-(K-1)/2}^{(K-1)/2} x(n+k). \tag{6.33}$$

The perturbation factor is just the average normalized absolute first-order difference, whereas the perturbation quotient involves subtracting a local running average $\mu_K(x_n)$ (obtained with a K-point rectangular window), and normalizing by it before taking the average.

Jitter may be quantified by directly applying any of these two measures to the time series $\{T_k\}$ consisting of intra-peak intervals (the distance between adjacent instances of maximum excitations returned by Algorithm 6.3), normalized by the sampling frequency. To quantify shimmer, we can do the same with the sequence of glottal excitation amplitudes. However,

because the effect of the vocal tract over the span of a few periods can be assumed to be more slowly varying than the glottal fluctuations we are trying to describe, it is admissible (and possibly more robust) to work directly with the speech waveform over the cycles of interest since the vocal tract influence remains approximately constant. In this work, we will take E_k to be the sequence of period-to-period energy values (sum of squared samples) of the speech signal, where the boundaries of each cycle are determined by the values in $\{T_k\}$, and use that sequence as inputs to Eqs. 6.31 and 6.32 to model shimmer values.

Another proposed way to analyze voice quality involves an assessment of the degree to which the periodic glottal waveform is affected by a noise component. Different parameters that try to capture the proportion of harmonicity and noise in the excitation signal include the Harmonics-to-Noise Ratio (HNR) (Boersma, 1993), the Cepstrum-based Harmonics-to-Noise Ratio (CHNR) (de Krom, 1993), and the Glottal-to-Noise Excitation Ratio (GNE) (Michaelis et al., 1997). We consider next an implementation of the latter measure, as it has been shown in experiments with synthetic data to be less linearly dependent on measures of jitter and shimmer, which we are already quantifying separately.

The algorithm is summarized in Fig. 6-7. The general motivation behind this approach to quantifying the excitation noise level lies in the observation that a pure noise-free glottal pulse simultaneously excites different frequency bands of the vocal tract. In such case, the bandpass filtered signals corresponding to different bandwidths, though oscillating at different frequencies, should retain similar overall shape, and their envelopes should be highly correlated. If, on the other hand, the vocal tract is excited by a turbulent noise-like signal, each frequency channel should be excited by a narrowband noise, and the output signals (and their envelopes) should be uncorrelated. After whitening the signal to emphasize the excitations (steps 1 and 2), the algorithm then proceeds by defining a series of frequency bands (step 3), finding the Hilbert envelopes of the signal on different frequency channels (steps 3.1 and 3.2) and searching for the maximum of the pairwise correlations between these envelopes to define the glottal-to-noise excitation ratio (steps 3.3 and 3.4). In our implementation, Algorithm 6.4 is applied to 40-msec. windows of voiced speech in steps of 10 msecs. to build a sequence of GNE values $\{GNE_k\}$.

The voice quality parameters we have discussed so far are measured from some time-domain features of the speech signal (after possibly some spectral manipulations). We next consider a parameter that is measured directly from a spectral representation. Spectral

Algorithm 6.4 Let $s(n)$ be a speech waveform.

1. Downsample $s(n)$ to 8kHz.

2. Estimate a $10th$ order LPC $a(n)$ model from $s(n)$ (e.g., using the autocorrelation method with a $30msec$ Hanning window in steps of $10msecs.$) and inverse-filter the entire signal to whiten the signal and enhance the excitations: $u(n) = s(n) * a(n)$.

3. Divide $u(n)$ into (possibly overlapping) frames containing several pitch periods. Apply steps 3.1. through 3.4. to each short-time voiced frame $u_f(n)$ (voicing probability can be obtained from, e.g., the pitch extraction algorithm described in 5.I):

   3.1. Find the spectrum $U(k)$.

   3.2. Define the frequency bands $f_1 = (0 - 2kHz)$, $f_2 = (1kHz - 3kHz)$ and $f_3 = (2kHz - 4kHz)$, and find the Hilbert envelopes of the bandpass signals corresponding to these frequency ranges:

      3.2.1. Finding Hilbert envelopes: Multiply the spectrum $U(k)$ by a Hanning window $H^{(f_n)}(k)$ defined over the positive frequencies corresponding to the band of interest, and zero elsewhere on the spectrum. Take the inverse fast Fourier transform of the (asymmetrically) windowed spectrum to obtain a complex signal. The Hilbert envelope is the absolute value of the resulting signal: $u_h^{(f_n)}(n) = |\mathcal{IFFT}\{U(k)H^{(f_n)}(k)\}|$

   3.3. Find the normalized cross-correlations $R_{jk}(\tau)$ and their maxima R_{jk}^{max} for pairs of Hilbert envelopes $u_h^{(f_j)}, u_h^{(f_k)}; j \neq k; j,k = 1,2,3$ in the lag range $-0.3ms \leq \tau \leq 0.3ms$.

   3.4. Define the Glottal to Noise Excitation Ratio as the maximum of all maxima of the cross-correlation functions:

$$GNE = \max R_{jk}^{max} \tag{6.34}$$

Figure 6-7: Algorithm for finding the Glottal-to-Noise Excitation (GNE) ratio.

variations associated with voice quality have been studied and found to be reflected in the general distribution of the spectrum at low versus high frequencies (spectral tilt) and in the relative height of the first two harmonics (Hanson, 1997). Alku et al. (1997) have proposed a single parameter for quantifying the glottal flow based on the decay of the glottal pulse spectrum at low frequencies, and shown that it correlates with different phonation types. The parameter, known as Parabolic Spectral Parameter (PSP), is based on describing the decay of the glottal pulse spectrum over low frequencies with a second-order polynomial. We apply this algorithm directly to the glottal volume velocity waveform since the differentiation process only introduces a spectral zero, and hence differences in phonation type captured by the spectrum of the glottal pulse will also be apparent in the spectrum of its derivative.

The algorithm is summarized in Fig. 6-8. It is a pitch-synchronous algorithm where a parameter is estimated for each phonatory cycle. After computing the spectrum of the glottal flow $G(k)$ (steps 1 and 2), the algorithm defines an initial region $[0, N]$ over low frequencies, and finds a second order polynomial $G(k) = \hat{a}k^2 + \hat{b}$ match that minimizes the sum-square error. The optimal parameters $\hat{a}$ and $\hat{b}$ can be solved for in close-form solution (Eqs. 6.35 and 6.36). This initial region is adaptively extended to the right until the normalized error (Eq. 6.37) exceeds a certain threshold, at which point the spectrum is assumed not to be suitably described by a parabola. The final value of $\hat{a}_{opt}$, governing the decay of the glottal flow spectrum over this low-frequency region, is saved and the above procedure repeated for a DC flow signal of equal period. The PSP is then defined as the ratio of the a parameters (Eq. 6.38) (the normalization is intended to factor out the scaling effect of the fundamental period on the shape of the spectrum). This algorithm is applied, as mentioned, to every phonatory cycle defined by the landmarks found with the algorithms described in previous sections. Let $\{PSP_k\}$ then represent the sequence of PSP values obtained for a portion of speech over its voiced sections.

Next we consider the performance of the parameters presented in this section on the task of discriminating affect categories. Jitter and shimmer features are obtained from the sequences of peak-to-peak intervals ($\{T_k\}$) and peak-to-peak energy ($\{E_k\}$) by evaluating the perturbation factor and the perturbation quotient, the latter with a window j spanning 3, 9, 15, and 21 periods. From the set of perturbation quotients, the minimum and maximum

Algorithm 6.5 Let $g_c(n)$ represent one cycle of the volume velocity waveform containing N_0 samples.

1. Adjust $g_c(n)$ such that its first and last samples are equal to ensure periodicity, and subtract the minimum value from the resulting signal. Normalize the signal to unit energy. Let $g_p(n)$ denote the signal thus adjusted and N_0 its length.

2. Compute the 1024-point FFT of $g_p(n)$. Let $G(k)$ be the square magnitude of the Fourier transform.

3. Parabolic Fit: Let $N \leftarrow 3$. Compute

   3.1 Compute the parabolic parameters and the normalized fitting error:

$$\hat{a} \leftarrow \frac{N \sum_{k=0}^{N-1} G(k)k^2 - \left(\sum_{k=0}^{N-1} G(k)\right)\left(\sum_{k=0}^{N-1} k^2\right)}{N \sum_{k=0}^{N-1} k^4 - \left(\sum_{k=0}^{N-1} k^2\right)^2} \tag{6.35}$$

$$\hat{b} \leftarrow \frac{1}{N} \sum_{k=0}^{N-1} \left(G(k) - \hat{a}k^2\right) \tag{6.36}$$

$$E_n \leftarrow \frac{\sum_{k=0}^{N-1} \left(G(k) - \hat{a}k^2 - \hat{b}\right)^2}{\sum_{k=0}^{N-1} G(k)^2}. \tag{6.37}$$

   3.2 If $E_n < E_{thresh}$ ($E_thresh \cong 0.015$): $N \leftarrow N + 1$; go to step 3.1.

   3.3 Let $a_{opt} = \hat{a}$ be the optimal parabolic parameter after convergence describing the decay of the glottal spectrum.

4. Repeat steps 1 through 3 for a DC flow signal of length N_0. Let a_{DC} be the optimal parameter found for the decay of the DC spectrum. Define the PSP as

$$PSP = \frac{a_{opt}}{a_{DC}} \tag{6.38}$$

Figure 6-8: Algorithm for finding the Parabolic Spectral Parameter (PSP).

value are retained. Eqs. 6.39-6.43 describe these features:

$$Jitt_{PF} = PF(\{T_k\}) \tag{6.39}$$

$$Jitt_{PQ}^{min} = \min_{j=3,9,15,21}\big(PQ_j(\{T_k\})\big) \tag{6.40}$$

$$Jitt_{PQ}^{max} = \max_{j=3,9,15,21}\big(PQ_j(\{T_k\})\big) \tag{6.41}$$

$$Shimm_{PF} = PF(\{E_k\}) \tag{6.42}$$

$$Shimm_{PQ}^{min} = \min_{j=3,9,15,21}\big(PQ_j(\{E_k\})\big) \tag{6.43}$$

$$Shimm_{PQ}^{max} = \min_{j=3,9,15,21}\big(PQ_j(\{E_k\})\big). \tag{6.44}$$

The GNE and PSP features are defined by applying the same set of statistics defined in Eq. 6.29 to the $\{GNE_k\}$ and $\{PSP_k\}$ sequences to parsimoniously describe the distribution of these parameters for the speech fragment under analysis. Altogether, the following 14-D feature vector is obtained:

$$\begin{aligned} FS^{VS} = \big[&Jitt_{PF}, \quad Jitt_{PQ}^{min}, \quad Jitt_{PQ}^{max}, \cdots \\ &Shimm_{PF}, \quad Shimm_{PQ}^{min}, \quad Shimm_{PQ}^{max} \cdots \\ &S(\{GNE_k\})^T, \quad S(\{PSP_k\})^T\big]^T. \end{aligned} \tag{6.45}$$

The result of submitting this feature vector to discriminant analysis is shown in Tables 6.4 and 6.5. For the most part, these results confirm the results obtained with parameters derived from the LF parametrization: significant discrimination is obtained for most subjects, with subject 8 exceptionally attaining significant recognition rates for only two pairwise comparisons; good discrimination is shown for subject 4 between the *Afraid* categories and the rest; and the *Happy* category is significantly discriminated from the rest of the set for subjects 3, 4, 5, 7, 9, and 11. Additionally, this discrimination is now also observed for subject 10, for whom significant results are obtained with every pair of discriminants (particularly, the distinction between the *Sad* category and the rest is now significant for this subject).

Table 6.6 shows the generalization error of the same classifier structure trained to discriminate between each affect category and the remaining categories in the set. The data have again be sampled to reflect equal priors during training (chance-level performance re-

	Ang	Hap	Ntr	Sad
Afr	**0.19**†	**0.29**†	0.38‡	0.35‡
Ang		**0.18**†	**0.28**†	**0.20**†
Hap			0.48	**0.16**†
Ntr				**0.26**†

Subject 1

	Ang	Hap	Ntr	Sad
Afr	**0.25**†	**0.12**†	**0.23**†	**0.24**†
Ang		**0.26**†	**0.25**†	**0.16**†
Hap			0.41	**0.25**†
Ntr				**0.31**†

Subject 2

	Ang	Hap	Ntr	Sad
Afr	**0.12**†	**0.29**†	**0.16**†	0.44
Ang		**0.13**†	**0.29**†	**0.15**†
Hap			**0.25**†	**0.34**†
Ntr				**0.24**†

Subject 3

	Ang	Hap	Ntr	Sad
Afr	**0.10**†	**0.17**†	**0.12**†	**0.09**†
Ang		**0.19**†	**0.27**†	**0.19**†
Hap			**0.32**†	**0.15**†
Ntr				**0.32**†

Subject 4

	Ang	Hap	Ntr	Sad
Afr	**0.15**†	**0.24**†	**0.11**†	**0.14**†
Ang		**0.24**†	**0.33**†	**0.28**†
Hap			**0.23**†	**0.31**†
Ntr				**0.30**†

Subject 5

	Ang	Hap	Ntr	Sad
Afr	**0.30**†	0.37‡	**0.30**†	0.39
Ang		**0.30**†	**0.30**†	**0.22**†
Hap			0.37‡	**0.33**†
Ntr				0.43

Subject 6

Table 6.4: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Bayesian classifiers (with pooled covariance matrix) trained on emotions pairs using feature set derived from vocal source parameters (jitter, shimmer, PSP, and GNE) for subjects 1 through 6 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

mains at 50%). For this feature set, it is notable that the *Anger* category is consistently discriminated from the rest for all subjects. Compared with the results obtained with the LF parametrization feature set, (see Table 6.3), the remaining vocal source features considered in this section provide improved discrimination for the *Sad* category. These results suggest that the features are not only providing useful discrimination, but also offering in some cases provide complementary voice quality information to discriminate between these states. In the next section we motivate one last kind of analysis of voice quality, based on a consonance-based model of the harmonic content of voice, and evaluate its ability to predict affective categories.

	Ang	Hap	Ntr	Sad
Afr	**0.11**†	**0.23**†	**0.23**†	**0.32**†
Ang		**0.13**†	**0.31**†	**0.19**†
Hap			**0.11**†	**0.06**†
Ntr				**0.34**†

Subject 7

	Ang	Hap	Ntr	Sad
Afr	**0.33**†	0.42	0.42	0.45
Ang		**0.33**†	0.42	0.39
Hap			0.45	0.61
Ntr				0.42

Subject 8

	Ang	Hap	Ntr	Sad
Afr	**0.26**†	**0.23**†	**0.24**†	0.40
Ang		**0.19**†	0.39	**0.30**†
Hap			**0.24**†	**0.26**†
Ntr				0.39

Subject 9

	Ang	Hap	Ntr	Sad
Afr	**0.06**†	**0.19**†	**0.19**†	**0.29**†
Ang		**0.24**†	**0.19**†	**0.15**†
Hap			**0.23**†	**0.24**†
Ntr				**0.26**†

Subject 10

	Ang	Hap	Ntr	Sad
Afr	**0.10**†	**0.22**†	**0.18**†	0.42
Ang		**0.16**†	**0.31**†	**0.15**†
Hap			**0.16**†	**0.18**†
Ntr				**0.30**†

Subject 11

Table 6.5: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Bayesian classifiers (with pooled covariance matrix) trained on emotions pairs using feature set derived from vocal source parameters (jitter, shimmer, PSP, and GNE) for subjects 7 through 11 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

6.6 Consonance-Based Analysis of the Voice Source

This last section describes a very different approach to quantifying voice quality. The focus of it is the analysis of spectral harmonic patterns and some of their perceptual implications. The method and the tools we will describe here are borrowed from research in the area of psychoacoustics where they have been primarily motivated by the goal of quantifying a phenomenon that was already discussed in Chapter 5, and which, following Sethares (1998), we term *sensory consonance*. The goal of this section is to extend the application of these tools in modeling harmonic signals by applying them directly to voiced speech, and to propose some measures which may prove fruitful in modeling the harmonic variation of the voice source associated with affective changes. The application of these techniques to vocal source analysis is, we believe, a novel contribution.

Subject	Afr	Ang	Hap	Ntr	Sad
1	0.39	**0.23**†	0.36‡	0.42	**0.31**†
2	**0.26**†	**0.31**†	**0.29**†	0.37‡	**0.33**†
3	**0.20**†	**0.25**†	**0.31**†	**0.31**†	0.36‡
4	**0.19**†	**0.25**†	**0.27**†	0.46	**0.24**†
5	**0.12**†	**0.29**†	**0.34**†	**0.32**†	**0.33**†
6	0.46	**0.32**†	0.37‡	0.39	0.38‡
7	**0.29**†	**0.22**†	**0.19**†	0.37‡	**0.30**†
8	0.40	**0.31**†	0.50	0.43	0.50
9	0.39	**0.33**†	**0.27**†	0.39	0.38‡
10	**0.20**†	**0.18**†	0.36‡	**0.25**†	**0.28**†
11	**0.32**†	**0.17**†	**0.23**†	**0.34**†	0.38‡

Table 6.6: Generalization error (estimated by leave-one-out cross validation) for subject-dependent Bayesian classifiers (with pooled covariance matrix) trained to discriminate between each emotion category and the remaining ones pooled together, and using feature set derived from vocal source parameters (jitter, shimmer, PSP, and GNE) (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

The notion of perceptual consonance was already introduced in Chapter 5 while trying to find a perceptually motivated method to quantify the intervalic distance between the clusters of frequency values obtained from the F0 contour. That analysis will be revisited here, though the goal is not to describe the "chordal" structures we built out of F0 by clustering together F0 values across a contour, but rather to describe some of the perceptual implications of a particular distribution of frequencies over the spectrum.

The basic perceptual result relevant to this analysis, studied by Plomp & Levelt (1965) and already described in previous chapters, is the rising-and-falling curve quantifying how listeners judge the amount of dissonance/consonance as a function of the interval between two tones (see Fig. B-1 in Appendix B) When there is a more complex tone (i.e., two sets of frequencies sounding simultaneously), the different frequencies interact with each other, depending on their relative strengths and separation, to produce a more complex pattern of rising and falling dissonance. Sethares (1998) has proposed a simple linear model, reviewed in Appendix B, to describe this variation in sensory consonance as a function of frequency interval. The result, known as a dissonance curve, is particularly enlightening in that it shows intervals at which a given spectral pattern can sound particularly consonant or dissonant. One of the major implications of this work has been a challenge to the notion of a *scale* since it states that there is no such thing as an absolute restful or dissonant interval.

Rather, the consonance associated with playing a sound at different intervals is directly influenced by the harmonic content of the sound. It can be shown that one can produce particular spectral patterns to achieve dissonance at arbitrary intervals of the musical scale, even at intervals as commonly accepted as consonant as the octave.

This kind of analysis has been put to use to obtain suitable (consonant) scales for playing arbitrary sounds, to synthesize sounds that match desired arbitrary scales, and even to compare the overall dissonance of different performances of a music piece. In this work, we are interested in the particular quantification of the harmonic content of the voice, and to the degree to which irregularities and affective variations can be captured by the perceptual model described. Although the voice may not operate like a musical instrument during the regular production of connected speech, the intrinsic dissonance of each spectral pattern may be a salient perceptual component of voice quality. It remains purely speculative, but it is nonetheless reasonable to consider whether the perception of certain voice qualities as *harsh* might not regulated by the same psychoacoustic underpinnings regulating the perception of certain intervals as *rough*.

The extension to voice analysis is straightforward: As shown in panels (a) and (b) of Fig. 6-9, different voiced sections of speech show different distributions of the spectral harmonics, showing variations of harmonic strength (relative heights) and location (the degree to which the upper partials are multiple of the fundamental). If we remove from the power spectral densities shown in Fig. 6-9 (b) any spurious non-harmonic structure, then the spectral pattern of the idealized harmonics may be described by the dissonance curves shown in Fig. 6-9 (c). Repeating this procedure for near-stationary short-segments of speech then yields a time-varying representation which we can represent by the dissonance diagram shown in Fig. 6-10, where each vertical slice of this figure is a dissonance curve, such as the ones shown in Fig. 6-9 (c). The algorithm to build a dissonance diagram is summarized in Fig. 6-11.

To build a feature set, let us first identify a series of landmarks and features on the dissonance curve $D_m(\alpha_n)$ associated with the mth time slice, where n is an index spanning the frequency intervals ($n = 1, \cdots, N$). Let us assume that each interval at which there is maximum dissonance is indexed by α_k^d and every interval of minimum dissonance (excluding unison) by α_j^c. Without losing any generality, let us order them such that $D(\alpha_k^d) \geq D(\alpha_{k+1}^d)$ and $D(\alpha_j^c) \leq D(\alpha_{j+1}^c)$, and assume there are K maxima and J minima. Implicit in the

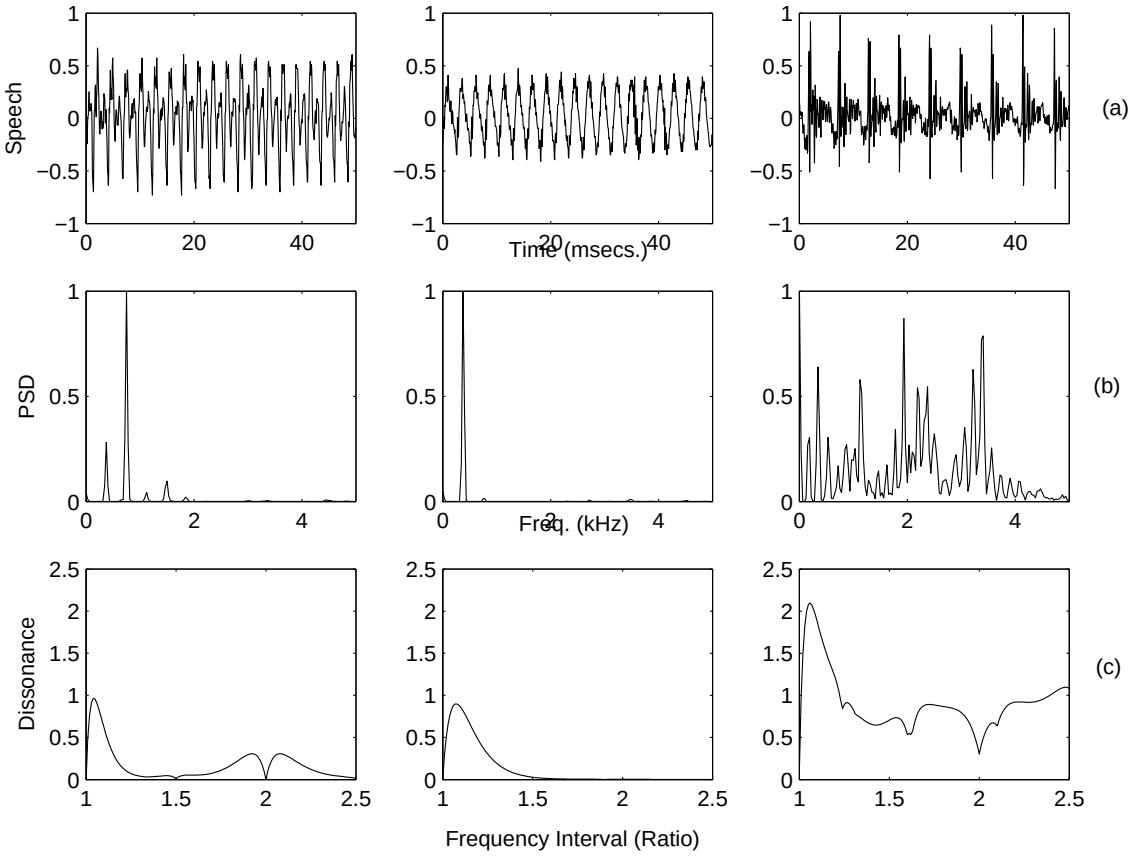

Figure 6-9: (a) Short nearly-stationary speech segments. (b) Power spectral densities. (c) Dissonance curves.

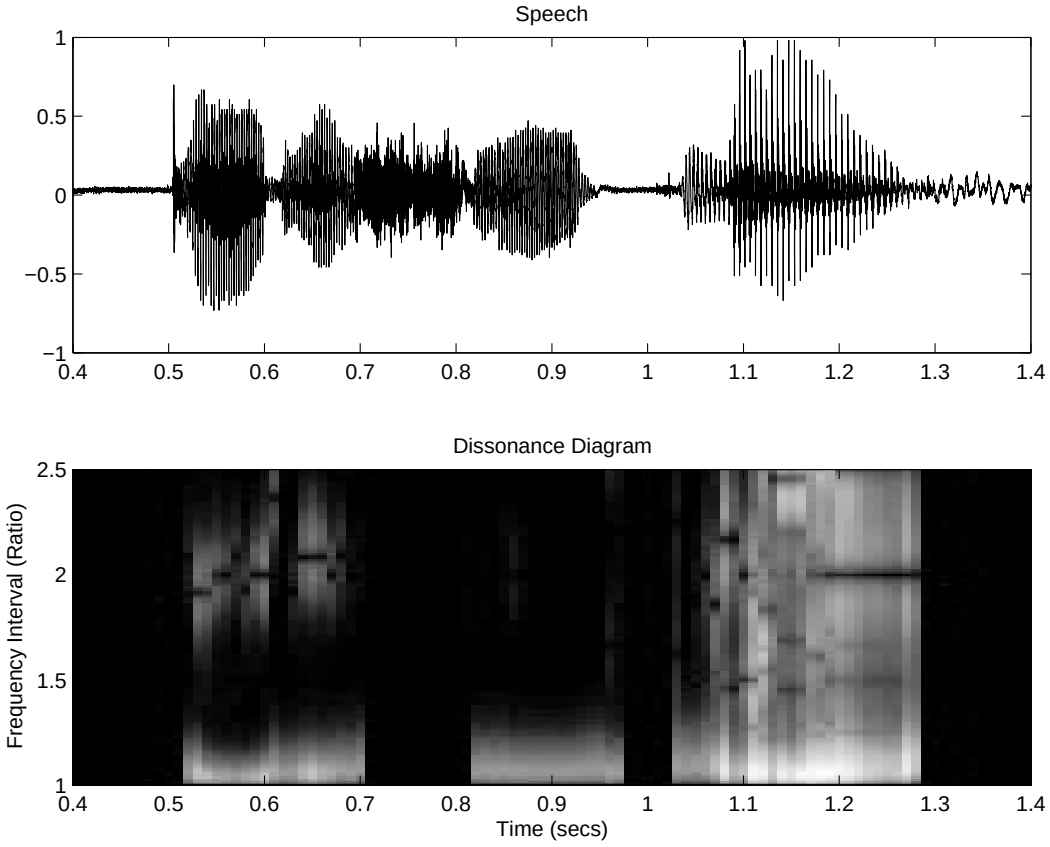

Figure 6-10: Speech waveform and dissonance diagram.

Algorithm 6.6: Let $s^{seg}(n)_l$ be the lth segment of voiced speech, 40 $msecs.$ long, computed every 10 $msecs.$

1. Let $l \leftarrow 1$
2. Compute the power spectral density S_k from $s^{seg}(n)_l$ using the method of averaged periodograms and an FFT size N_{fft} (e.g., $N_{fft} = 512$); $k = 1, \cdots, N_{fft}$.
3. Find the first N harmonics (e.g., $N = 6$) locations and amplitudes $\{f_n\}$ and $\{a_n\}$ from S_k.
4. Normalize the harmonics: $a_n \leftarrow \frac{a_n}{\max_n a_n}$.
5. Convert frequency samples to Hz: $f_n \leftarrow \frac{f_n - 1}{N_{fft}}$. Let $F = \{f_n, a_n\}$
5. Compute the dissonance curve $D_F(\alpha)$ (described in Appendix B) as a function of the frequency interval α
6. Let $DG_{l,\alpha} \leftarrow D_F(\alpha)$ be the value of the dissonance curve for the kth speech segment.
7. Let $l \leftarrow l + 1$; go to 2.

Figure 6-11: Algorithm for computing dissonance diagrams.

dissonance curve is the quantity known as the inherent (or intrinsic) dissonance of the spectral harmonic pattern (defined by Eq. B.9); that is, the sum of the dissonances between all pairs of harmonics within the pattern. Let us denote this quantity as D_{I_m} for the mth time slice. Let us then summarize the landmarks and shape of a curve by a 12-parameter vector containing:

$$\begin{aligned} D_m^{parm} = \Big[& \alpha_1^c, \quad D_m(\alpha_1^c), \quad \alpha_2^c, \quad D_m(\alpha_2^c) \cdots \\ & \alpha_1^d, \quad D_m(\alpha_1^d), \quad \alpha_2^d, \quad D_m(\alpha_2^d) \cdots \\ & \frac{1}{J}\sum_{j=1}^{J} D_m(\alpha_j^c), \quad \frac{1}{K}\sum_{k=1}^{K} D_m(\alpha_k^d) \cdots \\ & \frac{1}{N}\sum_{n=1}^{N} D_m(\alpha_n), \quad \frac{1}{N-1}\sum_{n=1}^{N-1} |D_m(\alpha_{n+1}) - D_m(\alpha_n)| \Big]^T. \end{aligned} \tag{6.46}$$

The vector contains the locations of the two strongest intervals of dissonance and consonance and their respective dissonance/consonance values, the average value of dissonance (consonance) peaks (valleys), and finally the average value of the dissonance curve and its

	Ang	Hap	Ntr	Sad
Afr	**0.28**†	**0.34**†	0.48	**0.30**†
Ang		**0.19**†	**0.31**†	**0.26**†
Hap			0.40	**0.28**†
Ntr				0.42

Subject 1

	Ang	Hap	Ntr	Sad
Afr	**0.29**†	0.37‡	**0.28**†	**0.26**†
Ang		**0.31**†	**0.32**†	0.50
Hap			0.37‡	**0.33**†
Ntr				0.47

Subject 2

	Ang	Hap	Ntr	Sad
Afr	**0.11**†	0.36‡	**0.10**†	**0.25**†
Ang		**0.28**†	**0.29**†	**0.18**†
Hap			**0.26**†	**0.31**†
Ntr				**0.18**†

Subject 3

	Ang	Hap	Ntr	Sad
Afr	**0.16**†	**0.15**†	**0.15**†	**0.14**†
Ang		**0.22**†	**0.34**†	**0.23**†
Hap			**0.17**†	**0.05**†
Ntr				**0.32**†

Subject 4

	Ang	Hap	Ntr	Sad
Afr	**0.11**†	**0.34**†	**0.08**†	**0.29**†
Ang		0.41	0.42	**0.29**†
Hap			**0.22**†	0.50
Ntr				**0.24**†

Subject 5

	Ang	Hap	Ntr	Sad
Afr	**0.25**†	0.45	**0.27**†	**0.34**†
Ang		**0.27**†	**0.29**†	**0.34**†
Hap			**0.33**†	0.42
Ntr				0.37‡

Subject 6

Table 6.7: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent linear classifiers trained on emotions pairs using harmonic consonance feature set of F0 for subjects 1 through 6 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

absolute first difference. Together these values summarize major landmarks on the curve, as well as information about the level and rate of change of the curve. For each such curve in the dissonance diagram, we obtain a value of its intrinsic dissonance D_{I_m} and its parametric summary D_m^{parm}, from which now we define an observation feature vector

$$FS^{cons} = \left[\text{median}_m\{D_{I_m}\}, \quad \text{range}_m\{D_{I_m}\}, \quad \text{median}_m\{D_m^{parm}\}^T \right]^T \tag{6.47}$$

summarizing the entire dissonance diagram. This 14-D feature vector is used next to investigate discrimination between affect categories.

6.6.1 Evaluation and Discussion

Table 6.7 and 6.8 show the results of using the feature set derived from the consonance-based analysis of the harmonic spectral content of voice to build emotion-pair discriminants.

	Ang	Hap	Ntr	Sad
Afr	**0.14**†	**0.26**†	**0.16**†	**0.26**†
Ang		**0.21**†	0.51	0.41
Hap			**0.26**†	0.35‡
Ntr				0.39

Subject 7

	Ang	Hap	Ntr	Sad
Afr	0.56	**0.33**†	0.54	0.55
Ang		0.40	0.45	0.48
Hap			0.36‡	0.42
Ntr				0.45

Subject 8

	Ang	Hap	Ntr	Sad
Afr	**0.31**†	0.43	**0.27**†	0.40
Ang		**0.24**†	0.47	0.36‡
Hap			**0.27**†	0.35‡
Ntr				0.37‡

Subject 9

	Ang	Hap	Ntr	Sad
Afr	**0.13**†	**0.30**†	**0.27**†	0.37‡
Ang		**0.22**†	**0.30**†	**0.31**†
Hap			**0.30**†	0.38‡
Ntr				0.54

Subject 10

	Ang	Hap	Ntr	Sad
Afr	**0.10**†	0.36‡	**0.12**†	**0.33**†
Ang		**0.17**†	**0.28**†	**0.15**†
Hap			**0.12**†	**0.26**†
Ntr				**0.26**†

Subject 11

Table 6.8: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent linear classifiers trained on emotions pairs using harmonic consonance feature set of F0 for subjects 7 through 11 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

Table 6.9 summarizes the generalization errors obtained with classifiers trained instead to discriminate between each emotion category and the remaining members of the set in an equal prior task (chance-level performance is 50%).

The results show that the set FS^{cons} yields significant discrimination for several of the pairs considered, a result that is repeated for the data of several subjects. The general trends observed with the other feature sets considered in this chapter are replicated to a certain extent, although individual rates vary above and below the results already presented for specific emotion combinations. In terms of specific emotions, the consonance features mostly contribute to separate the *Afraid* and *Angry* categories from the remaining ones. For Subject 4, these features provide consistent discrimination in the pairwise combination and in the 1-vs-rest discrimination tasks. It can also be seen that the features provide in some instances (e.g., subject 11) complementary information, and that therefore they do not constitute a redundant set beyond those that we have already presented in this chapter. The

significant discrimination obtained suggests that the consonance-based analysis introduced in this chapter provides a viable representation of the harmonic spectral content of voice, and represents a suitable framework for modeling differences of affective content.

