Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency

Jackson J. Liscombe

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2007

ABSTRACT

Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency

Jackson J. Liscombe

Prosody—suprasegmental characteristics of speech such as pitch, rhythm, and loudness—is a rich source of information in spoken language and can tell a listener much about the internal state of a speaker. This thesis explores the role of prosody in conveying three very different types of speaker state: paralinguistic state, in particular emotion; pragmatic state, in particular questioning; and the state of spoken language proficiency of non-native English speakers.

Paralinguistics. Intonational features describing pitch contour shape were found to discriminate emotion in terms of positive and negative affect. A procedure is described for clustering groups of listeners according to perceptual emotion ratings that foster further understanding of the relationship between acoustic-prosodic cues and emotion perception.

Pragmatics. Student questions in a corpus of one-on-one tutorial dialogs were found to be signaled primarily by phrase-final rising intonation, an important cue used in conjunction with lexico-pragmatic cues to differentiate the high rate of observed declarative questions from proper declaratives. The automatic classification of question form and function is explored.

Proficiency. Intonational features including syllable prominence, pitch accent, and boundary tones were found to correlate with language proficiency assessment scores at a strength equal to that of traditional fluency metrics. The combination of all prosodic features further increased correlation strength, indicating that suprasegmental information encodes different aspects of communicative competence.

Contents

List of Figures

List of Tables

List of Emotion Hypotheses

Acknowledgments

It goes without saying that huge achievements are not possible without the help and guidance of others. And yet we feel the need to say it nevertheless. The foundational work for this thesis would not have been possible without the help and guidance of many people during my five years at Columbia University. Additionally, friends and relatives have provided emotional support so that I could persevere when I felt like quitting.

First and foremost, I would like to thank my advisor Julia Hirschberg. They say that one's graduate experience is determined, in large part, by one's advisor and I can say unequivocally that under her supervision I had one of the best graduate experiences that one could have. Equally as important has been the love and support of my partner Johnny Linville. Without him I surely would have lost my way. I would also like to thank the following people: Fadi Biadsy, Frank Enos, Agustín Gravano, Sameer Maskey, Andrew Rosenberg, and Jeniffer Venditti (The Speech Lab); Derrick Higgins and Klaus Zechner (The ETS Group); Kate Forbes-Riley, Diane Litman, and Mihai Rotaru (The ITSPOKE Group); Dilek Hakkani-Tür and Giuseppe Riccardi (The AT&T Group); Ellen Eide and Raul Fernandez (The IBM Group); and Dale Liscombe, Lee Ann Liscombe, and Max Liscombe (The Family).

Part I

INTRODUCTION

Prosody—suprasegmental characteristics of speech such as pitch, rhythm, and loudness—is a rich source of information in spoken language. It signifies a myriad of things, some of which are intentional, some of which are not, including the emotion of a speaker, the pragmatic force behind an utterance, and demographic and cultural information about a speaker. In short, prosody can tell a listener a lot about the internal state of a speaker, whether these states be short-term or long-term processes. In this thesis we explore the role of prosody in conveying three very different types of speaker state: emotion, question-status, and non-native language proficiency.

In the first part of this thesis, we examine affective speaker state by characterizing the prosodic cues present in emotional speech. Discrete emotions were found to be most successfully characterized when they were defined using a perceptual, polythetic labeling typology. As per past studies, global acoustic characteristics were found to be characteristic of the activation level of emotion. Intonational features describing pitch contour shape were found to further discriminate emotion by differentiating positive and negative emotions. A procedure is described for clustering groups of listeners according to perceptual emotion ratings that was found to foster further understanding of the relationship between prosodic cues and emotion perception.

The role of prosody in signaling the form and function of questions is explored in the second part of the thesis. Student questions in a corpus of one-on-one tutorial dialogs were found to be signaled primarily by phrase-final rising intonation. This finding was particularly important because over half of all student questions were found to be syntactically identical to declarative statements and intonation can be viewed as tool speakers might use to differentiate the pragmatic force of each sentence type. Lexico-pragmatic and lexico-syntactic features were found to be crucial for further differentiating the form and function of student questions.

In the final part of the thesis we analyze how prosodic proficiency can be indicative of the current level of communicative competence of non-native English speakers. Intonational features including syllable prominence, pitch accent, and boundary tones were found to correlate with human assessment scores to the same degree that more traditional fluency-based metrics have been shown to do in the past. Furthermore, it was found that combination

of all information (fluency, intonation, and rhythm) further increased the correlation, indicating that various types of suprasegmental information convey the level of proficiency of non-native speakers.

In each part of the thesis, we describe how prosodic cues can be successfully used to automatically predict the internal state of a speaker by utilizing statistical machine learning algorithms. Our goal in this thesis was to predict the emotion, question-status, and proficiency of speakers for potential use in artificially intelligent applications, such as Spoken Dialog Systems, Intelligent Tutoring Systems, and language learning software. This work is seen as an ongoing effort to construct a model of the prosodic indicators of the broader concept of metacognitive speaker state.

Part II

PARALINGUISTICS

To most linguists—and, more broadly, to anyone who thinks about language—speech is a fascinating phenomenon because with it we can communicate meaning both by the words we say and also by *how* we say them. *Paralanguage* is a term that, generally, describes nonverbal communication in human interaction, though it has actually been more loosely defined than that in the past. Nöth (1990) identified several definitions of paralanguage proposed in previous literature, including some that contradict each other in that they comprise both human and nonhuman vocalization, vocal and nonvocal communication, suprasegmental and segmental features, and certain communicative functions including emotion and personality. In this thesis we use the latter sense—primarily as a term that encompasses affective communication or, more simply, emotion. Furthermore, though suprasegmental information is often considered to be an important component of paralanguage, the two are not synonymous. *Suprasegmental* describes language use that has a broader scope than a single segment such as a sound or a word. It is a term often used to describe prosodic information, such as pitch, intonation stress, rhythm, and duration. However, paralanguage can be conveyed via both suprasegmental and segmental information. In the first part of this thesis we present the findings of experiments designed to investigate the communication of paralinguistic meaning via suprasegmental information. We paid particular attention to the differences between intended and perceived emotion.

We introduce the topic by first describing the previous literature in Chapter 1. In Chapter 2 we detail the corpora used for analysis and in Chapter 3 we describe the acoustic cues examined. The remaining chapters are largely devoted to experiments designed to explore the difference between intended emotions (Chapter 4) and perceived emotions (Chapter 5). In the latter chapter we also present the results of listener clustering based on responses to a survey in which listeners were asked to rate the emotional content of utterances. The last two chapters are reserved for auxiliary investigations. In Chapter 6 we describe experiments classifying emotion in two dimensions (activation and valency) and in Chapter 7 we report on the correlation between abstract intonational units and emotions.

Chapter 1

Previous Research on Emotional Speech

Emotion is something intrinsically human and, as such, is part of our everyday interaction. It has proved to be a fascinating subject to researchers of all persuasions—including Socrates, Charles Darwin, and William James, to name just a very few—so much so that tomes have been devoted to theoretical frameworks describing it (cf. Cornelius, 1996). Despite such extensive exploration, or perhaps because of it, emotion remains somewhat ambiguous to define. Kleinginna & Kleinginna (1981) conducted a meta-study of theoretical descriptions of emotion and found nearly one hundred different definitions. The goal of this thesis is not to choose the "best" definition nor is it to provide the one hundred and first definition. Despite an absence of agreement on an exact definition of the concept of emotion or on the manifestations of different emotional states, researchers have been able to successfully conduct research on emotion. This is partially because, when it comes to emotion, people "know it when they feel it" even if they cannot provide a concise or complete definition of it. Imprecise conceptualizations of emotion are usually sufficient for describing precise phenomena related to emotion. For example, even if the emotion anger cannot be defined satisfactorily in the classical sense, research has shown that certain physiological changes occur in its presence, such as increased heart palpitation and breathing rate.

We have situated our work in a rather narrow sub-domain of emotion research. We

empirically explored the nonverbal acoustic manifestations of emotion in a corpus of acted speech. Even in this restricted domain, past research has been quite extensive, some dating as far back as 70 years (Fairbanks & Pronovost, 1939), and we cannot provide an account of all previous studies here. Instead, this section highlights several past studies that either have made the most impact on the field or that are directly related to our own research.

Traditionally, empirical studies of the acoustic cues of emotional speech have favored corpora designed to elicit emotion that is acted or simulated in some way. Corpora of this type have been used for several reasons. First, hand-crafted corpora are rich in emotions of interest to the researcher, precisely because they are designed to be so. Second, hand-crafted corpora can be constructed in such a way as to control for certain variables. For example, several past studies have instructed subjects to read or act a vignette with a particular target emotion and, in the process, produce utterances for discrete emotions under equivalent linguistic and contextual conditions (e.g., Davitz, 1964; Williams & Stevens, 1972; Scherer et al., 1984; Johnstone & Scherer, 1999; Kienast & Sendlmeier, 2000; Batliner et al., 2003; Liscombe et al., 2003; Väyrynen, 2005; Laukka et al., 2005, inter alia). In this way, contextual information is controlled and one is able to isolate the role of acoustic information in conveying emotion. Several studies have even made use of nonsensical utterances to factor out all possible lexical or semantic effects (e.g., Scherer, 2000; Tato et al., 2002; Oudeyer, 2002; Bänziger & Scherer, 2005).

There is a well-worn criticism of using acted or elicited corpora for the study of nonverbal acoustic communication of emotion. The argument is that acted emotion is not the same as real-life emotion, and there appears to be some truth to this. Williams & Stevens (1972) found that acted emotion tended to be more exaggerated than real-life emotion. However, they also found that the relationship between acoustic correlates and acted emotions, though accordingly exaggerated, were not contradictory to those found for real-life emotions. In other words, while the absolute values of gradient acoustic cues for acted speech differed from those found in a non-acted scenario, their relative values across emotions were equivalent. In summary, we do not contest the criticism that acted emotion is not the same as non-acted emotion. However, we believe that the relationship between acoustic cues and emotions remains constant in both scenarios and that acted speech is a useful medium for drawing

inferences about the acoustic cues of spoken emotion.

Global acoustic measurements of speech were found to play a role in discriminating emotions in the earliest experiments (Fairbanks & Pronovost, 1939; Davitz, 1964) and these results have been reduplicated time and time again in subsequent years. The global acoustic measurements most often cited are generally those that quantify speaking rate and mean or range of fundamental frequency ($f0$) and intensity.[1] Most studies have found that anger, fear, and happiness are characterized by high mean $f0$, large $f0$ variability, high mean intensity, large intensity variability, and fast speaking rate; whereas, boredom and sadness are characterized by the opposite (low $f0$ mean/variability, low intensity mean/variability, and slow speaking rate). There have been several meta-studies in recent years comparing past findings of empirical research in nonverbal communication of emotion (e.g., Johnstone & Scherer, 2000; Cowie, 2000; Schröder, 2001), though Juslin & Laukka (2003) put forth one of the most comprehensive. They composed a meta-analysis of over one hundred studies conducted between 1939 and 2002. Table 1.1 generalizes the dominant findings across studies.

Though significant acoustic correlates have been observed for discrete emotions, there is a general consensus that acoustic features of this sort fall short of fully discriminating discrete emotions. Emotion is generally thought to be composed of at least two dimensions:

	Acoustic Feature				
Emotion	speaking rate	mean intens.	intens. variability	mean $f0$	$f0$ variability
anger	fast	high	high	high	high
fear	fast	low/high	high	high	low/high
happiness	fast	high	high	high	high
sadness	slow	low	low	low	low

Table 1.1: Acoustic correlates of emotion based on a meta-analysis of over one hundred empirical studies. Take from Juslin & Laukka (2003).

[1] Fundamental frequency and intensity are known correlates of perceived pitch and intensity, respectively. Both will be defined in subsequent chapters.

activation and valency. **Activation** describes the amount of energy expenditure involved in handling a situation and/or its consequences. Global acoustic measurements are thought to correlate with this dimension because it is intrinsically related to biological phenomena and physiological state. As a person's emotional state becomes more activated, his or her breathing often increases and this has an effect on the how fast air is expelled (intensity) and how fast the vocal folds vibrate ($f0$) (Williams & Stevens, 1972). In other words, some acoustic correlates of emotion can be considered evolutionary reflexes (Johnstone & Scherer, 2000). However, describing emotion in terms of activation cannot fully differentiate even the most colloquially familiar emotional states. The most notorious examples of this are anger and happiness, both of which are highly activated and are known to correlate with high $f0$, intensity, and speaking rate (see Table 1.1). They differ, of course, in that one conveys negative affect and the other positive affect. This dimension is what is referred to as **valency** and there is almost no consensus in the research community on how, or even if, acoustic cues correlate with this dimension. Some have suggested that voice quality may play a role (e.g, Scherer et al., 1984; Ladd et al., 1985; Zetterholm, 1999; Tato et al., 2002; Gobl & Chasaide, 2003; Fernandez, 2004; Turk et al., 2005), while others have suggested categorical intonation units (e.g., Uldall, 1964; O'Connor & Arnold, 1973; Scherer et al., 1984; Mozziconacci & Hermes, 1999; Wichmann, 2002; Pollermann, 2002).