Subject	Afr	Ang	Hap	Ntr	Sad
1	0.43	**0.23**†	0.37‡	0.52	0.39
2	**0.32**†	0.41	**0.33**†	0.38‡	0.50
3	**0.32**†	**0.24**†	0.43	**0.30**†	**0.28**†
4	**0.18**†	**0.22**†	**0.16**†	**0.33**†	**0.22**†
5	**0.20**†	**0.24**†	0.53	**0.23**†	0.60
6	**0.37**†	**0.29**†	0.48	0.44	0.49
7	**0.25**†	**0.33**†	0.35‡	0.38‡	0.39
8	0.54	0.52	0.32	0.52	0.50
9	0.45	0.38‡	0.38‡	0.36‡	0.47
10	**0.31**†	**0.25**†	**0.30**†	0.41	0.50
11	**0.23**†	**0.22**†	**0.21**†	0.36‡	0.41

Table 6.9: Generalization error (estimated by leave-one-out cross validation) for subject-dependent Bayesian classifiers (with pooled covariance matrix) trained to discriminate between each emotion category and the remaining ones pooled together, and using feature set derived from harmonic consonance analysis (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

6.7 Chapter Summary

In this chapter we have analyzed the contribution of voice quality measures to the discrimination of affective categories from speech. Variations in voice quality have been proposed and studied as an important correlate of affective variation in speech, a result which has been corroborated by the discriminative performance of the proposed features in this work. We have discussed a variety of supporting algorithms for obtaining different parameters related to the voice source, and proposed three different sets of features based on this processing. The first approach focused on recovering the volume velocity flow directly and describing it with the parametric model proposed by Liljencrants and Fant, an approach that led to a series of features directly related to the parametric representation. The second proposed feature set investigated features of the short-term period-to-period irregularities of fundamental period and energy, as well as parameters describing the spectral decay of the glottal flow and the level of noise in the excitation. Finally, we presented a consonance-based approach to modeling vocal quality information which we believe constitutes a novel

way of looking at the problem. The performance of all feature sets have been shown to provide significant discrimination for a good number of the emotion pairs considered in the data set.

Although the Liljencrants-Fant model has been explored for modeling affective variations, we have made novel use of a much wider set of features derived from this parametric model. Likewise, the use of an extended feature set derived from acoustic parameters most often encountered in the study of vocal pathology represents a novel application of such parameters. Finally, the application of a dissonance model as a tool to analyze the affective variations in harmonicity of the voice source represents, we believe, an entirely novel contribution of the work described in this thesis.

Chapter 7

Analysis of Rhythm

7.1 Introduction

In this chapter we turn to the analysis of one last set of acoustic parameters that are affected by the prosodic organization of speech. Under the header of *rhythm*, we aim to investigate parameters that capture information about alternation patterns in *prominence* and *duration* along an utterance, and whether such parameters exhibit correlations with changes in affective tone. Before turning to a particular proposal of these parameters, we will provide a short overview of existing accounts of linguistic rhythm in an attempt to draw a proposal of rhythmic features to quantify this aspect of speech.

7.2 Rhythm, Isochrony and Categorization

The notion of *rhythm*, as applied to spoken language, has, in principle, a very intuitive appeal: We clearly perceive that, in a language like English, different groupings of speech material are accorded different patterns of prominence and duration. Poets are supposed to show a very fluid command of particular rhythmic patterns (meters), but even the non poet can be a good judge of when a minor change to a text can make an entire rhythmic structure collapse. Linguistically speaking, however, an objective agreed-upon definition of what constitutes rhythm and what is the range of variation in rhythmic parameters has proved more elusive. One of the first attempts to formalize this aspect of speech was provided by the work of Pike and Abercrombie, where rhythm in language was viewed as the isochronous recurrence of some speech unit. Pike and Abercrombie proposed that all languages exhibit

some isochronous units, and described in particular two kinds of isochrony: languages could be *stress-timed* if the intervals between stressed syllables tended to be similar, or *syllable-timed* if adjacent syllables tended to be of near equal length. More strongly, Abercrombie asserted that this distinction was categorical, and that any language could be categorized in terms of its rhythmic properties into one of these two exclusive groups (Cummins & Port, 1998; Ramus, 2002; Cummins, 2002; Grabe & Low, 2002).

Much subsequent work has tried to seek acoustic verification for either of these two claims, namely, that there is such a thing as isochrony, and that this isochrony defines categories rather than two points in a continuum. The latter remains an open research question (Ramus et al., 1999; Ramus, 2002). Failure to produce acoustic parameters that verified the first claim led to a revision, and relaxation in some cases, of strict isochrony (Grabe & Low, 2002). However, recent work has proposed some acoustic measurements which produce good separation between classically stress-timed (e.g., Germanic) and syllable-timed (e.g., Romance) languages, leading to a re-appraisal of the hypothesis (Ramus, 2002; Galves et al., 2002).

We are not concerned with linguistic typology in this work; however, we are interested in studying whether there is rhythmic variability associated with affective factors. While investigating the existence of linguistic rhythm categories, research in this area has had to contend with performance factors that influence the distribution of the acoustic measures under examination (Grabe & Low, 2002). While the challenge in that research program is to abstract from this variability in order to discover underlying linguistic categories, this variability may provide, in our case, a potential source of interesting information if it turns out that speakers use it paralinguistically and encode affective information with such variations. In order to pursue this idea further, we have investigated the feasibility of using a feature set containing similar features as have been examined in the literature for rhythmic modeling, and examined their contribution to affect discrimination.

7.3 Rhythm Features

In line with basic assumptions about what constitutes rhythm in language, our description is based on a representation in terms of syllabic units, as facilitated by the segmentation algorithms described in Chapter 3. The automatic segmentation provides information about

the boundaries of syllable nuclei, as well as any intervening pauses or breaths. It does not provide, however, an assignment of the intra-nuclear material to adjacent syllables. Given this representation, we will assume that the speech waveform has been pre-processed such that all silences and breaths have been removed, and that the resulting signal can be represented as an alternating sequence of nuclear and inter-nuclear segments $\{N_k\}$ and $\{IN_k\}$.

We will find it useful to define the following auxiliary measures on an arbitrary sequence x_n of length N:

$$\mu_g(x_n) = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{7.1}$$

$$\mu_l(x_n) = \frac{x_n + x_{n+1}}{2} \tag{7.2}$$

$$\Delta_{ab}(x_n) = |x_{n+1} - x_n| \tag{7.3}$$

$$\Delta_{norm}(x_n) = \frac{\Delta_{ab}(x_n)}{\mu_l(x_n)} = 2\frac{|x_{n+1} - x_n|}{x_n + x_{n+1}} \tag{7.4}$$

$$\mu_{dev}(x_n) = \mu_l(x_n) - \mu_g(x_n) = |x_{n+1} - x_n| - \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{7.5}$$

Eqs. 7.1 and 7.2 are, respectively, the global mean of x_n and a sequence of the pairwise means of the samples. Eq. 7.3 is the absolute first-order difference of the elements in x_n, and Eq. 7.4 this absolute first difference normalized by the "local" mean $\mu_l(x_n)$. Finally, Eq. 7.5 is the sequence containing the difference between the local pairwise means and the global mean. Assume now that the *kth* nuclear interval N_k encompasses the indices $n_1^k, \cdots, n_K^k$, and, likewise, the *kth* inter-nuclear interval is defined over the indices $i_1^k, \cdots, i_K^k$. Let s_n be a speech signal sampled at F_s Hz, and define the following:

$$Dur_k^N = \frac{n_K^k - n_1^k}{F_s} \tag{7.6}$$

$$Dur_k^{IN} = \frac{i_K^k - i_1^k}{F_s} \tag{7.7}$$

$$RMS_k^N = \left(\frac{1}{n_K^k - n_1^k + 1}\sum_{n=n_1^k}^{n_K^k} s_n^2\right)^{\frac{1}{2}} \tag{7.8}$$

$$RMS_k^{IN} = \left(\frac{1}{i_K^k - i_1^k + 1}\sum_{n=i_1^k}^{i_K^k} s_n^2\right)^{\frac{1}{2}}. \tag{7.9}$$

The first two equations are the duration, in seconds, of the kth nucleus and the kth inter-nuclear segment whereas Eqs. 7.8 and 7.9 are the root-mean-square value of the kth nuclear and inter-nuclear speech intervals. From the sequences of duration and energy defined by Eqs. 7.6-7.9, together with Eqs. 7.1-7.5, we define the following observation vector:

$$\begin{aligned} FS_k^{rhythm} = [&\Delta_{ab}(Dur_k^N), \quad \Delta_{norm}(Dur_k^N), \quad \mu_{dev}(Dur_k^N), \cdots \\ &\Delta_{ab}(RMS_k^N), \quad \Delta_{norm}(RMS_k^N), \quad \mu_{dev}(RMS_k^N), \cdots \\ &\Delta_{ab}(Dur_k^{IN}), \quad \mu_{dev}(Dur_k^{IN}), \cdots \\ &\Delta_{ab}(RMS_k^{IN}), \quad \mu_{dev}(RMS_k^{IN})]^T. \end{aligned} \tag{7.10}$$

The quantities gathered in Eq. 7.10 are intended to provide a time-varying account of the alternations in patterns of duration and prominence (the latter roughly estimated by an energy measure averaged along the speech interval under consideration).[1] In the next section we evaluate the performance of these features on predicting affect categories.

7.4 Evaluation and Discussion

The feature set in Eq. 7.10 defines a time series of 10-dimensional observations. It is assumed that for modeling the alternations of prominence and duration the time dimension will be essential.[2] This conjecture was tested by first evaluating the performance of several Bayesian classifiers on a time-averaged version of the feature vector: by taking the mean value of the features along k, and by representing FS_k^{rhythm} with its median value along k. No significant results were obtained in either case. Although this may simply be a consequence of implementing a simple classification scheme (i.e., Bayesian classifiers with pooled covariance matrices, or class-dependent full or diagonal covariances), it may also point to the need for the time dimension to capture dynamical dependencies. In the next stage, therefore, we fitted Hidden Markov Models (HMM) to explicitly model the dynamic evolution of the observations. The generalization errors, estimated through leave-one-out cross-validation, are shown in Tables 7.1 and 7.2 for a two-state HMM with arbitrary co-

[1] The quantities $\Delta_{norm}(Dur_k^{IN})$ and $\Delta_{norm}(RMS_k^{IN})$ were originally included in the feature vector of Eq. 7.10. However, they were found to cause numerical instability during parameter estimation (see next section) due to colinearity, and were omitted from further processing.

[2] The variable k does not, of course, index uniform time (sampled or continuous), but rather the non-uniform structural units (e.g., syllable nuclei, etc.) we have discussed

	Ang	Hap	Ntr	Sad
Afr	$0.38^{\ddagger}$	$0.37^{\ddagger}$	$\mathbf{0.34}^{\dagger}$	0.49
Ang		0.41	0.46	0.46
Hap			0.44	0.40
Ntr				0.43

Subject 1

	Ang	Hap	Ntr	Sad
Afr	0.47	0.45	0.51	0.43
Ang		0.41	0.43	0.44
Hap			0.44	$0.37^{\ddagger}$
Ntr				$0.38^{\ddagger}$

Subject 2

	Ang	Hap	Ntr	Sad
Afr	0.44	0.46	$0.38^{\ddagger}$	0.44
Ang		$0.37^{\ddagger}$	$0.37^{\dagger}$	0.39
Hap			$0.37^{\dagger}$	0.49
Ntr				$\mathbf{0.34}^{\dagger}$

Subject 3

	Ang	Hap	Ntr	Sad
Afr	$0.36^{\dagger}$	$0.38^{\dagger}$	0.39	$\mathbf{0.33}^{\dagger}$
Ang		0.45	$0.37^{\dagger}$	0.54
Hap			0.53	0.52
Ntr				0.46

Subject 4

	Ang	Hap	Ntr	Sad
Afr	$\mathbf{0.25}^{\dagger}$	0.42	$\mathbf{0.32}^{\dagger}$	0.43
Ang		0.43	0.61	$0.38^{\ddagger}$
Hap			0.49	0.47
Ntr				0.38

Subject 5

	Ang	Hap	Ntr	Sad
Afr	0.43	0.42	$0.38^{\ddagger}$	0.42
Ang		0.47	0.41	0.48
Hap			0.44	0.46
Ntr				0.45

Subject 6

Table 7.1: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Hidden Markov Models trained on emotions pairs using feature set containing only rhythm features for subjects 1 through 6 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

variance matrices and full connectivity.

As the result tables show, the proposed rhythm features only constitute weak predictors of affect categories. In fact, this feature set provides the weakest prediction of all the different feature types considered in this thesis (loudness, intonation, and voice quality features). Of the 110 total pairs of classifiers trained, only 28 achieve significant discrimination at the $p < 0.05$ level, and of these only 8 reach it at the $p < 0.01$ level. Since this represents an improvement over the performance obtained with the time averaged values of these features, however, it suggests that the dynamic evolution of the feature values reveal some affect-dependent effect, albeit a weak one.

Table 7.3 shows the generalization error of classifiers trained to discriminate between each affective category and its complements. As suggested by the results obtained from the pairwise comparisons between emotion pairs, no single affective state is consistently well predicted and discriminated from the rest by the use of the proposed rhythm features. Only

	Ang	Hap	Ntr	Sad
Afr	**0.25**†	0.49	0.35‡	0.42
Ang		0.37‡	0.38‡	0.35‡
Hap			0.45	0.50
Ntr				0.44

Subject 7

	Ang	Hap	Ntr	Sad
Afr	0.52	0.51	0.60	0.66
Ang		0.43	0.48	0.50
Hap			0.55	0.52
Ntr				0.56

Subject 8

	Ang	Hap	Ntr	Sad
Afr	0.47	0.51	0.52	0.55
Ang		0.53	0.47	0.53
Hap			0.56	0.58
Ntr				0.56

Subject 9

	Ang	Hap	Ntr	Sad
Afr	0.50	0.45	0.54	0.44
Ang		0.54	0.58	0.48
Hap			0.65	**0.37**†
Ntr				0.45

Subject 10

	Ang	Hap	Ntr	Sad
Afr	0.35†	0.50	0.51	0.49
Ang		**0.32**†	0.35‡	0.35‡
Hap			0.43	0.45
Ntr				0.49

Subject 11

Table 7.2: Generalization error (estimated by leave-one-out cross-validation) for subject-dependent Hidden Markov Models trained on emotions pairs using feature set containing only rhythm features for subjects 7 through 11 (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

the *Angry* category achieves a better than chance performance for only 3 subjects of the 11 considered. It may be the case that affective speech does not influence these rhythmic variables to a large extent, and that the weak effect we are observing represents the limit of the contribution of rhythm features to signaling affect. It can be that changes in rhythmic features are only likely to occur when pronounced changes in speaking style accompany a change of affective tone (e.g., when overemphasizing a sentence with marked irritation). It can also be the case that an accurate identification of nuclear and non-nuclear intervals is needed, not to mention a more precise representation of syllable structure, and that therefore the results may be sensitive to errors introduced by the automatic segmentation algorithm on which this modeling is based. The representation proposed here may also have limited modeling power, and more comprehensive rhythmic features may also be needed to capture affective differences. At any rate, the performance figures obtained here suggest that further work should be carried out to better understand which correlations might exist

Subject	Afr	Ang	Hap	Ntr	Sad
1	0.46	0.58	0.59	0.57	0.54
2	0.58	0.48	0.55	0.59	0.47
3	0.47	$0.37^{\ddagger}$	0.58	0.41	0.54
4	0.39	0.51	0.57	0.55	0.52
5	$0.38^{\ddagger}$	$0.35^{\ddagger}$	0.51	0.47	0.50
6	0.46	0.39	0.47	0.40	0.53
7	0.48	$0.35^{\ddagger}$	0.56	0.43	$0.35^{\ddagger}$
8	0.51	0.48	0.56	0.58	0.55
9	0.58	0.58	0.58	0.50	0.52
10	0.43	0.68	0.50	0.53	0.44
11	0.61	0.39	0.51	0.43	0.51

Table 7.3: Generalization error (estimated through 10-fold cross validation) for subject-dependent Hidden Markov Models trained to discriminate between each emotion category and the remaining ones pooled together, and using feature set containing rhythm features only (†and ‡denote error rates that are significantly smaller than chance at the $p < 0.01$ and $p < 0.05$ level respectively).

between phonetic rhythmic variation and affect categories, as well as to propose better features that might reflect this correlation.

7.5 Chapter Summary

This chapter has taken a look at the performance of a set of proposed rhythm features in predicting affect categories. Traditional descriptions of rhythm have tended to view it as a linguistic property of a language, and some recent work has gone into exploring feature spaces that might allow us to describe different languages in terms of the explored rhythmic dimensions. In this work, the emphasis has been on exploring within-language variations of a set of rhythm features and whether this source of variation can be consistently linked to affective variations. The results show that these features do not predict affect categories as well as the feature sets previously considered and containing loudness, fundamental frequency, and voice quality information. However, we have shown enough significance to suggest that there might an interaction between the features and affective categories, and that further modeling ought to target this source of variation as well. The application of the features described in this chapter to the task of automatic discrimination of affective categories constitutes a novel contribution of this thesis. However, further work is needed to explore what other rhythm-related features contribute more significantly to affect encoding,

and to investigate the boundaries of their contributions.

Chapter 8

Bayesian Networks for Modeling Prosodic Phenomena

8.1 Introduction

In this chapter we address the machine learning component of this thesis. The main contribution is the motivation of a graphical model for modeling data that is dynamical and hierarchical in nature, and can thus be suited for modeling acoustic prosodic parameters from speech. This model is applied to classification tasks, where we seek to learn and predict affect categories from observations of the acoustic parameters that have been examined in previous chapters.

8.2 Directed Acyclic Graphs

A directed acyclic graph is a pair $G = \{Y, E\}$ where $Y = \{Y_1, \cdots, Y_n\}$ is a set of nodes, each of which is associated with a random variable, and $E = \{(Y_i, Y_j) : i \neq j\}$ is a set of directed edges *from* Y_i to Y_j encoding local conditional probabilities associated with the graph G. In order for the graph to be acyclic no loops are allowed. Every node Y_i has a (possibly empty) set of parent nodes denoted by $\pi(Y_i)$. A node Y_j is said to be a parent of Y_i if there is a directed edge $(Y_j, Y_i) : i \neq j$. A graph G encodes a factorization of the joint

probability distribution over all variables in terms of these local conditional distributions:

$$p(y_1, \cdots, y_n) \doteq \prod_i p(y_i|\pi(y_i)). \tag{8.1}$$

Directed Graphs, also known as Bayesian Networks, offer a very powerful framework for statistical modeling since they allow us to encode arbitrary dependencies between variables. Furthermore, there exist general purpose algorithms that allow us to do learning on such models irrespective of the structure (i.e., the dependencies that may exist between variables). While directed graphs offer flexibility to the designer in establishing arbitrary dependencies between variables, they become particularly useful when used somewhat sparsely to encode structure about the problem. That is, a heavily connected graph reveals many dependencies but also (i) makes a very general statement about the joint probability distribution on the graph and (ii) may be computationally intractable. When structured properly, Bayesian networks subsume many very well known models such as principal component analysis, independent component analysis, mixture of experts, hidden Markov models (HMM), and linear dynamical systems.

Suppose we are interested in modeling stochastic processes that can be described temporally across several scales, and we wish to introduce some dependencies between scales. Suppose, furthermore, that an observation at a given scale *spans* or *dominates* a subset of the observations at a scale below it, and consider the structure of the graph shown in figure 8-1. For this and the remaining graphs shown, we will adopt the convention of representing continuous valued nodes with circles and discrete nodes with squares. Unless otherwise indicated, shaded nodes will correspond to observation nodes, and unshaded nodes to hidden or latent nodes.

Let i, j and k be index variables used to encode the hierarchical (vertical) position of a variable (from levels 1, 2 and 3) as well as its "temporal" position within a level of the hierarchy, such that the following is true: The number of indices signals the level of the hierarchy where the variable is arranged, and the first $L-1$ indices trace a vertical path describing the dependencies between the latent variables. For instance, the variable x_{ijk} is the kth variable occurring at the third level whose "vertical" parent is x_{ij} (which, recursively, has x_i as a parent). Each latent variable at the first $L-1$ levels is therefore a parent to a set of ordered latent variables at the level below it. Furthermore, this set of

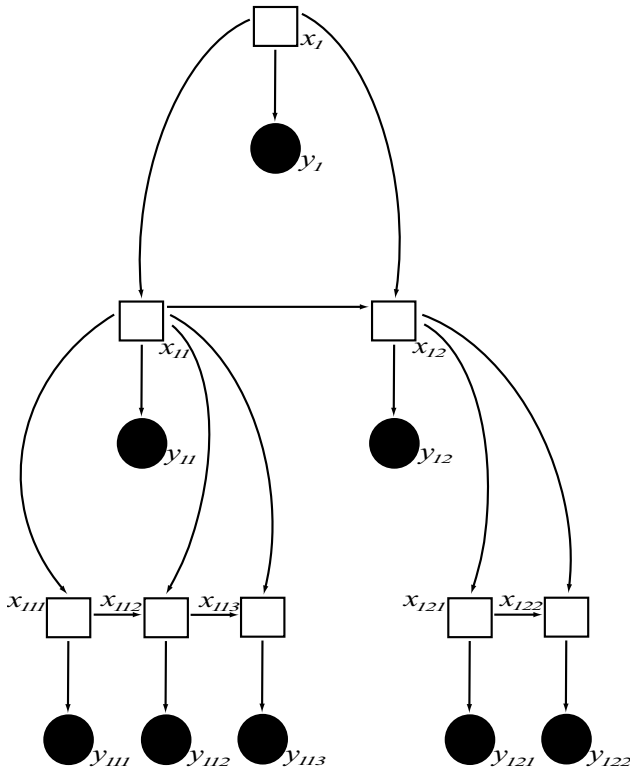

Figure 8-1: Directed Acyclic Graph (DAG). The shaded nodes correspond to observation variables, and the clear nodes to latent (or hidden) variables.

parent latent variables obeys a first-order Markovian relation between them. In addition to the latent variables x_i, x_{ij}, x_{ijk}, the model contains a set of observation variables y_i, y_{ij}, y_{ijk}, such that $x_i = \pi(y_i)$, $x_{ij} = \pi(y_{ij})$, and $x_{ijk} = \pi(y_{ijk})$. The model can be summarized as a hierarchical structure of hidden variables where there is a first-order Markov property describing dependencies between levels, as well as between variables within a level, and where each subsequence of hidden states is associated with a series of observations (i.e., a time series defined at that particular scale).

The next important feature to note about the structure shown in Fig. 8-1 is that the relationship between observations across adjacent scales does not follow a homogeneous pattern. That is, a state at level l does not span a fixed number of states at level $l+1$ (a hidden state does not have a fixed number of children). In fact, we do not want to impose this restriction on this class of models, as the data we are interested in modeling share this inhomogeneity across scales. Up to this point we have described the kinds of probabilistic dependencies that exist between the variables in the model (i.e., the connectedness of the graph). We will next impose further structure by grouping different nodes into groups, or

equivalence classes. Before doing that, however, it is convenient to introduce a straightforward generalization of the model discussed so far. This is shown by the model in Fig. 8-2, which contains an additional set of latent variables. The meaning of these variables will become clearer when we describe the particular choice of probability density functions associated to model the observation nodes. However, this extension is brought at this point to carry out the level of generality throughout the rest of the discussion. Suffice it for now to describe the $\{z\}$ variables as mixture variables that partition each of the state variables in $\{x\}$ into substates. The full network just described corresponds then to the following

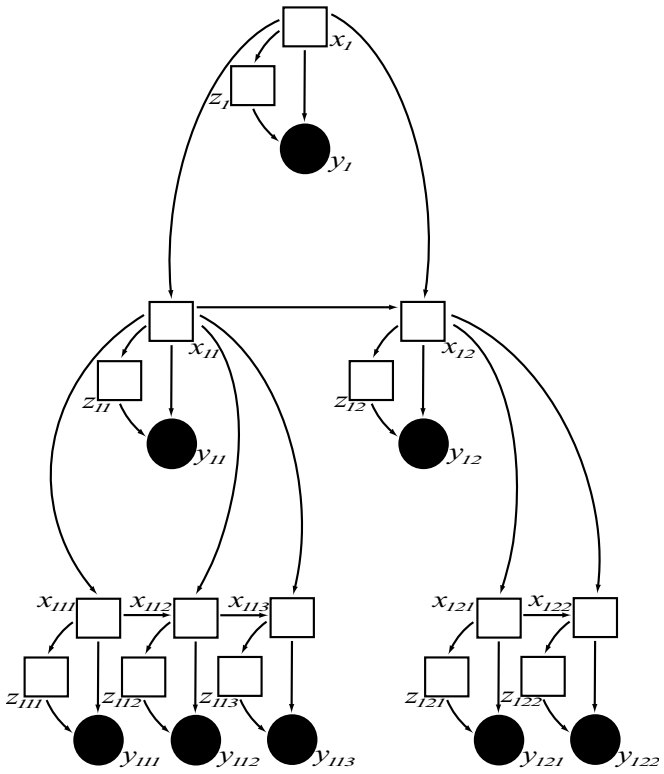

Figure 8-2: Hierarchical Bayesian network augmented with mixture nodes.

factorized probability distribution:

$$
\begin{aligned}
p(\{x\},\{y\},\{z\}) = & \Big[p(x_1)\prod_{i=2}^{I} p(x_i|x_{i-1})\Big] \\
& \Big[\prod_{i=1}^{I} p(x_{i1}|x_i)\prod_{j=2}^{J_i} p(x_{ij}|x_{ij-1},x_i)\Big] \\
& \Big[\prod_{i=1}^{I}\prod_{j=1}^{J_i} p(x_{ij1}|x_{ij})\prod_{k=2}^{K_{ij}} p(x_{ijk}|x_{ijk-1},x_{ij})\Big] \\
& \Big[\prod_{i=1}^{I} p(z_i|x_i)p(y_i|x_i,z_i)\prod_{j=1}^{J_i} p(z_{ij}|x_{ij})p(y_{ij}|x_{ij},z_{ij}) \\
& \prod_{k=1}^{K_{ij}} p(z_{ijk}|x_{ijk})p(y_{ijk}|x_{ijk},z_{ijk})\Big]. \qquad (8.2)
\end{aligned}
$$

An equivalence class is a set of nodes which share the same conditional probability distribution. It follows, therefore, that only nodes of the same type, size, with the same number of parents (and parents of commensurate dimensions) can be in the same class. Another way of restating this is that an equivalence class is a set of nodes which have their parameters tied. One can, for instance, define an equivalence class to group nodes that have a similar semantic interpretation, but which only change in one particular respect (e.g., different observations of a time series). This kind of interpretation is readily available for the model under consideration if we take the nodes along the horizontal chains to represent a dynamical evolution. Before specifying the equivalence classes, define the sets of variables X^l, Z^l and Y^l to be the sets containing the latent, mixture, and observation nodes at level l. (For instance, in Fig. 8-2, $X^1 = \{x_1\}$, $Z^2 = \{z_{11}, z_{12}\}$ and $Y^3 = \{y_{111}, y_{112}, y_{113}, y_{121}, y_{122}\}$.) For each hierarchical level, we will then define four equivalence classes (i.e., tie parameters) as follows:

- an equivalence class for each initial discrete node of X^l (an initial node is one which does not have a discrete parent in X^l)
- an equivalence class for each non-initial discrete node of X^l
- an equivalence class for each discrete mixture node in Z^l
- an equivalence class for each observation node in Y^l

This is graphically illustrated in Fig. 8-3 for two different networks with different structure. Nodes that share a similar shading pattern are in the same equivalence class. Equivalence classes can therefore be defined across networks to create an ensemble that is generated by a common model (i.e., the set of parameters associated with the equivalence classes just defined). The structure that we have further imposed on the networks by tying parameters

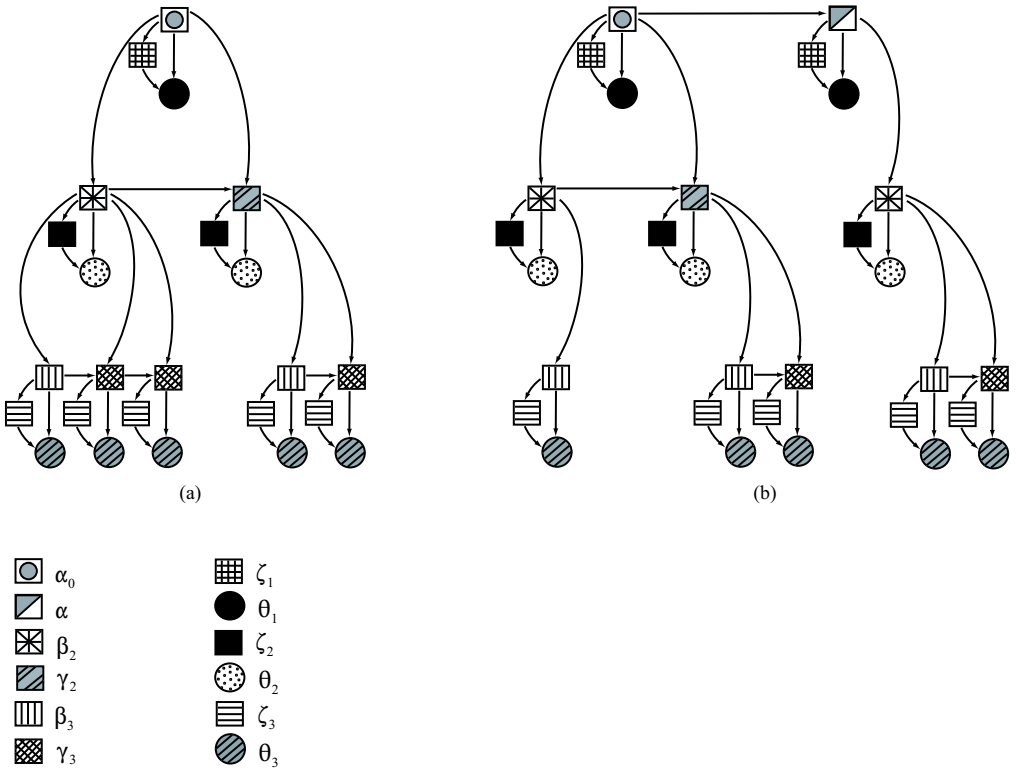

Figure 8-3: Equivalence classes

affords us a different kind of interpretation: each level of a hierarchy, disregarding vertical connections, represents one or more Dynamic Bayesian Networks (DBN). DBNs are a class of Bayesian networks that consist of repetitions of a basic structure with parameters tied after the second slice (for a network with first-order Markov dynamics). This is shown in Fig. 8-4, where each set of nodes circled is the Bayesian network representation of a hidden Markov model with mixture densities at its output. The structures we have presented here, together with the particular equivalence classes we have defined, therefore, can be thought of as a series of embedded hidden Markov models on which we have imposed some hierarchical dependencies. The evolution of the state is no longer just a function of the previous state, but also of the hidden state at the level above it.

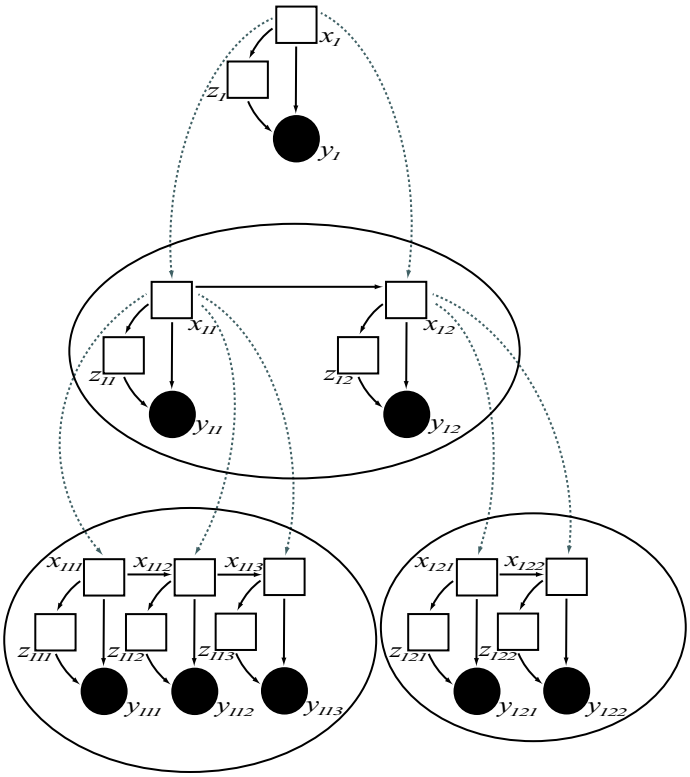

Figure 8-4: HMMs embedded within the hierarchical structure.

Up to this point we have described these graphical models in general terms. We now narrow down the choices by assigning particular types of probability density functions when we next address the issue of how to estimate the network parameters from a set of learning data.

8.3 Parameter Estimation

Let us make the notation more precise at this point by letting $\{x^n\}$, $\{y^n\}$, and $\{z^n\}$ be the set of nodes associated with a particular network n. Each network contains 12 equivalence classes or sets of parameters

$$\psi \doteq \{\alpha_0, \alpha, \beta_2, \beta_3, \gamma_2, \gamma_3, \zeta_1, \zeta_2, \zeta_3, \theta_1, \theta_2, \theta_3\}, \tag{8.3}$$

and its complete likelihood is given by the following expression, where the dependencies on the parameters has been made explicit:

$$
\begin{aligned}
p(\{x^n\},\{y^n\},\{z^n\}|\psi) = & \Big[p(x_1^n|\alpha_0)\prod_{i=2}^{I^n} p(x_i^n|x_{i-1}^n;\alpha)\Big] \\
& \Big[\prod_{i=1}^{I^n} p(x_{i1}^n|x_i^n;\beta_2)\prod_{j=2}^{J_i^n} p(x_{ij}^n|x_{ij-1}^n,x_i^n;\gamma_2)\Big] \\
& \Big[\prod_{i=1}^{I^n}\prod_{j=1}^{J_i^n} p(x_{ij1}^n|x_{ij}^n;\beta_3)\prod_{k=2}^{K_{ij}^n} p(x_{ijk}^n|x_{ijk-1}^n,x_{ij}^n;\gamma_3)\Big] \\
& \Big[\prod_{i=1}^{I^n} p(z_i^n|x_i^n;\zeta_1)p(y_i^n|x_i^n,z_i^n;\theta_1)\prod_{j=1}^{J_i^n} p(z_{ij}^n|x_{ij}^n;\zeta_2)p(y_{ij}^n|x_{ij}^n,z_{ij}^n;\theta_2) \\
& \prod_{k=1}^{K_{ij}^n} p(z_{ijk}^n|x_{ijk}^n;\zeta_3)p(y_{ijk}^n|x_{ijk}^n,z_{ijk}^n;\theta_3)\Big]. \qquad (8.4)
\end{aligned}
$$

The variables $\alpha_0, \alpha, \beta_2, \beta_3, \gamma_2$, and γ_3 denote the parameters of the discrete nodes associated with the "backbone" chains running horizontally across each network. Since all these variables and their parents are discrete, a multinomial probability density function is a natural choice for describing the distributions. The observable nodes are continuous valued and may be modeled with any continuous PDF, of which we will choose a Gaussian distribution. Furthermore, the mixture variables $\{z\}$, which we motivated by describing them as a partition of the states variables $\{x\}$ into substates, allow us to assign more than one Gaussian distribution per state. That is, in the general case in which we consider a network augmented with the $\{z\}$ variables, we effectively model the observable nodes with a mixture of Gaussian distribution. This is the primary reason for considering the mixture variables since any arbitrary distribution may be approximated by a mixture of Gaussians, so in principle we are not committed to the Gaussianity of the data by this particular choice of fitting distribution. The parameters $\{\eta\}$ then represent another multinomial distribution to model the discrete mixture variables, and $\{\theta\} = \{\mu, \Sigma\}$ are the sets of means and variances of each Gaussian equivalence class.