Acoustic cues of the type we have discussed thus far can be considered phonetic descriptors of speech. Mean $f0$, for example, is computed by measuring the rate of vibration of the vocal folds and, as such, is a gradient variable whose specific values are not considered to map directly to linguistic meaning. It is believed, however, that intonation—the shape of the pitch contour—can be abstracted in such a way that it does, in fact, represent linguistic (phonological) meaning. Of the one hundred studies addressed by Juslin & Laukka (2003), fourteen explored the role of intonation in nonverbal communication of emotion, though the authors noted that most of the studies in this latter group were fairly impressionistic and hard to compare due to different intonation annotation schemes. They summarized their findings by noting that anger, fear, and happiness are associated with "rising" intonation, whereas sadness is associated with "falling" intonation. This analysis implies that intonation adds no discriminative information beyond the simpler measurements of global

$f0$ behavior. However, a closer look at several of these studies reveals that there is contention with respect to the role of context regarding intonation. We will examine a few empirical experiments to highlight some of the findings, but note that there is by no means a consensus.

Scherer et al. (1984) found that low and high intonational phrase boundary tones were indicative of emotion in a corpus of 66 German utterances, but only when conditioned by sentence type. Yes-no questions were found to be **challenging** when they co-occurred with low boundary tones and **agreeable** when they co-occurred with high boundary tones. The opposite pattern was found for wh-questions. When not conditioned by sentence type, phrase-final intonation was found to have little effect on emotion perception. Wichmann (2002) agreed with this assessment, in part. Intonation, claimed Wichmann, can convey emotional meaning that is both independent of *and* dependent on context. **Expressive intonation** is not conditioned on linguistic context because it describes the speaker's internal state, whereas **attitudinal intonation** is defined in terms of speaker interaction and therefore must be conditioned on things such as speech acts, events, prior knowledge, etc. Wichmann agreed with Scherer et al.'s findings that intonation is interpreted emotionally based on marked and unmarked expectations of sentence type, though offered no empirical evidence of this.

In a more recent study, Bänziger & Scherer (2005) eschewed out of hand the usefulness of categorical intonation labels and instead represented pitch contour information by quantitatively measuring the slope of the pitch rises to, and falls from, pitch peaks. In addition, they measured the slope from the final pitch peak to the end of the phrase. It was found that such information was useful solely in terms of activation. **Uptrend**—a progressive increase in pitch until the final pitch peak—was found to signify despair and elation (high activation); **downtrend**—a gradual fall preceded by an early high pitch peak—was indicative of sadness and happiness (low activation). In addition, pitch falls from the last pitch peak to the end of the phrase were found to be steeper for high activation emotions (anger and joy) than for low activation emotions (fear and happiness).

Mozziconacci & Hermes (1999) labeled a corpus of Dutch acted speech with perceptual, categorical intonation labels using IPO intonation grammar ('t Hart et al., 1990). Though

the authors did not find a one-to-one relation between emotion and intonation pattern, they did find that utterances with an early prominence-lending rise followed by a very late non-prominence-lending rise were indicative of indignation. Furthermore, they found that a very late non-prominence-lending fall was generally indicative of non-neutral affect. Though their results were far from definitive, their findings were not conditioned by linguistic context, as Scherer et al.'s were.

The experiments described in this thesis are similar to the studies cited above in several respects. First, we extracted global acoustic cues from emotional utterances and correlated their values with labels corresponding to emotional states and dimensions. We also conducted machine learning experiments to predict emotional labels. However, whereas most past studies have profiled emotions using all statistically significant correlations, our approach was novel in that we sought to identify the most important cues and represented them as a set of minimally distinct features. Also, we explored the role of abstract intonation labels in emotion classification, which few previous studies have done.

Our approach differed from past studies in another important way. We compared the acoustic-prosodic cues associated with emotional states both as they were intended by the speakers and as they were perceived by listeners. Though it has been quite common for other research studies to evaluate perceived emotion by asking human listeners to judge the emotional content of utterances, we know of few research studies that have explored rater differences beyond reporting agreement among raters. Toivanen et al. (2005) is a notable exception, but they limited their analysis to the confusion of intended and perceived emotions based on gender. The authors reported that, on average, female listeners were 5% more accurate than male listeners at perceiving the intended emotion of speakers, though they performed no statistical analyses and were thus perhaps too broad in their conclusion that women are more attuned than mean to the emotional state of others. In our analysis, we looked at how listeners might be clustered into different groups based on their assessment of emotional speech. In the process, we examined several potential ways that raters might differ, including gender, though we did not segregate raters by gender *a priori*.

Chapter 2

The EPSAT and CU_EPSAT Corpora

The Emotional Prosody Speech and Transcripts corpus (henceforth referred to as **EPSAT**) was collected by a group of researchers at the University of Pennsylvania over an eight month period in 2000-2001 (Liberman et al., 2002) and was made publicly available through the Linguistic Data Consortium in 2002 (catalog number LDC2002S28; ISBN 1-58563-237-6). The corpus was modeled after an earlier study conducted for German (Banse & Scherer, 1996).

EPSAT comprises recordings of 8 professional actors (5 female, 3 male), each a native speaker of Standard American English, all of whom read semantically neutral phrases designed to convey several emotions for research on the prosodic indicators of emotion. Phrase length was normalized by providing actors with four syllable phrases only. The data were further controlled for semantic content by restricting the domain to only dates and numbers. In this way, the researchers hoped to isolate emotional prosody by disallowing linguistic cues to emotion. An example of one of the phrases presented to the actors was: "Two-thousand four."

The actors were instructed to read the phrases aloud such that 14 distinct emotions were conveyed, and were further provided with real-world contexts in which such an emotion might be felicitous. The 14 emotions (plus `neutral` to indicate no overt conveyance of

anxiety	boredom	cold-anger
contempt	despair	disgust
elation	happiness	hot-anger
interest	neutral	panic
pride	sadness	shame

Table 2.1: Emotion labels of the EPSAT corpus.

emotion) represented in the corpus are shown in Table 2.1. The data in EPSAT were recorded directly on two channels with a sampling rate of 22,050 Hz. The two microphones used were a stand-mounted boom Shure SN94 and a headset Seinnheiser HMD 410.

Each actor was encouraged to repeatedly utter a given phrase until he or she was satisfied with the production, though all uttered phrases were included in the corpus, along with their intended emotions. Consequently, the EPSAT corpus is quite large (approximately 2,400 utterances) and includes many utterances that may or may not convey the emotion listed as the intended emotion. For these reasons—as well as empirical curiosity about the difference between intended and perceived emotion, which we address at length in this chapter—we identified a subset of the corpus for experimentation. In this chapter we refer to this smaller corpus as **CU_EPSAT** because selection was performed at Columbia University.

For the CU_EPSAT corpus, the set of emotions was modified slightly in an attempt to balance negative and positive emotions; these labels are shown in Table 2.2. In addition, we selected only 2 male and 2 female speakers to balance for gender. The process by which

Positive	**Negative**
confident	angry
encouraging	anxious
friendly	bored
happy	frustrated
interested	sad

Table 2.2: Emotion labels of the CU_EPSAT corpus.

utterances were chosen was through agreement among three researchers (including the author). We listened to every utterance produced by each of the chosen speakers and selected the best exemplars for each of the emotions in the new emotion label set. In addition, an utterance that most convincingly conveyed no overt emotional content (`neutral`) was chosen for each speaker. Given 11 utterances for 4 speakers, CU_EPSAT contained 44 sound tokens. Information uniquely identifying the utterances selected from EPSAT for use in CU_EPSAT can be found in Appendix A on page 194.

Chapter 3

Extraction of Emotional Cues

The EPSAT corpus was designed to explore suprasegmental aspects of emotional speech. In other words, since each token was an isolated number or date, no lexical, pragmatic or otherwise contextual information was considered to be present. Accordingly, each actor had to rely on acoustic-prosodic information alone for emotional conveyance. There has been extensive quantitative exploration of acoustic-prosodic cues to emotion, dating back at least as far as Davitz's controlled experiments in the 1960s (Davitz, 1964). For the experiments presented in the first part of the thesis, we extracted acoustic-prosodic features that have most often been used in empirical emotion experiments in the past and and that have been shown to most reliably correlate with spoken emotion. Table 3.1 on page 16 catalogs the acoustic-prosodic features that were automatically extracted from each utterance in the EPSAT corpus.

By and large, the feature set was divided into two main classes: measurements of intensity and measurements of fundamental frequency ($f0$). Both information streams can be reliably extracted from digitized speech and techniques for extracting them are implemented in all modern speech analysis software programs. All features that relied on intensity and fundamental frequency were extracted using PRAAT, a freely-available program for speech analysis and synthesis.[1]

Intensity refers to the amplitude of sound waveforms (measured in decibels) and is

[1]For an overview of PRAAT functionality we refer the reader to Boersma (2001) and for information on the extraction algorithms implemented in PRAAT, we refer the reader to Boersma (1993).

Name	Description
`f0-mean`	mean $f0$
`f0-min`	minimum $f0$
`f0-max`	maximum $f0$
`f0-range`	`f0-max` minus `f0-min`
`f0-slope`	slope of linear regression line of $f0$ against time
`f0-curve`	coefficient of quadratic regression curve of $f0$ against time
`f0-rising`	the percentage of rising $f0$ between consecutive $f0$ samples
`f0-voiced`	for all samples, the percentage that are voiced
`db-mean`	mean intensity
`db-min`	minimum intensity
`db-max`	maximum intensity
`db-range`	db-maximum minus db-minimum

Table 3.1: Acoustic-prosodic features extracted from all EPSAT utterances.

a well-known correlate of loudness. **Fundamental frequency ($f0$)** describes the rate at which the vocal folds vibrate (measured in Hertz) and is a well-known correlate of pitch. The mean, minimum, maximum, and range (maximum - minimum) of the intensity (`db-mean`, `db-min`, `db-max`, `db-range`) and $f0$ (`f0-mean`, `f0-min`, `f0-max`, `f0-range`) measurements for each utterance in the EPSAT corpus were calculated automatically. Intonation, or pitch contour, was approximated in various ways. Several automatic measurements were calculated to indicate the global shape of the pitch contour. These features included the slope of linear regression line of $f0$ against time (`f0-slope`), the coefficient of quadratic regression curve of $f0$ against time (`f0-curve`), and the percentage of rising $f0$ in the utterance (`f0-rising`). The latter measurement was calculated by first recording the differences between all time-contiguous pairs of $f0$ measurements ($f0$ at frame t subtracted from $f0$ at time $t + 1$). The percentage of positive differences—each indicating a pitch "rise"—was then calculated against the total number of differences.

A final acoustic-prosodic feature—speaking rate (`f0-voiced`)—was approximated using the output of PRAAT's $f0$ extraction algorithm by calculating the percentage of voiced

samples in each utterance in the following manner. Only voiced sounds produce reliable $f0$ measurements because this is when the vocal folds are vibrating. Silence and voiceless sounds do not provide reliable output and undefined values are indicated in these instances. We generally had no internal silences in our data and so assumed that the ratio of voiceless to voiced sounds was roughly constant across all utterances, especially considering that the lexical content of the utterances was controlled. For these reasons, we believed `f0-voiced` to be a reliable approximation of speaking rate.

Though raw features values are sometimes used in acoustic analyses, we z-score normalized all feature values by actor. This was done so that feature values could be compared and generalized across speakers. Furthermore, the aforementioned features were all considered to be automatically extracted from the speech source because no hand-labeling—in the form of annotation or segmentation—was performed. These features were used in the bulk of the experiments in this chapter. In Section 7.1 we report on experiments that used hand-labeled intonational features in the form of ToBI phonemic tone labels.

Chapter 4

Intended Emotion

The emotion labels of the EPSAT corpus corresponded to the intention of the actors. As such, the labels did not *necessarily* correspond to the emotions that would be perceived by a listener. This is generally the case when actors lack the skill to successfully convey the emotions they intend, but may also be true in everyday life, even when a speaker truly does feel the emotion they intend to convey (consciously or subconsciously). Identifying the intended emotion of a speaker is a valid research endeavor, though in practice it is often conflated with perceived emotional content, to detrimental effect. If the distinction between intention and perception is not made then it is hard to generalize across studies. If one corpus has been labeled with intended emotion and another with perceived emotion, and if this distinction is not taken into account, then conclusions about emotional cues can be drawn that are actually artifacts of the mismatch between what is intended and what is perceived, rather than something inherent in emotion itself. In this chapter we identify the acoustic-prosodic characteristics of **intended** emotions in the EPSAT corpus, which we then compare and contrast with **perceived** emotions in the smaller CU_EPSAT corpus later on in Chapter 5. Through these analyses, we hoped to address the following hypothesis:

Hypothesis 4.1 (EmoAcoust) *Intended and perceived emotions may be characterized by different acoustic-prosodic cues.*

The intended emotion labels of the EPSAT corpus were **monothetic**. In other words, the utterances associated with each emotion were considered to carry information that was both necessary and sufficient for that emotion and, by implication, no other emotion. For example, if an utterance was labeled as `angry` then it was also considered to convey no other emotion. Such a typology of emotion labels is true of most corpora of acted emotion. A monothetic label set eradicates the notion of label similarity, even though it is quite natural for emotions to be grouped by similarity, most notoriously in terms of valency (positive vs. negative affect). For example, `elation` and `happiness` are generally considered to be more similar than are `elation` and `sadness`. A **polythetic** label typology, on the other hand, is capable of capturing label similarity because the labels are defined in terms of a broad set of criteria that are neither necessary nor sufficient. Instead, each utterance associated with a label need only possess a certain minimal number of defining characteristics, but none of the features has to be found in each member of the category. It was our hypothesis that a monothetic labeling typology for emotion would obscure important characteristics of emotions; namely, that emotions are inter-dependent and simultaneously present in an utterance. We state this belief formally with the following hypothesis:

Hypothesis 4.2 (EmoIndpt) *Emotions are not independent; they are inter-related.*

Additionally, a monothetic typology imposed on a polythetic phenomenon introduces the risk of improperly training computational models because similar classes cannot be discriminated. If two emotions are quite similar then the acoustic cues can be expected to be similar as well. Again, consider `elation` and `happiness`. Let us assume that an utterance that conveys `elation` is often perceived to convey `happiness` as well. With a monothetic labeling typology, the labels have been assigned to utterances that convey both emotions somewhat randomly. A machine learning approach cannot be expected to learn significant differentiating cues between the two, and classification performance can be expected to suffer. In this chapter we report on machine learning experiments using the EPSAT intended, monothetic feature set and, in Chapter 5, compare these results with the results of similar experiments that used a perceived, polythetic emotion label set. We hypothesized the following:

Hypothesis 4.3 (EmoClass) *A monothetic emotion label set performs worse in automatic classification experiments than does a polythetic label set.*

4.1 Automatic classification

We conducted a 15-way classification task using all utterances in the EPSAT corpus and the intended, monothetic emotion label set. Ten-fold cross-validation was run using J4.8 decision trees on 2,050 utterances from 6 actors.[1] We used the full set of acoustic-prosodic features, as described in Table 3.1 on page 16. Table 4.1 below lists the classification performance as F-measure (F) per emotion.[2] The emotion most often classified correctly

Emotion	**F-measure**
`contempt`	0.18
`pride`	0.19
`interest`	0.20
`sadness`	0.20
`cold-anger`	0.21
`shame`	0.22
`despair`	0.25
`disgust`	0.25
`anxiety`	0.28
`elation`	0.30
`happiness`	0.30
`panic`	0.39
`neutral`	0.40
`boredom`	0.41
`hot-anger`	0.53

Table 4.1: Classification performance of the EPSAT intended monothetic emotion labels.

[1] All machine learning experiments in this dissertation were conducted under the WEKA software package (Witten et al., 1999). The J4.8 decision tree algorithm is a Java-implemented version of the well-know C4.5 decision tree algorithm (Quinlan, 1993).

was `hot-anger` (F = 0.53); the least was `contempt` (F = 0.18). Average F-measure across all emotions was 0.29.

Though F-measure performance based on automatic classification indicated how predictable the intended emotions were given our classifier and feature set, it did not provide insight into the most predictive features for emotion discrimination, nor did it allow us to profile emotions based on the most significant acoustic-prosodic correlates of each emotion. One way we gained insight into these matters was to examine the decisions learned when training the decision tree. However, the resulting decision tree was much too complex to analyze in any meaningful way (over 1,000 decisions were made). Notwithstanding, we could still take note of the most informative features, which we did by using the correlation-based feature subset selection algorithm of Hall (1998).[3] The algorithm evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between features. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. The preferred subset observed using this technique correlated with the emotion labels with a strength of 0.234 and comprised the following features: {`f0-mean`, `f0-min`, `f0-rising`, `db-min`, `db-range`}. In other words, mean and minimum pitch, as well as the proportion of rising pitch, were the most important for distinguishing among emotions, as were the minimum and range of intensity.

4.2 Emotion profiling

In order to discern the profile of each intended emotion according to its acoustic-prosodic characteristics, we performed statistical tests to determine whether the mean feature values associated with an emotion were statistically different from those of other emotions. Table 4.2 lists, for each emotion, the mean values of all features under consideration. Since feature values were z-score normalized, a value close to 0 indicated that the raw value was near the mean value for this feature. A value of 1 or -1 indicated that the raw value was

[2]F-measure was calculated using the standard method: $(2 \cdot \text{precision} \cdot \text{recall})/(\text{precision} + \text{recall})$.

[3]This algorithm is implemented as `CfsSubsetEval()` in WEKA.

	db-max	db-mean	db-min	db-range	f0-max	f0-mean	f0-min	f0-rising	f0-range	f0-curve	f0-slope	f0-voiced
anxiety	-0.69	-0.43	0.14	-0.64	0.02	-0.43	-0.23	-0.28	0.09	0.18	0.21	-0.23
boredom	-0.42	0.03	0.51	-0.72	-0.21	-0.83	-0.68	0.40	0.03	0.14	0.12	-0.08
cold-anger	0.16	-0.23	-0.22	0.30	-0.27	-0.20	-0.31	-0.13	-0.11	-0.11	0.12	-0.31
contempt	-0.29	-0.39	-0.07	-0.17	-0.25	-0.54	-0.48	-0.37	-0.08	0.23	0.00	-0.13
despair	-0.12	0.09	0.29	-0.31	0.15	-0.02	-0.06	-0.21	0.13	0.21	-0.07	-0.03
disgust	0.39	0.34	0.10	0.23	-0.04	-0.30	-0.34	0.05	0.08	-0.06	-0.15	-0.18
elation	0.73	0.62	-0.44	0.81	0.27	1.30	1.11	0.15	-0.14	-0.28	-0.02	0.60
happy	-0.09	-0.34	-0.53	0.35	-0.08	0.36	0.37	0.15	-0.19	-0.32	-0.16	0.26
hot-anger	0.81	-0.08	-1.32	1.69	0.40	1.36	1.02	0.01	0.07	-0.36	-0.37	-0.02
interest	-0.17	0.08	0.30	-0.35	-0.01	-0.10	-0.18	0.45	0.07	0.10	0.23	-0.02
neutral	-0.27	0.30	0.64	-0.72	-0.11	-0.79	-0.60	-0.63	0.07	0.11	-0.05	0.45
panic	0.80	0.41	-0.19	0.78	0.29	1.46	1.66	-0.17	-0.34	-0.52	-0.57	0.19
pride	-0.05	-0.19	-0.05	-0.04	-0.11	-0.31	-0.27	0.09	-0.01	0.17	0.03	0.04
sadness	-0.54	-0.14	0.31	-0.63	0.16	-0.38	-0.45	0.07	0.31	0.31	0.29	-0.06
shame	-0.12	0.17	0.65	-0.60	-0.17	-0.62	-0.56	0.12	0.05	0.14	0.28	-0.27

Table 4.2: Mean feature values (as z-scores) for each emotion in the EPSAT corpus.

one standard deviation above or below the mean, respectively. Mean z-score normalized feature values were useful for painting an over-all picture of the relative acoustic-prosodic properties of each emotion.

To examine significant differences between emotions using the mean feature values of Table 4.2, we ran unpaired t-tests between all emotion pairs for each feature.[4] Figures 4.1 and 4.2 (on pages 24 and 25, respectively) show graphically where each emotion lay in the z-score space of each feature. Emotions were grouped together based on statistical differences between mean feature values ($p < 0.01$). The mean feature values of all emotions in a group were statistically different from the means of all emotions in all other groups. Each group is color-coded and separated by white space. For example, in Figure 4.1 (a), the result of running unpaired t-tests produced four statistically distinct groups with respect to mean pitch (`f0-mean`). The emotions in the group with the highest mean pitch—`panic`, `hot-anger`, and `elation`—had statistically different means from all other emotions. However, the results of running similar tests for mean maximum pitch (`f0-max`)

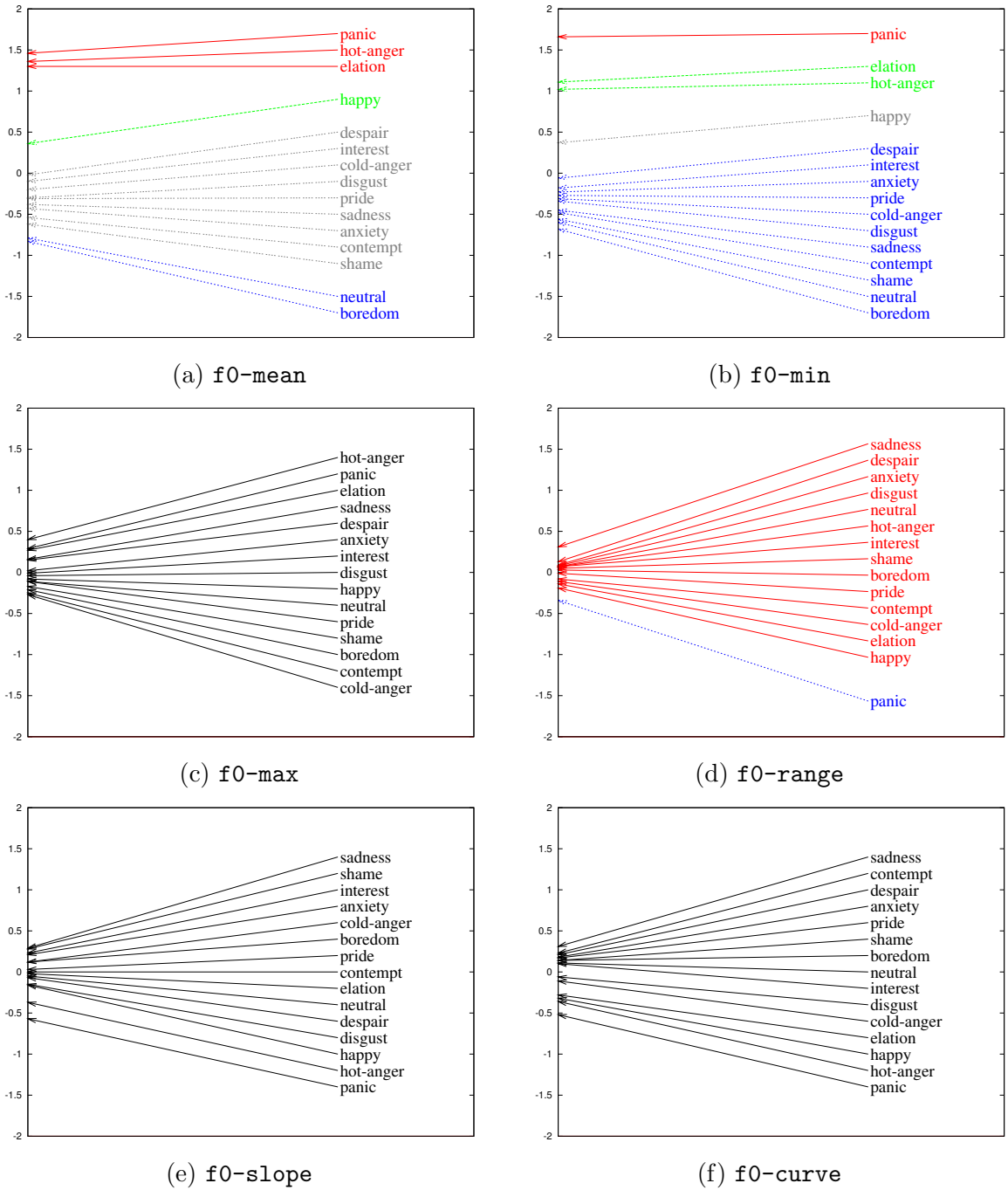
[4]T-tests were unpaired because intended emotion labels were associated with different utterances and therefor constituted independent sets of various sizes.

showed no emergent groupings, as shown in Figure 4.1 (c). This indicated that `f0-max` was a feature that did *not* help discriminate among the intended EPSAT emotions, given the acoustic-prosodic evidence.

We considered each emergent group to be a **natural classification** of emotion given a feature. Using this framework, we were able to quantize each group based on where the feature values lay in z-score space. In other words, if there were two emergent groups for a feature, then the group with the higher means was assigned a feature value of `H` for that feature and the group with the lower means was given a feature value of `L`. When we observed three statistically different groups we added a feature value of `M` for the group that resided in the mid range of feature means. We also introduced `MH` and `ML` for features that partitioned the emotion set into five natural classes for emotions that lay in the mid-high and mid-low range, respectively. For example, we can see in Figure 4.2 (e) that maximum intensity (`db-max`) could be considered a distinctive feature that partitioned the emotion set into two classes: `hot-anger`, `panic`, and `elation` were in the group with high means (`H`) and the remaining emotions were in the group with the low means (`L`). Features that did not divide the emotions into different groups were not considered to be distinctive and were not quantized. A view of the quantized values for all distinctive features given the intended labels in the EPSAT corpus can be seen in Table 4.3 on page 26.

An ideal distinctive feature set would allow us to uniquely identify all emotions. The dashed lines in Table 4.3 delineate emotions—or groups of emotions—that are discriminated by their sequences of quantized feature values. Six emotions were uniquely identified in this manner: `hot-anger`, `elation`, `happy`, `boredom`, `panic`, and `neutral`. The remaining emotions, though not uniquely identified, could nevertheless be grouped into natural classes of their own. One such group comprised {`cold-anger`, `disgust`}; the second, {`despair`, `interest`, `pride`, `contempt`}; and the third, {`sadness`, `anxiety`, `shame`}.