Learning in these models can be approached in a maximum likelihood framework by using the Expectation Maximization (EM) algorithm. If we have an ensemble of networks (each of which is supposed to model a set of independent identically distributed data), with likelihood as in (8.4), then we can proceed to estimate the parameters by writing the

expected complete log likelihood of the ensemble, and by differentiating with respect to each of the parameters in ψ:

$$\mathcal{L}_\psi = E\Big\{\sum_n \log p\big(\{x^n\},\{y^n\},\{z^n\}|\psi\big)\Big\}. \tag{8.5}$$

Notice that Eq. 8.5 consists of a sum of terms, each one involving one particular set of parameters from the equivalence classes, and that therefore the differentiation consists of a sum of localized contributions (i.e., the log likelihood decouples).

Since all the parameters of the networks fall under one of two types of distributions, multinomial and Gaussian, we look at the derivation of the maximum likelihood estimates of these parameters and then make the suitable assignment to the parameters in ψ. First we consider the general case of deriving the parameters of a conditional multinomial distribution/

8.3.1 Derivation of Maximum-Likelihood Estimates of a Multinomial Distribution

Let us describe the multinomial distribution by defining the set I_M of M orthogonal M-ary vectors as

$$I_M = \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \cdots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\}, \tag{8.6}$$

and letting x_p take values on the set I_{M_p}. The conditional multinomial distribution is then given by

$$p(x_1|x_2,\cdots,x_P) = \prod_{m_1=1}^{M_1} \cdots \prod_{m_P=1}^{M_P} \eta_{m_1\cdots m_P}^{\left([x_1]_{m_1}\times[x_2]_{m_2}\times\cdots\times[x_P]_{m_P}\right)}, \tag{8.7}$$

where

$$\eta_{m_1\cdots m_P} \doteq p([x_1]_{m_1}=1|[x_2]_{m_2}=1,\cdots,[x_P]_{M_P}=1), \tag{8.8}$$

and the notation $[x_p]_{m_p}$ refers to the *pth* entry of the binary vector x_p. We further apply the following constraint to ensure we have a legal probability mass function:

$$\sum_{m_1=1}^{M_1} \eta_{m_1,m_2,\cdots m_P} = 1 \quad \forall m_2, \cdots, m_P. \tag{8.9}$$

The parameters of this distribution therefore consists of a P dimensional table containing the values in Eq. 8.8.

If we have a set of independent identically distributed multinomial samples generated by η, then its complete log likelihood is given by

$$\begin{aligned}\mathcal{L}_\eta &= E\left\{\log \prod_n^N p(x_1^n | x_2^n, \cdots, x_P^n; \eta)\right\} \\ &= E\left\{\sum_n^N \sum_{m_1=1}^{M_1} \cdots \sum_{m_P=1}^{M_P} ([x_1^n]_{m_1} \cdots [x_P^n]_{m_P}) \log \eta_{m_1 \cdots m_P}\right\}. \end{aligned} \tag{8.10}$$

We now apply a Lagrangian multiplier to enforce the constraint in 8.9, to obtain the modified complete log likelihood $\tilde{\mathcal{L}}$:

$$\tilde{\mathcal{L}} = E\left\{\sum_n^N \sum_{m_1=1}^{M_1} \cdots \sum_{m_P=1}^{M_P} ([x_1^n]_{m_1} \cdots [x_P^n]_{m_P}) \log \eta_{m_1 \cdots m_P}\right\} + \lambda\left(1 - \sum_{m_1=1}^{M_1} \eta_{m_1,m_2,\cdots m_P}\right). \tag{8.11}$$

Taking derivatives with respect to η, setting to zero, and solving for λ, we obtain:

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \eta_{m_1,\cdots,m_P}} = \sum_n^N E([x_1^n]_{m_1} \cdots [x_P^n]_{m_P}) \frac{1}{\eta_{m_1,m_2,\cdots m_P}} - \lambda = 0, \tag{8.12}$$

$$\lambda = \frac{\sum_n^N E([x_1^n]_{m_1} \cdots [x_P^n]_{m_P})}{\eta_{m_1,m_2,\cdots m_P}}, \tag{8.13}$$

$$\lambda \sum_{m_1=1}^{M_1} \eta_{m_1,m_2,\cdots m_P} = \sum_{m_1=1}^{M_1} \sum_{n=1}^{N} E([x_1^n]_{m_1} \cdots [x_P^n]_{m_P}), \tag{8.14}$$

$$\lambda = \sum_{m_1=1}^{M_1} \sum_{n=1}^{N} E([x_1^n]_{m_1} \cdots [x_P^n]_{m_P}). \tag{8.15}$$

Combining Eqs. 8.13 and 8.15, we obtain

$$\hat{\eta}_{m_1,m_2,\cdots m_P} = \frac{\sum_n^N E([x_1^n]_{m_1} \cdots [x_P^n]_{m_P})}{\sum_{m_1=1}^{M_1} \sum_n^N E([x_1^n]_{m_1} \cdots [x_P^n]_{m_P})}. \tag{8.16}$$

Eq. 8.16 is the maximum likelihood estimate of the parameters of the multinomial, and it has the natural interpretation of a ratio of expected counts.

8.3.2 Derivation of Maximum-Likelihood Estimates of a Gaussian Distribution

Let us, without losing any generality, consider the case of optimizing the parameters of the Gaussian distribution described by θ_3 (i.e., the parameters of the third hierarchical level in the network); the extension from this to the other Gaussians is straightforward. The conditional Gaussian distribution of a node y_{ijk}^n (i.e., the ijk node in the nth network) with discrete parents x_{ijk}^n and z_{ijk}^n is given by

$$p\Big(y_{ijk}^n | [x_{ijk}^n]_r = 1, [z_{ijk}^n]_s = 1; \theta_3\Big) = (2\pi)^{-d_{ijk}/2} |\Sigma_{rs}|^{-1/2} \exp\left\{ -\frac{1}{2}(y_{ijk}^n - \mu_{rs})^T \Sigma_{rs}^{-1} (y_{ijk}^n - \mu_{rs}) \right\}, \tag{8.17}$$

where we have again assumed the notational conventions used in the previous section to describe multinomials. The complete log likelihood of N i.i.d. sets is given by

$$\mathcal{L}_{\theta_3} = E\left(\sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{i=1}^{I^n} \sum_{j=1}^{J_i^n} \sum_{k=1}^{K_{ij}^n} \log p(y_{ijk}^n | [x_{ijk}^n]_r = 1, [z_{ijk}^n]_s = 1; \theta_3)^{[x_{ijk}^n]_r [z_{ijk}^n]_s} \right) \tag{8.18}$$

$$\propto E\left(\sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{i=1}^{I^n} \sum_{j=1}^{J_i^n} \sum_{k=1}^{K_{ij}^n} \frac{1}{2} [x_{ijk}^n]_r [z_{ijk}^n]_s \Big(\log |\Sigma_{rs}^{-1}| - (y_{ijk}^n - \mu_{rs})^T \Sigma_{rs}^{-1} (y_{ijk}^n - \mu_{rs}) \Big) \right) \tag{8.19}$$

Where we have made use of the identity

$$\log |\Sigma| = -\log |\Sigma^{-1}| \tag{8.20}$$

Next we can differentiate $\mathcal{L}_{\theta_3}$ with respect to the parameters μ_{rs} and Σ_{rs}^{-1}, set to zero, and solve for the optimal values. To differentiate with respect to μ_{rs} we will find the

following relation useful

$$\frac{\partial}{\partial X}(Xa+b)^T Y(Xa+b) = (Y+Y^T)(Xa+b)a^T \tag{8.21}$$

Letting $Y = \Sigma_{rs}^{-1}$, $X = \mu_{rs}$, $b = -y_{ijk}^n$ and $a = 1$, and taking derivatives with respect to μ_{rs}, we obtain:

$$\frac{\partial \mathcal{L}_{\theta_3}}{\partial \mu_{rs}} = \frac{\partial}{\partial \mu_{rs}} E\left(\sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n}\frac{1}{2}[x_{ijk}^n]_r[z_{ijk}^n]_s(y_{ijk}^n-\mu_{rs})^T\Sigma_{rs}^{-1}(y_{ijk}^n-\mu_{rs})\right. \tag{8.22}$$

$$\left. = E\left(\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n}[x_{ijk}^n]_r[z_{ijk}^n]_s\Sigma_{rs}^{-1}(y_{ijk}^n-\mu_{rs})\right), \right. \tag{8.23}$$

and setting the derivative to zero:

$$\Sigma_{rs}^{-1} E\left(\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n}[x_{ijk}^n]_r[z_{ijk}^n]_s(y_{ijk}^n-\hat{\mu}_{rs})\right) = 0 \tag{8.24}$$

$$\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)y_{ijk}^n - E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)\hat{\mu}_{rs} = 0 \tag{8.25}$$

$$\hat{\mu}_{rs} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)y_{ijk}^n}{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)}, \tag{8.26}$$

where we have used the fact that y_{ijk}^n is observed and can be pulled out of the expectation.

To differentiate with respect to Σ_{rs}^{-1}, we will find the following identities useful

$$\frac{\partial \log|X|}{\partial X} = (X^T)^{-1}$$
$$\frac{\partial a^T X b}{\partial X} = ab^T.$$

Differentiating 8.19 with respect to Σ_{rs}^{-1}, we obtain:

$$\frac{\partial \mathcal{L}_{\theta_3}}{\partial \Sigma_{rs}^{-1}} = E\left\{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n}\frac{1}{2}[x_{ijk}^n]_r[z_{ijk}^n]_s\Sigma_{rs} - \frac{1}{2}[x_{ijk}^n]_r[z_{ijk}^n]_s(y_{ijk}^n-\mu_{rs})(y_{ijk}^n-\mu_{rs})^T\right\}, \tag{8.27}$$

and setting the derivative to zero:

$$\hat{\Sigma}_{rs}\left(\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E[x_{ijk}^n]_r[z_{ijk}^n]_s\right) - \sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\left(y_{ijk}^n y_{ijk}^{n^T} - y_{ijk}^n\mu_{rs}^T - \mu_{rs}y_{ijk}^{n^T} + \mu_{rs}\mu_{rs}^T\right)\right) = 0 \quad (8.28)$$

$$\begin{aligned}\hat{\Sigma}_{rs} =& \frac{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)y_{ijk}^n y_{ijk}^{n^T}}{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)} \\ &- \left(\frac{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)y_{ijk}^n}{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)}\right)\mu_{rs}^T \\ &- \mu_{rs}\left(\frac{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)y_{ijk}^n}{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)}\right) + \mu_{rs}\mu_{rs}^T. \end{aligned} \quad (8.29)$$

Letting $\mu_{rs} = \hat{\mu}_{rs}$, and recognizing that the terms in parentheses in Eq. 8.29 correspond to the mean estimates from Eq. 8.26, we obtain

$$\hat{\Sigma}_{rs} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)y_{ijk}^n y_{ijk}^{n^T}}{\sum_{n=1}^{N}\sum_{i=1}^{I^n}\sum_{j=1}^{J_i^n}\sum_{k=1}^{K_{ij}^n} E\left([x_{ijk}^n]_r[z_{ijk}^n]_s\right)} - \hat{\mu}_{rs}\hat{\mu}_{rs}^T. \quad (8.30)$$

Eqs. 8.26 and 8.30 provide ML estimates of the parameters of the Gaussian in Eq. x. Because of the symmetry of the structure of the observations in the network, the derivation of the Gaussian parameters θ_1 and θ_2 follow the same procedure and result, with the subindices i and ij replacing ijk respectively.

8.4 Inference

Evaluation of the ML estimates in Eqs. 8.16, 8.26, and 8.30 requires the evaluation of conditional expectations (conditioned on the observations $\{y\}$) between pairs of discrete nodes. This evaluation, corresponding to the Expectation step of the EM algorithm and known also as inference, can be carried out using the junction tree algorithm for performing general inference on graphical models (Jensen, 1996). The junction tree algorithm operates

by transforming the directed graphical representations that we have described thus far into an equivalent undirected singly connected graph (i.e., a tree) with certain properties, entering the conditioning evidence into this structure, and then running a message passing algorithm to obtain the posterior probability of any variables conditioned on this evidence.

A junction tree is a tree built out of cliques (i.e., sets of nodes in the graph that are maximally connected) that in addition has the running intersection property. That is, the unique path between any two nodes in the junction tree contains the intersection of the two nodes. Transforming an arbitrary graph to an equivalent junction tree representation (i.e., a representation that respects the same probability distribution over its variables) can be advantageous since it effectively rewrites the join probability distribution in terms of localized terms, each of which holds a local marginal over the terms.

Arbitrary graphs may generally need additional transformations before a junction tree can be assembled from it. However, the structures that we have introduced in the previous sections exhibit a graph-theoretical property known as triangulation, which effectively guarantees that a junction tree exists.[1] A graph is said to be triangulated if any cycle in its undirected graph version contains a chord. This existence result allows us to construct junction trees directly from the graphs by (i) identifying sets of maximally connected sets of nodes (i.e., cliques) and (ii) searching for a path (guaranteed to exist in this case, though not necessarily unique) between the cliques that, while obeying the running intersection property, links the nodes in a tree structure (i.e., singly linked). After a junction tree has been found, separator nodes are introduced between any adjacent cliques containing the intersection of the cliques. Once a junction tree has been built, potential functions ϕ_A are assigned to each clique A in the tree by multiplying all the conditional PDFs associated with each node in the tree:

$$\phi_A = \prod_{a \in A} p(a|\pi(a)) \tag{8.31}$$

Notice that if a node belongs to a clique, then so will its parents by construction. If a clique contains both continuous and discrete nodes, then its potential is a collection of PDFs for each combination of discrete parents. Since we are considering Gaussian continuous nodes, this amounts to keeping a table with the means and variances of each Gaussian.

The next step in carrying out inference conditioned on some observations is to introduce

[1] Triangulation, in fact, is not only a sufficient, but also a necessary condition for a junction tree to exist.

the observations as evidence into the junction tree. Evidence can be entered by selecting *one* clique containing the variable in question, and updating its potential to reflect this observation. In the case of observed discrete variables, this amounts to zeroing any configurations of the conditional multivariate multinomial not consistent with the observation. In the case of continuous valued variables, evidence may be introduced into the tree by setting the entries of the clique's potential (which, in the case treated here, contains the continuous observations and its discrete parents) to hold the probability of the observation under the different configurations of its discrete parents. Having entered evidence into the tree, the remaining clique potentials are updated until the tree is globally consistent. The details of propagating probabilities in a junction tree are detailed in Appendix C. Having outlined the main steps involved in making inference in graphical models, training the graphical models consists of iteratively introducing evidence in the model, obtaining the posterior probabilities over the latent variables and adjusting the parameters according to the update equations derived in the previous section. It can be shown (Dempster et al., 1977) that estimating the model parameters according to this Expectation-Maximization algorithm leads to an increase of the log likelihood of the training set at every iteration. The training procedure is summarized in Fig. 8-5.

8.5 Classification

Classifying a test observation can be carried out by evaluating the likelihood of that test sequence under various models and assigning it to the model that yields the maximum likelihood. As described in Appendix C, after an observation has been entered as evidence and the belief propagation algorithm has been run, each clique of the junction tree holds the joint probability of the evidence and the clique constituents. Evaluating the probability of an observation, therefore, does not require any additional algorithmic machinery. Given a trained model, we can enter the test sequence as evidence, run the belief propagation algorithm, and marginalize the clique constituents out of the joint local likelihood to obtained the desired sequence likelihood.

Given a data set of N sets of observations $D = \{\{y^n\}\}$

- Graphical Model Preparation

  1. Synthesize N graphical models and establish equivalence classes among nodes of the networks as outlined in the previous sections.
  2. Initialize the parameters of each equivalence class $\psi^{(0)}$.
  3. Construct a junction tree for each network.
  4. Enter each observation as evidence into a clique containing the observation node.

- For $k = 1, 2, \cdots$
  E Step:

  5. Evaluate the sets of posterior probabilities of the latent nodes given the current parameters and conditioned on the entered evidence needed to evaluate Eqs. 8.16, 8.26 and 8.30. This step also yields (see Appendix C) the log likelihood of each observation $\mathcal{L}_n^{(k)} = \log p(\{y^n\}|, \psi^{(k-1)}, D)$ given the current parameters and the data set , and hence the total log likelihood $\mathcal{L}^{(k)} = \sum_n \mathcal{L}_n^{(k)}$ of the training set D.

- M Step:

  6. Update the parameters of the model according to Eqs. 8.16, 8.26 and 8.30 to obtain $\psi^{(k)}$.

- Diagnose Convergence

  7. If $\mathcal{L}^{(k)} - \mathcal{L}^{(k-1)} > \epsilon_{conv}$, $k \leftarrow k + 1$; go to step 5. Otherwise, stop.

Figure 8-5: General algorithm for training Graphical Models.

8.6 Modeling Prosodic Phenomena with Bayesian Networks

Different aspects of speech take place at different time scales. Some of them may be described over longer spans, like the volume setting over a whole utterance, whereas others are meaningful over a finer scale, like information related to syllable stress. The hierarchical nature of prosodic structure was already discussed in Chapter 2. This view about the organization of speech helps us motivate applying the hierarchical-dynamic networks we have been describing to modeling prosodic-acoustic parameters. The dynamic dimension allows for modeling changes in these parameters over time whereas the hierarchical dimension allows us to model parameters observed or analyzed at different time scales.

Figure 8-6 illustrates a simple proposal of how we can use the acoustical analysis that we have developed in the previous chapters to provide inputs to the model. The hierarchi-

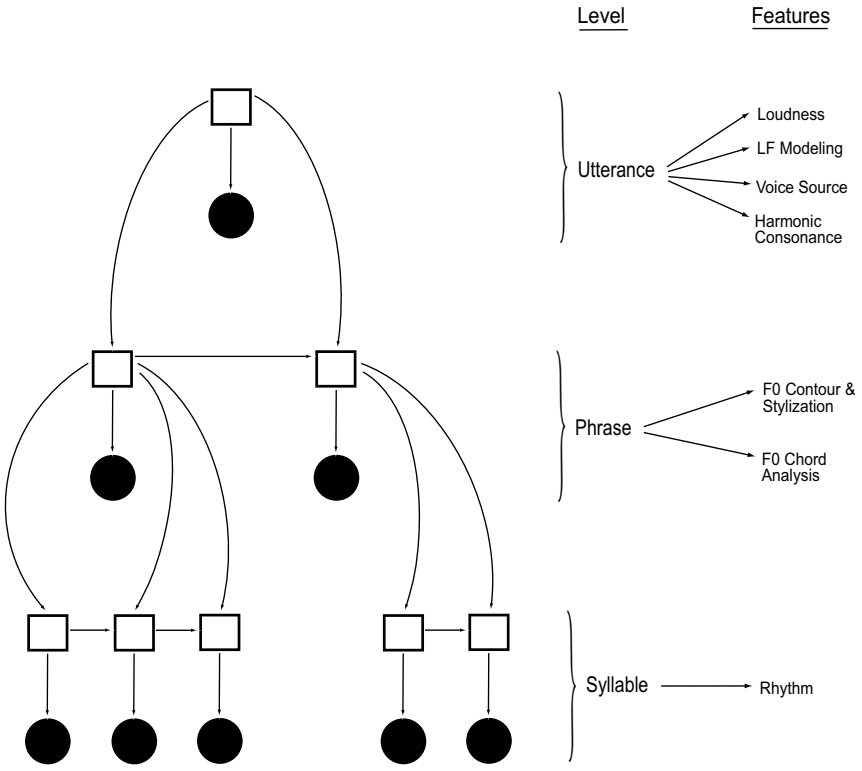

Figure 8-6: Representation of prosodic-acoustic parameters using Bayesian networks.

cal model shown attempts to model three different kinds of scales in speech, as shown by the second column: a more global scale models settings that are well described in terms of aggregate measures over a longer time span such as an entire utterance; a middle scale models intermediate phenomena, describable in terms of phrases; finally, a finer scale models

shorter-term phenomena, such as those taking place over syllables. It should be clear now that when we discuss the dynamic evolution of acoustic measures over time, we do not mean a uniform time window, but rather a unit in terms of the structural category that a particular level of the hierarchy models. It should also be clear that since different utterances differ structurally, we do not obtain homogeneous networks (one having the same nodal representation at every level). For that reason, our treatment of the model in the previous section allowed different networks to differ in structure, and only grouped nodes meaningfully by establishing equivalence classes between them (within and across networks). The third column in the figure summarizes a particular proposal to assign different acoustic measurements to structural time scales. We have included measurements of loudness and all the voice quality measurements at this level. Although voice quality parameters can definitely change and be perceived at smaller time scales, a whole utterance might represent a suitable unit for the description of paralinguistic voice quality (Laver, 1994). To the intermediate time scales we have assigned all features related to the intonational description of the speech utterance since phrases provide a natural structural unit over which to analyze these phenomena. Finally, the syllable level is used to represent rhythmic features since they have been defined over that domain. In this particular illustrative example, a single utterance is composed of two phrases, each of which holds three and two syllables respectively.

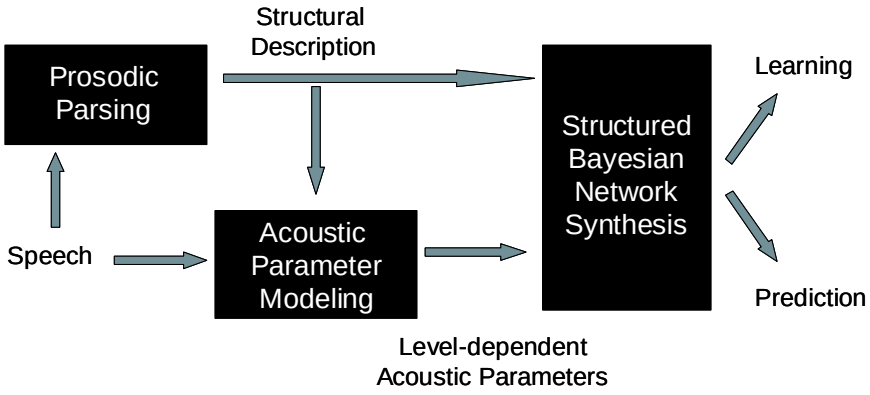

Figure 8-7: High-level architecture of the model.

The high level description of the architecture of the system is shown in Fig. 8-7: A structural representation of speech is first obtained since it informs the subsequent processing blocks in the system. This structural description defines spans of time over which we

restrict the acoustical analysis, and so this representation, together with the acoustic waveform, becomes the inputs to an acoustical analysis module which outputs level dependent acoustic features. A graphical model is then synthesized to match the structural description of speech, that is, to allocate hidden state nodes and link them in a way that follows the particular parsing of the waveform. The network also allocates observation nodes to store the acoustic measurements extracted from each unit. If the network is being generated for learning, its parameters may be randomized or set to specific initial values and then submitted to the parameter adjustment algorithms described earlier. If the network is being synthesized for classification (i.e., we wish to evaluate the likelihood of the observations under a particular set of equivalence classes), then its parameters are matched to those of a model already learned, and the likelihood evaluated.

8.7 Chapter Summary

In this chapter we have proposed a family of Bayesian networks or graphical models suitably designed for modeling observations which are best described in terms of a temporal as well as a hierarchical dimension. We have addressed the issue of learning the parameters of these models, and motivated their application to modeling distributions of prosodic-acoustic features. Different prosodic phenomena take place at different scales governing the structure of spoken language. The contribution advanced in this chapter lies in the application of these structured graphical models as a general and natural formalism to integrate time-evolving hierarchically governed acoustic parameters into a single tractable model. The graphical structures introduced in this chapter are attractive from a design point of view since, as is often the case with graphical models, the particular structuring and linking of nodes offer an amenable and semantic interpretation that summarizes dependencies in an atractive way. The generality of the formalism behind these models also allow a direct extension to modeling phenomena taking place at different scales from those considered here (they could easily be extended to include higher discourse-level structure, for instance).

Part III

Experiments and Evaluation

Chapter 9

Perceptual Experiments

9.1 Introduction

In this chapter we address the issue of how to annotate corpora of affective speech with perceptual labels. The contents of this chapter represent a momentary departure from the modeling topics which have been the object of this thesis thus far. However, the work described here has been a substantial component of this thesis research given the need to work with data annotated with affective content for implementing the automatic learning schemes which lie at the heart of this research, and given the lack of standardized speech corpora for achieving this purpose. In what follows we address a series of experiments carried out for that purpose.

We are interested in gaining grounding for tokens of spontaneously occurring affective speech, where no scripting or affect-elicitation procedures have been used to procure the speech. We are furthermore interested in labels which reflect the perceptual interpretation (given perhaps a constrained *emotion vocabulary*) from the listener's point of view, regardless of the speaker's intentions. The focus on the *expression* of affect, rather than the *affective experience* of the speaker, answers not only to a strong practical constraint but also to an ethical consideration. The practical issue avoids requiring access to a report of the speaker's own annotations. Even if self-reported labels are accepted as reliable indicators of the affective content of an utterance, this approach rules out working with already assembled datasets for which the speaker's self report is no longer available. The ethical consideration bears on the focus of this research on building computational systems that aim to perceive the range of human expression that humans intentionally use for

communicating[1], irrespective of the affective experience they may attempt to disguise or conceal.

In addition to the experiments we are about to describe in this chapter, we conducted an additional perceptual experiment with the goal of investigating how the intended labels annotating the actors data were perceived by human listeners. This experiment was more informal in nature, and its goal was not to re-annotate the data set with perceptual labels, but rather to investigate how well humans performed on a task for which there was a more "objective" external ground truth (i.e., intended emoted labels) for the purpose of establishing comparison with the performance of automatic recognition algorithms. The details of this experiment are deferred until Chapter 10, when human and machine recognition performance are discussed.

9.2 Experimental Setup

In order to gain an understanding of how listeners perceive affect, we have decided to adopt the valence-arousal dimensional model of affect already introduced in Chapter 2. Although this model has its shortcomings, it offers nonetheless a tractable tool to describe the affective content of speech tokens while skirting the issues associated with categorical models, specifically how to select a set of candidate categories *a priori*. Furthermore, its validity has already been tested in other experiments in the context of evaluating affect in the speech modality (Pereira, 2000; Cowie et al., 2001; Cowie & Cornelius, 2003). Although the final aim of this work is to build models for the classification of affect categories, we find it prudent not to introduce any categorical distinctions at this stage, and to provide listeners with an *unquantized* representational framework on which we can impose any discretization at a later stage for modeling purposes.

A graphical user interface was designed in Java to give users an intuitive tool to navigate through the space to locate a position that reflected the perceived relative valence and arousal ratings for each sound clip. Screenshots of the interface are provided in Figs. 9-1 through 9-4 to illustrate the flow of the experiment. After reading a series of instructions, the subject is taken to the experiment page looking like the screenshot in Fig. 9-1. The page consists of a square grid (648 by 648 pixels) representing the valence-arousal space

[1]This is a daunting challenge by itself!

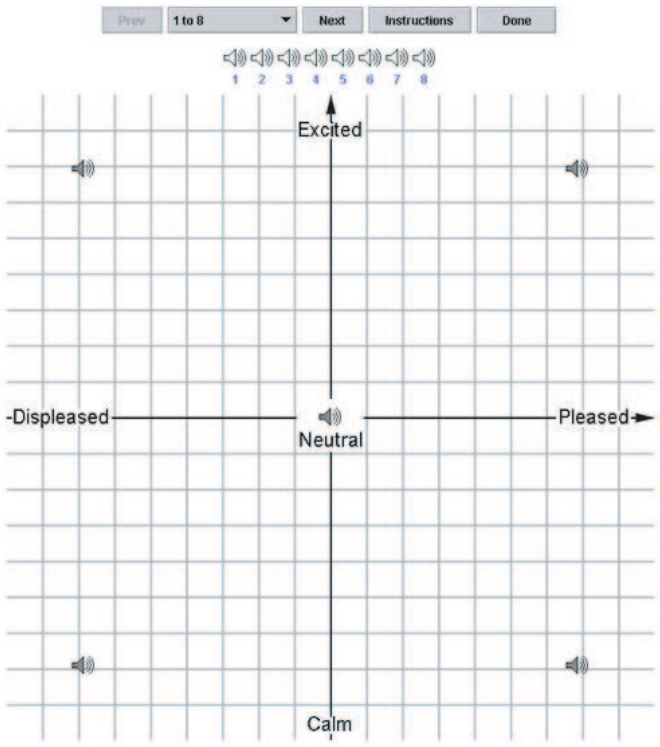

Figure 9-1: GUI I: Appearance of the graphical interface used in the perceptual experiment. Items to rate are aligned on top of the grid below a tool to navigate through the experiment.

and a series of icons linked to audio recordings. The five icons on the center and extrema of the quadrants are provided to relatively calibrate the space by providing anchor points which the experimenters thought serve as exemplars of the the extrema of the perceptual space for a given data set (and of the neutral position). The sound items linked to the five reference icons can be specified by the experimenter externally in a parameter file, and can therefore be adjusted for each data set. To allow for the possibility that a subject may find a test item to exceed the rating typified by the reference icon on any given quadrant, the references have been placed with some indentation from the edges. It is therefore possible to place test items outside of the boundary defined by the reference icons. The subjects are instructed to rate the test items with respect to these five points. Throughout the experiment, a subject may click on the reference icons as needed. Since *valence* and *arousal* were thought to be somewhat technical terms with which not every listener may be familiar, the meaning of each axis was conveyed by recasting the valence dimension in terms of a continuous *displeased-pleased* scale. Similarly, arousal was recast in terms of a *calm-excited* dimension.

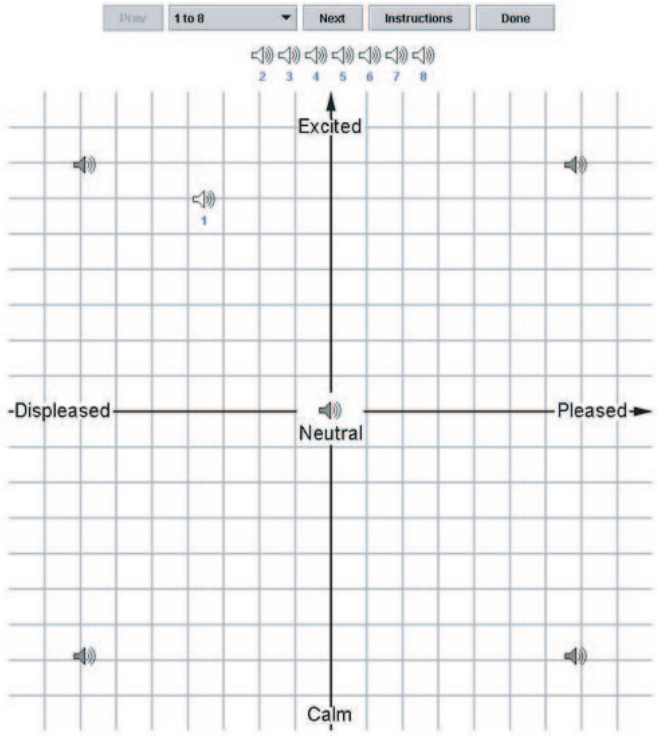

Figure 9-2: GUI II: The user can place the icons on the grid to reflect his perception of the valence and arousal ratings with respect to the reference points.

The test items are lined up at at the top of the grid, each one represented by a numbered icon linked to a sound clip. Listeners can click on an icon as needed, and drag it to the valence-arousal map to a desired position. They can also refine the position of a test item as wished throughout the experiment. The entire experiment consists of repeating this basic rating task for several "pages" containing a number of distinct sound clips. The number of sound clips per page is one of several parameters which the experimenter can control through the external parameter file. A subject can use the navigation tool at the top of the interface to advance through the pages, either sequentially by clicking on the *Prev* or *Next* buttons, or randomly by using the pull-down menu containing a list of icons (see Fig. 9-3). A subject may go back and re-read the instructions at any point during the experiment by using the *Instructions* button on the navigation tool.

A subject does not need to place all the icons within a page to advance to the next, and can freely move forward and backward as needed to alter the position of each test item. Once the subject has rated all the items, he may exit the experiment by pressing the *Done* button. If the subject, however, attempts to exit the experiment before all items have been

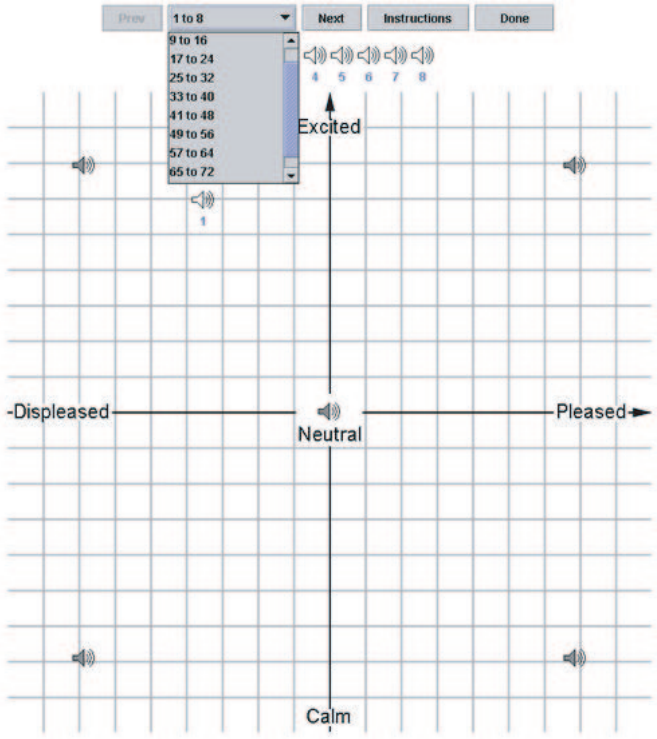

Figure 9-3: GUI III: The user can use the navigation tool to advance to a different screen containing a different set of test items.

rated (i.e., placed within the confines of the graph), an error window alerts him about this and prevents him from exiting the experiment. This is shown in Fig. 9-4:

After finishing the experiment, the subject is taken to a questionnaire where he is asked to disclose several self-report variables and is prompted for general feedback. A copy of the questionnaire is shown in Appendix E. Besides collecting information regarding the subject's status as a native English speaker and any known hearing problems, each subject is asked for three ratings on a 7-point scale. The questions prompt the user to rate how complex he found the task, how confident he is with the ratings provided, and how perceptive he thinks of himself at decoding the affect of others. There is also a blank space provided for any general free-form comments.