In fact, it was not necessary to use the full set of distinctive features to achieve the partitioning we observed in Table 4.3. In other words, it was not a **minimally distinctive set**. A minimally distinctive set was found that required only four features. These sets were either {`f0-mean`, `f0-min`, `f0-rising`, `db-range`} or {`f0-mean`, `f0-range`, `f0-rising`, `db-range`}. These features corresponded almost exactly to the set of most influential fea-

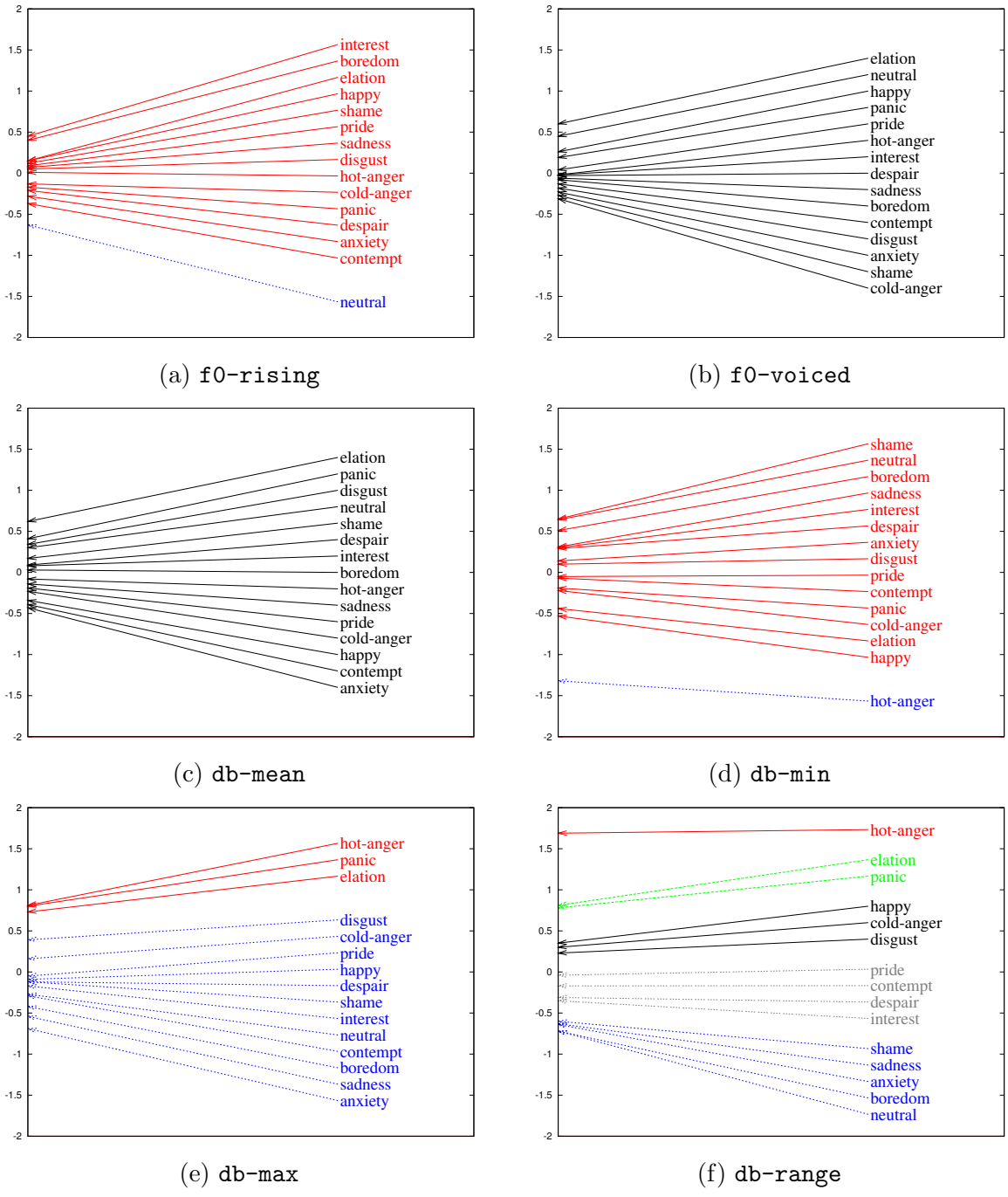


(a) `f0-mean` (b) `f0-min`

(c) `f0-max` (d) `f0-range`

(e) `f0-slope` (f) `f0-curve`

Figure 4.1: Statistically different groups of intended emotions based on t-tests of z-scored feature means.



(a) `f0-rising` (b) `f0-voiced`

(c) `db-mean` (d) `db-min`

(e) `db-max` (f) `db-range`

Figure 4.2: Statistically different groups of intended emotions based on t-tests of z-scored feature means.

	f0-mean	f0-min	f0-range	f0-rising	db-min	db-max	db-range
hot-anger	H	MH	H	H	L	H	H
elation	H	MH	H	H	H	H	MH
happy	MH	ML	H	H	H	L	M
cold-anger	ML	L	H	H	H	L	M
disgust	ML	L	H	H	H	L	M
despair	ML	L	H	H	H	L	ML
interest	ML	L	H	H	H	L	ML
pride	ML	L	H	H	H	L	ML
contempt	ML	L	H	H	H	L	ML
sadness	ML	L	H	H	H	L	L
anxiety	ML	L	H	H	H	L	L
shame	ML	L	H	H	H	L	L
boredom	L	L	H	H	H	L	L
panic	H	H	L	H	H	H	MH
neutral	L	L	H	L	H	L	L

Table 4.3: Quantized feature values per intended emotion determined by statistically different means.

	hot-anger	elation	happy	cold-anger	disgust	despair	interest	pride	contempt	sadness	anxiety	shame	boredom
f0-mean	H	H	MH	ML	ML	ML	ML	ML	ML	ML	ML	ML	L
db-range	H	MH	M	M	M	ML	ML	ML	ML	L	L	L	L

	panic	other
f0-range	L	H

	neutral	other
f0-rising	L	H

Table 4.4: Emotion profiles using the minimally distinctive acoustic-prosodic feature set.

tures found in our machine learning experiments. The only exception was db-min, which was found to be important in the machine learning experiments but not in our descriptive analysis here.

Emotion profiles based on minimally-distinctive features are shown in Table 4.4 on page 26. The feature f0-rising was the only feature required to uniquely identify neutral utterances, which were found to have less rising pitch relative to all other emotions. Similarly, both f0-min and f0-range uniquely identified panic, which was the only emotion to have high minimum pitch and low pitch range relative to the other emotions. The remaining features—f0-mean and db-range—were by far the most distinctive features since they alone were responsible for partitioning all other emotions in our set.

Utterances intended to convey hot-anger and elation had high mean pitch, but they were distinguishable in that hot-anger had a larger intensity range than elation, though both had a higher intensity range than the other emotions. Happy utterances tended to have a mean pitch in the mid-high region and an intensity range in the mid region. Boredom was uniquely identified as having both a low mean pitch and a narrow intensity range. The remaining emotions lay in the mid-low mean pitch region, though intensity range divided this group into three smaller groups.

4.3 Discussion

Given our feature set of automatically-derived acoustic features, we were able to uniquely profile six emotions using statistical analyses of their mean feature values: hot-anger, elation, happy, boredom, panic, and neutral. The remaining nine emotions, however, were *not* able to be uniquely profiled, though we were able to partition them into three groups: {cold-anger, disgust}; {despair, interest, pride, contempt}; and {sadness, anxiety, shame}. While the group that comprised cold-anger and disgust seemed to be quite coherent with respect to affect, the other two groups appeared less coherent in terms of affective similarity.

The reasons for the observed emotion groupings were difficult to determine, though it has been suggested by past studies of emotional speech that acoustic information is insufficient

for complete emotion discrimination. For example, it has been claimed that that global acoustic information of the sort we explored is only capable of discriminating along the activation dimension of emotion (e.g., Davitz, 1964; Schröder et al., 2001). Nevertheless, this notion was not entirely supported by our results. We observed that some emotions that were similar in activation but quite different with respect to valency—`hot-anger` and `elation`, for example—were in fact distinguishable by their acoustic cues, as identified by our analytic approach.

Furthermore, it has also been suggested that the emotional force of utterances with identical acoustic cues can change given different lexical, pragmatic, or discoursal contexts (e.g., Cauldwell, 2000; Oudeyer, 2002). An explanation of why we did not find distinguishing acoustic cues for some emotions might be due to the fact that there are not any; in other words, these emotions can *only* be disambiguated by context. This claim suggests that listeners would not be able to reliably identify some of the emotional content of the utterances in the EPSAT corpus because, by design, all such contextual cues were absent. In the next section we describe listener perception studies and will explore this topic further.

Another factor that might have contributed to our inability to fully differentiate the EPSAT emotion labels is the degree to which the actors were consistent with their intentions and vocalizations. It is impossible to know whether this was the case given only the intentional labels of the EPSAT corpus, and this issue will also be revisited in Chapter 5, where we report our findings of the analysis of perceived emotion.

The findings of the machine learning experiments and t-test quantification appear to be somewhat inconsistent in that the former exhibited relatively low F-measures, while the latter proved largely successful at discriminating emotions. However, the experimental frameworks of the analyses were different and thus the implications drawn from each should also be different. The descriptive analysis took a global view of the data. With automatic classification, though, computational models must be tested on unseen data. Such an approach introduces error in favor of presenting reasonable measurements of expected classification accuracy on unseen data. In other words, the goal of automatic classification is prediction rather than description. Also, our descriptive analysis looked only at mean feature values, but there was considerable variation within utterances. So, even when mean

feature values were statistically different, there would still be utterances associated with an emotion that had feature values deviating from the mean, contributing to incorrect prediction of the emotion of these utterances. Nevertheless, we noticed that the results of our descriptive analysis were analogous to our automatic classification results. `hot-anger` had the highest F-measure (0.53); an expected outcome given that there were several features that uniquely identified it and the mean z-score values were most extreme for this emotion. We also saw that classification of `panic` and `neutral` obtained some of the highest F-measures and showed significant differences from other emotions using the t-test analysis. In fact, we observed that the emotions that were uniquely identified via our quantized partitioning of the feature space were also the most correctly classified emotions in our machine learning experiments. The least correctly classified emotions were those which could not be uniquely identified by our feature space quantization.

There is a final point to make concerning the monothetic emotion labeling scheme of the EPSAT corpus. Our descriptive analysis indicated that actor intentions were relatively stable given that we were able to uniquely profile many emotions. However, the labeling scheme forced the decision tree learner to make distinctions between potentially similar emotions on somewhat arbitrary grounds. It was our hope that by conducting a perception experiment that eliminated a monothetic label typology (and adopted a polythetic one) that we would be able to train more accurate prediction models. We turn now to this study.

Chapter 5

Perceived Emotion

In this chapter we describe a study we conducted using a subset of the EPSAT corpus—referred to as the CU_EPSAT corpus—which was described in detail in Chapter 2. We had several goals, most of which were designed to answer the hypotheses put forth in Chapter 4, though in general we sought to examine the differences between intended, monothetic and perceived, polythetic emotion label sets and what effect, if any, different labeling schemes had on the acoustic cues of spoken emotions. Additionally, in this chapter we describe a technique we used for determining groups of listeners that, we believe, perceived emotion in systematically different ways. We formulate our hypothesis thusly:

Hypothesis 5.1 (RaterClust) *There are groups of raters who perceived emotion in systematically different ways.*

5.1 Perception survey

We designed an Internet-based survey to elicit perceived emotional ratings of the CU_EPSAT corpus. Each subject participated in the study remotely using their personal computers. After answering introductory questions about their language background and hearing abilities, subjects were given written instructions describing the procedure. Subjects were asked to rate each utterance—played out loud over headphones or speakers—on 10 emotional scales: `angry`, `anxious`, `bored`, `confident`, `encouraging`, `friendly`, `frustrated`, `happy`, `interested`, `sad`. For each emotion x, subjects were asked, "How x does this person

sound?" Subject responses could include: "not at all," "a little," "somewhat," "quite," or "extremely." These responses were later converted to the following ordinal scale:

- 0 ⇔ "not at all"
- 1 ⇔ "a little"
- 2 ⇔ "somewhat"
- 3 ⇔ "quite"
- 4 ⇔ "extremely"

At the onset of the experiment, subjects were presented with 3 practice stimuli in fixed order. Then the remaining 44 test stimuli of the CU_EPSAT corpus were presented one by one in random order. For each stimulus trial, a grid of blank radio-buttons appeared, as depicted in Figure 5.1 on page 32. The sound file for each trial played repeatedly every two seconds until the subject selected one response for each emotional scale. Subjects were not allowed to skip any scales.

The order in which the emotional scales were presented was rotated among subjects. Two randomized orders and their reverse orders were used. Each listener was presented with one of these fixed orders, shifted by one at each new token in a cyclic fashion to avoid any ordering effects.

All subjects were unpaid volunteer participants. Forty (40) native speakers of Standard American English with no reported hearing impairment completed the survey: 17 female and 23 male. All were 18 years of age or older, with a fairly even distribution among the following age groups: 18-22 years old (10.0%), 23-27 years old (20.0%), 28-32 years old (12.5%), 33-37 years old (7.5%), 38-42 years old (12.5%), 43-47 years old (20.0%), and over 48 years old (17.5%).

5.2 Rating correlation

The survey was designed to elicit polythetic perception ratings such that each utterance was rated for all emotions simultaneously. Since perception ratings lay on an ordinal scale, it was possible to examine the extent to which perceived emotions correlated with one

Emotion Recognition Survey: Sound File 1 of 47

	not at all	a little	somewhat	quite	extremely
How **frustrated** does this person sound?					
How **confident** does this person sound?					
How **interested** does this person sound?					
How **sad** does this person sound?					
How **happy** does this person sound?					
How **friendly** does this person sound?					
How **angry** does this person sound?					
How **anxious** does this person sound?					
How **bored** does this person sound?					
How **encouraging** does this person sound?					

Play Next Item

User ID: 8668462401
Having trouble with the survey? Please email the webmaster and include your user ID listed above.

Figure 5.1: Sample page from the CU_EPSAT web-based perception experiment.

another given the subject ratings as evidence. In other words, if emotion x was perceived, we investigated what this could tell us about the perception of emotion y.

Table 5.1 presents the Pearson product moment correlation matrix for all emotions using ratings from every subject in the study. Only correlations significant at $p < 0.001$ are shown. Of the 45 comparisons, only four were found to be non-significant. This alone supports Hypothesis 4.2: *Emotions are not independent; they are inter-related.* However, we wanted to say more about exactly *how* emotions were correlated and to do so we will discuss the correlations in more detail.