The interface went through a series of iterative designs before reaching the final version shown here. Informal pilot studies were conducted with several subjects at various points to gather feedback on the interface design. It became clear during the early trials that subjects needed some idea of the meaning associated with the extremes of each axis, and therefore we provided sound clips as anchor points on subsequent versions. The neutral

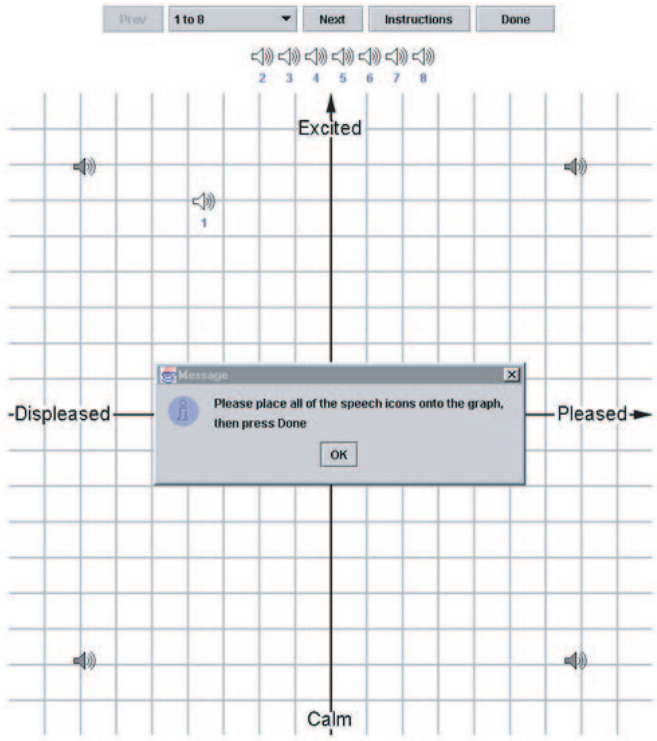

Figure 9-4: GUI IV: The user must finish rating all the test items before exiting the experiment.

icon, however, was a later addition suggested by the first round of experiments, after a few subjects expressed the need for further clarification of what it was meant by *neutral*. Of the two experiments carried out with this system, and discussed in the next section, only the second one exploited the full interface; the first one was similar in every respect except for the missing link to a neutral-sounding token.

Since typically a data set under evaluation would contain more items than it would be deemed reasonable for a single subject to rate in a single session, the interface accepts an experiment file at the beginning of the session. This experiment file contains a listing of the subset of the data set that a subject is to evaluate. An entire corpus can then be divided a priori into subsets to have each subject rate a portion of it. The total number of test items to consider within a single experiment was investigated in the pilot studies. Subjects seemed rather flexible with respect to this option, offering sometimes to rate as many as one hundred clips during a pilot study in which they were presented with a data set of acted speech. The number of sound clips displayed on each page was determined by a compromise between several factors: a larger number of test items on a single page

could help provide internal consistency between the items' ratings since the subjects are encouraged to position the icons not only with respect to the reference points but also with respect to each other. On the other hand, an unreasonably large set to work with at once would also make the task more tedious to the user. Based on users' feedback, it was decided to set this number at 8. Before a subject arrives, the experimenter loads the program with a suitable experiment file containing the subset to rate and a corpus-dependent parameter file containing the specifications discussed above. It was found that most users "enjoyed" the task, finding the interface "easy to use," "straightforward," and "intuitive." A few thought there were too many test items and found the task tiresome after a while. Some subjects remarked on the difficulty of rating test items for which they lacked the "context," or "background"; some found it difficult to separate the semantic or pragmatic meaning from the prosodic expression (they were instructed to pay attention foremost to *how* the speakers sounded regardless of *what* they were saying). One subject, a Mechanical Engineering student, expressed a preference for a system that would use polar coordinates instead.[2] A large majority of subjects liked the idea of a continuous map where they were free to move around and place their ratings without categorical constraints.

9.3 Data Sets

Two data sets were submitted to the perceptual rating experiment just described. The first set consisted of 304 tokens assembled from the OASIS database, a set derived from conversations recorded by British Telecom (BT) between operator service agents and customers in the United Kingdom (Durston et al., 2001). This database was collected by BT for use as a research resource to encourage the development of technology to enhance its customer care. The classification of emotional state in call center speech is an issue that is expected to be useful to manage agent performance, and also to improve the usability of automated spoken dialogue systems for customer care. The final set used in the experiment was narrowed down from an initial set of tokens pre-screened at the company for potential affective content. Starting from this initial set, two other judges, working separately and in tandem, iteratively selected tokens to assemble a set of manageable working size

[2]This is a very sensible suggestion, indeed. See Cowie et al. (2001) and Cowie & Cornelius (2003) for a discussion of Plutchik's "emotion wheel" and of an interface similar to ours that uses this idea by encoding ratings primarily as distances from the center.

which in their opinion spanned a good portion of the valence-arousal affective space. The recordings were originally made available in 8kHz 16-bit encoded files in WAV format. A root-mean-square normalization procedure was applied to compensate for minor recording differences, and this final set (which we shall denote as the Call Center dataset) was used in the perceptual tests.

The second data set was assembled from recordings taken from the CallHome American English Speech corpus available through the Linguistic Data Consortium (LDC Catalog No. LDC97S42). Though originally conceived primarily with speech recognition applications in mind, the CallHome corpus contains a good sampling of affective discourse, as it mostly consists of calls made between friends and relatives. The same procedure implemented with the call center data was used to construct a data set of affective tokens from this corpus. Two judges listened to several hours of audio and screened potential segments of conversation, trying to produce a sampling which in their opinion covered as much of the affective valence-arousal map as possible. The fragments were iteratively refined, discarded, or further ones were added until a final set comprising 254 tokens was obtained. The original dual-channel, 8kHz, 8-bit μ-law-encoded recordings were re-encoded in a 16-bit linear format, and the root-mean-square value normalized across recordings, as one of the two channels showed significant attenuation in many cases. This procedure yielded more balanced recording across the set, with better audio quality properties. It was the set thus normalized (and which we shall henceforth denote as CHE) that was used in the perceptual experiment.

For each of the data sets above, a set of reference icons were determined to help guide the subjects through the rating task. They were selected as follows: First a judge familiar with the goal of the experiment pre-screened the speech corpus from which the test corpus was being assembled to come up with a list of candidates for the reference positions. These were submitted to other judges, also familiar with the task, who narrowed down the set and in some cases suggested additional members from the corpus that they thought were better candidates. To resolve any remaining discrepancies, a few (typically about 3) items were submitted to other judges not directly working on the experiment design to gather extra input concerning which one would serve as a better exemplar of an extremum (or neutral) position on the map. The reference icons associated with each data set (four for the Call Center data and five for the CHE set) were not replicated in the set of items to be rated.

As mentioned earlier, subjects in pilot studies were willing to rate up to 100 items from

a corpus of acted speech. Since these final studies, however, contained spontaneous non-acted speech data where the signaled affect was thought to be potentially more subtle and require more introspection and paused analysis from the listeners, it was decided to set the total number of test items per experiment around 75 for the first experiment. Based on the feedback from the users carrying out this task and the time they devoted to completing each session, it was felt that it would be safe to ask subjects to rate 85 items on the second task.

9.4 Data Analysis

9.4.1 Performance Statistics

In addition to the self-report variables collected from each subject at the end of his experimental session, we recorded the subject's age and gender. In addition to these variables, we are interested in defining some measures that allow us to quantify each subject's ratings with respect to the pooled ratings from all subjects. Such a measure would allow us, for instance, to detect any outlier subjects, subjects that consistently produce ratings that deviate from the distribution of the collective ratings averaged over the entire dataset. By experimental design, each subject produced a partial annotation of the data set. Before postulating a set of performance statistics, let us first define the following auxiliary quantities involved in the analysis. Let N_s be the number of subjects from whom valid ratings are available and N_t the number of speech tokens contained in the corpus. Let I be a $N_s \times N_t$ matrix of binary entries, where $[i_{jk}]$ is one if the j-th subject provided a rating for the k-th token, and zero otherwise. Let V and A be similarly defined, with $[v_{jk}]$ and $[a_{jk}]$ containing the valence and arousal ratings from the j-th subject for the k-th token (if the subject evaluated the token) and zero otherwise. This encoding conveniently allows us to evaluate the rows and columns of the matrices to evaluate the performance of a subject, or the results for a particular speech token.

For the k-th token, let us then calculate the centroid and scatter of the ratings distri-

bution for each dimension (valence and arousal) as

$$\mu_{v_k}^{50} = \underset{j:i_{jk}\neq 0}{\text{median}}(v_{jk}) \tag{9.1}$$

$$\sigma_{v_k} = \left(\frac{1}{\sum_j i_{jk} - 1} \sum_{j:i_{jk}\neq 0} (v_{jk} - \mu_{v_k}^{50})^2\right)^{\frac{1}{2}} \tag{9.2}$$

$$\mu_{a_k}^{50} = \underset{j:i_{jk}\neq 0}{\text{median}}(a_{jk}) \tag{9.3}$$

$$\sigma_{a_k} = \left(\frac{1}{\sum_j i_{jk} - 1} \sum_{j:i_{jk}\neq 0} (a_{jk} - \mu_{a_k}^{50})^2\right)^{\frac{1}{2}}, \tag{9.4}$$

where the median, instead of the mean, has been used as a robust estimate of each cluster's centroid to ameliorate the effect of outliers.

Let us then transform each rating to a z-score measuring the relative deviation from the cluster's center (for each dimension):

$$z_{v_{jk}} = i_{jk}\frac{v_{jk} - \mu_{v_k}^{50}}{\sigma_{v_k}} \tag{9.5}$$

$$z_{a_{jk}} = i_{jk}\frac{a_{jk} - \mu_{a_k}^{50}}{\sigma_{a_k}} \tag{9.6}$$

And finally, define the following subject-dependent statistics over the collection of absolute normalized z-scores from each subject:

$$\mu_{z_{vj}} = \frac{1}{\sum_k i_{jk}} \sum_{k:i_{jk}\neq 0} |z_{v_{jk}}| \tag{9.7}$$

$$\sigma_{z_{vj}} = \left(\frac{1}{\sum_k i_{jk} - 1} \sum_{k:i_{jk}\neq 0} (|z_{v_{jk}}| - \mu_{z_{vj}})^2\right)^{\frac{1}{2}} \tag{9.8}$$

$$\mu_{z_{aj}} = \frac{1}{\sum_k i_{jk}} \sum_{k:i_{jk}\neq 0} |z_{a_{jk}}| \tag{9.9}$$

$$\sigma_{z_{aj}} = \left(\frac{1}{\sum_k i_{jk} - 1} \sum_{k:i_{jk}\neq 0} (|z_{a_{jk}}| - \mu_{z_{aj}})^2\right)^{\frac{1}{2}}. \tag{9.10}$$

The statistics in Eqs. 9.7 and 9.9 measure how much, on the average, a subject's ratings deviate from a cluster's center, along the valence and arousal dimensions, taking into account the spread of the cluster. Intuitively, they measure how much a subject tends to provide prototypical (with respect to the other subjects) ratings, since prototype ratings exhibit a

low absolute z-score. The statistics in Eqs. 9.8 and 9.10, on the other hand, measure the spread of the absolute normalized scores. Intuitively, they reflect how consistently a subject tends to give ratings of the same typicality, with consistently typical or consistently atypical ratings achieving low variance of absolute z-scores. The statistics in Eqs. 9.7-9.10 define the four subject-dependent performance statistics which will be considered in the analysis of the subject's responses.

9.4.2 Call Center Data

The four statistics just described were considered in addition to three personal statistics (age, gender and native speaker status), and three self-report variables related to the task (self-assessment in perceptively decoding the affect of others; confidence in the ratings provided; and task complexity) from an evaluation of a corpus of 304 speech tokens assembled from a set of call center quality-monitoring recordings. Twenty-nine subjects participated in the experiment, 15 of them female, ranging from 17 to 60 years of age, with a median age of 22. The reported performance-related measures, all on 7-point scale, were as follows: The mean reported task complexity was 3.62 with a standard deviation of 1.54; values ranged from 1 to 6. Reported perceptivity ranged from 1 to 5, averaging 2.93 with a standard deviation of 1.33. Subjects' confidence on their ratings ranged from 1 to 6, with a mean of 3.1 and a standard deviation of 1.3. Noteworthy is the fact that no subject chose the extreme value of 7 to characterize the task as very difficult, or themselves as very unperceptive or not at all confident of their ratings.

Multiple analysis of variance between the variables showed only a statistically significant dependency between gender and the statistic σ_{z_v} at the $p = 0.028$ level, as shown in Fig. 9-5. This effect is not too strong. However, it tells us that female subjects tended to provide more consistent ratings (whether typical or not) for the valence dimension than male subjects did. No other effects were found through ANOVA between the performance statistics and the groups defined by each discrete value in the 7-point scale for the self-report variables. However, a correlation analysis between the variables produced a few results significant at the $p < 0.05$ level. One of these effects consisted of a correlation between the reported perceptivity and confidence. As shown in Fig. 9-6, subjects who tended to rate themselves on the endpoints of the range of the scale (1 - 5) of perceptivity also tended to rate themselves on the extrema of the confidence scale.

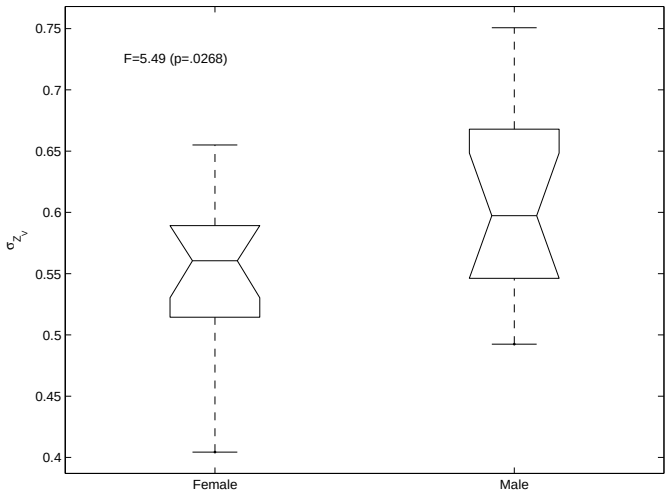

Figure 9-5: Boxplot of the distribution of the statistic σ_{z_v} according to gender (Call Center experiment).

A stronger effect, at the $p = 0.01$ level, is shown in Fig. 9-7 concerning an inverse correlation between subjects' reported perceptivity and the statistic μ_{z_a}. The result states that subjects who rated themselves as very perceptive tended to give less typical arousal ratings than subjects who rated themselves as less perceptive. Since the typicality is judged with respect to other subjects' ratings whereas the perceptivity is a self-reported variable, the result is not particularly surprising (or informative?) in principle.

Finally, Fig. 9-8 shows plots of significant correlations obtained between various pairs of the performance statistics. The most significant of these, shown in the upper-right and lower-left panels of the figure, established a positive correlation (at the $p < 10^{-8}$ level) between the mean and the standard deviation of the normalized z-scores for each subject, both for the valence and arousal dimensions. In other words, subjects who provided typical ratings tended to give consistent ratings as well, the converse holding as well.

Since the ultimate goal of the experiment is to provide annotations to the corpus, it is of interest, therefore, to assess in particular the degree to which the collected ratings can be taken as reliable indicators of a point in the valence-arousal space for each token. We will return to this point in section 9.5, but at least intuitively, we would expect good rating agreements to occur when the distribution of values do not differ greatly from a prototype rating. If we take the prototype to be defined by the median of the rating ensemble (Eqs. 9.1 and 9.3) for each token, then Eqs. 9.2 and 9.4 model the scatter of points around this prototype and provide an initial suitable measure of the goodness of representing the

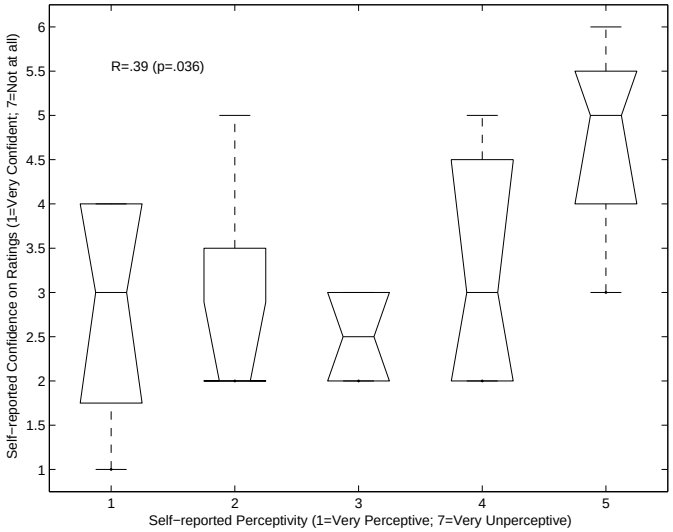

Figure 9-6: Boxplot of the distribution of self-reported rating confidence according to self-reported perceptivity (Call Center experiment).

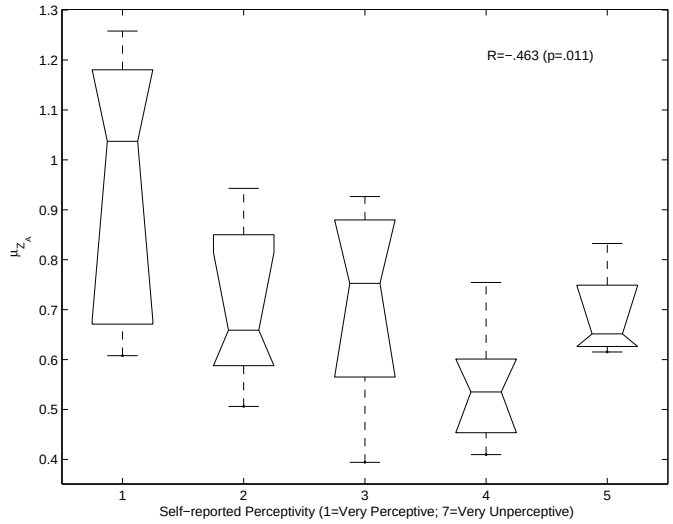

Figure 9-7: Boxplot of the distribution of the statistic μ_{z_a} according to self-reported perceptivity (Call Center experiment).

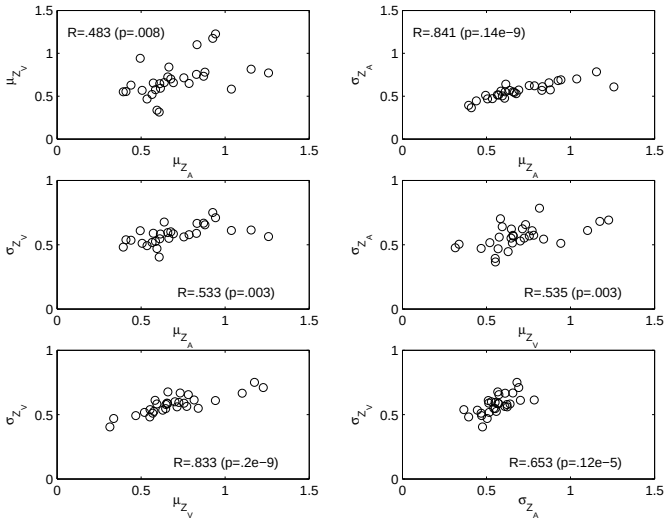

Figure 9-8: Plots of pairwise statistics significantly correlated at $p \leq 0.05$ for the Call Center experiment. The correlation coefficient and p-value are indicated for each case.

ensemble by the prototype. The two panels in Fig. 9-9 show the distribution of $\sigma^2_{v_k}$ and $\sigma^2_{a_k}$ for all tokens of the Call Center database.

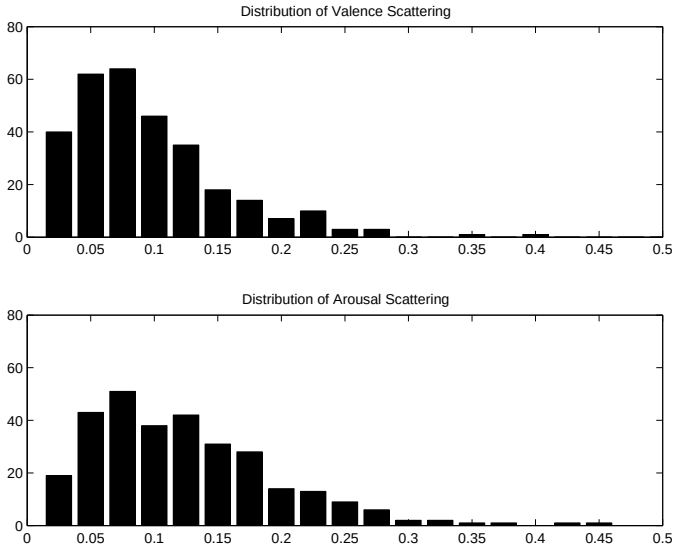

Figure 9-9: Distribution of valence and arousal scatters of rated tokens from the Call Center database.

9.4.3 CHE Data Set

Twenty subjects participated in a perceptual experiment to rate the CHE set. Fourteen of these subjects were female, and all but one reported to be native speakers of English.

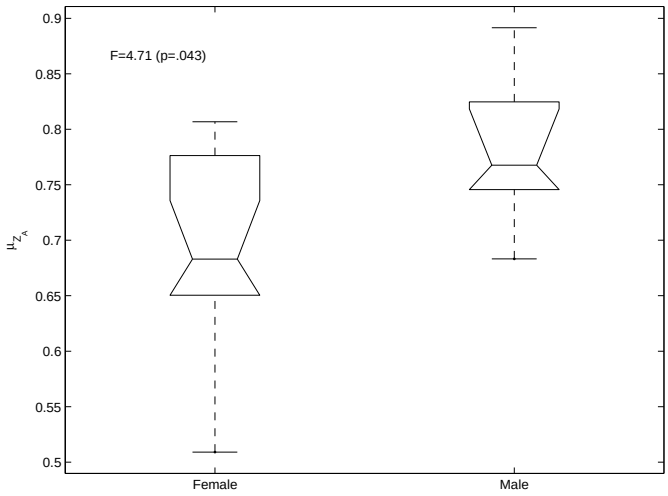

Figure 9-10: Boxplot of the distribution of the statistic μ_{z_a} according to gender (CHE experiment).

Subjects ranged from 18 to 43 years of age, with the median age at 27. The following descriptive statistics summarize the self-report variables: Perceptivity was reported at a mean value of 2.37, ranging from 1 to 6. Subjects reported an average confidence level of 2.9 on the ratings they provided, ranging from 1 to 6, and an average task complexity of 3, ranging from 1 to 5. As was the case with the Call Center experiment, no subject chose to rate any variable on the "negative" extremum of the 7-point scale.

Analysis of variance showed only effects between gender and the statistic μ_{z_a}, as shown in Fig. 9-10. Although the effects are only significant at the $p = 0.043$ level, they reveal a tendency for females to provide more typical arousal ratings (in the former experiment, the observed tendency was to provide more consistent valence ratings).

Correlation analysis also revealed some effects at a significance level $p < 0.05$. Subjects showed a positive correlation between their self-reported perceptivity and the confidence on their ratings (Fig. 9-11), a result which is consistent with the previous experiment and which suggests they may be addressing the same issue asked by these two different questions. Self-reported perceptivity also strongly interacted with task complexity (Fig. 9-12), an effect that was unobserved previously. Fewer correlations between pairs of the performance statistics (summarized in Fig. 9-13) were obtained in this experiment, but they are all consistent with the results previously obtained. As before, typicality correlated positively with consistency in both dimensions (top panels of Fig. 9-13). In addition, consistency of valence judgments correlated positively with consistency of arousal judgments (lower panel).

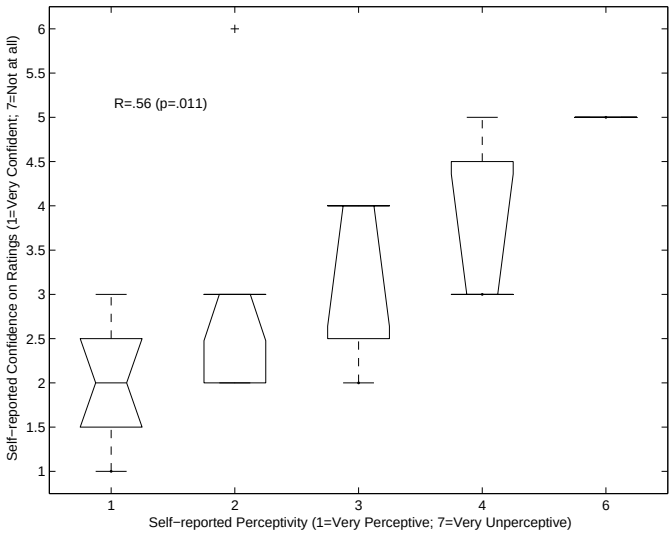

Figure 9-11: Boxplot of the distribution of self-reported rating confidence according to self-reported perceptivity (CHE experiment).

Finally, the distribution of the scatters of the collected valence and arousal ratings, which we are proposing to use as a measure of the fitness of the ratings, is shown in Fig. 9-14.

9.5 Discussion

Figs. 9-9 and 9-14 show the distributions of of the scatter values obtained for each of the independent perceptual experiments with the two data sets described in the previous section. It can be noted that in the first experiment the affective sound samples tended to be rated more tightly along the valence dimension, as evidenced by the decay rate of the histograms in Fig. 9-9: the wider tail in the arousal scatter histogram indicates a larger number of tokens that received a less coherent rating along that dimension. This effect was somewhat reversed in the second experiment (Fig. 9-14), although the difference between histograms for the CHE set, discounting outliers, is not that significant.

The wider variance along the valence dimension for a larger number of tokens in the first experiment set may be explained, at least partially, by appealing to dialectal differences between the speakers and the subjects performing the ratings. The Call Center data set consisted, almost exclusively, of recordings of speakers of various representative dialects of the United Kingdom (some speakers also showed a notable foreign influence). The raters, on the other hand, were for the most part native speakers of North American English, or

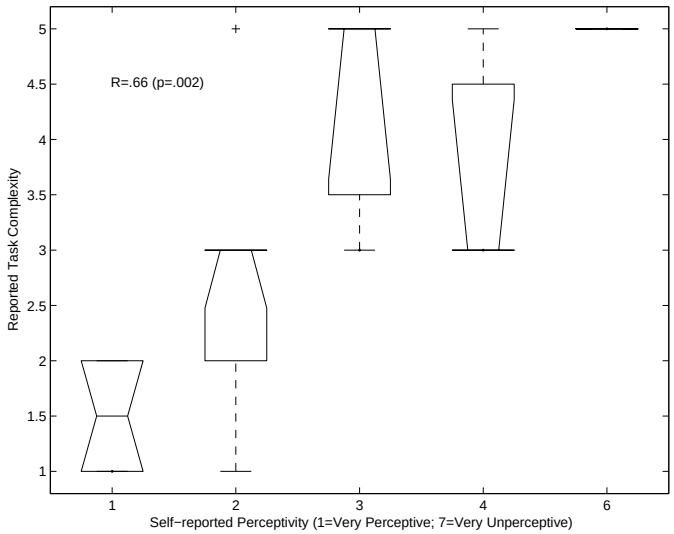

Figure 9-12: Boxplot of the distribution of self-reported rating confidence according to self-reported task complexity (CHE experiment).

non-native speakers residing in North America. This mismatch did not occur for the CHE, where most of speakers and listeners employed some North American dialect. It is possible that valence decoding may be more susceptible to the influence of such factors as cultural background, and that arousal ratings are perhaps more "universal," more transparent to listeners from a different language or dialect.

Although certain studies have looked at cross-linguistic interpretation of affect from speech (Abelin, 2000; Tickle, 2000), and affective states involving higher cognitive appraisal are thought to be more language-dependent (in their encoding and perception) than the more prototypical (or so-called basic) emotions (e.g., fear, anger, sadness, etc.), it remains to be further investigated whether the valence and arousal dimensions fare differently, and how differently, when there is a mismatch between speaker and listener. Whatever the cause for the effect described, it is certainly the case that many subjects remarked on the fact that they were unfamiliar with the particular dialect of the speakers, and that this made the task of judging the tokens more difficult.

As pointed out earlier, the scatter values provide us with an initial estimate of the fit of a prototypical judgment for a given token to the collection of perceptual ratings obtained for that token. We could arguably use this measure to select a subset of reliable tokens for which the scatters do not exceed a certain threshold. However, when inspecting the subjects' ratings to obtain some qualitative patterns, it becomes clear that ratings are not

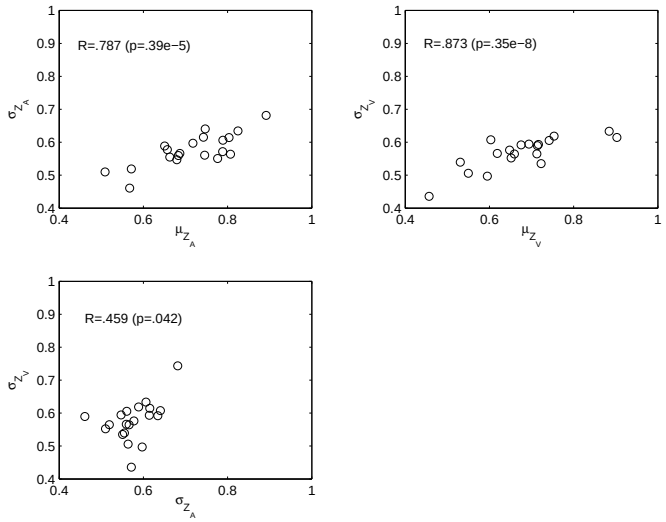

Figure 9-13: Plots of pairwise statistics significantly correlated at $p \leq 0.05$ for the CHE experiment. The correlation coefficient and p-value are indicated for each case.

always randomly distributed around the centroids defined by Eqs. 9.1 and 9.3 (see Fig. 9-15). There are several cases where the presence of outliers makes the statistics defined by 9.2 and 9.4 unsuitable for distinguishing between a spread cluster, and a tight cluster with outliers. To compensate for outliers, therefore, we will use a more robust statistic of the spread of a cluster by computing the second central moment (around the mean) of the points lying within the lower and upper p-percentile of the data set; that is, we trim the data set before computing the spread statistic: Let x_p stand for the p-th percentile value of a data set x, and define the following trimmed scatter measures for the valence and arousal ratings:

$$\sigma_{v_k}^{tr} = \left(\frac{1}{\sum_{j:v_{15} \leq v_{jk} \leq v_{85}} i_{jk} - 1} \sum_{j:\left\{ \begin{array}{l} i_{jk} \neq 0 \\ v_{15} \leq v_{jk} \leq v_{85} \end{array} \right.} \left(v_{jk} - \mu_{v_k}^{50} \right)^2 \right)^{\frac{1}{2}} \tag{9.11}$$

$$\sigma_{a_k}^{tr} = \left(\frac{1}{\sum_{j:a_{15} \leq a_{jk} \leq a_{85}} i_{jk} - 1} \sum_{j:\left\{ \begin{array}{l} i_{jk} \neq 0 \\ a_{15} \leq a_{jk} \leq a_{85} \end{array} \right.} \left(a_{jk} - \mu_{a_k}^{50} \right)^2 \right)^{\frac{1}{2}}. \tag{9.12}$$

This trimmed scatter measure proves to be more robust to the effects of a few outlier ratings. The value of p is chosen low enough so that only the effects of a few outlier ratings are ignored (setting p higher can yield values that are too forgiving since only a few

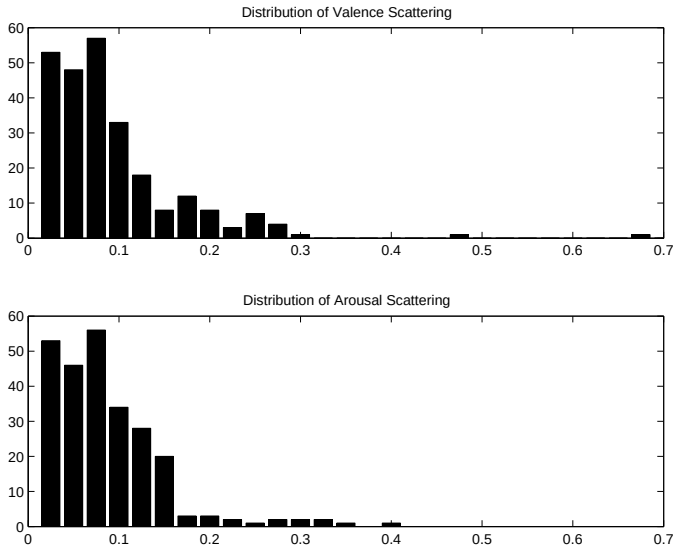

Figure 9-14: Distribution of valence and arousal scatters of rated tokens from the CHE database.

judgments are available for each token, and a considerable portion of the ratings could be thus discarded with a high p value). The statistics from Eqs. 9.11 and 9.12 are shown at work in Fig. 9-15 for several qualitatively different sets of ratings. The figures are arranged so that the coherence of the ratings roughly decreases sequentially from the upper-left to the lower-right panels. The top row shows a series of judgments that are fairly consistent with each other; the middle panel shows cases with few outliers, and the lower row shows cases with a fair amount of disagreement between subjects.

One way to quantitatively evaluate the results of the perceptual experiments is to define the coherence of a series of ratings for the k-th token as

$$\sigma^{tr}_{max_k} = \max(\sigma^{tr}_{v,k}, \sigma^{tr}_{a_k}), \tag{9.13}$$

a variable used to retain or discard judgments. Judgments that exceed a certain threshold can be considered too unreliable and left out of further analysis, re-evaluated with a larger number of subjects, etc. For each of the data sets evaluated using this experimental approach, Fig. 9-16 shows the percentage of the data set which falls below each threshold value and hence the amount of data discarded for each desired level of coherence. It is notable that the CHE set contains a larger percentage of tokens with coherent ratings for any given threshold value than the Call Center data. For instance, for a value of $\sigma^{tr}_{max_k} = 0.0875$,

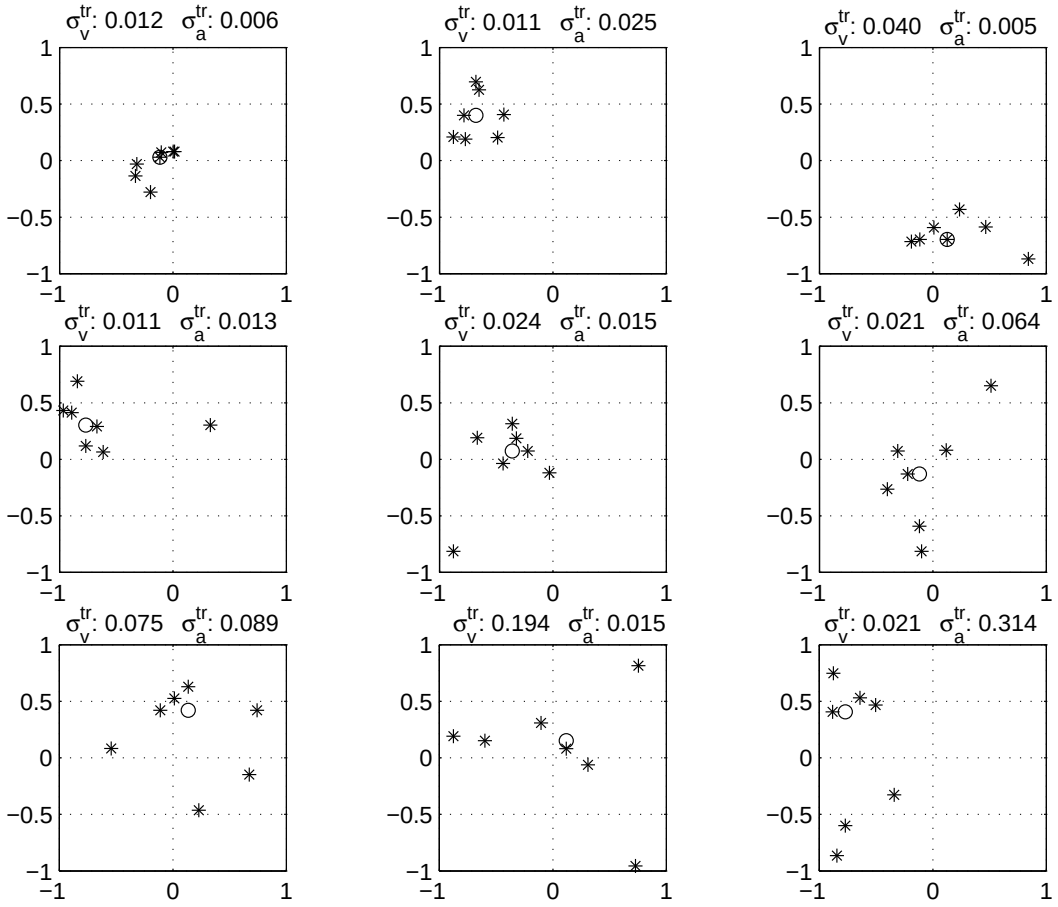

Figure 9-15: Ratings of tokens from Call Center database.

which was found to qualitatively reflect good agreement between judges while excluding few outliers, about 90% of the CHE set is retained compared to 76% of the Call Center data. Some of the possible reasons for this disparity have already been discussed, but as the complexity of each data set varies, we should expect to obtain different shapes for these monotonically increasing curves. Complexity here is understood to reflect the level of affective ambiguity that a given set poses for a set of listeners, a rough measure of which can be quickly assessed by the rise-time of the curves in 9-16.

9.6 Chapter Summary

In this chapter we have provided an account of two experiments conducted to gain perceptually-based affective annotations for two independent speech corpora showing a range of affective variability. Following much of the work in the literature, we have adopted the valence-

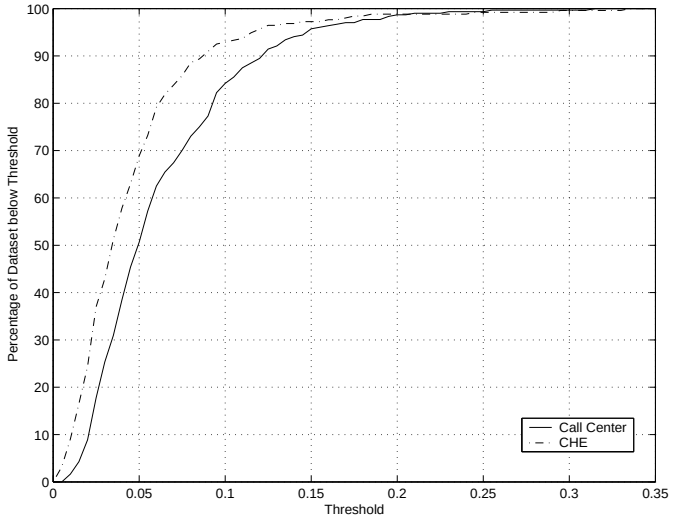

Figure 9-16: Percentage of data sets accepted as a function of the rejection threshold.

arousal model to capture differences between affective utterances. We feel that this model, although simplified and lacking the power to account for many subtle differences that may be perceptible to some listeners, offers nonetheless a good initial tool to obtain large trends in the data that are perceptually salient and relatively consistent across listeners. The sets under study were constructed by careful selection of tokens that were thought to cover the dimensions of valence and arousal. We have examined the existence of inter-variable relations between self-report measures gathered during the experiment and a set of external measures defined to capture the notions of typicality and consistency of judgments. Among the results discussed is a statistically significant interaction between gender and the typicality/consistency judgments, as well as a strong linear correlation between self-reported perceptivity and self-reported confidence on ratings. The agreements of judges for each token rated has been investigated by defining a measure of cluster spread resilient to the presence of few outliers. The results discussed show that this perceptual paradigm provides a good framework to obtain reliable annotations for a large percentage of the datasets under investigation.

Chapter 10

Evaluation

10.1 Introduction

In this chapter we apply the learning model introduced earlier to the classification of affect categories from prosodic acoustic parameters extracted from speech. This statistical model, described in detail in Chapter 8, attempts to model the hierarchical distribution of these parameters across scales as well as their dynamic evolution. The outline of the present chapter is as follows. We first provide a summary of all the features used as input to the model; these features were described in detail in Chapters 4-6 and are reviewed here in tabular form to provide a concise summary of the model's input. We then use the proposed model to build subject-dependent classifiers of affective categories from the actors data. At this stage we also consider more simplified versions of this model, and evaluate the performance of models with a shallower hierarchical structure showing that a decrease in model complexity can lead to an increase in performance. We also address in this chapter the issue of feature relevance, and apply a feature selection algorithm to explore the contribution of each individual feature to the discrimination task. Finally, we investigate whether the system can be extended to learn affect categories from a corpus of natural spontaneous speech, independent of the speaker, and explore the two data sets described earlier in Chapter 9, the CallHome English dataset and the Call Center data.

10.2 Review of Features and Model Summary

The main architecture of the system proposed in this thesis is summarized in Fig. 8-7: An acoustic waveform is used as input to the model. Then the prosodic parsing module, described in detail in Chapter 3, first produces a multi-tiered segmentation of this acoustic waveform in terms of structural units that are used provide different domains of analysis (i.e., the temporal span over which a particular acoustical analysis is performed on the waveform). The waveform, and its prosodic annotation, is then presented to the feature extraction algorithms that were covered in Chapters 4-6 to extract different feature vectors over different analysis windows. The following relation (also made explicit by Fig. 8-6 in Chapter 8) shows how different feature types are associated with different hierarchical domains:

- Utterance-level domain:
  - Loudness features
  - Liljencrants-Fant model features
  - Other voice source features
  - Harmonic consonance features
- Phrase-level domain:
  - F0 contour (raw and stylized) features
  - F0 chordal analysis
- Syllable-level domain:
  - Rhythm features

The components of each of these feature subsets are summarized in detail in Tables 10.1 through 10.4. Each table includes the general types of features, the particular features treated within each category, and the reference in the text where a more detailed description, and relevant equations, can be found. We therefore see that the highest-level domain (the utterance) is used as the analysis domain for a majority of the features explored in this work (features 1 through 76 in Tables 10.1 through 10.3). The intermediate level claims 19

No.	Type	Description	Eq. refs.
1	Loudness (See § 4.2.4, p. 83)	Mean perceived loudness	(4.5), (4.9)
2		25^{+}-percentile perceived loudness	(4.6), (4.9)
3		50^{+}-percentile perceived loudness	(4.6), (4.9)
4		75^{+}-percentile perceived loudness	(4.6), (4.9)
5		25^{+}-percentile RMS	(4.7), (4.9)
6		50^{+}-percentile RMS	(4.7), (4.9)
7		75^{+}-percentile RMS	(4.7), (4.9)
8		Mean specific loudness (Band 1)	(4.8), (4.9)
9		Mean specific loudness (Band 2)	(4.8), (4.9)
10		Mean specific loudness (Band 3)	(4.8), (4.9)
11		Mean specific loudness (Band 4)	(4.8), (4.9)
12		Mean specific loudness (Band 5)	(4.8), (4.9)
13		Mean specific loudness (Band 6)	(4.8), (4.9)
14		Mean specific loudness (Band 7)	(4.8), (4.9)
15		Mean specific loudness (Band 8)	(4.8), (4.9)
16		Mean specific loudness (Band 9)	(4.8), (4.9)
17		Mean specific loudness (Band 10)	(4.8), (4.9)
18		Mean specific loudness (Band 11)	(4.8), (4.9)
19		Mean specific loudness (Band 12)	(4.8), (4.9)
20		Mean specific loudness (Band 13)	(4.8), (4.9)

Table 10.1: Summary of features used as inputs to the model (1).

features (labeled as features 77 through 95 in Table 10.4), and finally, the lowest level is used to extract and represent 10 features (features 96 through 105 in Table 10.4).

The prosodic structure output by the parsing module in turn determines the basic structure of the hierarchical Bayesian network, as explained in Chapter 8. Once this basic structure is defined, a Bayesian network can be synthesized to store the observation variables at each scale (i.e., the feature vectors of dimensionality 76, 19 and 10, respectively) for each acoustic waveform under analysis. We can then submit an ensemble of Bayesian networks created for commonly labeled acoustic waveforms to the algorithms described in Chapter 8 to carry out learning and classification.

10.3 Performance on Actors Data

In this first set of evaluation experiments, we have built subject-dependent models of the affective categories *Afraid*, *Angry*, *Happy*, *Neutral* and *Sad*, from the set of actors data we have used repeatedly in the previous chapters of this thesis to test acoustic correlates of affective tone of voice. Structured Bayesian networks have been built for each speaker's

No.	Type	Description	Eq. refs.
21	Liljencrants-Fant Model (See § 6.4, p. 134)	25-percentile of E_e	(6.14), (6.29), (6.30)
22		Median value of E_e	(6.14), (6.29), (6.30)
23		75-percentile of E_e	(6.14), (6.29), (6.30)
24		IQR of normalized ΔE_e	(6.14), (6.29), (6.30)
25		25-percentile of γ	(6.14), (6.27), (6.29), (6.30)
26		Median value of γ	(6.14), (6.27), (6.29), (6.30)
27		75-percentile of γ	(6.14), (6.27), (6.29), (6.30)
28		IQR of normalized $\Delta\gamma$	(6.14), (6.27), (6.29), (6.30)
29		25-percentile of α	(6.14), (6.29), (6.30)
30		Median value of α	(6.14), (6.29), (6.30)
31		75-percentile of α	(6.14), (6.29), (6.30)
32		IQR of normalized $\Delta\alpha$	(6.14), (6.29), (6.30)
33		25-percentile of β	(6.14), (6.29), (6.30)
34		Median value of β	(6.14), (6.29), (6.30)
35		75-percentile of β	(6.14), (6.29), (6.30)
36		IQR of normalized $\Delta\beta$	(6.14), (6.29), (6.30)
37		25-percentile of OQ	(6.14), (6.28), (6.29), (6.30)
38		Median value of OQ	(6.14), (6.28), (6.29), (6.30)
39		75-percentile of OQ	(6.14), (6.28), (6.29), (6.30)
40		IQR of normalized ΔOQ	(6.14), (6.28), (6.29), (6.30)
41		25-percentile of ϵ_o	(6.14), (6.28), (6.26), (6.30)
42		Median value of ϵ_o	(6.14), (6.28), (6.26), (6.30)
43		75-percentile of ϵ_o	(6.14), (6.28), (6.26), (6.30)
44		IQR of normalized $\Delta\epsilon_o$	(6.14), (6.28), (6.26), (6.30)
45		25-percentile of ϵ_c	(6.14), (6.28), (6.26), (6.30)
46		Median value of ϵ_c	(6.14), (6.28), (6.26), (6.30)
47		75-percentile of ϵ_c	(6.14), (6.28), (6.26), (6.30)
48		IQR of normalized $\Delta\epsilon_c$	(6.14), (6.28), (6.26), (6.30)

Table 10.2: Summary of features used as inputs to the model (2).

No.	Type	Description	Eq. refs.
49	Other Voice Source Parameters (See § 6.5, p. 139)	Jitter (Pert. Factor)	(6.31), (6.39), (6.45)
50		Minimimum Jitter (Pert. Quotient)	(6.32), (6.40), (6.45)
51		Maximum Jitter (Pert. Quotient)	(6.32), (6.41), (6.45)
52		Shimmer (Pert. Factor)	(6.31), (6.42), (6.45)
53		Minimum Shimmer (Pert. Quotient)	(6.32), (6.43), (6.45)
54		Maximum Shimmer (Pert. Quotient)	(6.32), (6.44), (6.45)
55		25-percentile of GNE	(6.29), (6.34), (6.45)
56		Median value of GNE	(6.29), (6.34), (6.45
57		75-percentile of GNE	(6.29), (6.34), (6.45
58		IQR of normalized ΔGNE	(6.29), (6.34), (6.45
59		25-percentile of PSP	(6.29), (6.38), (6.45)
60		Median value of PSP	(6.29), (6.38), (6.45
61		75-percentile of PSP	(6.29), (6.38), (6.45
62		IQR of normalized ΔPSP	(6.29), (6.38), (6.45
63	Spectral Dissonance - Consonance (See § 6.6, p. 147)	Median of intrinsic diss. D_I	(6.47)
64		Range of intrinsic diss. D_I	(6.47)
65		Median of highest cons. interval α_1^c	(6.47)
66		Median of cons. values at interval α_1^c	(6.47)
67		Median of second highest cons. interval α_2^c	(6.47)
68		Median of cons. values at interval α_2^c	(6.47)
69		Median of highest diss. interval α_1^d	(6.47)
70		Median of diss. values at interval α_1^d	(6.47)
71		Median of second highest diss. interval α_2^d	(6.47)
72		Median of diss. values at interval α_2^d	(6.47)
73		Median of average cons. peak values	(6.47)
74		Median of average diss. peak values	(6.47)
75		Median of average diss.	(6.47)
76		Median of average diss. derivative	(6.47)

Table 10.3: Summary of features used as inputs to the model (3).

No.	Type	Description	Eq. refs.
77	F0 (Raw and Stylized) (See § 5.4, p. 101)	Sknewss of voiced $F0$	(5.12), (5.17)
78		Percentage of voiced $F0$ above the mean	(5.13), (5.17)
79		IQR of voiced $F0$	(5.14) , (5.17)
80		Range of voiced $F0$ above the mean	(5.15), (5.17)
81		Range of voiced $F0$ below the mean	(5.16), (5.17)
82		F0 declination slope	(5.24)
83		Proportion of pitch accented syllables	(5.18), (5.24)
84		Proportion of slope changes at turning points	(5.19), (5.24)
85		Normalized height of strongest pitch accent	(5.20), (5.21), 5.24)
86		Normalized range of strongest pitch accent	(5.20), (5.22), 5.24)
87		Slope of last stylized section	(5.23), (5.24)
88	F0 (Chordal Analysis) (See § 5.5, p. 107)	Chord instability	(5.29), (5.38)
89		Height of mode	(5.32), (5.38)
90		Normalized height of weakest mode	(5.33), (5.38)
91		Normalized height of medium mode	(5.34), (5.38)
92		Normalized spread of weakest mode	(5.35), (5.38)
93		Normalized spread of medium mode	(5.36), (5.38)
94		First interval ratio	(5.37), (5.38)
95		Second interval ratio	(5.37), (5.38)
96	Rhythm (See § 7.3, p. 158)	Absolute difference of nuclei duration	(7.3), (7.6), (7.10)
97		Normalized abs. diff. of nuclei duration	(7.4), (7.6), (7.10)
98		Local-to-global mean diff. of nuclei duration	(7.5), 7.6), (7.10)
99		Absolute difference of nuclei RMS	(7.3), (7.8), (7.10)
100		Normalized abs. diff. of nuclei RMS	(7.4), (7.8), (7.10)
101		Local-to-global mean diff. of nuclei RMS	(7.5), (7.8), (7.10)
102		Absolute difference of internuclei duration	(7.3), (7.7), (7.10)
103		Local-to-global mean diff. of internuclei duration	(7.5), (7.7), (7.10)
104		Absolute difference of internuclei RMS	(7.3), (7.9), (7.10)
105		Local-to-global mean diff. of internuclei RMS	(7.5), (7.9), (7.10)

Table 10.4: Summary of features used as inputs to the model (4).

Subject 1. ORR:38.8

	Afr	Ang	Hap	Ntr	Sad
Afr	**44**	8	6	16	26
Ang	4	**38**	8	22	28
Hap	12	8	**42**	30	8
Ntr	12	10	26	**34**	18
Sad	24	18	10	12	**36**

Subject 2. ORR:40

	Afr	Ang	Hap	Ntr	Sad
Afr	**42**	14	20	12	12
Ang	14	**44**	8	26	8
Hap	16	8	**54**	14	8
Ntr	14	16	22	**26**	22
Sad	20	14	14	18	**34**

Subject 3. ORR:47.6

	Afr	Ang	Hap	Ntr	Sad
Afr	**40**	6	20	8	26
Ang	6	**50**	16	22	6
Hap	20	8	**52**	12	8
Ntr	4	22	10	**56**	8
Sad	26	4	10	20	**40**

Subject 4. ORR:50.4

	Afr	Ang	Hap	Ntr	Sad
Afr	**62**	6	8	18	6
Ang	2	**48**	6	38	6
Hap	8	4	**68**	16	4
Ntr	4	22	10	**34**	30
Sad	4	6	2	48	**40**

Table 10.5: Generalization of subject-dependent classifiers on actor's dataset for subjects 1-4 for all five emotion categories considered. Recognition rates have been estimated through 15-fold cross-validation.

data as described in the previous section. Level-dependent principal component analysis is applied to each of the 3 components of the 105-dimensional vector to reduce the dimensionality. Enough components are retained to account for 90% of the variance in the data. This typically reduces the dimensionality from a total of 105 to approximately 60 dimensions. Two and three hidden states have been used to model the level-dependent hidden states, and the output distributions have been modeled using unimodal Gaussians. The results for the subject-dependent model with best generalization are shown in the confusion matrices shown in Tables 10.5-10.7. Each entry (i, j) of the matrices represents the proportion of tokens from the ith category classified as the jth category (each row of the matrices is normalized to add up to 100). The performance of the classifiers has been estimated through 15-fold cross-validation on a data set with equal priors. Each matrix shows in bold typeface the percent of correct classifications, as well as the overall recognition rate on the header.

Recognition rates significantly greater than random are obtained for 10 out of 11 subjects. with the exception being subject 8. The results reveal some of the subject-dependency we have been observing throughout this work. However, we can still observe a few trends from these results: *fear*, *anger* and *happiness*, for instance, tend to be better predicted than the *sad* and *neutral* tokens. More striking, however, is the fact that a great source

Subject 5. ORR: 49.8

	Afr	Ang	Hap	Ntr	Sad
Afr	**62**	12	22	0	4
Ang	6	**54**	24	6	10
Hap	18	22	**44**	2	14
Ntr	2	29	10	**45**	14
Sad	22	16	8	10	**44**

Subject 6. ORR:36

	Afr	Ang	Hap	Ntr	Sad
Afr	**36**	12	20	16	16
Ang	12	**40**	20	12	16
Hap	20	8	**38**	16	18
Ntr	22	10	10	**26**	32
Sad	18	14	8	20	**40**

Subject 7. ORR:51.2

	Afr	Ang	Hap	Ntr	Sad
Afr	**38**	2	18	24	18
Ang	6	**50**	12	28	4
Hap	6	0	**74**	14	6
Ntr	8	18	8	**48**	18
Sad	14	6	4	30	**46**

Subject 8. ORR:28

	Afr	Ang	Hap	Ntr	Sad
Afr	**36**	16	20	12	16
Ang	20	**34**	12	14	20
Hap	34	16	**30**	14	6
Ntr	14	20	20	**20**	26
Sad	32	8	10	30	**20**

Table 10.6: Generalization of subject-dependent classifiers on actor's dataset for subjects 5-8 for all five emotions considered. Recognition rates have been estimated through 15-fold cross-validation.

of confusion lies between the *neutral* category and the rest. Much of the reported work in the literature tends to compare the "basic" emotion categories, often failing to include a *neutral* category as a control. This source of confusion, however, suggests that recognition performance can degrade considerably if we take this added descriptor into account. This is a result that we might have expected outside any experimental confirmation since there should be greater overlap between neutral categories and the rest than between separate members of the emotion set.

If we only consider the set of "full-blown" emotions without the neutral category, the performances change noticeably as evidenced by the confusion matrices in Tables 10.8 - 10.10. If we restrict ourselves to this set, then it is possible to make some tentative comparisons with some benchmarks published in the literature regarding human performance on emotion recognition. Banse & Scherer (1996), for instance, present results of different studies that suggest humans attain about 50% on average, reaching as high as 70% for *Anger* and *Happiness*.[1] A more relevant study to this dataset is mentioned in Polzin (2000), who

[1]The set of emotions that Banse & Scherer (1996) discuss also include the *Disgust* category, so a direct comparison is not possible; however, the point being raised here is that the *Neutral* category may be a stronger confounding variable than any other full-blown emotion which we might add to the set given the presumed overlap between this and other categories.

Subject 9. ORR:33.33

	Afr	Ang	Hap	Ntr	Sad
Afr	**31**	8	18	14	29
Ang	4	**26**	12	30	28
Hap	24	12	**46**	6	12
Ntr	8	38	8	**32**	14
Sad	14	20	10	24	**32**

Subject 10. ORR:49.2

	Afr	Ang	Hap	Ntr	Sad
Afr	**50**	10	8	16	16
Ang	2	**68**	12	8	10
Hap	16	4	**56**	10	14
Ntr	16	20	12	**32**	20
Sad	20	16	8	16	**40**

Subject 11. ORR: 48.4

	Afr	Ang	Hap	Ntr	Sad
Afr	**58**	2	14	2	24
Ang	2	**56**	6	26	10
Hap	14	4	**60**	14	8
Ntr	14	26	6	**44**	10
Sad	36	4	16	20	**24**

Table 10.7: Generalization of subject-dependent classifiers on actor's dataset for subjects 9-11 for all five emotions considered. Recognition rates have been estimated through 15-fold cross-validation.

used a portion of the same set of actor's data to gain some insight on human categorical perception of its affective content. The study also limited itself to the four basic emotions, and it reports human performance in the $60\% - 80\%$ range. However, this perceptual test included a "training" period where subjects were presented with labeled exemplars of what each emotion meant, and it is therefore reasonable to expect that these figures might be too optimistic.

With these general caveats and observations in mind, we can suggest that the model is beginning to perform within the scale of human recognition for some subjects and in the case when we restrict ourselves to the set of full-blown emotions. In the cases of subjects 7 or 4, the performance, we could argue, is competitive with the benchmarks discussed. This observation should not detract, however, from the weakness of having the recognition rates impeded when we consider the neutral category as well. Clearly a system that recognizes between different affective states with good accuracy should be able to detect the lack of any full blown emotion with reasonable performance as well.

Subject 1. ORR:44

	Afr	Ang	Hap	Sad
Afr	**50**	8	14	28
Ang	12	**44**	8	36
Hap	20	14	**46**	20
Sad	28	24	12	**36**

Subject 2. ORR:50.5

	Afr	Ang	Hap	Sad
Afr	**44**	20	30	6
Ang	10	**62**	16	12
Hap	22	16	**56**	6
Sad	6	26	28	**40**

Subject 3. ORR:53

	Afr	Ang	Hap	Sad
Afr	**46**	6	22	26
Ang	8	**60**	20	12
Hap	22	10	**60**	8
Sad	30	12	12	**46**

Subject 4. ORR:68

	Afr	Ang	Hap	Sad
Afr	**64**	14	12	10
Ang	6	**58**	10	26
Hap	10	6	**76**	8
Sad	4	12	10	**74**

Table 10.8: Generalization of subject-dependent classifiers on actor's dataset for subjects 1-4 for four emotion subset. Recognition rates have been estimated through 15-fold cross-validation.

Subject 5. ORR:51.5

	Afr	Ang	Hap	Sad
Afr	**62**	12	22	4
Ang	6	**56**	26	12
Hap	18	22	**44**	16
Sad	22	18	16	**44**

Subject 6. ORR:44

	Afr	Ang	Hap	Sad
Afr	**38**	18	22	22
Ang	12	**46**	22	20
Hap	24	10	**46**	20
Sad	18	20	16	**46**

Subject 7. ORR:65

	Afr	Ang	Hap	Sad
Afr	**54**	4	20	22
Ang	8	**66**	16	10
Hap	6	8	**76**	10
Sad	20	10	6	**64**

Subject 8. ORR:39

	Afr	Ang	Hap	Sad
Afr	**42**	18	20	20
Ang	24	**36**	14	26
Hap	36	16	**42**	6
Sad	38	16	10	**36**

Table 10.9: Generalization of subject-dependent classifiers on actor's dataset for subjects 5-8 for four emotion subset. Recognition rates have been estimated through 15-fold cross-validation.

Subject 9. ORR:41.71

	Afr	Ang	Hap	Sad
Afr	**35**	8	20	37
Ang	4	**46**	16	34
Hap	24	14	**48**	14
Sad	18	32	12	**38**

Subject 10. ORR:57.5

	Afr	Ang	Hap	Sad
Afr	**54**	12	14	20
Ang	8	**68**	12	12
Hap	18	4	**62**	16
Sad	26	20	8	**46**

Subject 11. ORR:56.5

	Afr	Ang	Hap	Sad
Afr	**58**	2	16	24
Ang	6	**68**	12	14
Hap	18	6	**68**	8
Sad	40	8	20	**32**

Table 10.10: Generalization of subject-dependent classifiers on actor's dataset for subjects 9-11 for four emotion subset. Recognition rates have been estimated through 15-fold cross-validation.

10.3.1 Trimming the Models

The results tabulated in the previous section are based on the full representation of the 105-dimensional feature set, including measures related to loudness, voice quality, intonation, and rhythm. Although the feature set is reduced at each level by first applying principal component analysis, it may be that the resulting model is still too complex for the amount of data being modeled. In this section we consider an alternative to cope with the model complexity by reducing the number of hierarchical levels in the structured Bayesian networks. Recall that the networks grow disproportionately denser as more hierarchical levels are added since each hidden state node at a given level "spawns" one or more nodes at the level below it. This suggests that we may want to consider shallower models where lower levels have been successively trimmed. Recall also that the lowest level of the hierarchical representation is responsible for modeling the rhythm-related features, which were found to be less predictive of affective categories than the remaining feature types considered. It is also possible that the complexity gained by the model in accommodating this last level is not offset by the modeling power of those features, and that a sparser model will perform better even by discarding features.

Tables 10.11 through 10.13 show the generalization results of training the Bayesian networks with data from levels 1 and 2 (i.e., rhythm-related features have been omitted).

Subject 1. ORR:49.5

	Afr	Ang	Hap	Sad
Afr	**52**	8	20	20
Ang	14	**50**	6	30
Hap	6	14	**52**	28
Sad	24	20	12	**44**

Subject 2. ORR:54.5

	Afr	Ang	Hap	Sad
Afr	**48**	24	12	16
Ang	22	**52**	8	18
Hap	14	12	**66**	8
Sad	22	14	12	**52**

Subject 3. ORR:54

	Afr	Ang	Hap	Sad
Afr	**54**	6	14	26
Ang	12	**74**	6	8
Hap	18	16	**44**	22
Sad	34	10	12	**44**

Subject 4. ORR:75.5

	Afr	Ang	Hap	Sad
Afr	**72**	10	10	8
Ang	2	**76**	8	14
Hap	16	10	**68**	6
Sad	8	4	2	**86**

Table 10.11: Generalization of subject-dependent classifiers on actor's dataset for subjects 1-4 and four emotion subset using two-level hierarchical models. Recognition rates have been estimated through 20-fold cross-validation.

PCA has been applied as before to the feature vectors at the remaining levels. One, two, and three hidden states have been used to model each level-dependent discrete state variable, and the outputs have been modeled using diagonal-covariance unimodal Gaussian distributions. The tables include the results of the model with the best generalization properties for each subject. Tables 10.14 through 10.16 show the results when the networks are trained using only the top level (i.e., intonational and rhythm features are omitted). A similar training procedure is used, and the best subject-dependent results are tabulated.

These results confirm the hypothesis that simpler models perform better even at the expense of discarding subsets of features. The mean performance increases from 52.9% (when the full representation is used) to 54% (when only levels 1 and 2 are used) and finally to 57.9% (when only the top level is used). This represents an increase of 2.2% and 9.5% respectively. As noted, there is a trade-off between adding extra measurements to the feature set and adding the extra connectivity to the model to accommodate those features. The increased performance of the simplest models suggests that perhaps the features used at the highest level (i.e., the loudness and different voice quality measures) may be the most predictive features in this discrimination task, and that therefore the modeling power of the features at lower levels do not offset the complexity introduced into the model. We will address this issue further in Section 10.4 when we examine feature relevance.

Subject 5. ORR:54

	Afr	Ang	Hap	Sad
Afr	**60**	14	24	2
Ang	8	**66**	18	8
Hap	12	26	**44**	18
Sad	22	12	20	**46**

Subject 6. ORR:46

	Afr	Ang	Hap	Sad
Afr	**36**	22	20	22
Ang	16	**62**	6	16
Hap	30	8	**48**	14
Sad	26	18	18	**38**

Subject 7. ORR:66

	Afr	Ang	Hap	Sad
Afr	**54**	4	16	26
Ang	8	**70**	14	8
Hap	8	6	**76**	10
Sad	20	10	6	**64**

Subject 8. ORR:33.5

	Afr	Ang	Hap	Sad
Afr	**46**	22	12	20
Ang	30	**28**	12	30
Hap	42	14	**26**	18
Sad	32	18	16	**34**

Table 10.12: Generalization of subject-dependent classifiers on actor's dataset for subjects 5-8 and four emotion subset using two-level hierarchical models. Recognition rates have been estimated through 20-fold cross-validation.

Subject 9. ORR:42.53

	Afr	Ang	Hap	Sad
Afr	**41**	10	18	31
Ang	8	**54**	10	28
Hap	24	10	**52**	14
Sad	24	30	12	**34**

Subject 10. ORR:58

	Afr	Ang	Hap	Sad
Afr	**68**	6	10	16
Ang	4	**60**	10	26
Hap	22	10	**58**	10
Sad	26	14	14	**46**

Subject 11. ORR:61

	Afr	Ang	Hap	Sad
Afr	**58**	2	18	22
Ang	4	**88**	4	4
Hap	12	10	**64**	14
Sad	46	4	16	**34**

Table 10.13: Generalization of subject-dependent classifiers on actor's dataset for subjects 9-11 and four emotion subset using two-level hierarchical models. Recognition rates have been estimated through 20-fold cross-validation.

Subject 1. ORR:58

	Afr	Ang	Hap	Sad
Afr	**52**	12	14	22
Ang	4	**64**	12	20
Hap	6	12	**66**	16
Sad	20	12	20	**48**

Subject 2. ORR:60

	Afr	Ang	Hap	Sad
Afr	**66**	12	14	8
Ang	22	**54**	8	16
Hap	8	12	**64**	16
Sad	20	18	6	**56**

Subject 3. ORR:60

	Afr	Ang	Hap	Sad
Afr	**62**	2	8	28
Ang	4	**80**	6	10
Hap	26	8	**40**	26
Sad	24	4	16	**56**

Subject 4. ORR:77

	Afr	Ang	Hap	Sad
Afr	**80**	2	6	12
Ang	2	**76**	8	14
Hap	10	22	**68**	0
Sad	4	8	4	**84**

Table 10.14: Generalization of subject-dependent classifiers on actor's dataset for subjects 1-4 and four emotion subset using one-level models. Recognition rates have been estimated through 20-fold cross-validation.

Subject 5. ORR:60

	Afr	Ang	Hap	Sad
Afr	**70**	4	14	12
Ang	4	**60**	24	12
Hap	10	14	**46**	30
Sad	6	6	24	**64**

Subject 6. ORR:47

	Afr	Ang	Hap	Sad
Afr	**50**	12	22	16
Ang	8	**62**	10	20
Hap	32	14	**42**	12
Sad	30	20	16	**34**

Subject 7. ORR:68

	Afr	Ang	Hap	Sad
Afr	**62**	0	18	20
Ang	2	**82**	8	8
Hap	20	4	**62**	14
Sad	12	10	12	**66**

Subject 8. ORR:34

	Afr	Ang	Hap	Sad
Afr	**26**	16	20	38
Ang	26	**34**	14	26
Hap	18	10	**34**	38
Sad	24	14	22	**40**

Table 10.15: Generalization of subject-dependent classifiers on actor's dataset for subjects 5-8 and four emotion subset using one-level models. Recognition rates have been estimated through 20-fold cross-validation.

Subject 9. ORR:51

	Afr	Ang	Hap	Sad
Afr	**18**	6	16	60
Ang	2	**68**	6	24
Hap	10	10	**64**	16
Sad	14	24	8	**54**

Subject 10. ORR:59

	Afr	Ang	Hap	Sad
Afr	**66**	0	20	14
Ang	6	**66**	12	16
Hap	12	24	**52**	12
Sad	16	10	24	**50**

Subject 11. ORR:63

	Afr	Ang	Hap	Sad
Afr	**46**	6	10	38
Ang	0	**86**	6	8
Hap	12	8	**66**	14
Sad	32	4	12	**52**

Table 10.16: Generalization of subject-dependent classifiers on actor's dataset for subjects 9-11 and four emotion subset using one-level hierarchical models. Recognition rates have been estimated through 20-fold cross-validation.

10.3.2 Comparison to Human Performance Benchmarks

One of the notable features of the system's performance summarized in Tables 10.5 through 10.16 is its subject dependency. The recognition rates for subjects 4 and 7, for example, clearly exceed those for subjects 6 and 8 irrespective of the classifier employed or the complexity of the model explored. In order to explore whether this behavior is an artifact of the models, or whether there's an inherent limitation in the data (e.g., some speakers may not consistently encode the affective difference using the acoustic cues we are exploiting), we decided to assess how human listeners would perform on this recognition task in order to establish a comparison benchmark.

We designed an informal experiment where listeners were presented with a random subject-dependent subsample of the speech tokens that were part of this database and asked to evaluate them in a forced-choice experiment. The speech tokens were chosen such that the semantic content was as ambiguous as possible in order to guide the listener's attention as much as possible to the prosodic aspects of the utterances. Aside from this pre-selection criterion, however, tokens were randomly selected. Each listener was presented with a balanced sample consisting of ten tokens from each of the five emotions in a randomized order. The experiment took place over a web interface, so subjects were never present in the laboratory. The interface consisted of a sound icon (on which the listener could click as

Subject 1. ORR:60

	Afr	Ang	Hap	Ntr	Sad
Afr	**25**	0	0	5	70
Ang	0	**90**	0	10	0
Hap	5	5	**65**	25	0
Ntr	5	15	15	**65**	0
Sad	0	15	0	30	**55**

Subject 2. ORR:82

	Afr	Ang	Hap	Ntr	Sad
Afr	**80**	0	0	0	20
Ang	0	**85**	5	10	0
Hap	5	0	**75**	15	5
Ntr	0	5	5	**90**	0
Sad	10	5	0	5	80

Subject 3. ORR:79

	Afr	Ang	Hap	Ntr	Sad
Afr	**70**	5	0	10	15
Ang	0	**90**	0	10	0
Hap	5	10	**65**	10	10
Ntr	0	5	0	**95**	0
Sad	0	0	5	20	75

Subject 4. ORR:70

	Afr	Ang	Hap	Ntr	Sad
Afr	**70**	0	10	10	10
Ang	0	**80**	0	10	10
Hap	0	0	**70**	30	0
Ntr	10	0	0	**70**	20
Sad	0	10	0	30	**60**

Table 10.17: Human recognition rates on actors data for subjects 1-4. All emotions considered.

often as needed to listen to the sample), five buttons each labeled with an emotion category, and a button labeled *Next* to enable the user to advance to the next selection. In order to ensure that a choice was made for each token, a reminder message would be displayed if the user attempted to advance without having made a selection; the message would remain active until the user selected a category.