Table 5.1 is divided into quadrants by valency. The top right quadrant (red text) shows correlation between what are traditionally considered to be negative and positive emotions, whereas the top left quadrant (green text) contains correlation of negative emotions and the bottom right quadrant (blue text) displays significant correlations for positive emotions. A pattern quickly emerges: negative emotions negatively correlated with positive emotions, whereas positive correlation was found between emotions of equivalent valency. The only exception to this was the negative correlation observed between `bored` and `anxious`. Even though both are considered to be negative emotions, the presence of one implied the absence of the other.

	angry	**bored**	**frustrated**	**anxious**	**friendly**	**confident**	**happy**	**interested**	**encouraging**
sad	ns	**0.44**	**0.26**	**0.22**	**-0.27**	**-0.32**	**-0.42**	**-0.32**	**-0.33**
angry		ns	**0.70**	**0.21**	**-0.41**	ns	**-0.37**	**-0.09**	**-0.32**
bored			**0.14**	**-0.14**	**-0.28**	**-0.17**	**-0.32**	**-0.42**	**-0.27**
frustrated				**0.32**	**-0.43**	**-0.09**	**-0.47**	**-0.16**	**-0.39**
anxious					**-0.14**	**-0.25**	**-0.17**	ns	**-0.14**
friendly						**0.44**	**0.77**	**0.59**	**0.75**
confident							**0.45**	**0.51**	**0.53**
happy								**0.58**	**0.73**
interested									**0.62**

Table 5.1: Perceived emotion correlations ($p < 0.001$) based on CU_EPSAT ratings.

One might argue that strong negative correlations are indicative of a monothetic relationship between those emotions. For example, the fact that `happy` and `frustrated` were negatively correlated (-0.47) implies that when one was perceived the other was not perceived. In these cases it would be safe to adopt a monothetic labeling typology. However, the wealth of strong positive correlations between the remaining emotions clearly suggests that an utterance was rarely perceived to convey only one emotion, and thus motivates a polythetic labeling scheme overall.

Another observation is that positive emotions inter-correlated with a higher degree of strength than negative emotions did. Average correlation of positive emotions was 0.60; average correlation for negative emotions was much lower at 0.27. This implies that our set of negative emotions was more discrete than our set of positive emotions, a finding suggested as well by Juslin & Laukka (2003) in their meta-analysis. In fact, when developing the label set we found it much easier to arrive at a diverse set of colloquial terms for negative emotions than for positive ones. It would not be unreasonable to suggest that—at least in modern American culture—that there is a richer inventory of distinct negative emotions than there are positive ones. It remains to be seen whether such a statement could be made for all cultures, but it is an interesting finding nonetheless.

Correlations that were extremely strong—say, above 0.70—may suggest that that the labels themselves were redundant. We observed such high correlation with all pairwise combinations of `happy`, `encouraging`, and `friendly`. Similarly, `angry` and `frustrated` had a correlation coefficient of 0.70. Does this imply that the labels themselves were not discriminative, or is this an artifact of the way in which the actors produced the utterances, and/or a limitation of using only an acoustic information channel? The answers to these questions remain unclear, but regardless of the reasons, we expected to find similar acoustic cues for these highly correlated emotions.

5.3 Label correspondence

We hypothesized that intended emotion labels would not necessarily correspond to perceived emotion labels. In order to examine this, we recorded the frequency of co-occurrence

between intended and perceived emotion labels for all 44 utterances in the CU_EPSAT corpus. Each cell in Table 5.2 lists the number of times an intended label co-occurred with a perceived label divided by the total number of utterances with the intended label (N). Though counts were normalized by row, the rows do not sum to 1 because an utterance could be perceived as any number of our ten polythetic emotion labels. A perceived emotion was considered to be present for an utterance if the majority of subject ratings were between 1 and 4 for that emotion, indicating some degree of emotion perception. In the cases where there were no correspondences, the table cells have been left blank. When the intended and perceived emotion labels were the same, the proportion of correspondence is indicated in red. An intended emotion that was correctly perceived on all utterances would have a cell value of 1.0.

Overall, it seems, the actors tended to hit their marks: `sadness`, `interest`, `hot-anger`, and `boredom` were perceived as such 100% of the time, `anxiety` was perceived 75% of the

Intended Emotions	**angry**	**anxious**	**bored**	**confident**	**encouraging**	**friendly**	**frustrated**	**happy**	**interested**	**sad**	**N**
	Perceived Emotions										
anxiety		.75		.25		.25	.75		.75		4
boredom			1.0	.50			.25			.75	4
cold-anger	.50			1.0					1.0		2
contempt	.66		.33	.33			.66				3
elation		.66		.33		.66		.33	1.0		3
happy		.11		1.0	.67	.89		.78	1.0		9
hot-anger	1.0			1.0			1.0		.33		3
interest		.25		.25		.75			1.0		4
neutral			.50	.75		.25			.50	.25	4
pride				1.0	.75	1.0		.75	1.0		4
sadness		.75	.50				.50			1.0	4

Table 5.2: Proportion of intended emotion labels that corresponded to perceived labels.

time, and happiness 78% of the time. However, cold-anger was only perceived as angry half the time. The fact that all intended emotions corresponded to more than one perceived emotion is not unexpected given that we have shown in Section 5.2 that emotions were simultaneously perceived by raters. For example, the fact that hot-anger was perceived as both angry and frustrated further supports the polythetic nature of emotions.

What was problematic was when we observed label correspondences that were contrary to the perceived emotion correlations we reported earlier. For example, since anxiety and confident were negatively correlated we would not expect both of them to be simultaneously perceived given an intended emotion. However, we note that this occurred a quarter of the time. Anxiety was also perceived as friendly a quarter of the time as well. Other troubling correspondences were boredom/confident (.50), elation/anxious (.66), and happy/anxious (.11). It is clear, then, that the emotion that a speaker intended to convey was not always the emotion that was perceived. Recall that we only compared EPSAT utterances that were hand-picked precisely because they were perceived by three researchers to be unambiguous and prototypical examples of the emotions in the perceived emotion label set. The fact that we observed mismatch between speaker intention and listener perception cannot be attributed to bad tokens, and if we were to consider the entire EPSAT corpus we would likely observe even more mismatch.

There are a few other interesting points about the correspondences. The first is the abundant frequencies of perceived confident and interested labels. These occurred for almost every emotion and would seem to suggest that these are not full-blown emotions, but instead are representative of another sort of cognitive state, possibly attitude (Wichmann, 2000) or second-order emotions (Cowie, 2000).

Finally, it is very interesting to note that intended neutral utterances had a wide range of perception. They were perceived as bored 50% of the time, confident 75% of the time, friendly 25% of the time, interested 50% of the time, and sad 50% of the time. This is a widely disparate group of perceived emotions and lends credence to the commonly-held notion that there is no such thing as an emotionally-neutral utterance; rather, emotion is omnipresent, intended or not.

5.4 Rater behavior

We turn now to the rating behavior of 39 of the 40 subjects who participated in the perception study.[1] Not only were we interested in overall rating distribution, but we were also concerned with the degree to which raters agreed with one another.

In total, we collected 17,160 ratings (44 utterances · 10 emotions · 39 raters) via our perception study. The count of each rating is listed per emotion in Table 5.3. We observed that roughly half of all ratings were considered to "not at all/' convey the emotion surveyed (rating 0), while the other half was split more evenly between the other four ratings, though the counts generally declined as ratings became more extreme. The rating distributions for individual emotions followed this pattern as well, with a few exceptions. The label `angry` actually had a higher frequency of ratings with a value of 4 than those with a value of 3. Additionally, rating distributions for `interested` and `confident` were drastically

		Ordinal rating/Label				
		0/not at all	1/a little	2/somewhat	3/quite	4/extremely
Emotion	`angry`	1173	245	125	75	98
	`anxious`	921	327	200	178	90
	`bored`	1164	244	136	113	59
	`confident`	438	299	444	411	124
	`encouraging`	965	267	239	184	61
	`friendly`	768	378	313	191	66
	`frustrated`	918	307	244	162	85
	`happy`	1014	246	218	143	95
	`interested`	503	371	350	380	112
	`sad`	1092	261	170	103	90
Total	count	8956	2945	2439	1940	880
	percentage	52.2	17.2	14.2	11.3	5.1

Table 5.3: Rating distribution of all raters over all utterances in the CU_EPSAT corpus.

[1] One rater was excluded from all analyses reported here because of severely aberrant behavior.

different than the distributions of the other emotions. For both, the rating frequency was distributed much more evenly among the ratings than were the other emotions. This might be indicative of the fact that these labels better represented attitude or mood, rather than full-blown emotional states, as suggested previously. As we saw in the previous section, most intended emotions were perceived as `confident` and `interested`, even when they were simultaneously perceived to be other emotions, and irrespective of valency.

The preponderance of rating 0 can be explained in part by the design of the experiment itself. We chose what we felt were particularly unambiguous examples of each of the ten emotions, and chose an equal number of utterances per emotion. In the experiment, though, subjects were asked to rate each utterance for *every* emotion, independent of other emotions. Therefore, we might expect to find that any given utterance did not convey most emotions. Since the experiment was designed such that each of the ten emotions was equally represented, we would have expected that roughly 10% of all ratings for those utterances would be in the range of 1 to 4. Stated differently, if the emotion labels were monothetic we would predict that 90% of the ratings would be 0. This estimate is much higher that the 52% we observed and further supports Hypothesis 4.2, that perceived emotions labels were, in fact, polythetic.

In addition, our rating scale was designed to record the degree of perceived presence of emotions, and not the converse. Therefore, ratings 1, 2, 3, and 4 were all used when an emotion was perceived, whereas rating 0 was reserved for all possible degrees of the *absence* of emotion.[2] Using such a design produced an unbalanced scale and explains why the distributions were so heavily skewed in favor of rating 0. We addressed this issue by analyzing the data in a dichotomous fashion. In subsequently reported analyses we often treated rating 0 differently from the non-zero ratings.

Figure 5.2 on page 40 shows the histogram of the frequency, per subject, of all ratings with a value of 0. We noticed that while most raters tended to assign rating 0 half of the time (M = 229.6, SD = 54.6, N = 440), the frequency range was actually quite large. The

[2]One could imagine a different study with the following four point scale, for example: (-2) *extreme absence of emotion* ... (-1) *mild absence of emotion* ... (1) *mild presence of emotion* ... (2) *extreme presence of emotion*.

most conservative rater assigned rating 0 only 30% of the time, whereas the most generous chose it 78% of the time. There appeared to be a sizable subset of raters who chose rating 0 about 75% of the time. The distribution was not shown to be significantly different from the expected normal distribution given the Kolmogorov-Smirnov Test of Composite Normality (ks = 0.127, p = 0.5).

When we considered only ratings between 1 and 4, we observed that mean rating per subject ranged from 1.5 to 2.7. Observed mean rating over all speakers was 2.1, with a standard deviation of 0.3. Here, too, we found the distribution to be normal (ks = 0.120, p = 0.5). A histogram of the mean positive ratings is presented in Figure 5.3 on page 40. When an emotion was perceived, its strength was relatively weak. To use the parlance of the survey itself, if an emotion was detected, it was judged, on average, to be "somewhat" (rather than "a little", "quite", or "extremely") present.

To summarize the general behavior of the subjects of the CU_EPSAT perception study, then, we conclude that it was relatively consistent. Half of the time emotions were not perceived and, when they were, they were considered by subjects to be "somewhat" present. This type of global analysis is contrary to our earlier hypothesis that raters perceived emotions differently. What we can say is that raters were consistent with respect to the frequency with which they perceived the presence of emotions and that—when an emotion was perceived—they were consistent with respect to strength of that emotion. However, the present analysis says nothing about whether the decisions made at the utterance level were consistent across raters. We turn to this now.

5.5 Rater agreement

Kappa (κ) is a commonly used measurement for comparing the agreement of two labelers on the same task. It is especially well regarded because it computes the probability that two raters agree on a set of labels conditioned by what chance would afford given the label distribution (Cohen, 1968). Traditionally, κ considers only exact label matches to be correct, an assumption best suited for nominal data. For ordinal or ranked data, weighted κ is often used because it assigns partial credit for ratings that are close on the ordinal scale.

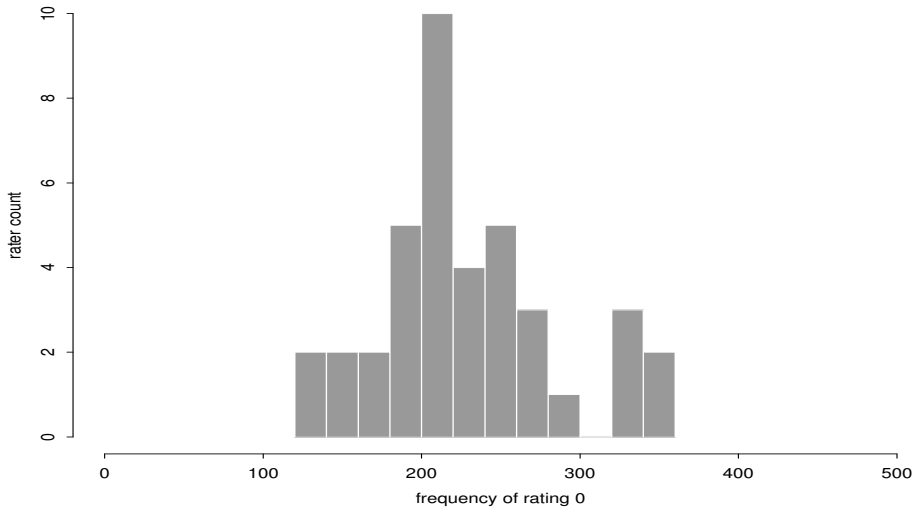


Figure 5.2: Histogram of the number of times each rater chose rating 0 (no emotion present) in the CU_EPSAT perception survey.