Tables 10.17 through 10.19 show the confusion matrices and overall recognition rates obtained for each subject. The results for subjects 1, 2 and 3 are based on the evaluation from two listeners each; the remaining subjects were evaluated by one listener each. Since the task included all five emotions and we have also been considering the special case where only the full-blown categories are considered, the performance figures are re-estimated in Tables 10.20 through 10.22 for the case when the *Neutral* category is not considered by eliminating from the count those tokens which were labeled as such.[2]

This experiment, though informal and based on a few data points collected, affords us some interesting observations. Listeners produce their lowest recognition rates when presented with data from the speakers for whom the automatic systems reach the poorest

[2]One must bear in mind, however, that this is done *a posteriori*, and that listeners' "allocation" of speech tokens to category labels may have differed if the set of categories had lacked the *Neutral* label from the start.

Subject 5. ORR:78

	Afr	Ang	Hap	Ntr	Sad
Afr	**50**	20	10	20	0
Ang	0	**90**	0	10	0
Hap	0	0	**80**	20	0
Ntr	0	0	0	**100**	0
Sad	10	20	0	0	**70**

Subject 6. ORR:50

	Afr	Ang	Hap	Ntr	Sad
Afr	**40**	10	0	30	20
Ang	10	**70**	0	10	10
Hap	0	10	**60**	20	10
Ntr	10	0	10	**50**	30
Sad	40	10	0	20	**30**

Subject 7. ORR:72

	Afr	Ang	Hap	Ntr	Sad
Afr	**60**	0	10	10	20
Ang	0	**60**	10	30	0
Hap	0	0	**90**	10	0
Ntr	0	20	0	**80**	0
Sad	10	0	0	20	**70**

Subject 8. ORR:26

	Afr	Ang	Hap	Ntr	Sad
Afr	**30**	30	0	30	10
Ang	10	**30**	10	40	10
Hap	10	10	**20**	30	30
Ntr	10	50	0	**30**	10
Sad	10	30	0	40	**20**

Table 10.18: Human recognition rates on actors data for subjects 5-8. All emotions considered.

recognition (subjects 8, 9 and 6). Likewise, human listeners tend to attain highest recognition rates evaluating the data of those speakers for whom the systems offer their top performance (e.g., subjects 4, 7 and 2). Another general trend is that human performance surpasses the automatic recognition rates. This gap is not significant for the lower performers. In fact, the automatic systems surpass the human performance rates for subjects 8 and 11 (though more data points would be needed to discern if this difference is significant). However, as the human and machine performance rates increase, the human performance quickly tends to exceed that of the automatic system. This behavior would be compatible with the hypothesis that those speakers may be employing a larger inventory of acoustic cues than the automatic systems are exploiting for recognition, cues that presumably would be available to human listeners. It may also be the case that, in the case of compatible recognition rates between listeners and machine, the system is exploiting the inventory of cues employed by the speakers more fully. At any rate, this experiment confirms that the subject dependency is not an erratic behavior of the automatic models, and that human performance, though generally exceeding automatic recognition, tends to exhibit a similar trend.

Subject 9. ORR:48

	Afr	Ang	Hap	Ntr	Sad
Afr	**0**	20	40	30	10
Ang	10	**70**	0	20	0
Hap	0	0	**80**	20	0
Ntr	10	10	20	**60**	0
Sad	0	20	20	30	**30**

Subject 10. ORR:62

	Afr	Ang	Hap	Ntr	Sad
Afr	**50**	0	0	40	10
Ang	0	**40**	30	30	0
Hap	10	0	**70**	20	0
Ntr	10	10	0	**80**	0
Sad	0	10	10	10	**70**

Subject 11. ORR:42

	Afr	Ang	Hap	Ntr	Sad
Afr	**10**	0	10	30	50
Ang	10	**60**	10	0	20
Hap	10	10	**50**	30	0
Ntr	10	10	0	**60**	20
Sad	20	20	10	20	**30**

Table 10.19: Human recognition rates on actors data for subjects 9-11. All emotions considered.

Subject 1. ORR:71

	Afr	Ang	Hap	Sad
Afr	**26**	0	0	74
Ang	0	**100**	0	0
Hap	7	7	**86**	0
Sad	0	21	0	**79**

Subject 2. ORR:87

	Afr	Ang	Hap	Sad
Afr	**80**	0	0	20
Ang	0	**94**	6	0
Hap	6	0	**88**	6
Sad	11	5	0	**84**

Subject 3. ORR:86

	Afr	Ang	Hap	Sad
Afr	**78**	5	0	17
Ang	0	**100**	0	0
Hap	6	11	**72**	11
Sad	0	0	6	**94**

Subject 4. ORR:88

	Afr	Ang	Hap	Sad
Afr	**78**	0	11	11
Ang	0	**89**	0	11
Hap	0	0	**100**	0
Sad	0	14	0	**86**

Table 10.20: Human recognition rates on actors data for subjects 1-4. Only full-blown emotions considered.

Subject 5. ORR:83

	Afr	Ang	Hap	Sad
Afr	**62**	25	13	0
Ang	0	**100**	0	0
Hap	0	0	**100**	0
Sad	10	20	0	**70**

Subject 6. ORR:62

	Afr	Ang	Hap	Sad
Afr	**57**	14	0	29
Ang	11	**78**	0	11
Hap	0	12	**75**	13
Sad	50	12	0	**38**

Subject 7. ORR:85

	Afr	Ang	Hap	Sad
Afr	**67**	0	11	22
Ang	0	**86**	14	0
Hap	0	0	**100**	0
Sad	12	0	0	**88**

Subject 8. ORR:38

	Afr	Ang	Hap	Sad
Afr	**43**	43	0	14
Ang	17	**50**	17	16
Hap	14	14	**29**	43
Sad	17	50	0	**33**

Table 10.21: Human recognition rates on actors data for subjects 5-8. Only full-blown emotions considered.

Subject 9. ORR:60

	Afr	Ang	Hap	Sad
Afr	**0**	29	57	14
Ang	12	**88**	0	0
Hap	0	0	**100**	0
Sad	0	29	29	**42**

Subject 10. ORR: 77

	Afr	Ang	Hap	Sad
Afr	**84**	0	0	16
Ang	0	**57**	43	0
Hap	12	0	**88**	0
Sad	0	11	11	**78**

Subject 11. ORR: 47

	Afr	Ang	Hap	Sad
Afr	**15**	0	14	71
Ang	10	**60**	10	20
Hap	14	15	**71**	0
Sad	25	25	12	**38**

Table 10.22: Human recognition rates on actors data for subjects 9-11. Only full-blown emotions considered.

10.4 Feature Relevance

An important question that we would like to address is the extent to which the features proposed in this work contribute to the task of discriminating between affective categories. Given that the size of the original feature set is considerable (105 features), we would like to investigate which are the most relevant features, and whether the performance can be retained or improved with a subset of the full set (i.e., whether the curse of dimensionality is affecting the ability to model a finite-size set defined in a large dimensional space). With the exception of cases where exhaustive search of the $2^D - 1$ combinations of features in a D-dimensional space is tractable, feature selection is typically a difficult problem to which different suboptimal approaches have been proposed. In order to explore the questions we have introduced in this section, we propose to use the Sequential Forward Floating Selection algorithm for feature selection (Pudil et al., 1994; Jain & Zongker, 1997), given its documented empirical success in solving feature selection problems.

Let $Y = \{y_i : 1 \leq i \leq D\}$ be the full set of D features, and let $X_k = \{x_j : 1 \leq j \leq k, x_i \in Y$ be a subset of k features from Y. Finally, let $J(X_k)$ represent a criterion used to assess the performance of the subset X_k (a natural choice, for classification, might be a classifier's generalization error when trained with the subset of features X_k). The Sequential Forward Floating Selection algorithm is an interactive procedure to build a feature subset of size $k+1$ starting with a feature subset of size k. Assume that a set X_k, with corresponding criterion evaluation $J(X_k)$ has already been built by the algorithm, and that the value $J(X_i)$ for $i = 1, \cdots, k-1$ is also available for all preceding subsets. Then the algorithm involves evaluating the following set of steps

1. Inclusion: Iterate through every feature remaining in the set of available measurements $(Y \setminus X_k)$ to find the most significant feature x_{k+1}, and add it to the set to form $X_{k+1} \leftarrow X_k \cup x_{k+1}$. (The most significant feature is the one that, when added to X_k, optimizes $J(X_{k+1})$.

2. Conditional Exclusion: Find the least significant feature x_r in the newly built set X_{k+1}. If the least significant feature is the recently added feature ($x_r = x_{k+1}$), then let $k \leftarrow k+1$ and go to step 1; otherwise, exclude x_r to build a new set $X'_k \leftarrow X_{k+1} \setminus x_r$. If $k = 2$, set $X_k \leftarrow X'_k$ and $J(X_k) = J(X'_k)$ and go to step 1; otherwise, continue to step 3.

3. Continuation: Find the least significant feature x_s in the set X'_k. If $J(X'_k \setminus x_s) \leq J(X_{k-1})$, then set $X_k \leftarrow X'_k$ and $J(X_k) \leftarrow J(X'_k)$ and return to step 1. Otherwise, exclude x_s to form the newly reduced set to form $X'_{k-1} \leftarrow X'_k \setminus x_s$, and assign $k \leftarrow k-1$. If $k = 2$, set $X_k \leftarrow X'_k$ and $J(X_k) \leftarrow J(X'_k)$ and return to step 1; otherwise, repeat step 3.

One of the properties (and strengths) of this algorithm lies in the fact that it need not create a sequentially nested subset of features. That is, the "optimal" set of size $k+1$ need not contain the "optimal" set of size k. This is a result of repeatedly evaluating steps 2 and 3 when the conditional exclusion test becomes true in step 1. The algorithm is suboptimal. However, it has been shown to exhibit good properties in empirical benchmarks (Jain & Zongker, 1997).

We have applied this algorithm to explore the contribution of the set of 105 features described earlier. Although ideally we would like to use the generalization error of the Bayesian network classifiers as the evaluative criterion in the algorithm, the cost of fully training graphical models for all classes of interest with subsets of the data at every evaluation of steps 1-3 above becomes computationally prohibitive. Instead, we have opted to use the leave-one-out generalization error of a K-nearest neighbor classifier as a simpler (computationally tractable) criterion to explore the feature set. Since the feature set exhibits a dynamic, and hierarchical, structure that is not directly modeled by a nearest neighbor classifier, we have first converted the feature vectors to a static representation by summarizing the time series at each hierarchical level with its sample mean, and then stacking the observations from each level to obtain a 105-dimensional vector associated with each analysis utterance.

Fig. 10-1 shows the generalization performance, estimated through leave-one-out cross-validation, of the best subset of size k found by the forward floating selection algorithm for each value of the training set size. Since the leave-one-out estimate of the recognition rate is a stochastic variable, the figure shows fluctuations over the different set sizes considered. However, a clear trend is discernible in the plot illustrating the well-known effect of increasing the dimensionality of the feature space on a fixed sample (the curse-of-dimensionality effect): the average performance increases up to a certain maximum value and then decreases as more features are added.

As pointed out, the floating selection algorithm does not produce a strictly incremental

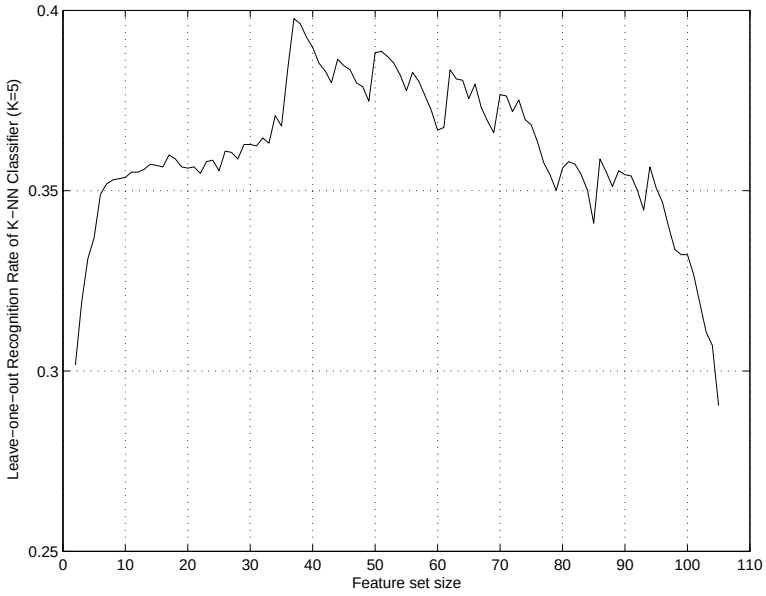

Figure 10-1: Recognition rate (estimated through leave-one-out cross-validation) of nearest-neighbor classifier as a function of training set size.

ordering of the features (which would lead to a natural ranking of the features in terms of their contribution) since features are added or deleted as the algorithm cycles through different set sizes. In order to explore the contribution of different features to the performance of the classifier, however, we could take into account the number of times that a feature was selected for the best set given a certain set size. Furthermore, we may want to weigh the contribution of features that were incorporated earlier differently from those that the algorithm only favored toward the end of the cycle. (Intuitively, we want to capture features that are used often, and those that the algorithm turns to initially, as those capture most of the modeling power.) Based on these observations, we would like to propose the following figure of merit. Let $f_{i,k}, 1 \leq i, k \leq D$ be an indicator variable which is defined to be 1 if the $i-th$ feature was selected when building the best set of size k, and zero otherwise. Define then

$$w_k = \frac{k - D}{1 - D} \tag{10.1}$$

$$FOM_i = \sum_k w_k f_{i,k}. \tag{10.2}$$

Eq. 10.2 is a weighted count of the number of times a feature was selected, where the weights w_k linearly favor those features selected earlier. We can therefore evaluate the

Rank	Feat. No.	Type	Description
1	37	LF Parms.	25-percentile of OQ
2	25	LF Parms.	25-percentile of γ
3	67	Spectral Diss.	Median of second highest cons. interval α_2^c
4	75	Spectral Diss.	Median of average diss.
5	42	LF Parms.	Median value of ϵ_o
6	9	Loudness	50^+-percentile RMS
7	11	Loudness	Mean specific loudness (Band 4)
8	41	LF Parms.	25-percentile of ϵ_o
9	18	Loudness	Mean specific loudness (Band 11)
10	17	Loudness	Mean specific loudness (Band 10)
11	63	Spectral Diss.	Median of intrinsic diss. D_I
12	16	Loudness	Mean specific loudness (Band 9)
13	15	Loudness	Mean specific loudness (Band 8)
14	14	Loudness	Mean specific loudness (Band 7)
15	13	Loudness	Mean specific loudness (Band 6)
16	8	Loudness	Mean specific loudness (Band 1)
17	66	Spectral Diss.	Median of cons. values at interval α_1^c
18	20	Loudness	Mean specific loudness (Band 13)
19	51	Other VS Parms.	Maximum Jitter (Pert. Quotient)
20	19	Loudness	Mean specific loudness (Band 12)
21	12	Loudness	Mean specific loudness (Band 5)
22	10	Loudness	Mean specific loudness (Band 3)
23	39	LF Parms.	75-percentile of OQ
24	73	Spectral Diss.	Median of average cons. peak values
25	38	LF Parms.	Median value of OQ

Table 10.23: Ranking of features (1 through 25).

figure of merit for all the features in the set, and rank them according to their score. The results of the top ranked 50 features based on the forward floating selection algorithm are summarized in Tables 10.23 and 10.24. The first column of the tables indicates the rank (with 1 representing the "best" feature according to the figure of merit just motivated). The second, third and last columns reference the feature identity number, the type of feature, and the short description used in the feature summary tables (Tables 10.1 through 10.4).

As can be seen by inspecting Tables 10.23 and 10.24, the more global features (measured across utterances) describing aspects of loudness and voice quality prove to be the most fruitful in discrimination between affective categories. The loudness features selected include the specific loudness in several bands of the Bark scale, as well as the mean of the integrated perceived loudness across the spectrum. The voice quality features with highest rank include

Rank	Feat. No.	Type	Description
26	62	Other VS Parms	IQR of normalized ΔPSP
27	69	Spectral Diss.	Median of highest diss. interval α_1^d
28	1	Loudness	Mean perceived loudness
29	59	Other VS Parms.	25-percentile of PSP
30	40	LF Parms.	IQR of normalized ΔOQ
31	45	LF Parms.	25-percentile of ϵ_c
32	60	Other VS Parms.	Median value of PSP
33	58	Other VS Parms.	IQR of normalized ΔGNE
34	54	Other VS Parms.	Maximum Shimmer (Pert. Quotient)
35	53	Other VS Parms.	Minimum Shimmer (Pert. Quotient)
36	55	Other VS Parms.	25-percentile of GNE
37	68	Spectral Diss.	Median of cons. values at interval α_2^c
38	33	LF Parms.	25-percentile of β
39	65	Spectral Diss.	Median of highest cons. interval α_1^c
40	61	Other VS Parms.	75-percentile of PSP
41	57	Other VS Parms.	75-percentile of GNE
42	2	Loudness	25^+-percentile perceived loudness
43	71	Spectral Diss.	Median of second highest diss. interval α_2^d
44	3	Loudness	50^+-percentile perceived loudness
45	43	LF Parms.	75-percentile of ϵ_o
46	76	Spectral Diss.	Median of average diss. derivative
47	64	Spectral Diss.	Range of intrinsic diss. D_I
48	26	LF Parms.	Median value of γ
49	21	LF Parms.	25-percentile of E_e
50	52	Other VS Parms.	Shimmer (Pert. Factor)

Table 10.24: Ranking of features (26 through 50).

several parameters derived from the Liljencrants-Fant parametrization of the glottal volume velocity waveform, as well as several parameters derived from the consonance-based analysis of the spectral harmonics. Other voice source features favored in the selection include measures related to shimmer, glottal-to-noise excitation and parabolic spectral parameters.

The value of the generalization error (or its complement plotted in Fig. 10-1) as a function of set size can be used to find an optimal feature subset (i.e., the subset which maximizes the generalization recognition rate in this case). Although we trained the Bayesian network models using this optimal feature subset, the results failed to reach the significance of those obtained with the top single-layer models already discussed in section 10.3.1. We can cite at least two reasons why this can be so. The value of the subset with highest recognition rate (a set of size 37) includes only utterance-level features, and these features are already used as inputs to the trimmed model implemented earlier. The trimmed model uses, in addition, a set of utterance-level features which are not selected by the nearest-neighbor criterion, but which could nonetheless be contributing to the performance of the Bayesian network models (which are, in this case, just simple HMMs since the two lower hierarchical levels have been trimmed). The second reason is related to this mismatch between the feature selection evaluation criterion (a nearest neighbor criterion) and the final performance criterion (generalization performance under the Bayesian network models): the feature set selected by the first criterion can very well represent a suboptimal set for the final models implemented. It should be reiterated that the nearest neighbor criterion has not been chosen for its better ability to model the data, but rather for its computational tractability, in order to evaluate the performance of features with respect to each other.

10.5 Comparison with Other Classification Approaches: Support Vector Machines

The objective of this section is to assess the performance of an alternative model on the data with the purpose of obtaining a benchmark against which to evaluate the performance of the structured Bayesian networks proposed in this thesis for modeling prosodic phenomena. Because it is not straightforward to find a competing model that takes into account the hierarchical and dynamic structure of the data, we have resorted to working with the static representation of the data that we explored in the previous section for feature selection.

As we saw then, a simple classification scheme such as a nearest-neighbor classifier does not produce improved classification rates (the curve in Fig. 10-1 only attains less than 40% for the optimal feature set). Instead, the nearest-neighbor criterion was used for its computational tractability in a computationally intensive feature selection algorithm. In this section we examine the performance of a more promising classification scheme.

A support vector machine (SVM) implements an approximation to the structural risk minimization principle in which both the empirical error and a bound related to the generalization ability of the classifier are minimized. Because of these discriminative training techniques, they have been reported to attain excellent generalization performance on many practical tasks. The SVM fits a hyperplane that achieves maximum margin between two classes, and its decision boundary is determined by the discriminant

$$f(\mathbf{x}) = \sum_i y_i \lambda_i K(\mathbf{x}, \mathbf{x}_i) + b, \tag{10.3}$$

where $\mathbf{x}_i$ and $y_i \in \{-1, 1\}$ are the input-output pairs, $K(\mathbf{x}, \mathbf{y}) \doteq \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$ is a kernel function which computes inner products, and $\phi(\mathbf{x})$ is a transformation from the input space to a higher dimensional space. In the linearly separable case, $\phi(\mathbf{x}) = \mathbf{x}$. An SVM is generalizable to non linearly separable cases by first applying the mapping $\phi(\cdot)$ to increase dimensionality and then applying a linear classifier in the higher-dimensional space. The parameters of this model are the values λ_i, non-negative constraints that determine the contribution of each data point to the decision surface, and b, an overall bias term. The data points for which $\lambda_i \neq 0$ are the only ones that contribute to (10.3) and are known as support vectors.

Fitting an SVM consists of solving the optimization (Vapnik, 1995; E.E. Osuna & Girosi, 1997):

$$\begin{aligned} \max \quad F(\Lambda) &= \Lambda \cdot \mathbf{1} - \frac{1}{2}\Lambda \cdot D\Lambda \\ \text{subject to} \quad \Lambda \cdot \mathbf{y} &= 0 \\ \Lambda &\leq C\mathbf{1} \\ \Lambda &\geq \mathbf{0} \end{aligned} \tag{10.4}$$

where $\Lambda = [\lambda_1 \cdots \lambda_l]'$ and D is a symmetric matrix with elements $D_{i,j} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and

C is a non-negative constant that bounds each λ_i, and which is related to the width of the margin between the classes. Having solved Λ from the equations in (10.4), the bias term can be found:

$$b = -\frac{1}{2}\sum_i \lambda_i y_i \Big(K(\mathbf{x}_-, \mathbf{x}_i) + K(\mathbf{x}_+, \mathbf{x}_i) \Big), \tag{10.5}$$

where $\mathbf{x}_-$ and $\mathbf{x}_+$ are any two correctly classified support vectors from classes -1 and $+1$ respectively (Gunn, 1998). In the work reported here we have use a Gaussian kernel of the form $K(\mathbf{x}, \mathbf{y}) = e^{\frac{||\mathbf{x}-\mathbf{y}||^2}{\sigma}}$, where σ is a free parameter that has been explored over a range of values. In the N-class problem, N binary SVMs are fit to binary classes constructed from data sets constructed from the *nth* class (class label $+1$) and its complement (class label -1). At classification time, a test token is assigned to the class modeled by the SVM outputting the highest value.

Subject No.	1	2	3	4	5	6	7	8	9	10	11	Mean
Human Perf.	60	82	79	70	78	50	72	26	48	62	42	60.8
Bayesian Nets	45	46	52	61	55	40	60	29	39	56	58	49.2
SVM	43	48	58	57	51	41	66	31	31	62	56	49.4

Table 10.25: Comparison of human performance, Bayesian networks, and SVM overall recognition rates for subject-dependent 5-emotion classification task. Last column includes averages over all subjects.

Subject-dependent SVMs were trained on the data, and their generalization error estimated through 15-fold cross-validation. After applying PCA to the data, SVMs were fit using a Gaussian kernel of variable width ($\sigma = 1, 2.5, 7.5, 10$). Table 10.5 compares the overall recognition rates attained by the best subject-dependent SVM with the recognition rate of the best structured Bayesian network, as well as the performance of human listeners. The Bayesian networks attain an overall recognition rate of 49.1% compared to 49.5% provided by the SVMs. This marginal increase is not statistically significant. Nonetheless, it is important to point out that a state-of-the-art method performs comparably to the alternative method proposed in this work. The good generalization property of SVMs can be partly responsible for the comparable performance which is attained with a representationally simpler model (recall that the dynamic structure of the data was collapsed onto the sample mean to convert to a static representation). On the other hand, the Bayesian networks remain an attractive modeling tool because of the more intuitive semantic interpretation they afford us: connections between different hierarchical levels and different nodes allow

us to embed some prior knowledge about the way we belief spoken language may be structured. An obvious future research direction that these results suggest would be to combine the representational advantage of these generative models with some discriminative training techniques which can lead to better trained models with better generalization properties.

10.6 Performance on Natural Spontaneous Speech

Having demonstrated the performance of the recognition system on subject-specific speech data from actors, we would now like to investigate whether we can relax these two restrictions and extend the applicability of the model to spontaneous, unscripted, naturally occurring speech data from several speakers. We next turn to the analysis of the CHE and Call Center datasets described earlier in Chapter 9. we have used the results of the perceptual experiments described in detail in that chapter in order to assign categories to the data. Recall that the *Valence* and *Arousal* ratings were obtained on a continuous scale, and that any categorization was deferred to the analysis stage. In order to come up with a categorical description, we have used some alternative quantizations of the continuous *Valence-Arousal* map. This is shown in Fig. 10-2. Training was combined with principal component analysis, as before, using two and three discrete states per level respectively, and unimodal Gaussian distributions with diagonal covariance matrices to model the output observations at leach level.

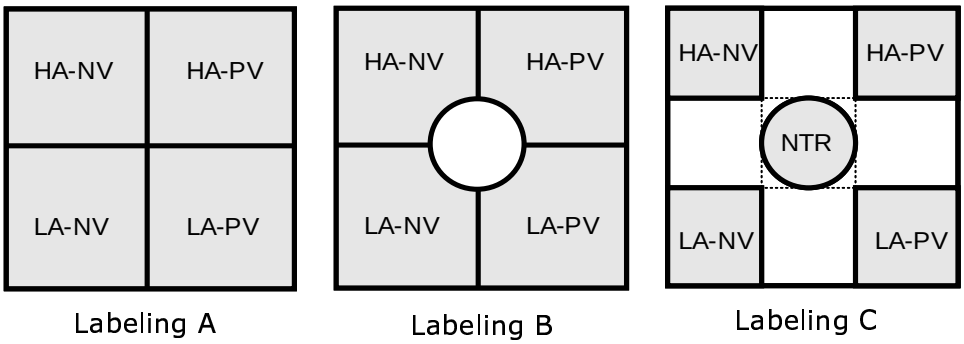

Figure 10-2: Tilings of the *Valence-Arousal* map.

Tables 10.26 - 10.27 show an application of the model to predicting affect categories from the CHE corpus. Categories have been obtained in this case by quantizing the valence arousal map into its four natural quadrants. The categories represented in this simulation are *Positive Valence-High Arousal* (PV-HA), *Negative Valence-High Arousal* (NV-HA),

CHE-A. ORR:43.12

	PV-HA	NV-HA	NV-NA	PV-LA
PV-HA	36.86	1.27	0.42	1.27
NV-HA	18.22	0.42	1.69	0.85
NV-NA	18.64	0.85	4.66	1.27
PV-LA	10.59	0.85	0.85	1.27

Table 10.26: Generalization of subject-independent classifiers on CHE data set using labeling scheme A (See Fig. 10-2) for four-category discrimination task. (Chance-level performance is 29%.) Results are normalized such that the sum of all entries is 100%

CHE-A (Equal prior subset). ORR: 42.19

	PV-HA	NV-HA	NV-NA	PV-LA
PV-HA	34.38	40.62	21.88	3.12
NV-HA	59.38	31.25	3.12	6.25
NV-LA	18.75	15.62	53.12	12.5
PV-LA	25	15.62	9.38	50

Table 10.27: Generalization of subject-independent classifiers on CHE data set using labeling scheme A (See Fig. 10-2) for four-category discrimination task using an equal-prior subsampling. (Chance-level performance is 25%.)

Negative Valence-Negative Arousal (NV-NA), and *Positive Valence-Low Arousal* (PV-LA). In table 10.26 the results have been normalized differently to illustrate a different effect obtained with this data set: All entries in the matrix in this case have been divided by the number of points in the set, so the sum of all entries yields 100% of the set. Fig. 10.26 clearly shows that the priors are not equal in this set. Although the overall recognition rate is 43.12%, the performance is skewed, so that most of the points are classified in the PV-HA quadrant. Figs. 10.27 and 10.28 show the results of two alternative training strategies to reduce this effect. Fig. 10.27 shows the generalization performance of an equal prior subsampling of the four categories. Overall performance in this case decreases slightly, but the matrix is less skewed to the first quadrant.

Fig. 10.28 shows the results of another subsampling of the four quadrants (using labeling B), this time unequally done to avoid including points close to the origin (i.e., the neutral category). The valence-arousal map has been calibrated to produce values between ± 1, and in this case only points outside the 1/4 radius from the origin are included in training. The performance rate increases in this case to 44.05. However, there are some interesting trends that still remain on Figs. 10.27 and 10.28: There is considerable confusion between the first

CHE-B (Equal prior subset). ORR: 44.05

	PV-HA	NV-HA	NV-NA	PV-LA
PV-HA	8.33	8.33	4.76	3.57
NV-HA	4.76	13.10	4.76	2.38
NV-LA	3.57	4.76	14.29	2.38
PV-LA	3.57	2.38	10.71	8.33

Table 10.28: Generalization of subject-independent classifiers on CHE data set (estimated through 15-fold cross-validation) using labeling scheme B (See Fig. 10-2) for four-category discrimination task using an equal-prior subsampling. (Chance-level performance is 25%.) Results have been normalized such that the sum of all entries is 100%.

CHE-C (Equal prior subset). ORR: 37

	Neutral	PV-HA	NV-HA	NV-NA
Neutral	30	20	40	10
PV-HA	15	60	20	5
NV-HA	29.41	23.53	41.18	5.88
NV-NA	22.22	11.11	50	16.67

Table 10.29: Generalization of subject-independent classifiers on CHE data set (estimated through 10-fold cross-validation) using labeling scheme C (See Fig. 10-2) four-category discrimination task using a near equal-prior subsampling. (Chance-level performance is 25%.)

two quadrants, so the high arousal dimension is identified correctly, but the system fails to discriminate between positive and negative valence when the arousal is high. The second observation is that the performance degrades notably with respect to the figures obtained for actors data and subject-dependent classifiers, although still attaining significance.

We have also investigated the performance of the system on the Call Center dataset described previously in Chapter 9. Unlike the CHE set, which included calls between friends and family members and showed a range of topics and affective expression, the Call Center

CHE-C (Equal prior subset). ORR: 51

	Neutral	PV-HA	NV
Neutral	30	20	50
PV-HA	15	60	25
NV	25.71	17.14	57.14

Table 10.30: Generalization of subject-independent classifiers on CHE data set (estimated through 10-fold cross-validation) using labeling scheme C (See Fig. 10-2), a near equal-prior subsampling, and combining negative valence classification results. (Chance-level performance is 36%.

dataset is a domain-specific collection of recordings showing more limited affective examples. The most typical of such examples tend to occur when callers are showing relatively strong and negative affect (such as when lodging a complaint). Not surprisingly, given the nature of the interaction, the tokens to which the listeners attributed positive valence do not typify positive affect they way that the high arousal and negative valence tokens did. In fact, automatically discriminating categories derived from the *Valence* and *Arousal* ratings that listeners had provided for this dataset proved to be more challenging. With one exception, no differences were consistently found between the different categorizations of the data that we explored. The only significant difference was obtained when the tokens with a more extreme negative valence and high arousal ratings were modeled separately from the rest, as graphically illustrated by labeling C in Fig. 10-2. Intuitively, this corresponds to those utterances from potentially irritated, frustrated, or annoyed callers. Although this is a limited result from this dataset, it is nonetheless an important one for human-computer interaction applications: much of human interaction with technology is often described as frustrating or annoying, and therefore, systems that can detect the occurrence of these episodes and respond to them have begun to receive attention in the human-machine interaction literature (Scheirer et al., 2002).

Based on these observations, and on the binary nature of the distinction the system was able to make, we have treated the problem as a detection task and summarized the results with the receiver-operating curves show in Figs. 10-3 and 10-4. Each of the curves represents the trade-off between the hit rate (the probability that a negative valence, high arousal token was correctly identified when it occurred) vs. the false alarm (the probability that the token was incorrectly identified when it was not present). Discrete states have taken values of $(1, 2)$, $(2, 3)$ and $(2, 3, 4)$ for each of the three levels respectively. As before, principal component analysis has been applied to the data, and the output distributions have been modeled using unimodal Gaussians. The best performance obtained on this data set estimates an approximately 70% detection rate at the expense of a 30% false alarm rate. Although this is a result that can certainly be improved upon, the current performance might be acceptable in systems where there might be a lower cost associated with incorrectly perceiving the negative affect of a user than in failing to detect it. Customer relationship management provides an example of an area where this can be further exploited.

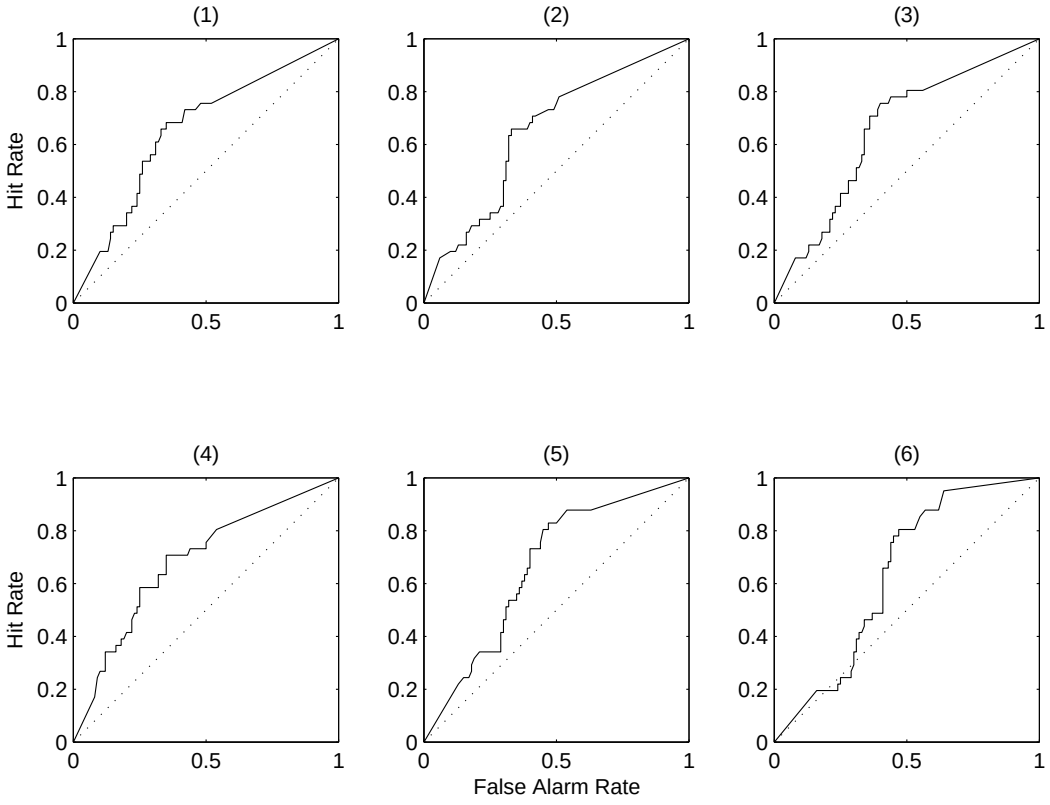

Figure 10-3: Receiver Operating Characteristics of systems trained to detect High Arousal-Negative Valence (HA-NV) tokens in Call Center database.