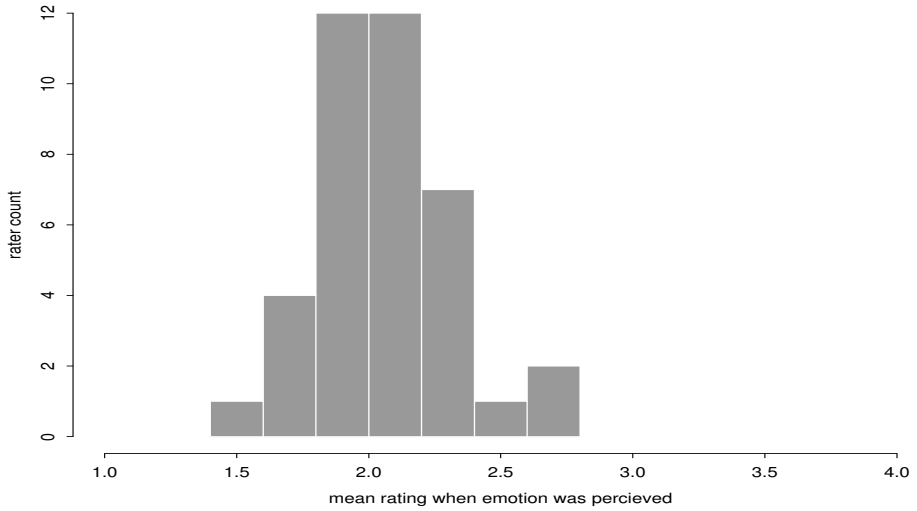


Figure 5.3: Histogram of the mean rating considering only ratings between 1 and 4 (indicating the degree of emotion present) for each rater in the CU_EPSAT perception survey.

We adopted **quadratically-weighted** κ, though we measured the agreement between two raters in a slightly non-standard way. This algorithm is shown in pseudocode below.

Algorithm 1: ComputeKappaWeight(**r1**, **r2**)

```
r1 = rating of first rater
r2 = rating of second rater
\\ special case:  emotion was perceived by one rater but not the other
if (r1 = 0 and r2 > 0) or (r1 > 0 and r2 = 0) then
  return 0
end if
\\ normal case:  quadratic weighting
diff = |r1 − r2|
if diff = 0 then
  return 1
else if diff = 1 then
  return 0.89
else if diff = 2 then
  return 0.56
else
  return 0
end if
```

For pairs of ratings in the range of (1,4), agreement was calculated using standard quadratic weighting: two ratings that did not differ at all were assigned an agreement score of 1; two that differed by one rank received an agreement score of 0.89; two that differed by two ranks were assigned an agreement score of 0.56, and two that differed by three ranks received an agreement score of 0. We felt it was important to respect the theoretical distinction between the perception of the *absence* of an emotion and the perception of the *presence* of an emotion. Therefore, in the case where one rater chose rating 0 and the other did not, the agreement score assigned would be 0, even if the non-zero rating was very close to 0 (for example, if the rating were 1). If both raters chose "no at all" then they received an agreement score of 1 for that utterance. In this manner, we computed semi-quadratically-weighted κ (henceforth referred to as κ_{qw}) between every pair of raters who participated in the survey.

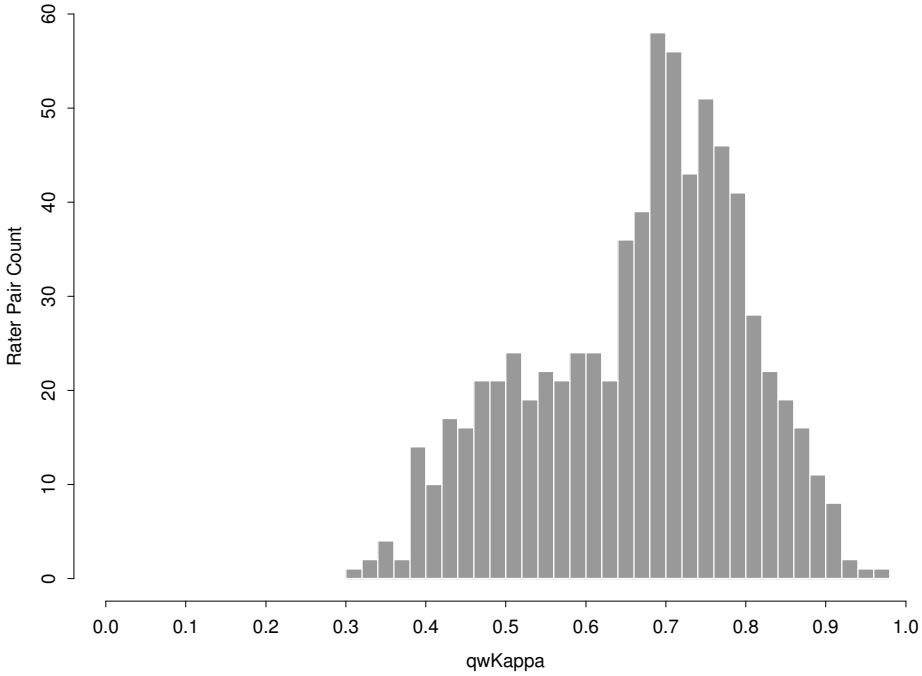


Figure 5.4: Histogram of inter-rater κ_{qw} measurements (M = 0.67, SD = 0.13, N = 780).

A histogram of all observed κ_{qw} scores (M = 0.67, SD = 0.13, N = 780) is shown in Figure 5.4. The distribution was found to be statistically different from the normal distribution (ks = 0.098, $p < 0.001$), indicating a certain amount of heterogeneity in the distribution of κ_{qw} scores. In other words, there appeared to be a certain amount of disagreement among the ratings assigned to the utterances in the CU_EPSAT corpus. Though most of the raters seemed generally to agree, as seen by the largest mass centering around $\kappa_{qw} \approx 0.75$, there was a second mass observed around $\kappa_{qw} \approx 0.55$. We interpret this to indicate that there was a considerable number of raters who did not agree with one another. In the next section we present results of automatic clustering experiments designed to determine whether rater disagreements were random or systematic, thus addressing Hypothesis 5.1.

5.6 Rater clustering

We chose to explore automatic clustering techniques as a way to determine whether raters could be grouped together based on systematic emotion perception behavior. We were motivated to do so based on our observation of the non-normal distribution of inter-rater κ_{qw} measurements. As with all clustering techniques, a method of evaluating the similarity between two items is necessary. We chose as our similarity metric κ_{qw}, described above. The clustering algorithm is shown in pseudocode below.

Algorithm 2: ClusterRaters

```
R = set of raters
C = set of clusters
Chose initial cluster centroids; add these raters to C and remove from R.
for each rater r in R do
  maxRater = 0
  for each cluster c in C do
    mean = mean κ_qw between r and all raters in c
    if mean > maxMean then
      maxMean = mean
      maxCluster = c
    end if
  end for
  add r to C[maxCluster]
end for
```

The algorithm considered all raters who had not yet been added to a cluster. For each existing cluster, the mean κ_{qw} scores between the rater and all raters in each cluster were calculated. The rater was added to the cluster with which it had the highest mean κ_{qw}. This process was continued until there were no raters left to cluster.

The results of most clustering techniques are heavily influenced by the initial centroids of each cluster. We adopted an approach by which we chose the number of clusters *a priori* and took special care to seed the clusters with centroids that were as dissimilar from one another as possible. We did so by first identifying the two raters with the lowest inter-rater

κ_{qw}. Each of these was made the centroid of its own cluster. If more than two clusters were desired, we chose as the centroid of the next cluster the rater whose average κ_{qw} when compared with the current cluster centroids was smallest. In this way we ensured that the initial clusters represented maximally dissimilar raters.

We report here experiments with two, three, and four clusters. One reason for limiting the number of clusters was because, with only 39 raters, we felt that we could not properly generalize any findings using more than four clusters. Table 5.4 describes the properties of each cluster under each experimental design. We observed that under all conditions, the number of raters in each cluster was uneven. Under each condition there was one cluster that contained the overwhelming majority of the raters (`a1`, `b1`, and `c1`). From this we confirmed what we put forth in the previous section—that most raters displayed similar rating behavior. When clustering with two initial centroids, we observed that cluster `a1` had a high mean inter-rater κ_{qw} (0.70) and that cluster `a2` had a lower mean κ_{qw} (0.56), but it also showed very small standard deviation (0.01), signifying a high degree of cluster-internal homogeneity. The standard deviation of the raters in cluster `a1`, however, was much higher (0.11) and thus motivated increasing the number of clusters. When we ran

no. of clusters	**cluster id**	**mean κ_{qw}**	**stdv κ_{qw}**	**no. of raters**
1	—	0.67	0.13	39
2	`a1`	0.70	0.11	36
	`a2`	0.56	0.01	3
3	`b1`	0.73	0.09	34
	`b2`	0.56	0.01	3
	`b3`	0.57	NA	2
4	`c1`	0.76	0.08	28
	`c2`	0.70	0.07	6
	`c3`	0.56	0.01	3
	`c4`	0.57	NA	2

Table 5.4: Results of automatic clustering of raters using κ_{qw} as the comparison metric for two, three, and four clusters.

the algorithm with three initial centroids we saw that the cluster with the largest amount of raters (`b1`) had a lower standard deviation (0.09) and a higher mean κ_{qw} (0.76) than did cluster `a1`. This trend continued when we clustered with four initial centroids; however, here we noticed that the largest cluster had in fact been split into two clusters (`c1` and `c2`), each with a cluster-internal mean κ_{qw} greater than or equal to `b1`'s and a standard deviation less than `b1`'s. In addition, both `c1` and `c2` were recorded to have higher inter-rater κ_{qw} scores and lower standard deviation than did the entire group of raters together (M = 0.67, SD = 0.13). What this indicated was that when four clusters were chosen *a priori*, a clustering was found that divided the dominant rater behavior into two more homogeneous groups of raters, albeit still unbalanced (`c1` contained 28 raters while `c2` only had 6). When experiments were run with more than four clusters, each new centroid was always taken from cluster `c1` and segregated to its own centroid with no other raters added to it. The implication here was that clustering equilibrium had been reached at four clusters.

Based on these observations, we felt that the clustering that resulted when running the algorithm with four initial centroids was the best way to cluster the raters. This is the rater grouping that we adopted and the one we will be discussing for the remainder of this chapter. Unfortunately, the number of raters for clusters `c2` and `c3` were too few to run statistical tests on, so we cannot discuss the possible implications of the behavior of these raters, except for the fact that they were most similar to the other raters in their respective clusters.

Given the high means and low standard deviations of clusters `c1` and `c2`, we suspected that the clusters were homogeneous and were indicative of a high rate of inter-cluster rater agreement. In order to test this, we examined the histograms of the pairwise inter-rater κ_{qw} within each cluster, shown in Figure 5.5 on page 46. Both clusters were, in fact, normally distributed according to the Kolmogorov-Smirnov Test of Composite Normality (`c1`: ks = 0.036, p = 0.5; `c2`: ks = 0.152, p = 0.5), which was not true of the unclustered distribution (cf. Figure 5.4). This further supported our claim that clustering was appropriate for the raters based on their perceptual ratings of emotions, because doing so resulted in groups of raters that were more internally consistent than the entire group of raters was on the whole.

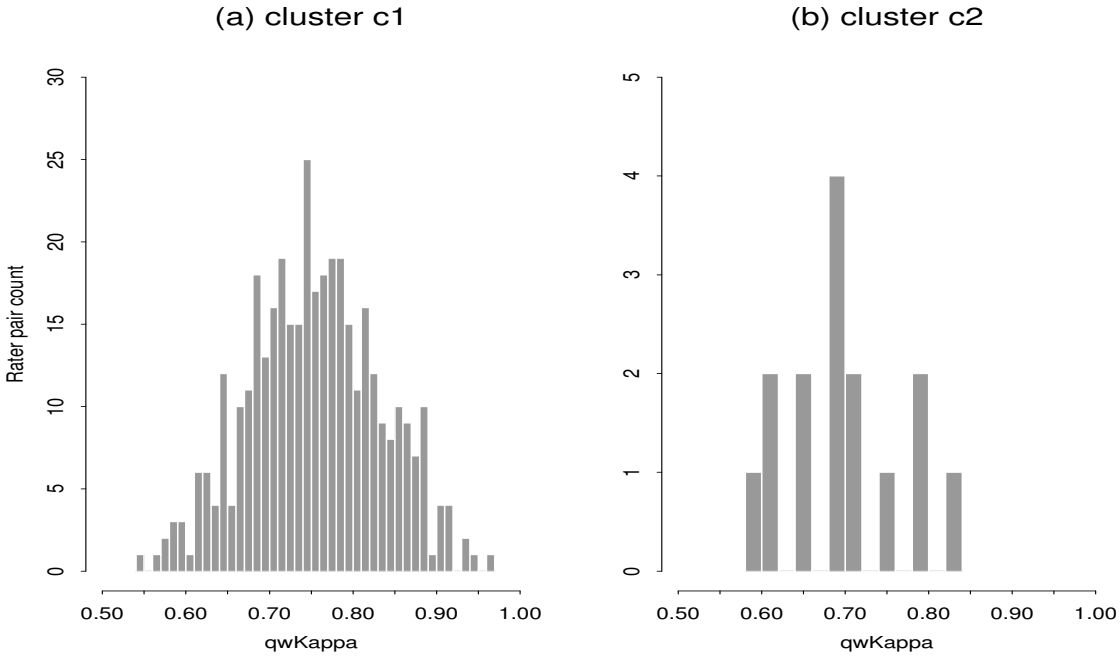


Figure 5.5: Histogram of pairwise rater κ_{qw} for those raters in clusters `c1` and `c2`.