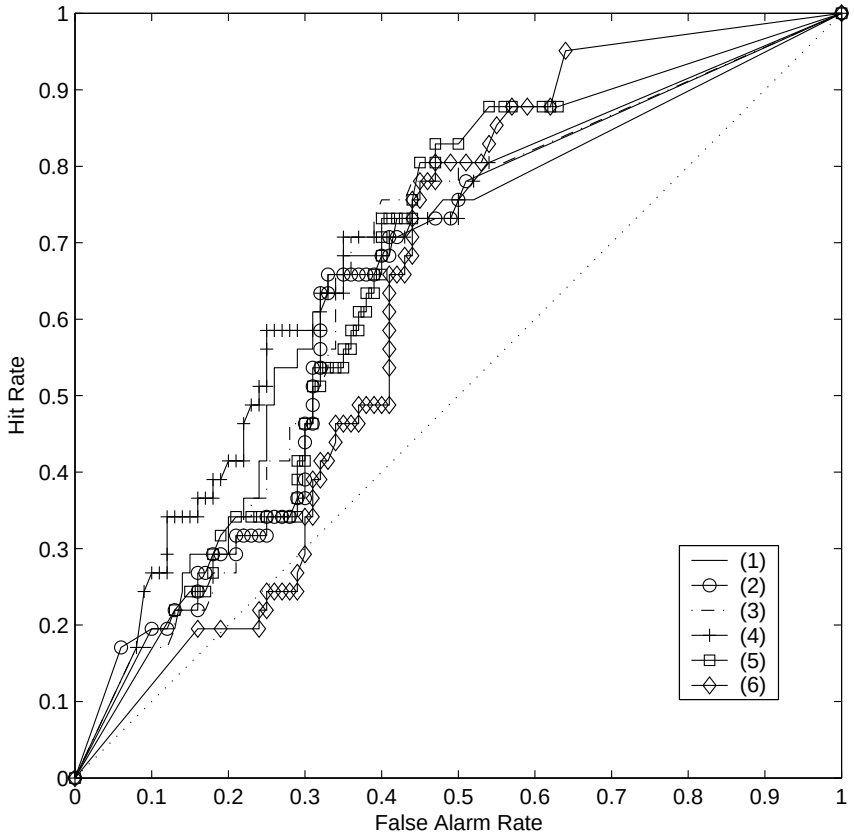

Figure 10-4: Combined Receiver Operating Characteristics of systems trained to detect HA-NV tokens in Call Center database.

10.7 Discussion and Comparison with Other Work

The variability that we have found in human decoding and machine recognition performance points to the difficulty of establishing comparisons across different research works in the literature. Even when we fix such variables as methodology or the lexical content of the speech tokens, and we collect the data consistently across different speakers, the variability that humans exhibit while emoting can often override any other normalization effects. Bearing this word of caution in mind, we review other results related to this research area.

The best speaker-dependent systems that we trained attained an average of 60% on a four emotion classification task on actors data, ranging from 34% to 77%. Because of the observed subject dependency on performance and the fact that human decoding experiments mirrored this trend, it is perhaps instructive to describe the performance in terms of groups: The average performance in the lower quartile group was 44% compared to 53.6% human performance. In the upper quartile group, performance averaged 69.3% compared to 73% human performance. Finally, in the inter-quartile group, machine recognition averaged 59.4% compared with a human decoding rate of 80.45%.

The work with which a most direct comparison might be possible is that by Polzin (2000), who reports an average recognition rate of 60.4% using prosodic cues for four subjects taken from this set of actors recordings. However, we should caution that it is not clear to which actors these results apply, and therefore a more specific comparison is not possible. Another substantial difference that impedes a more direct comparison lies in the nature of the use of prosody in that work. Polzin (2000) used a speech recognizer that incorporates prosodic information at training time. By supplying affect-specific prosodic cues as well as a textual transcription of the utterances, he constructs affect- and prosody-specific acoustic models of Mel spectral coefficients. He then uses the likelihood of the spectral coefficients *given* a textual transcription and an affective state to predict the class of the data by selecting the affective state that maximizes this likelihood. That approach departs from the work presented here, where no textual transcription of the text is assumed, and where the analysis is purely based on acoustic parameters extracted from the waveform.

Other research on basic emotion classification includes the work of Huang et al. (1998), who, using energy- and pitch-related features, report machine recognition rates of acted speech averaging 75% for one speaker of Spanish and 75% for one speaker of Sinhala for

the set of four affective categories treated in this work plus categories for *Dislike* and *Surprise*. This same work reports that human performance on this limited dataset, using only audio, averages 42% and 33% for each speaker respectively. It is noteworthy that machine performance significantly surpasses human ratings, a result that is not further addressed in the paper. This result suggests, nonetheless, that each of these speakers may be employing particular acoustic patterns distinctively and idiosyncratically as a function of intended affective expression (a feature which might aid in machine discrimination though not necessarily in human performance if the patterns are not observable in speakers at large). That is, the data may not be particularly *natural*, even for acted speech.

The work of Batliner et al. (2003) bridges the gap between performance on actors data and performance on spontaneous emotion. Employing Wizard-of-Oz scenarios designed to elicit emotion during human-computer interaction, Batliner et al. (2003) have analyzed the performance of a system designed to discriminate in a binary classification task between *emotional* and *non-emotional* tokens in German. *Emotional* utterances are described as those which contain primarily angry reactions to a malfunctioning interface. In the acted condition, they are imagined, or internalized, by an actor. In the spontaneous case, they are naturally elicited by the Wizard-of-Oz design. Using 52 utterances from one speaker who is described as a very experienced actor, the system is able to attain a 95% average recognition rate based on prosodic features derived from energy, duration, pitch, and pause occurrences. This performance degrades to approximately 73% when evaluating a set of 43 *emotional* utterances which occur spontaneously. The performance figures are estimated by leave-one-out cross validation.

Along these lines, Ang et al. (2002) have reported recognition rates for frustration detection in human-computer dialog tasks. Their work is different in that it distinguishes between *Frustration* and *Annoyance*. It should also be made clear that the nature of the frustration experienced by their users (i.e., typically errors made by automatic systems in human computer dialogs) is different from that examined for the Call Center database (which also included general complaints which were not resulting from the use of the technology they were employing while calling). It might therefore be expected that the acoustic signaling of such frustration is also different. Using acoustic features related to pitch, energy, spectral tilt, pausing as well as duration and speaking rate, they report a baseline recognition rate of 72.6% for the categories *Frustration* and *Annoyance* combined. This rate increases to

approximately 85% when considering only the tokens labeled as *Annoyed*, which constituted a small fraction (about 7%) of the data set.

Similar work has been carried out in discriminating *negative* from *non-negative* emotions in spoken dialogs by Lee & Narayanan (2002; Lee et al. (2002). Lee & Narayanan (2002) report a recognition rate of approximately 70% using only pitch- and energy-related acoustic features. An improvement to this rate using Support Vector Machines (74%) is reported by Lee et al. (2002). Their work benefits from using a portion of the dataset for which perfect agreement in labeling *negative* vs. *non-negative* was achieved by external judges. It is possible that this methodological difference (using such a categorical distinction *a priori* rather than the continuous *Valence-Arousal* model employed in this work) can isolate a more homogeneous set of tokens for automatic classification.

In all the work cited here, the prosodic features that have received attention consist of measures derived from the F0 contour, energy, and timing. A notable difference between that work and the research presented here is the inclusion of a large set of features related to voice quality, as well as a set of measures derived from a perceptual model of loudness. As we have shown, those features prove to be good predictors of affective categories, outperforming some of the more "classical" features, such as intonational or durational features.

10.8 Chapter Summary

In this chapter we have presented and discussed the results of conducting a series of evaluation tests to assess the performance of the computational model advanced in this work. We have applied the machine learning models described earlier to the task of classification of affective categories from speech. We have used two different types of data to carry out the evaluation. Using a set of actors data, we demonstrated the ability of the model to learn affective categories in a speaker-dependent fashion, and to generalize this categorization to significant levels for 10 out of 11 subjects. Performance was found to be particularly improved when only 'full-blown" emotions were considered, suggesting there is an inherent overlap, one that remains challenging to model, between the neutral condition and the remaining set of emotions.

We also tested the performance of simpler models, obtained by trimming the hierarchical structure of the Bayesian networks, and showed that performance benefits from a simpler

structure even at the expense of eliminating subsets of features. This behavior may be partially due to the fact that the most discriminative features (those related to voice quality and loudness), are modeled at the highest level of the hierarchy, and therefore most of the significant modeling power is retained in the simpler models while losing model complexity. The relevance of these features was another topic explored in this chapter using the forward floating search method of feature selection coupled with a simple nearest-neighbor evaluation criterion to assess the effectiveness of different feature subsets.

We have also tested whether the model is capable of learning affective categorizations when the training data occurred unelicited, spontaneously and in natural manner, and whether it could do so in a speaker-independent fashion. The performance in this case degrades. However, the results show that significant discrimination can still be obtained. This was applied to modeling categories derived from the *Valence-Arousal* descriptions we had introduced in earlier chapters.

Part IV

Conclusion

Chapter 11

Contributions, Future Directions and Concluding Remarks

This thesis has proposed a model for the automatic recognition of affect in speech which is based on the prosodic structuring of spoken language and on the distribution of acoustic parameters at different prosodic hierarchical levels to build learning machines. The contributions in this work can be found in several of the major components of the model, among which we can cite the following:

- Investigating the relative contribution of different aspects of spoken language (perceptual loudness, intonation, voice quality, rhythm) to affect classification, benchmarked through the generalization performance of automatic learning schemes.
- Introducing novel features and novel applications of established methods to define new features for affect analysis. In this area, we can highlight:
  - Defining new features derived from an established perceptual model of loudness for affect modeling.
  - Applying a model of perceptual consonance to characterizing the spectral harmonics patterns from voice and, based on this analysis, introducing novel features for affect modeling.
- Developing and implementing a prosodic parsing algorithm that makes no use of lexical or segmental information to blindly produce a decomposition of an acoustic waveform into phrases and syllables.

- Demonstrating the validity of the theoretical valence-arousal model as a viable tool for annotating speech corpora with affective labels.

- Proposing a graphical model framework that naturally integrates acoustic phenomena taking place at several time scales. This framework, though comparable in performance to some discriminative training methods, affords a natural and amenable representation of hierarchical and dynamic processes that is lacking in other methods. The generality of the representation also allows the designer to begin the analysis with a full (and more complex) model, and through some structured simplifications that we have addressed in the thesis, arrive at simpler models (like HMMs and standard Gaussian distributions) as specific cases of the general framework.

The work in this thesis has only scratched the surface of what remains an open and challenging research area. Much needs to be investigated and understood before the research agenda on which this work is founded can come to full fruition and allow the interaction between humans and machines to approach the level of humans acting as equals.

Many of these open questions are of a fundamental theoretical nature. It will no doubt help us to have a better understanding of the processes underlying the human encoding and perception of affective content, through its many modalities, and to better delineate the bounds of what is accomplishable through each. Future research should also aim for the development of efficient and robust algorithms for acoustic processing that can accommodate the kinds of eventualities that are likely to occur with speech (and with affective speech in particular).

Computationally, the models described here are of perhaps naive simplicity compared to what is needed to fully account for the subtlety and range of human expression. One ambitious but immediate extension is to link the prosodic representations obtained from speech, and the acoustic parameters extracted from the structurally viewed time domain, with lexical and semantic representations to build an affective lexicon of sorts, a structure that ties together lexical entries, semantic entries, acoustic patterns, prosodic structures, any knowledge of contextual and pragmatic factors, and affective representations.

From the point of view of streamlining the research and liberating it from laboratory constraints, there is a growing need for creating efficient data gathering and data annotation systems that bypass to some extent the need to rely on experimental paradigms to procure

some insights into the ground truth of the data we seek to model. Humans are constantly listening to speech and, very likely, making inferences a great deal of the time about its affective content. At least for a research agenda that seeks to emulate this perceptual mechanism, it would be good to tap into this process instead of simulating it in the laboratory with more limited data.

Finally, this research agenda can not only benefit, but in fact needs for its success, an interdisciplinary treatment. It has been a goal of this thesis to be informed by the knowledge lying at the intersection of several disciplines. It is hoped that it has learned from it, and that it has in return offered a humble contribution.

Appendix A

Nonlinear Minimization for Problems Subject to Bounds

A.1 A Trust-Region Minimization Algorithm for Nonlinear Problems Subject to Bounds

Trust-region algorithms are a general class of iterative algorithms that proceed with the minimization of an *objective* function by replacing it, at every step of the iteration, with a simpler *model* function resembling the objective function within a neighborhood (the trust region), and minimizing the model instead. If taking the step which minimizes the model also produces a significant decrease in the objective function, the step is accepted, the trust region is enlarged, and a new iterate is produced. Otherwise, the step is rejected, and the trust region is shrunk to produce a better match between the model and objective functions.

The algorithm discussed in this section, due originally to Coleman & Li (1996), considers the general problem of computing a local minimizer of a smooth nonlinear function subject to bounds on its variables:

$$\min_{\theta \in \mathcal{R}^n} \{f(\theta): \quad \theta_l \leq \theta \leq \theta_u\}.. \tag{A.1}$$

For a point θ^* to be a critical point (i.e., a local extremum or saddle point) of $f(\theta)$, it has to satisfy the following conditions (known as the Karush-Kuhn-Tucker conditions) on the

gradient $\mathbf{q}(\theta) = \nabla_\theta f(\theta)$:

$$\begin{cases} [\mathbf{q}(\theta^*)]_i = 0 & \text{if} \quad [\theta_l]_i < [\theta]_i < [\theta_u]_i \\ [\mathbf{q}(\theta^*)]_i \leq 0 & \text{if} \quad [\theta]_i = [\theta_u]_i \\ [\mathbf{q}(\theta^*)]_i \geq 0 & \text{if} \quad [\theta]_i = [\theta_l]_i \end{cases} \tag{A.2}$$

where $[\cdot]_i$ for $1 \leq i \leq n$ is the ith component of a vector. It can be shown that a point θ^* that satisfies these conditions, also satisfies the following system of nonlinear equations

$$h(\theta^*) = S(\theta^*)^2 \nabla_\theta f(\theta^*) = 0, \tag{A.3}$$

where $S(\theta)$ and $\mathbf{v}$ are defined as follows

$$S(\theta) \triangleq \text{diag}\left(\left|[\mathbf{v}(\theta)]_1\right|^{\frac{1}{2}}, \cdots, \left|[\mathbf{v}(\theta)]_n\right|^{\frac{1}{2}} \right) \tag{A.4}$$

$$[\mathbf{v}]_i \triangleq \begin{cases} [\theta]_i - [\theta_u]_i & \text{if} \quad [\mathbf{q}]_i < 0 \text{ and } [\theta_u]_i < \infty \\ [\theta]_i - [\theta_l]_i & \text{if} \quad [\mathbf{q}]_i \geq 0 \text{ and } [\theta_l]_i > -\infty \\ -1 & \text{if} \quad [\mathbf{q}]_i < 0 \text{ and } [\theta_u]_i = \infty \\ 1 & \text{if} \quad [\mathbf{q}]_i \geq 0 \text{ and } [\theta_l]_i = -\infty. \end{cases} \tag{A.5}$$

The problem of minimizing a function with bounded constraints, therefore, can be viewed as solving an unconstrained system of equations[1]. Following this reformulation of the minimization problem, Coleman and Li's algorithm proceeds by (i) formulating the Newton process associated with the solution of Eq. A.3, (ii) introducing a quadratic model function to replace the objective function $f(\theta)$, and (iii) recognizing that the resulting Newton process is associated with the minimization of a scaled quadratic problem, and hence the two solutions are related, and optimization of the auxiliary problem leads to optimization of the original.

In order to solve for the optimal value θ^* (that is, to find the roots of Eq. A.3) we can apply Newton's method. Newton's method consists of a series of iterations to solve for the

[1]The reader may be more familiar with the equivalent result for the unconstrained case (i.e.;$\theta_l = -\infty$ and $\theta_u = \infty$), where $S(\theta) = I$ and θ^* satisfies $\nabla_\theta f(\theta^*) = 0$)

roots of an equation $h(\theta) = 0$ by updating the current estimate θ_k with a step s_k obeying

$$\nabla_\theta h(\theta_k) s_k = -h(\theta_k). \tag{A.6}$$

The main quantity involved in this calculation is the derivative of the vector

$$h(\theta) = \begin{pmatrix} |v_1(\theta_1, \cdots, \theta_n)| & & \\ & \ddots & \\ & & |v_n(\theta_1, \cdots, \theta_n)| \end{pmatrix} \begin{pmatrix} \frac{\partial f(\theta_1, \cdots, \theta_n)}{\partial \theta_1} \\ \vdots \\ \frac{\partial f(\theta_1, \cdots, \theta_n)}{\partial \theta_n} \end{pmatrix}. \tag{A.7}$$

The (i, j) entry of the derivative of this matrix is given by

$$\frac{\partial [h(\theta)]_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left(|v_i(\theta_1, \cdots, \theta_n)| \frac{\partial f(\theta_1, \cdots, \theta_n)}{\partial \theta_i} \right) \tag{A.8}$$

$$= |v_i(\theta_1, \cdots, \theta_n)| \frac{\partial^2 f(\theta_1, \cdots, \theta_n)}{\partial \theta_j \partial \theta_i} + \frac{\partial f(\theta_1, \cdots, \theta_n)}{\partial \theta_i} \frac{\partial |v_i(\theta_1, \cdots, \theta_n)|}{\partial \theta_j}, \tag{A.9}$$

which can be arranged into matrix notation to obtain

$$\nabla_\theta h(\theta) = S(\theta)^2 \nabla_{\theta\theta} f(\theta) + QJ, \tag{A.10}$$

where

$$Q = \mathrm{diag}\Big([\nabla_\theta f(\theta)]_1, \cdots, [\nabla_\theta f(\theta)]_n \Big) = \mathrm{diag}\Big([\mathbf{q}]_1, \cdots, [\mathbf{q}]_n \Big) \tag{A.11}$$

and J is the (diagonal) Jacobian matrix of the vector $|\mathbf{v}(\theta)|$ with entries

$$[J]_{ii} = \begin{cases} \mathrm{sgn}(\mathbf{q}(\theta)]_i) & \text{if} \quad [\theta_u]_i < \infty \quad \text{and} \quad [\theta_l]_i > -\infty \\ 0 & \text{otherwise.} \end{cases} \tag{A.12}$$

Combining Eqs. A.3, A.6 and A.10, we obtain the Newton process

$$\Big(S(\theta_k)^2 \nabla_{\theta\theta} f(\theta_k) + Q_k J_k \Big) \mathbf{s}_k = -S(\theta_k)^2 \mathbf{q}_k. \tag{A.13}$$

If the objective function $f(\theta)$ is now replaced with a quadratic model m_k^f, with gradient $\mathbf{q}$

and Hessian H

$$m_k^f(\theta_k + s) = f(\theta_k) + \langle \mathbf{q}_k, \mathbf{s}_k \rangle + \frac{1}{2}\langle \mathbf{s}, H_k \mathbf{s} \rangle, \tag{A.14}$$

then Eq. A.13 becomes

$$\Big(S_k^2 H_k + Q_k J_k\Big)\mathbf{s}_k = -S_k^2 \mathbf{q}_k, \tag{A.15}$$

and finally, by left-multiplying by S_k^{-1} and introducing the change of variable $\mathbf{s}_k = S_k \hat{\mathbf{s}}_k$:

$$\Big(S_k H_k S_k + Q_k J_k\Big)\hat{\mathbf{s}}_k = -S_k \mathbf{q}_k, \tag{A.16}$$

where we have made use of the fact that $Q_k J_k$ is diagonal and thus unaltered by left- and right-multiplication with matrices inverse of each other. We now make the observation that Eq. A.16 corresponds to Newton's equation for the minimization of the quadratic

$$\begin{aligned} \hat{m}_k(\theta_k + \hat{\mathbf{w}}) &= f(\theta_k) + \langle S_k \mathbf{q}_k, \hat{\mathbf{s}}_k \rangle + \frac{1}{2}\langle \hat{\mathbf{s}}, S_k(H_k + M_k)S_k \hat{\mathbf{s}} \rangle & \text{(A.17)} \\ &\triangleq f(\theta_k) + \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}_k \rangle + \frac{1}{2}\langle \hat{\mathbf{s}}, (\hat{H}_k + \hat{M}_k)\hat{\mathbf{s}} \rangle, & \text{(A.18)} \end{aligned}$$

where

$$M_k = S_k^{-2} Q_k J_k. \tag{A.19}$$

This result suggests that if $\hat{\mathbf{s}}_k$ is an (approximate) minimizer of A.18, then taking a step $\mathbf{s}_k = S_k \hat{\mathbf{s}}_k$ helps minimize the objective function, as long as the model m_k^f (from which $\hat{m}_k$ in Eq. A.18 was obtained) accurately resembles the objective function. That is to say, we wish to find a minimizer of $\hat{m}_k$ in a trust region $\{\hat{\mathbf{s}} \in \mathcal{R}^n : ||\hat{\mathbf{s}}|| \leq \Delta_k\}$. Before summarizing the algorithm in full, we last observe that the minimization of the scaled step $\hat{\mathbf{s}}_k$ within this spherical trust region is equivalent to the minimization of the original step $\mathbf{s}_k$ within an ellipsoidal trust region; in other words, the scaling is incorporated into the geometry of the trust region (see Conn et al. (2000) for details). Therefore, we restate the problem as one of minimizing the model

$$m_k(\theta_k + \mathbf{s}) = f(\theta_k) + \langle \mathbf{q}_k, \mathbf{s} \rangle + \frac{1}{2}\langle \mathbf{s}, (H_k + M_k)\mathbf{s} \rangle \tag{A.20}$$

within the scaled trust region $\{\mathbf{s} \in \mathcal{R}^n : ||S_k^{-1}\mathbf{s}|| \leq \Delta_k\}$. The algorithm is summarized below.

Algorithm A.1

1. Initialization: Select an initial point θ_0, an initial trust-region radius Δ_0, and constants η_1, η_2, γ_1, and γ_2 satisfying $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_1 \leq \gamma_2 < 1$. Let $\mathcal{F}$ be the feasible set (θ_l, θ_u), and set $k = 0$.

2. Model definition: Compute $f(\theta_k)$, its gradient $q(\theta_k)$, and (approximate) Hessian H_k. Using Eqs. A.4, A.5, A.11, A.12 and A.19, find S_k, M_k and set up the quadratic model

$$\psi_k(\mathbf{s}) = m_k(\theta_k + \mathbf{s}) - m_k(\theta) = \langle \mathbf{q}_k, \mathbf{s} \rangle + \frac{1}{2}\langle \mathbf{s}, (H_k + M_k)\mathbf{s} \rangle \tag{A.21}$$

3. Step calculation: Compute a step $\mathbf{s}_k$, with $\theta_k + \mathbf{s}_k \in \text{int}(\mathcal{F})$ based on the subproblem

$$\min_{\mathbf{s}} \{ \psi_k(\mathbf{s}) : ||S_k^{-1}\mathbf{s}||_2 \leq \Delta_k \} \tag{A.22}$$

4. Acceptance of trial point: Compute

$$\rho_k = \frac{f(\theta_k + \mathbf{s}_k) - f(\theta_k) + \frac{1}{2}\langle \mathbf{s}_k, M_k\mathbf{s}_k \rangle}{\psi_k(\mathbf{s}_k)} \tag{A.23}$$

   If $\rho_k \geq \eta_1$, define $\theta_{k+1} = \theta_k + \mathbf{s}_k$; otherwise, $\theta_{k+1} = \theta_k$.

5. Trust-region radius update: Set

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, \infty) & \text{if} \quad \rho_k \geq \eta_2, \\ [\gamma_2\Delta_k, \Delta_k] & \text{if} \quad \rho_k \in [\eta_1, \eta_2), \\ [\gamma_1\Delta_k, \gamma_2\Delta_k] & \text{if} \quad \rho_k < \eta_1. \end{cases} \tag{A.24}$$

   Let $k \leftarrow k + 1$ and go to step 2.

Figure A-1: Trust-region optimization algorithm.

It remains to be specified, however, how to actually compute a step $\mathbf{s}_k$ in Eq. A.22 with the given constrains (step 3 in the algorithm). We next turn to this question.

A.2 Computing a Step

It can be shown that finding the optimal solution to the problem

$$\min_{\mathbf{s}} \{ \langle \mathbf{b}, \mathbf{s} \rangle + \frac{1}{2}\langle \mathbf{s}, A\mathbf{s} \rangle : ||\mathbf{s}||_2 \leq \Delta \} \tag{A.25}$$

with A diagonalizable as $A = U\text{diag}(\lambda_i, \cdots, \lambda_n)U^T$ and $\gamma \doteq -U^T\mathbf{b}$ involves solving the following secular equation in λ (Conn et al., 2000)

$$\frac{1}{\Delta} - \frac{1}{\sum_{i=1}^{n}\left(\frac{[\gamma]_i}{\lambda_i+\lambda}\right)^{\frac{1}{2}}} = 0, \tag{A.26}$$

from which the optimal step can then be obtained from the relation $(A+\lambda I)\mathbf{s} = -\mathbf{b}$. Solving Eq. A.26 while maintaining numerical stability is computationally intensive. However, reliable algorithms exist when the dimension of the space (i.e., n) is small. This has motivated the introduction of subspace approaches where s is required to lie in a subspace $\mathcal{S} \in \mathcal{R}^n$ which is properly defined to guarantee convergence of the overall algorithm in Fig. 3 (Shultz et al., 1985; Byrd et al., 1988; Branch et al., 1999). The particular approach of Coleman and Li replaces the step-calculation subproblem in Eq. A.22 with

$$\min_{\mathbf{s}_k}\{\langle \mathbf{q}_k, \mathbf{s}_k\rangle + \frac{1}{2}\langle \mathbf{s}_k, C_k\mathbf{s}_k\rangle : ||S_k^{-1}\mathbf{s}||_2 \leq \Delta_k, \mathbf{s}_k \in \text{span}(S^2\mathbf{q}_k, \mathbf{v}_k)\}, \tag{A.27}$$

where we have used $C_k \doteq H_k + M_k$ and the subspace is spanned by (i) the direction of the gradient and (ii) a vector $\mathbf{v}_k$ defined to be either the exact Newton step ... or a vector pointing in a direction of negative curvature. The details of finding $\mathbf{v}_k$ are omitted here, but Branch et al. use a preconditioned conjugate gradient method which either converges to the Newton step or exits when a negative curvature direction is found (see Branch et al. (1999) for details). The authors have shown that restricting the search of the step to this subspace retains the overall convergence properties of algorithm III, and that therefore the approximation is valid.

By incorporating the scaling explicitly into the variables (and not into the geometry of the trust region), as discussed in 3.2.1., Eq. A.27 becomes

$$\min_{\hat{\mathbf{s}}_k}\{\langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}_k\rangle + \frac{1}{2}\langle \hat{\mathbf{s}}_k, \hat{C}_k\hat{\mathbf{s}}_k\rangle : ||\hat{\mathbf{s}}||_2 \leq \Delta_k, S_k\hat{\mathbf{s}}_k \in \text{span}(S^2\mathbf{q}_k, \mathbf{v}_k)\}, \tag{A.28}$$

where $\hat{\mathbf{q}}_k = S_k\mathbf{q}_k$ and $\hat{\mathbf{s}}_k = S_k^{-1}\mathbf{s}_k$. If we let V_k be the matrix containing as its columns the vectors used to define the subspace in A.27, and Y_k be an orthonormalization of the scaled version $S_k^{-1}V_k$, then we can rewrite the scaled step as $S_k^{-1}\mathbf{s}_k = S_k^{-1}V_k\mathbf{s}_{V_k} = Y_k\mathbf{s}_{Y_k}$, where $\mathbf{s}_{Y_k}$ is now a vector in the reduced two-dimensional space and $\mathbf{s}_k = S_kY_k\mathbf{s}_{Y_k}$. Eq. A.28 then

becomes

$$\min_{\mathbf{s}_{Y_k}}\{\langle \hat{\mathbf{q}}_k, Y_k \mathbf{s}_{Y_k}\rangle + \frac{1}{2}\langle Y_k \mathbf{s}_{Y_k}, \hat{C}_k Y_k \mathbf{s}_{Y_k}\rangle : ||\mathbf{s}_{Y_k}||_2 \leq \Delta_k\}. \tag{A.29}$$

Before fully specifying the algorithm to obtain a step, we need to introduce two additional notions exploited by this algorithm. One is needed in order to guarantee feasibility; the other is introduced to speed up convergence. Notice that in the formulation of Eq. A.29 we have not made explicit use of the boxed constrains, and therefore, the step may lie outside the boundary of the feasible set. The algorithm proposed by Branch et al. (1999) allows this kind of transgression, and then considers a modified truncated step version of the unconstrained solution. Let $\mathbf{s}_k$ be the optimal solution to the step problem, and let the truncated step be

$$\mathbf{s}_k^* = \begin{cases} \mathbf{s}_k & \text{if} \quad \theta_{\mathbf{k}} + \mathbf{s}_k \in \text{int}(\mathcal{F}) \\ \eta \mathbf{s}_k & \text{otherwise} \end{cases} \tag{A.30}$$

for some value $\eta_l < \eta \leq 1$ that enforces feasibility (once the search direction $\mathbf{s}$ is determined, a value for η can be found by interpolation).

The last element of the algorithm consists of applying a reflection transformation to a step. Given a step $\mathbf{s}_k$, let i denote the first (lower or upper) bound constraint crossed by the step, and define its reflection as $\mathbf{s}_k^r = \mathbf{s}_k$ except in the ith component where $[\mathbf{s}_k^r]_i = -[\mathbf{s}_k]_i$. Although applying a reflection to a step may seem non-intuitive, Coleman and Li (Coleman & Li, 1994; Coleman & Li, 1996) show that the transformation obeys convergence requirements and illustrate how it speeds up convergence. The algorithm proposed by Branch et al. (1999) to solve the trust-region subproblem proceeds by considering three candidate steps at every iterate and retaining the winner from (i) the (truncated) subspace solution to Eq. A.22, (ii) the (truncated) reflection of this step, and (iii) the (truncated) steepest descent step. The full algorithm is summarized in Fig. A-2.

Algorithm A.II Given S, H and M from Eq. A.21, let $C = H + M$ and $\hat{C} = SCS$. The subscript k is omitted, as it is understood the following steps correspond to a single iteration of the trust-region algorithm (i.e., step 3 in Algorithm A.I).

1. Set up the subspace: Given $\mathbf{q}$ and C in the quadratic model in Eq. A.21, use the preconditioned conjugate gradient (PCG) algorithm (Branch et al., 1999) to find a direction $\mathbf{v}$ (PCG returns a direction which reduces the Newton residual in A.13 or a direction of negative curvature). Let $V = [S^2\mathbf{q} \quad \mathbf{v}] \in \mathcal{R}^{n\times 2}$, and Y an orthonormalization of the columns in $S^{-1}V$.

2. Solve the subspace secular equation: Set up the model in Eq. A.29. Diagonalize $A = Y^T\hat{C}Y = U\,\mathrm{diag}(\lambda_1, \lambda_2)U^T$ and let $\gamma = -U^TY^TS\mathbf{q}$. Solve the secular equation A.26 for λ (see Appendix in (Coleman & Li, 1992)), and let $\mathbf{s}_{sec_Y} = -(A + \lambda I)^{-1}Y^T\mathbf{q}$ and $\mathbf{s}_{sec} = SY\mathbf{s}_{sec_Y}$.

3. Choose the minimum of three candidates. Let $\mathbf{s}_1 = (\mathbf{s}_{sec})^*$ $\mathbf{s}_2 = (\mathbf{s}^r_{sec})^*$, and $\mathbf{s}_3 = (S^2\mathbf{q})^*$. Then
$$\mathbf{s} = \underset{\mathbf{s}_i}{\mathrm{argmin}}\psi(\mathbf{s}_i). \tag{A.31}$$

Figure A-2: Finding a step.

Appendix B

A Model of Dissonance

In this appendix, we present the model of dissonance that underlies the chordal and harmonic consonance analyses that were discussed in Chapters 5 and 6. The germ of this model is the effect described by Plomp & Levelt (1965) concerning the perception of the interval between two pure tones. The dissonance of such a frequency pattern is minimum at unison, then quickly rises to a dissonance maximum, from where it decays toward consonance again. This effect may be described by a parametric model of the form in Eq. B.1:

$$d(f_1, f_2, a_1, a_2) = a_{12}\Big(\exp^{-b_1 s(f2-f1)} - \exp^{-b_2 s(f2-f1)}\Big), \tag{B.1}$$

where (f_1, a_1) and (f_2, a_2) are the frequency-amplitude pairs of each pure tone, and

$$a12 = a_1 a_2 \tag{B.2}$$

$$s = \frac{d^*}{s_1 f_1 + s_2}. \tag{B.3}$$

The remaining constants in Eq. B.1 are found by a least-squares fit between the parametric model and experimental data. Sethares (1998) provides the following values:

$$b_1 = 3.51 \tag{B.4}$$

$$b_2 = 5.75 \tag{B.5}$$

$$s_1 = 0.0207 \tag{B.6}$$

$$s_2 = 18.96 \tag{B.7}$$

$$d^* = 0.24 \tag{B.8}$$

Eq. B.1 not only describes the general shape of the rise-and-fall effect obtained by Plomp & Levelt (1965), it also incorporates a weighting proportional to the amplitude of the two tones to soften the effect of weak tones, and normalizes the curve so that different curves are described as a function of the lower tone. Fig. B-1 shows several such curves for a two-tone pattern of unit amplitudes and the indicated values of f_1, revealing sharper curves as the lower tone increases.

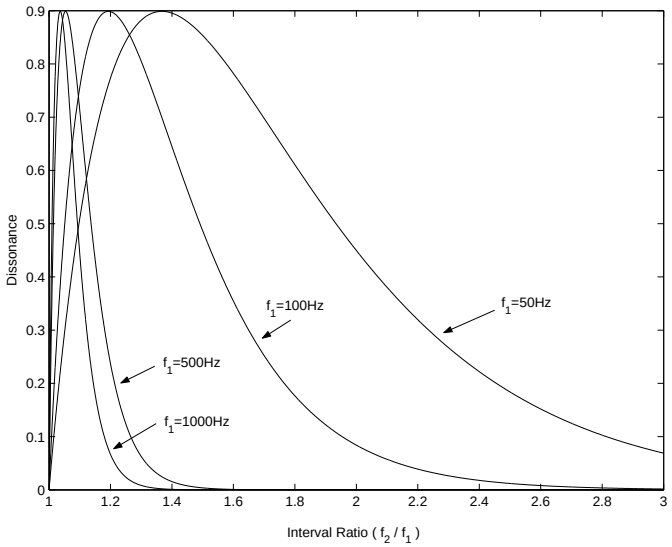

Figure B-1: Dissonance curves

Sethares (1998) extends this basic model to analyze the dissonance of a more complex frequency pattern $F = \{f_1 < f_2 < \cdots f_n\}$ with amplitudes $a_1, a_2, \cdots, a_n$ (what he calls a *timbre*), by allowing Eq. B.1 to describe the dissonance between all pairs of frequencies f_j and f_k resulting from the interaction of the pattern F and a pattern αF at an interval of ratio α. The intrinsic or inherent dissonance of the patterns F and αF is the linear interaction of all pairwise components in each pattern:

$$D_F = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} d(f_j, f_k, a_j, a_k) \tag{B.9}$$

$$D_{\alpha F} = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} d(\alpha f_j, \alpha f_k, a_j, a_k), \tag{B.10}$$

and the total dissonance, as a function of the interval α, is given by the sum of these two terms, and the cross-interactions of the frequencies from each pattern:

$$D_F(\alpha) = D_F + D_{\alpha F} + \sum_{j=1}^{n} \sum_{k=1}^{n} d(f_j, \alpha f_k, a_j, a_k). \tag{B.11}$$

The curve thus constructed no longer exhibits the simple rise-and-fall behavior of Eq. B.1. Instead, a pattern F can have points of maximum consonance and dissonance at arbitrary intervals based on the interaction of the constitutive frequencies and their amplitudes. One of the fundamental implications of this model is that the concept of (musical) scale is intimately tied up to the frequency components and amplitudes that make up a pattern since a pattern may be synthesized to achieve locally maximal consonance and dissonance at arbitrary intervals. From the analytic point of view, dissonance curves offer a tool to quantify the dissonance for a spectral pattern, whether the sequence of fundamental and upper partials produced by a musical instrument or the F0 harmonic pattern of a fragment of voiced speech.

Appendix C

Propagation of Probabilities in Graphical Models

In this appendix, we will review basic existing algorithms for updating probabilities in a graphical model with discrete and continuous nodes. It will be assumed that a junction tree T representation of the graphical model is available, and that we wish to update the probabilistic representations at each clique of the tree (i.e., the clique potentials) to reflect some newly introduced evidence at some nodes. The ultimate goal is that, after the algorithm's convergence, we have a tree that is globally consistent. That is, for two cliques A and B with intersection I, $\sum_{A\setminus I}\phi_A = \sum_{B\setminus I}\phi_B$ (where $\phi_{(\cdot)}$ are the potential functions associated with each clique, and the notation $A\setminus I$ denotes the subset of nodes in clique A excluding those in I). (Global consistency ensures that a marginal over any subset of variables can be obtained from any clique containing the variables.)

Although we will shortly be more explicit on what it means to pass probabilistic messages, let for now PassMessage(A, B) be an abstract procedure that operates on the probabilities of cliques A and B, such that A and B are locally consistent. It can be shown (Jensen, 1996) that global consistency can be achieved by an algorithm that passes messages in both directions between each pairs of nodes while respecting the following protocol: *a clique A can send a message to a neighbor B only when it, in turn, has received messages from all its neighbors.* HUGIN propagation is one such algorithm to implements this protocol based on the following recursive procedures:

- DistributeEvidence(A): Clique A executes PassMessage(A, B) on each of its neigh-

bors B, and each neighbor B recursively calls DistributeEvidence on its remaining neighbors (excluding A).

- CollectEvidence(A): Clique A asks each of its neighbors B to pass them a message (i.e., to run PassMessage(B, A)). If they are not allowed to do so (i.e., they have not received messages from their remaining neighbors yet), they recursively call CollectEvidence on all their neighbors except A.

HUGIN propagation then implements the following

1. Designate a node of the tree as the root Rt.
2. Call CollectEvidence(Rt).
3. Call DistributeEvidence(Rt).

At the conclusion of this procedure, the junction tree T is globally consistent.

Let us now define PassMessage(A, B) from clique A to B with potentials ϕ_A and ϕ_B respectively, and separator S with potential ϕ_S. We will assume that S contains only discrete variables, as that is the case for the graphical structures proposed in Chapter 8. It is possible, however, to consider more general settings with continuous variables or hybrid cases with continuous and mixed variables. For an extension of the algorithm to those cases, see (Lauritzen, 1992) and (Murphy, 2002). Passing a message from node A to node B consists of evaluating the following update equations:

$$\phi_S^* = \sum_{A \setminus S} \phi_A \tag{C.1}$$

$$\phi_B^* = \frac{\phi_S^*}{\phi_S} \phi_B . \tag{C.2}$$

After evaluating the above equations, it is said that A has absorbed from B. Passing a message from A to B followed by a message passing in the opposite direction will achieve local consistency between A and B (at least until another message passing disturbs it). However, passing messages through the cliques in the tree according to the protocol specified above will cause a message to be passed through each link twice leading to global consistency. At the end of this procedure, each clique potential has been modified to contain the join probability of the clique nodes and the evidence $\phi_A = p(A, E)$. The posterior probability of

the evidence (a quantity needed, for instance, for classification) can then be readily obtained by marginalization of any clique over its constituent nodes.

Appendix D

Significance Test

Consider the following probability density function, given count parameters k and n:

$$p(\epsilon|k,n) = \frac{1}{\alpha}\epsilon^k(1-\epsilon)^{n-k}, \tag{D.1}$$

where α is a proportionality factor to ensure the PDF integrates to one. The random variable ϵ in Eq. D.1 could represent, for instance, the distribution of the error rate of a classifier misclassifying k patterns out of n. Consider now two independent classifiers with error rates ϵ_1 and ϵ_2, and suppose we want to consider the event

$$p(\epsilon_1 < \epsilon_2|k_1,n_1,k_2,n_2) = \frac{\int_{\epsilon_2=0}^{1}\int_{\epsilon_1=0}^{\epsilon_2} p(\epsilon_1,\epsilon_2|k_1,n_1,k_2,n_2)d\epsilon_1 d\epsilon_2}{\int_{\epsilon_2=0}^{1}\int_{\epsilon_1=0}^{1} p(\epsilon_1,\epsilon_2|k_1,n_1,k_2,n_2)d\epsilon_1 d\epsilon_2}. \tag{D.2}$$

Assuming independence of the variables ϵ_1 and ϵ_2, Eq. D.2 becomes:

$$p(\epsilon_1 < \epsilon_2|k_1,n_1,k_2,n_2) = \frac{\int_{\epsilon_2=0}^{1}\epsilon_2^{k_2}(1-\epsilon_2)^{n_2-k_2}\left(\int_{\epsilon_1=0}^{\epsilon_2}\epsilon_1^{k_1}(1-\epsilon_1)^{n_1-k_1}d\epsilon_1\right)d\epsilon_2}{\left(\int_{\epsilon_2=0}^{1}\epsilon_2^{k_2}(1-\epsilon_2)^{n_2-k_2}d\epsilon_2\right)\left(\int_{\epsilon_1=0}^{1}\epsilon_1^{k_1}(1-\epsilon_1)^{n_1-k_1}d\epsilon_1\right)}. \tag{D.3}$$

Eq. D.3 can be conveniently expressed in terms of beta functions. Let us recall two relevant important relations. The beta function can be written as the integral relation

$$\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt. \tag{D.4}$$

Related to this expression is the so-called incomplete beta function, given by

$$\beta_{inc}(x, y, z) = \frac{\int_0^z t^{x-1}(1-t)^{y-1}dt}{\beta(x, y)}. \tag{D.5}$$

Using these expressions, Eq. D.3 then becomes

$$p(\epsilon_1 < \epsilon_2 | k_1, n_1, k_2, n_2) = \frac{\int_0^1 \epsilon_2^{k_2}(1-\epsilon_2)^{n_2-k_2}\beta_{inc}(k_1+1, n_1-k_1+1, \epsilon_2)d\epsilon_2}{\beta(k_2+1, n_2-k_2+1)}. \tag{D.6}$$

The functions $\beta(x, y)$ and $\beta_{inc}(x, y, z)$ are typically made available in numerical packages like Matlab or Mathematica, or can be found tabulated. The definite integral in Eq. D.6 may then be approximated through a variety of methods, like quadrature or importance sampling. Since the variables involved are unidimensional, a quadrature approximation scheme will typically produce a reasonable approximation:

$$\begin{aligned} p(\epsilon_1 < \epsilon_2 | k_1, n_1, k_2, n_2) \approx & \frac{1}{\beta(k_2+1, n_2-k_2+1)} \times \\ & \sum_i \epsilon_{2_i}^{k_2}(1-\epsilon_{2_i})^{n_2-k_2}\beta_{inc}(k_1+1, n_1-k_1+1, \epsilon_2)(\epsilon_{2_{i+1}} - \epsilon_{2_i}). \end{aligned} \tag{D.7}$$

Appendix E

Perceptual Experiment Protocol

This appendix includes documentation relevant to the perceptual experiments discussed at length in Chapter 9. Section E.1 includes the consent form that every subject was presented with prior to their participation in the experiment to gather their willful consent and signature in accordance with guidelines stipulated by the Committee on the Use of Humans as Experimental Subjects (COUHES), which reviewed and approved this study (COUHES ID No. 2755), at the Massachusetts Institute of Technology. The experiment began by asking the subjects to read the instructions summarized in section E.2, prior to carrying out the actual rating task. The questionnaire that subjects were asked to fill out at the end of each session is reproduced in section E.3.

E.1 Consent Form

Your participation in the following experiment is completely voluntary. You are free to withdraw this consent at any time, for any reason, and to request that any or all of the data collected be destroyed. If at any time you feel uncomfortable, or unsure that you wish your results to be part of the experiment, you may discontinue your participation with no repercussions.

In a few minutes you will be asked to interact with a computer interface which presents a series of icons linked to short speech utterances. Your task consists of listening to the speech samples by clicking on the icons and then registering your perception of (i) how calm/excited and (ii) how pleased/displeased the speaker sounds. You will find more specific details and instructions on how to navigate through the system once you begin the experiment. This

is a perceptual experiment, not a test. There are no correct or incorrect answers, and your answers are not being judged or compared against any standards.

A voucher for payment of $8 will be given to you prior to the participation in the study, along with instructions for its redemption.

Your participation in this study should take approximately one hour. Any responses that are collected during the experiment will be completely anonymous. From this point forward, you will be referred to only as the ID number that the experimenter will hand you.

If you have any questions, at any point during the experiment, the experimenter will gladly answer them.

Please read the following and sign on the lines below:

I, the undersigned, have read and understood the explanations of the following research experiment and voluntarily consent to my participation in it. I understand that my responses will remain confidential and that I may terminate my participation at any time.

In the unlikely event of physical injury resulting from participation in this research, I understand that medical treatment will be available from the MIT Medical Department, including first-aid emergency treatment and follow-up care as needed, and that my insurance carrier may be billed for the cost of such treatment. However, no compensation can be provided for medical care apart from the foregoing. I further understand that making such medical treatment available, or providing it, does not imply that such injury is the investigator's fault. I also understand that by my participation in this study, I am not waiving any of my legal rights.

I understand that I may also contact the Chairman of the Committee on the Use of Humans as Experimental Subjects at MIT (253-6787) if I feel that I have been treated unfairly as a subject.

Name: _____

Date: _____

Location: _____

Additionally, please read the following paragraph:

The data collected during this experiment will be used for research purposes only. After

the data collection is over, they will be permanently stored in a private archive. In the future, portions of this record may be published and/or presented in scientific journals and/or in scientific conference proceedings. No private information, such as subject's name, address or other private data, can be linked to the data gathered during the experiment. Only the experimental data is subject to being presented in a scientific venue. These data may also be made available to other researchers in the field for usage according to the same research guidelines stated here. Again, at any time during or after the experiment you may request that your records be destroyed.

Please sign on the lines below to give permission to the collection of this material.

Name: _____

Date: _____

Location: _____

E.2 Experiment Instructions

Experiment Summary:

Please read through these instructions before starting the experiment.

For this experiment you are asked to listen to a set of human speech samples and then try to rate how the speaker sounds on two scales: How **pleased** or **displeased** the speaker sounds, and how **excited** or **calm** the speaker sounds. A numbered sound icon represents each speech sample. You will rate the speech by moving the icons onto a two dimensional grid.

Occasionally, the speaker may repeat the same phrase, but the way the words are stated may be different.

Your job is to rate how the speaker sounds, not how the speaker makes you feel.

We appreciate your help. Good Luck!

How to use the Speech Graph:

1. When you start the experiment, you will see a group of sound icons followed by a picture of a grid. Your job is to place all of the sound icons onto the grid according

to how pleased, displeased, excited or calm the speaker sounds.

[Fig. 9-1 appears at this point in the instructions.]

2. The five dark gray sound icons are to be used as references, and you should place the draggable icons relative to them. Double-click on these icons to hear a speaker who would be placed in that section of the grid, and to acquaint yourself with how these speakers sound. They are not necessarily the maximum values, so you may place icons beyond these reference icons. The reference speakers remain constant throughout the experiment.

3. To rate a speech sample, double-click on one of the icons and you will then hear the sound it represents. Next, click on an icon and then drag and drop it to a desired location on the grid.

[Fig. 9-2 appears at this point in the instructions.]

4. To rate other speech samples in the set, which are not currently visible, click on the pull-down menu and select another group of icons from the list. You can also use the **Previous** and **Next** buttons to cycle through the groups of icons. You can go back and re-evaluate any of the speech samples by using the pull-down menu. The experiment will remember where you placed any earlier icons.

[Fig. 9-3 appears at this point in the instructions.]

5. All of the icons must be placed on the grid before the experiment is over. You finish the experiment by pressing the **Done** button. If any icons remain to be placed, you will see a message box indicating so.

[Fig. 9-4 appears at this point in the instructions.]

6. After you have placed all of the icons to your satisfaction on the grid, click the **Done** button, and you are done! Just fill out the final survey to let us know what you thought of the experiment.

7. Now press the **Go To Experiment** button at the bottom of this page to start the experiment.

E.3 Questionnaire

1. How would you rate the complexity of completing this task?

○	○	○	○	○	○	○
1	2	3	4	5	6	7
Very Easy						Very Difficult

2. How confident are you in the overall ratings you provided?

○	○	○	○	○	○	○
1	2	3	4	5	6	7
Very confident						Not confident at all

3. How would you rate yourself at decoding the affect of others?

○	○	○	○	○	○	○
1	2	3	4	5	6	7
Very perceptive						Very unperceptive

4. Do you suffer from any known hearing problems?

5. Are you a native speaker of English?

6. Do you have any suggestions to make this interface easier to use or navigate, or to make the task easier to complete?

References

Abelin, Å. (2000). Cross linguistic interpretation of emotional prosody. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *Proc. ISCA Workshop on Speech and Emotion*, (pp. 110–113)., Newcastle, Northern Ireland.

Alani, A. & Deriche, M. (1999). A novel approach to speech segmentation using the wavelet transform. In *Fifth Intnl. Symposium on Signal Proc. and its Applications (ISSPA)*, (pp. 127–130)., Brisbane, Australia.

Alku, P., Strik, H., & Vilkman, E. (1997). Parabolic spectral parameter - A new method for quantification of the glottal flow. *Speech Communication*, *22*, 67–79.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. Intnl. Conf. Spoken Language Processing (ICSLP)*, volume 3, (pp. 2037–2040)., Denver.

Bagshaw, P. C. (1994). *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching*. PhD thesis, The University of Edinburgh.

Banse, R. & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Personality and Social Psychology*, *70*(3), 614–636.

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. (2003). How to find trouble in communication. *Speech Communication*, *40*, 117–143.

Batliner, A., Huber, R., Niemann, H., Nöth, E., & Spilker, J. (2000). The recognition of emotion. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translations* (pp. 122–130). New York: Springer.

Beckman, M. E. (1996). The parsing of prosody. In P. Warren (Ed.), *Prosody and Parsing*, Special Issue of Language and Cognitive Processes (pp. 17–68). Psychology Press.

Bickmore, T. (2003). *Relational Agents: Effecting Change through Human-Computer Relationships*. PhD thesis, Media Arts and Sciences. Massachussets Institute of Technology.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Institute of Phonetic Sciences*, *17*, 97–110.

Branch, M. A., Coleman, T. F., & Li, Y. (1999). A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal of Scientific Computation*, *21*(1), 1–23.

Breazeal, C. & Aryananda, L. (2002). Recognizing affective intent in robot directed speech. *Autonomous Robots*, *12*(1), 83–104.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). *Classification and Regression Trees*. London: Chapman & Hall.

Byrd, R. H., Schnabel, R. B., & Shultz, G. A. (1988). Approximate solution of the trust region problem by minimization over two-dimensional subspaces. *Mathematical Programming*, *40*, 247–263.

Cahn, J. (1990). The generation of affect in synthesized speech. *J. of the American Voice I/O Soc.*, *8*, 1–19.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *PAMI-8*(6), 679–698.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied Conversational Agents*. MIT Press.

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., & Shneiderman, B. (2003). Determining causes and severity of end-user frustration. *International Journal of Human-Computer Interaction*, *To apper*.

Childers, D. & Lee, C. (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustic Society of America*, *90*(5).

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. MIT Press.

Coleman, T. F. & Li, Y. (1992). A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables. Technical report, Comp. Science Dept. and Advance Computing Research Institute. Cornell University, Ithaca, NY.

Coleman, T. F. & Li, Y. (1994). On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Mathematical Programming*, *67*, 189–224.

Coleman, T. F. & Li, Y. (1996). An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal of Optimization*, *6*(2), 418–445.

Conn, A. R., Gould, N. I. M., & Toint, P. L. (2000). *Trust-Region Methods*. MPS/SIAM Series on Optimization. MPS/SIAM.

Cook, N. D. (2002). *Tone of Voice and Mind. The Connections between Intonation, Cognition and Consciousness*, volume 47 of *Advances in Consciousness Research*. Amsterdam/Philadelphia: John Benjamins.

Cowie, R. & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1-2), 5–32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Cruttenden, A. (1997). *Intonation*. Cambridge Textbooks in Linguistics. Cambridge Univ. Press.

Cummings, K. E. & Clements, M. A. (1995a). Analysis of the glottal excitation of emotionally styled and stressed speech. *J. Acoust. Soc. Am.*, *1*, 88–98.

Cummings, K. E. & Clements, M. A. (1995b). Glottal models for digital speech processing: A historical survey and new results. *Signal Processing*, *5*, 21–42.

Cummins, F. (2002). Speech rhythm and rhythmic taxonomy. In Bel, B. & Marlien, I. (Eds.), *Proceedings 1st. International Conference on Speech Prosody*, (pp. 121–126)., Aix-en-Provence, France.

Cummins, F. & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*, 145–171.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *J. Speech and Hearing Research*, *36*, 254–266.

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. In *Proc. Intnl. Conf. Spoken Language Processing (ICSLP)*, volume 3, (pp. 1970–1973).

Dempster, A. P., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*, 185–197.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*, 423–444.

Durston, P. J., Farrell, M., Attwater, D., Allen, J., Kuo, H.-K. J., Afify, M., Fosler-Lussier, E., & Lee, C.-H. (2001). OASIS natural language call steering trial. In *Proceedings Eurospeech*, (pp. 1323–1326)., Aalborg, Denmark.

Eberman, B. & Goldenthal, W. (1996). Time-based clustering for phonetic segmentation. In *Proc. Intnl. Conf. Spoken Language Processing (ICSLP)*, volume 2, (pp. 1225–1228).

E.E. Osuna, R. F. & Girosi, F. (1997). Support vector machines: Training and applications. Technical Report A.I. Memo 1602/C.B.C.L. Paper 144, MIT.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392.

El-Jaroudi, A. & Makhoul, J. (1991). Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, *39*(2), 411–423.

Fell, J. (2003). *Emotion*, chapter The Phenomenological Approach to Emotion. Authors Choice Press.

Fernandez, R. & Picard, R. W. (2003). Modeling drivers' speech under stress. *Speech Communication*, *40*, 145–159.

Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In Bel, B. & Marlien, I. (Eds.), *Proceedings 1st. International Conference on Speech Prosody*, (pp. 115–120)., Aix-en-Provence, France.

Gee, J. P. & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*, 411–458.

Gobl, C. & Ní Chasaide, A. (2000). Testing affective correlates of voice quality through analysis and resynthesis. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *ISCA Workshop on Speech and Emotion*, (pp. 178–183)., Newcastle, Northern Ireland.

Gobl, C. & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1-2), 189–212.

Gordon, M. & Ladefoged, P. (2001). Phonation types: A cross-linguistic review. *Journal of Phonetics*, *29*, 383–406.

Grabe, E. & Low, E. L. (2002). Durational variability in speech and the rhythm class hyphothesis. In *Papers in Laboratory Phonology 7*. Mouton.

Greenwood, D. D. (1961). Critical bandwidth and the frequency coordinates of the basilar membrane. *Journal of the Acoustic Society of America*, *33*, 1344–1356.

Gunn, S. (1998). Support vector machines for classification and regression. Technical report, Image, Speech and Intelligent Systems Group. University of Southampton.

Hansen, J. H. L., Bou-Ghazale, S. E., Sarikaya, R., & Pellom, B. (1998). Getting started with the SUSAS: Speech Under Simulated and Actual Stress database. Technical Report RSPL-98-10 1.4, Duke University. Dept. of Electrical Engineering.

Hansen, J. H. L. & Womack, B. D. (1996). Feature analysis and neural network-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing.*, *IV*(4), 307–313.

Hanson, H. (1997). Glottal characteristics of female speakers: Acoustic correlates. *J. Acoust. Soc. Am.*, *101*(1), 466–481.

Hanson, H. M., Stevens, K. N., Kuo, H.-K. J., Chen, M. Y., & Slifka, J. (2001). Towards models of phonation. *Journal of Phonetics*, *29*, 451–480.

Hastings, A. (2002). ISO 532B/DIN 45631 loudness calculation (implemented in Matlab). http://widget.ecn.purdue.edu/~hastinga/Research.htm.

Hayes, B. (1989). The prosodic hierarchy in meter. In P. Kiparsky & G. Youmans (Eds.), *Phonetics and Phonology. Rythm and Meter*, volume 1. Academic Press.

Hieronymus, J. L. (1989). Automatic sentential vowel stress labelling. In *Proc. 1st. European Conference on Speech Communication and Technology*, volume 1, (pp. 226–229)., Paris.

Hirschberg, J. & Nakatani, C. (1998). Using machine learning to identify intonational segments. In *Proc. of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing.*

Huang, T. S., Chen, L., Tao, H., Miyasato, T., & Nakatsu, R. (1998). Bimodal emotion recognition by man and machine. ATR Workshop on Virtual Communication Environments.

Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V., & Niemann, H. (2000). Recognition of emotion in realistic dialogue scenario. In *Proc. Intnl. Conf. Spoken Language Processing (ICSLP)*, volume 1, (pp. 665–668)., Beijing, China.

Jain, A. & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(2), 153–158.

Jensen, F. (1996). *An Introduction to Bayesian Networks.* Springer.

Kasuya, H., Endo, Y., & Saliu, S. (1993). Novel acoustic measurements of jitter and shimmer characteristics from pathological voice. In Fellbaum, K. (Ed.), *Proceedings Eurospeech 1993*, volume 3, (pp. 1973–1976)., Berlin.

Klatt, D. H. & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. American*, *87*(2), 820–857.

Kleinginna Jr., P. R. & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definitions. *Motivation and Emotion*, *5*(4), 345–379.

Kompe, R., Batliner, A., Kießling, A., Kilian, U., Niemman, H., & Nöth, E. (1994). Automatic classificiation of prosodically marked phrase boundaries in German. In *Proceedings Intnl. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, (pp. 173–176).

Ladd, D. (1996). *Intonational Phonology.* Number 79 in Cambridge Studies in Linguisitics. Cambridge University Press.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergman, G., & Scherer, K. R. (1985).

Evidence for the independent function of intonation contour type, voice quality and F0 range in signalling speaker affect. *J. Acoust. Soc. Am.*, *78*, 435–444.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*(5), 372–385.

Lauritzen, S. L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, *87*, 1098–1108.

Laver, J. (1994). *Principles of Phonetics*. Cambridge Univ. Press.

Lee, C. M. & Narayanan, S. S. (2002). Combining acoustic and language information for emotion recognition. In *Proc. Intnl. Conf. Spoken Language Processing (ICSLP)*, Denver.

Lee, C. M., Narayanan, S. S., & Pieraccini, R. (2002). Classifying emotions in human-machine spoken dialogs. In *Proc. Intnl. Conf. on Multimedia Expo*, Lausanne, Switzerland.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Liscombe, J., Venditti, J., & Hirschberg, J. (2003.). Classifying subject ratings of emotinal speech using acoustic features. In *Proceedings of Eurospeech 2003*, (pp. To appear.).

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proc. IEEE*, *63*(4), 561–580.

McCarthy, J. (2000). From here to human-level intelligence. http://www-formal.stanford.edu/jmc/human.html.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., & Stroeve, S. (2000). Automatic recognition of emotion from voice: A rough benchmark. In *ISCA Workshop on Speech and Emotion*, (pp. 207–212)., Newcastle, Northern Ireland.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. America*, *58*(4), 880–883.

Michaelis, D., Fröhlich, M., & Strube, H. W. (1998). Selection and combination of acoustic features for the description of pathologic voices. *J. Acoust. Soc. Am.*, *103*(3), 1628–1639.

Michaelis, D., Grams, T., & Strube, H. (1997). Glottal-to-noise excitation ratio – a new measure for describing pathological voices. *ACUSTICA*, *83*, 700–706.

Mozziconacci, S. (2000). The expression of emotion considered in the framework of an intonational model. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *Proc. ISCA Workshop on Speech and Emotion*, (pp. 45–52)., Newcastle, Northern Ireland.

Mozziconacci, S. (2002). Prosody and emotions. In Bel, B. & Marlien, I. (Eds.), *Proceedings 1st. International Conference on Speech Prosody*, (pp. 1–9)., Aix-en-Provence, France.

Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning.* PhD thesis, Department of Computer Science. University of California, Berkeley.

Murray, I. R., Baber, C., & South, A. J. (1996). Towards a definition and working model of stress and its effects on speech. *Speech Communication*, *20*, 1–12.

Nöth, E., Batliner, A., Kießling, A., Kompe, R., & Niemann, H. (2000). VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system. *IEEE Trans. on Speech and Audio Proc.*, *8*(5), 519–532.

Oppenheim, A. V. & Schafer, R. W. (1989). *Discrete-Time Signal Processing.* Signal Processing Series. New Jersey: Prentice Hall.

Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus. Technical Report ECS-95-001, Boston University.

Paeschke, A., Kienast, M., & Sendlmeier, W. F. (1999). F0-contours in emotional speech. In *Proc. Intnl. Conf. of Phonetic Sciences (ICPhS)*, (pp. 929–932).

Paeschke, A. & Sedlmeier, W. F. (2000). Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *Proc. ISCA Workshop on Speech and Emotion*, (pp. 75–80)., Newcastle, Northern Ireland.

Paulus, E. & Zwicker, E. (1972). Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln. *Acustica*, *27*, 253–266.

Pereira, C. (2000). Dimensions of emotional meaning in speech. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *Proc. ISCA Workshop on Speech and Emotion*, (pp. 25–28)., Newcastle, Northern Ireland.

Pfitzinger, H. R., Burger, S., & Heid, S. (1996). Syllable detection in read and spontaneous speech. In *Proc. Intnl. Conf. Spoken Language Processing (ICSLP)*, volume 2, (pp. 1261–1264).

Picard, R. (1997). *Affective Computing.* MIT Press.

Plomp, R. & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *J. Acoust. Soc. Am.*, *38*, 548–560.

Plumpe, M. D., Quatieri, T. F., & Reynolds, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech and Audio Proc.*, 7(5), 569–586.

Polzin, T. (2000). *Detecting Verbal and Non-verbal Cues in the Communication of Emotions.* PhD thesis, School of Computer Science. Carnegie Mellon University.

Poyatos, F. (1993). *Paralanguage. A Linguistic and Interdisciplinary Approach to Interactive Speech and Sound.* Current Issues in Linguistic Series. Amsterdam: John Benjamins Publishing Company.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*.

Quast, H. (2001). Automatic recognition of nonverbal speech. Master's thesis, Georg August Universität Götingen.

Rabiner, L. R. & Schafer, R. W. (1978). *Digital Processing of Speech Signals.* Signal Processing Series. New Jersey: Prentice Hall.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. In Bel, B. & Marlien, I. (Eds.), *Proceedings 1st. International Conference on Speech Prosody*, (pp. 115–120)., Aix-en-Provence, France.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*, 265–292.

Redi, L. & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, *29*, 407–429.

Reeves, B. & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places.* Cambridge University Press.

Riegelsberger, E. L. & Krishnamurthy, A. K. (1995). Glottal source estimation: Methods of applying the LF-model to inverse filtering. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, (pp. 542–545).

Roach, P. (2000). Techniques for the phonetic description of emotional speech. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *Proc. ISCA Workshop on Speech and Emotion*, (pp. 53–59)., Newcastle, Northern Ireland.

Roy, D. & Pentland, A. (1996). Automatic spoken affect analysis and classification. In *Proc. 2nd. Intnl. Conf. Automatic Face and Gesture Recognition*, (pp. 363–367)., Killington, VT.

Sarikaya, R. & Gowdy, J. N. (1997). Wavelet based analysis of speech under stress. In *Southeastcon '97. Engineering New Century. Proceedings IEEE.*, (pp. 92–96).

Sarikaya, R. & Gowdy, J. N. (1998). Subband based classification of speech under stress. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume I, (pp. 569–572).

Scheffers, M. T. M. (1988). Automatic stylization of f0-contours. In Ainsworth, W. A. & Holmes, J. N. (Eds.), *Proceedings 7th FASE Symposium*, volume 3, (pp. 981–987).

Scheirer, J., Fernandez, R., Klein, J., & Picard, R. (2002). Frustrating the user on purpose: A step toward building an affective computer. *Interacting with Computers*, *14*(2), 93–118.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *J. Acoust. Soc. Am.*, *76*(5), 1346–1356.

Schlosberg, H. (1953). Three dimensions of emotion. *The Psychological Review*, *61*(2), 81–88.

Schötz, S. (2003). Prosody in relation to paralinguistic phonetics - earlier and recent definitions, distinctions, and discussions. Technical report, Department of Linguistics and Phonetics. Lund University.

Selkirk, E. O. (1984). *Phonology and Syntax: The Relation Between Sound and Structure.* Cambridge, MA: MIT Press.

Sethares, W. A. (1998). *Tuning, Timbre, Spectrum, Scale.* London: Springer-Verlag.

Sharma, M. & Mammone, R. (1996). "Blind" speech segmentation: Automatic segmentation of speech without linguistic knowledge. In *Proc. Intnl. Conf. Spoken Language Processing (ICSLP)*, volume 2, (pp. 1237–1240).

Shattuck-Hufnagel, S. & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (2001). *Emotions in Social Psychology*, chapter Emotion Knowledge: Further Exploration of a Prototype Approach, (pp. 26–56). Key Readings in Social Psychology Series. Psychology Press.

Shire, M. L. (1997). Syllable onset detection from acoustics. Master's thesis, Electrical Engineering and Computer Science. U.C. Berkeley.

Shultz, G. A., Schnabel, R. B., & Byrd, R. H. (1985). A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties. *SIAM Journal Numerical Analysis*, *22*(1), 47–67.

Smits, R. & Yegnanarayana, B. (1995). Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, *3*(5), 325–333.

Steeneken, H. J. M. & Hansen, J. H. L. (1999). Speech under stress conditions: Overview of the effect on speech production and of system performance. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume IV, (pp. 2079–2082).

t'Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation.* Cambridge: Cambridge University Press.

Tickle, A. (2000). English and Japanese speakers' emotion vocalisation and recognition: A comparison highlighting vowel quality. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *Proc. ISCA Workshop on Speech and Emotion*, (pp. 104–109)., Newcastle, Northern Ireland.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer.

Wang, M. & Hirschberg, J. (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, *6*, 175–196.

Wichman, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In Cowie, R., Douglas-Cowie, E., & Schröder, M. (Eds.), *Proc. ISCA Workshop on Speech and Emotion*, (pp. 143–147)., Newcastle, Northern Ireland.

Wightman, C. W. & Ostendorf, M. (1991). Automatic recognition of prosodic phrases. In *Intnl. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, volume 1, (pp. 321–324).

Wightman, C. W. & Ostendorf, M. (1992). Automatic recognition of intonational features. In *Intnl. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, volume 1, (pp. 221–224).

Wightman, C. W. & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Trans. Speech and Audio Processing*, *2*(4), 469–481.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *J. Acoust. Soc. Am.*, *91*(3), 1707–1717.

Williams, C. E. & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *J. Acoust. Soc. Am.*, *52*(4), 1238–1250.

Wong, D. Y., Markel, J. D., & Gray Jr., A. H. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-27*(4), 350–355.

Yegnanarayana, B. & Smits, R. L. H. M. (1995). A robust method for determining instants of major excitations in voiced speech. In *International Conference on Acoustic, Speech, and Signal Processing*, volume 1, (pp. 776–779).

Zhou, G., Hansen, J., & Kaiser, J. (1998). Classification of speech under stress based on features derived from the nonlinear Teager energy operator. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume I, (pp. 549–552).

Zwicker, E. & Fastl, H. (1999). *Psychoacoustics. Facts and Models* (2nd ed.)., volume 22 of *Springer Series in Information Sciences*. Berlin: Springer-Verlag.

Zwicker, E., Fastl, H., & Dallmayr, C. (1984). BASIC program for calculating the loudness of sounds from their 1/3-oct. band spectra according to ISO 532B. *Acustica*, *55*, 63–67.

Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., & Namba, S. (1991). Program for calculating loudness according to DIN 45631 (ISO 532B). *J. Acoust. Soc. Jpn.*, *12*(1), 39–42.