5.7 Cluster profiling

Having identified two internally coherent clusters of raters, we wished to explore the specific ways in which these clusters differed, motivated by the underlying hypothesis that clusters could be differentiated by their demographic and behavioral makeup. In this section we document our exploration of several hypotheses related to this notion and describe our findings with respect to each.

Hypothesis 5.2 (ClustRaterSex) *Clusters differed by the sex of the raters in each.*

In both clusters, there were fewer female raters than there were male raters: 46% of the raters in `c1` were female compared with 33% in `c2`. This was also true of the experiment overall (42.5% were female). However, the ratio of male to female raters in cluster `c2` was 2 to 1, much higher than it was in cluster `c1` (1.2 to 1). In other words, cluster `c2` represented a more male-dominated cluster than did cluster `c1`, though we were unable to verify the statistical significance of this finding due to the small number of raters in cluster `c2`. We must say, then, that these findings only tentatively support Hypothesis 5.2.

Hypothesis 5.3 (ClustRaterAge) *Clusters differed by the age of the raters in each.*

Table 5.5 lists the the age range distribution of the raters in each cluster. Again, due to the small size of cluster `c2` we were unable to test whether the distributions were statistically different. Though the relative age distributions were not equivalent, both clusters showed similar spreads of raters across all age ranges. For example, for both clusters approximately half were under 38 years of age and the other half were older. Therefore, Hypothesis 5.3 could not be supported; age range of raters did not appear to be a distinguishing factor in rater clustering.

Hypothesis 5.4 (ClustDegree) *Clusters differed in the degree of perceived emotion reported.*

What of the degree to which raters assigned ratings? Was it the case that raters in `c1` were more generous or conservative with respect to their assignment of ratings than were the raters in `c2`? In order to address this, we computed the mean rating of each

	c1		c2	
age	count	percent	count	percent
18-22	1	3.6%	1	16.7%
23-27	6	21.4%	1	16.7%
28-32	5	17.9%	0	0.0%
33-37	2	7.1%	1	16.7%
38-42	2	7.1%	2	33.3%
43-47	7	25.0%	0	0.0%
>48	5	17.9%	1	16.7%

Table 5.5: Age range distribution of clusters c1 and c2.

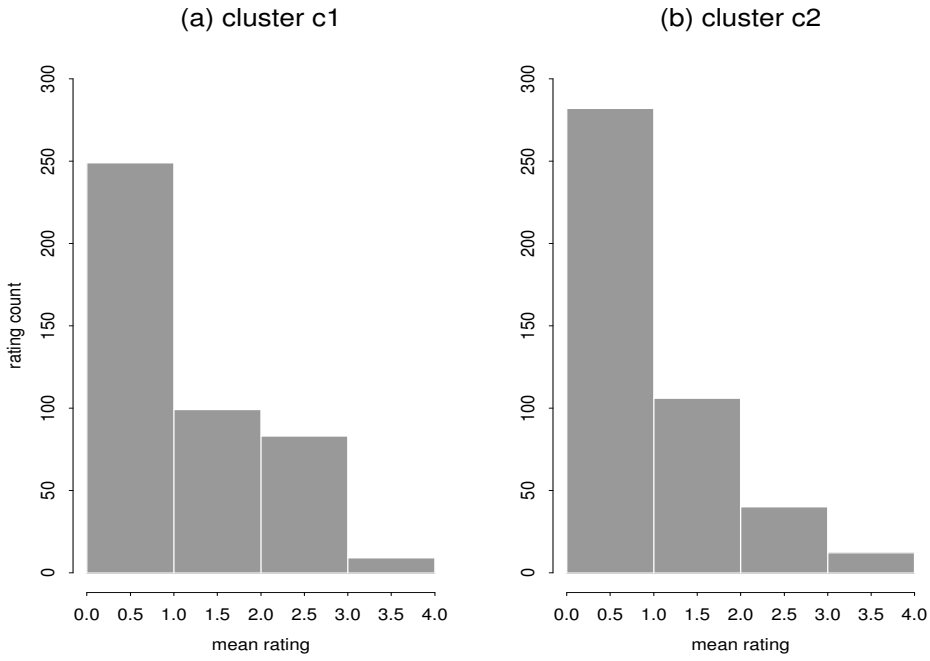


Figure 5.6: Histogram of mean ratings for clusters c1 and c2.

utterance for each emotion (N = 10 emotions · 44 utterances = 440). Histograms of these data are plotted for each cluster in Figure 5.6 on page 48. The distributions did not appear radically different under visual inspection, though a t-test showed that the means of the rating distributions were significantly different ($t = 5.9$, $df = 439$, $p < 0.001$). The Kolmogorov-Smirnov Test of Composite Normality also showed that the distributions were significantly different ($ks = 0.12$, $p = 0.002$). By visual analysis we can see that this was due to a higher proportion of ratings in the range of 2-3 for cluster `c1`. In other words, cluster `c1` perceived significantly more emotions that were 'quite' present than did cluster `c2`, thus lending support for Hypothesis 5.4.

Hypothesis 5.5 (ClustUtt) *Clusters differed in their ratings on only some utterances.*

Due to the fact that the cluster distributions of mean ratings, though statistically different, were not radically different, we suspected that raters in each cluster only differed on some utterances, not all. To this end, we ran an analysis of variance (ANOVA) for every utterance per emotion, comparing the ratings in both clusters. We observed that 19% (84/440) were significantly different ($df = 32$, $p < 0.05$) with respect to the ratings assigned by the raters in each cluster. Though most of time the raters in each cluster tended to agree on the emotion perceived, we found a number of utterances and emotions for which they disagreed. This finding supported Hypothesis 5.5 and subsequent analyses examined cluster differences with respect to these 84 samples.

Hypothesis 5.6 (ClustSpeakerSex) *The sex of both speakers and raters affected emotion perception.*

At this point we can revisit possible sex differences between the two clusters. We examined the distribution of speaker/rater sex pairings over the 84 statistically different utterances and emotions, a summary of which is shown in Table 5.6 on page 50. Surprisingly, the 84 samples were spoken by the exact same number of male and female speakers, a clear indication that Hypothesis 5.6 did not hold and that the combination of speaker and rater sex had no influence on rating behavior.

Hypothesis 5.7 (ClustEmo) *Clusters differed in rating behavior by emotion.*

We also examined whether the raters in the two clusters rated individual emotions differently. Table 5.7 shows each perceived emotion and the number of utterances for which mean ratings were significantly different between clusters. All emotions labels were represented, though the positive emotions far outweighed the negative ones by 2 to 1 (56 positive, 28 negative). This implied that raters differed in terms of how they perceived the valency of emotions and that, specifically, raters disagreed more on ratings for positive emotions than for negative ones. Though, it is important to note that the most conventionally prototypical emotion—`angry`—was also highly disagreed upon between the two clusters.

Sex		Cluster	
speaker	rater	`c1`	`c2`
F	F	546 (23.2%)	84 (16.7%)
M	F	546 (23.2%)	84 (16.7%)
F	M	630 (26.8%)	168 (33.3%)
M	M	630 (26.8%)	168 (33.3%)

Table 5.6: Distribution of speaker/rater gender pairings over all 84 statistically different inter-cluster utterances and over all emotions.

Positive		**Negative**	
Emotion	**Count**	**Emotion**	**Count**
`confident`	16	`angry`	10
`interested`	12	`bored`	6
`encouraging`	12	`anxious`	5
`friendly`	9	`frustrated`	4
`happy`	7	`sad`	3

Table 5.7: Number of utterances with statistically different inter-cluster mean ratings per perceived emotion.

Positive		Negative	
Emotion	**Count**	**Emotion**	**Count**
happy	31	anxiety	5
pride	16	boredom	4
interest	11	sadness	2
elation	8	hot-anger	2
		contempt	2
		cold-anger	2

Table 5.8: Number of utterances with statistically different inter-cluster mean ratings per intended emotion.

Also of interest was the patterning among clusters with respect to the *intended* emotions of the actors in the corpus. Table 5.8 above lists the number of utterances that were disagreed upon by the two clusters, per emotion. Once again, there was more disagreement on intended positive emotions than on intended negative ones (positive emotions outnumbered negative emotions by almost 4 to 1). This behavior might be related to our earlier reported finding of the high inter-correlation of positive emotions. These emotions might have been harder to differentiate because of such strong co-occurrence. Alternatively, less effort might have been made on the part of the actors to distinguish among positive emotions since they were all so similar. Regardless, our observations lent support for Hypothesis 5.7 in that rater differences did seem to be dependent on specific emotions, though we note that disagreements may have been due more to a specific dimension of emotionality—valency—rather than on an emotion-by-emotion basis.

To further explore how raters disagreed with respect to perceived emotion, we examined the mean cluster rating of each emotion for the 84 different samples, as illustrated in Table 5.9. A pattern emerged that was dependent on the valence of the perceived emotions. For negative emotions, cluster `c2` provided higher ratings than did cluster `c1` and the inverse was true for positive emotions. In other words, cluster `c2` rated negative emotions as more negative and positive emotions as less positive than did cluster `c1`. If one were inclined to talk about personality types then one might say that the raters in cluster `c2` appeared to

		Mean Rating		Difference
	Emotion	c1	c2	c1 - c2
negative	angry	0.2	0.9	-0.7
	sad	0.1	0.6	-0.5
	frustrated	0.4	1.2	-0.8
	bored	0.1	0.5	-0.4
	anxious	0.5	1.4	-0.9
positive	friendly	2.0	0.9	1.1
	confident	2.3	1.1	1.2
	interested	2.6	1.4	1.2
	happy	2.3	0.9	1.4
	encouraging	1.8	0.5	1.3

Table 5.9: A comparison of mean ratings of perceived emotion between clusters c1 and c2.

be pessimists or introverts, always ready to perceive negative emotion, whereas the raters of cluster c1 appeared to be optimists or extroverts, perceiving emotions—and the people who are conveying them—in a positive light. However, since no personality tests were administered for the subjects of the perception experiment, no such claims can be made with any degree of scientific rigor.

To summarize our findings on cluster profiling, we report that most of the time the raters in both clusters behaved consistently, though 19% of the time they showed significant differences based on the ratings they provided. Though age and speaker sex played no part in differentiating the clusters, cluster c2 did appear to be more male-dominated with respect to rater sex. This cluster also tended to rate positive emotions as less positive and negative emotions as more negative than did cluster c1. These findings indicate that, at least in our data, coherent groups of people could be identified who perceived emotions differently and that these differences were characterized largely in terms of valency.

5.8 Automatic classification

In this section we present the results of machine learning experiments for the automatic classification of the perceptual emotion labels. Our goal was to explore Hypothesis 4.3 that predicted that the automatic classification of perceived emotion labels would perform better than the automatic classification of intended emotion labels because the latter set was monothetic and forced unnatural distinctions between possibly very similar emotions. A second hypothesis we wished to explore arose from the results of the rater clustering experiment. Since two clusters of like-minded raters were found, we expected that automatic classification using the perceptual emotion labels of each cluster would perform better than when using the labels derived from all raters in aggregation. Stated formally, our hypothesis was as follows:

Hypothesis 5.8 (ClustEmoClass) *Automatic classification of perceptual emotion labels of each rater cluster will perform better than perceptual emotion labels of all raters taken together.*

In order to address Hypothesis 4.3 we ran a machine learning experiment analogous to the one that was undertaken for the intended emotion set, as reported in Section 4.1. However, we had to dichotomize the ordinal ratings of the perceived emotion labels for compatibility. If the rating for a particular emotion was 0 ("not at all") then we considered that that emotion was not perceived; if the rating was in the range of (1,4) then we considered that emotion to have been perceived. We ran 10-fold cross-validation experiments to automatically classify the perceived emotions of each rater. Classification performance is reported as F-measure and is shown in Table 5.10. As before, we used the J4.8 decision tree learning algorithm as implemented in the Weka software package. Also present in Table 5.10 are the F-measure performances of the intended emotion labels of the EPSAT corpus, for comparison. When emotion labels were the same across sets they are shown side-by-side.

When comparisons were possible between label sets, we noticed that using perceptual emotion labels lead to better performance for all but one emotion: `angry`. The mean F-measure of the six corresponding emotions (`angry`, `anxious`, `bored`, `sad`, `happy`,

Intended		Perceived	
Emotion	**F-measure**	**Emotion**	**F-measure**
hot-anger	0.53	angry	0.29
anxiety	0.28	anxious	0.46
boredom	0.41	bored	0.44
sadness	0.20	sad	0.38
happiness	0.30	happy	0.49
interest	0.20	interested	0.75
—	—	confident	0.80
—	—	encouraging	0.49
—	—	friendly	0.55
—	—	frustrated	0.43
contempt	0.18	—	—
despair	0.25	—	—
disgust	0.25	—	—
panic	0.39	—	—
shame	0.22	—	—
cold-anger	0.21	—	—
neutral	0.40	—	—
pride	0.19	—	—
elation	0.30	—	—
mean	0.29	**mean**	0.53

Table 5.10: Classification performance (F-measure) of intended and dichotomized perceived emotion labels.

`interested`) was 0.32 for the intended emotion label set and 0.47 for the perceived emotion label set. When considering all the emotions of each labeling paradigm, the mean performance for intended emotions was 0.29 and for perceived emotions was 0.51. The classification of perceived emotions was more successful than the classification of intended emotions, and we believe this was because of the polythetic nature of the former set. These findings supported Hypothesis 4.3.

As a final note, the corpus size of each paradigm should be taken into account. The corpus of intended emotions (EPSAT) was at least an order of magnitude larger than the corpus of perceived emotions (CU_EPSAT).[3] In fact, the utterances used for automatic classification of the EPSAT corpus outnumbered those of the CU_EPSAT corpus by 55 to 1. This is relevant because increasing corpus size tends to yield higher prediction accuracy due to the fact more training data contribute to more robust prediction models. So, it is much more significant that the perceived emotion models, trained on far fewer data, nevertheless outperformed the models trained on the more plentiful data of the EPSAT corpus.

Having established that perceived polythetic emotion labels begot better classification accuracy than intended monothetic emotion labels, we next turned our attention to classification of the degree of emotion provided by the two rater clusters. For these experiments we changed the experimental design to model the ordinal ratings provided by the subjects of the perception study. The pseudocode describing the classification algorithm is presented on page 56. The first step was to chose the majority rating provided by a group of raters (i.e, those of a cluster) for each emotion and each utterance. These majority ratings were considered to be the degree of perception for each emotion. A two-step classification algorithm was then administered in the following manner. First, with the same dichotomized design described earlier, J4.8 decision trees were used to predict whether an emotion was perceived or not. If an emotion was predicted, then a separate continuous classifier was used to predict the degree to which the emotion was perceived by assigning a rating of 1, 2, 3, or 4. Linear regression was used as the the continuous prediction algorithm and—though continuous predictions were originally made—predictions were rounded to the nearest allowable

[3]The EPSAT corpus contained 2,400 utterances; the CU_EPSAT corpus contained 44.

Algorithm 3: 2TierClassification

```
\\ make a dichotomized copy of the data:
for each rating r do
  if r ∈ (1, 4) then
    r ← emotion
  else if r = 0 then
    r ← no-emotion
  end if
end for
\\ run machine learning experiments using leave-one-out cross-validation
for each emotion e do
  for each utterance u do
    train binary emotion classifier and get prediction p for utterance u
    if p = no-emotion then
      p ← 0
    else
      \\ emotion was predicted so predict its degree
      train a continuous classifier using utterances with ratings ∈ (1, 4)
      get continuous prediction p for utterance u
      p ← p rounded to nearest integer ∈ (1, 4)
    end if
  end for
  compute κ_qw between actual and predicted labels
end for
```

integer rating in the range (1,4).

Performance was measured not by F-measure but by computing κ_{qw} between the actual and predicted ratings. This was done for several reasons. First, F-measure can only be reported for each label in a discrete label set. Since ordinal ratings are in fact related to each other by definition, we did not wish to disregard such relationships. Additionally, κ_{qw} was used previously to report rater agreement and we found it conceptually expedient to compare the actual and predicted ratings, as we would two raters. Though adopting κ_{qw} as the performance metric no longer afforded us the ability to compare our results with the intended emotion experiments, this was not problematic because here we wished only to compare classification performance of the rater clusters with performance without clustering.

Table 5.11 presents the results of the experiments using the label sets of clusters `c1` and `c2` and all the unclustered raters as well. The average F-measures across all emotions indicated that performance of cluster `c1` (0.45) and cluster `c2` (0.53) both outperformed the unclustered label set (0.31). This finding supported Hypothesis 5.8 and was most likely the result of the internal consistency of the label sets provided by clustering based on rater behavior. However, our findings were tempered slightly by the fact that clustering did not outperform the unclustered label in all cases. There were two emotions—`angry` and `encouraging`—for which the κ_{qw} of the unclustered label set was actually larger than the κ_{qw} of both `c1` and `c2`. Also, there were two other emotions—`sad` and `anxious`—for which performance was 0.00 for both the unclustered and cluster `c1` data, though it was higher for cluster `c2`. The same could be said of `frustrated` as well, whose performance was near 0.00 for cluster `c1`.

We noticed that cluster `c1` performed similarly to the unclustered raters overall, which was understandable given that cluster `c1` effectively contained most of the raters. All label sets performed better on the positive emotions than they did on the negative emotions. Cluster `c2`, though, was much more robust with respect to valency. Average performance on negative emotions (0.50) was only slightly less than on positive emotions (0.53). Contrast this with cluster `c1` where the observed mean F-measure for negative emotions (0.23) was far below that of the positive emotions (0.67).

		Rater Group		
	Emotion	unclustered	c1	c2
Negative	sad	0.00	0.00	0.53
	angry	0.40	0.39	0.34
	anxious	0.00	0.00	0.61
	bored	0.27	0.68	0.53
	frustrated	0.00	0.07	0.50
negative means		0.13	0.23	0.50
Positive	confident	0.55	0.92	0.94
	encouraging	0.68	0.48	0.45
	friendly	0.34	0.52	0.39
	happy	0.17	0.45	0.30
	interested	0.69	0.95	0.75
positive means		0.49	0.66	0.57
overall means		0.31	0.45	0.53

Table 5.11: Classification performance (κ_{qw}) of different perceived label sets.

All in all, given the results of our machine learning experiments, we conclude that the internal consistency created by automatic rater clustering produced perceptual emotion label sets that were easier to learn. Generalized across all experiments, we observed that classification performance for the clustered data outperformed the unclustered data. Furthermore, cluster c2 performed the best overall and was the most consistent across all emotions.

5.9 Emotion profiling

In this section we describe how we profiled the intended emotions of the CU_EPSAT corpus according to their acoustic-prosodic cues using the majority rating given by the raters in each cluster. We took the same approach that we took for the intended emotions of the EPSAT corpus, as described in Section 4.2. In other words, mean z-score feature values

were computed for each emotion and unpaired t-tests were performed to isolate significant natural classes of emotions. Since our data were considerably fewer in the CU_EPSAT corpus than in the EPSAT corpus, we lowered the level of significance from $p < 0.01$ to $p < 0.05$ for these t-tests.

The analytic results presented in this section were meant to compare emotion profiles given different labeling schemes addressed our initial hypothesis that:

Hypothesis 4.1 (EmoAcoust) *Intended and perceived emotions are characterized by different acoustic-prosodic cues.*

Having found that there were different clusters, we also proposed a new but related hypothesis:

Hypothesis 5.9 (PerceivedEmoAcoust) *Perceptual emotion labels assigned by different clusters of raters are cued by different acoustic-prosodic cues.*

We refer the reader to Appendix B on page 197 for exhaustive results, including all feature means and quantized feature profiles per emotion, per label set. As an overall analysis, we list here only the minimally distinctive quantized feature values corresponding to the natural classes that resulted from significant t-tests. In each experiment, the majority rating of each utterance given each set of raters was considered to be the emotional label. If most ratings were 0 then no emotion was considered to be perceived; otherwise the emotion in question was considered to be present.

Table 5.12 displays the minimally distinctive quantized feature values when no clustering was performed. We observed that `f0-min` and `db-range` were redundant for `bored`, `sad`, `frustrated`, and `angry`. Both of these features are known to distinguish emotions in terms of activation. Emotions with low activation tend to have low values for these features, and as activation increases, so too does minimum pitch and intensity range.

Quite striking was the observation that `f0-rising`—the percentage of overall pitch rise—separated emotions in terms of valency. Emotions with negative valency had a lower percentage of rising intonation than emotions with positive valency did. Specifically, `angry`, `bored`, `sad`, `frustrated`, and `anxious` each had a value of `L`, while both `happy` and

	bored	sad	frustrated	anxious	angry	interested	confident	happy	encouraging	friendly
f0-min	L	L	M	H	H	H	H	H	H	H
db-range	L	L	M	M	H	M	M	M	M	M
f0-rising	L	L	L	L	L	M	M	H	H	H
f0-curve	M	H	M	M	M	M	L	L	L	M

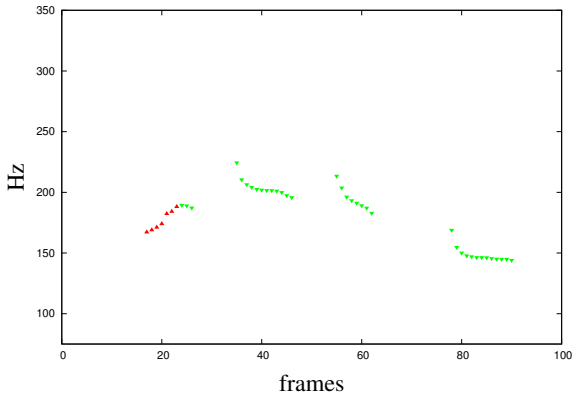
Table 5.12: Perceived emotion profiles (unclustered) using minimally distinctive feature set.

`encouraging` (which could not be distinguished given our feature set) had a value of `H`. Emotions that lay in the mid range (`M`) were `interested` and `confident`. Though a perfect dichotomy was not observed, we saw that no emotions that are commonly considered to be negative were assigned to the same natural class as any that are commonly considered to be positive, given the `f0-rising` quantization. This is something that could not be said for any other feature examined.

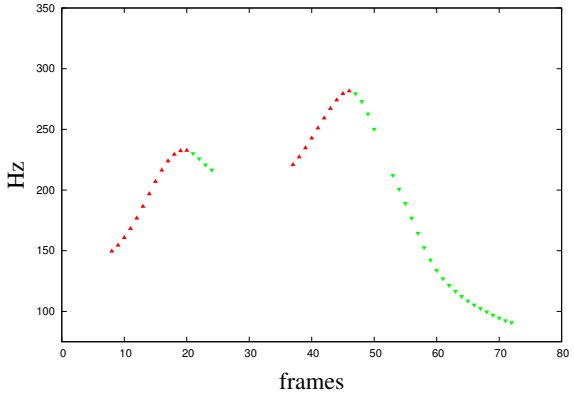
It was illuminating to examine exactly how `L` and `H` `f0-rising` values were manifested in the data. The percentages of rising pitch in the CU_EPSAT corpus ranged from 16.2% to 65.5%, had a mean of 38.2%, and a standard deviation of 12.5%. This indicated that most utterances tended to have more falling than rising pitch, on average. In fact, a z-score value of 0 indicated that roughly a third (38.2%) of the pitch in an utterance was falling, a value of -1 indicated that about a quarter (25.7%) was falling, and a value of 1 meant that half (50.7%) was falling. Figure 5.7 on page 61 shows how some quantized `f0-rising` values were manifested. As in (a), an `f0-rising` value of `L` showed little-to-no rising intonation; only sequences of pitch falls were present. We can claim that this was the general pattern for negative emotions. This was contrasted this with (b), in which such falls were present but they were preceded by rising pitch before each fall. In (c) we notice another pattern of `H`-valued `f0-rising`: a phrase-final rise. Both (b) and (c) were indicative of positive emotions.

The `f0-curve` feature is a less intuitive feature to interpret, though it is important to do so, since it was found to be a minimally distinctive feature for perceived emotion. The raw

(a) L



(b) H



(c) H

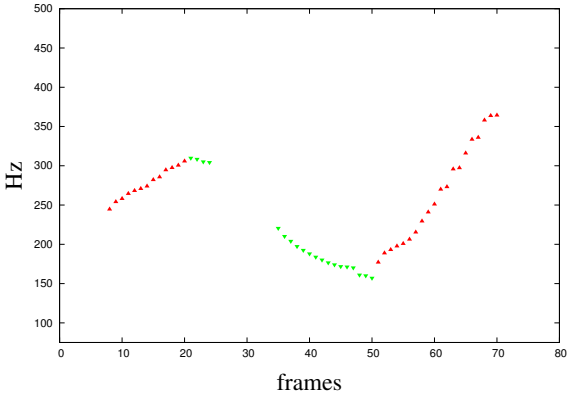


Figure 5.7: Examples of f0-rising quantizations.

feature values were calculated by running polynomial linear regression over the pitch slope and taking the coefficient of the polynomial term to describe the parabola of the fitted curve. When the value was near zero the curve was flat; the farther away from zero the coefficient became, the steeper the parabola. It was convex when negative and concave when positive. In our data, the raw values were symmetrical; we observed approximately as many negative values as positive ones. Therefore, z-score conversion was straightforward: 0 indicated no curve, negative values indicated convex curves, and positive values indicated concave curves. Through manual analysis, we noticed that large negative values were indicative of utterances with utterance-final falling intonation and large pitch excursion, something we might expect with emotions of high activation, which in fact we did observe considering that `happy`, `encouraging`, and `confident` all had `L` values for `f0-curve`. However, we would also expect `angry` to have such a value as well, which we did not observe.

What was most peculiar was the `H` value of the `f0-curve` for `sad`. It is critical to understand why this would be the case considering that `f0-curve` minimally differentiated `sad` and `bored`. It seemed counter intuitive, based on what we know about the acoustic properties of emotion, to expect `sad` to have a large pitch range and phrase-final rising intonation, as is suggested by the quantized feature value. Through manual analysis, we observed that this was due to irregular voicing that caused very high pitch readings at the beginning and end of the `sad` utterances. Such readings were actually pitch tracking errors—errors that `f0-curve` was particularly susceptible to—and they occurred in all the data on occasion. Sporadic mistracks lie well outside of average $f0$ behavior and don't generally affect global features based on it, such as mean pitch (`f0-mean`). For `sad` utterances, though, we observed that such mistracks occurred more often than chance would allow. It is an unintended yet beneficial consequence that `f0-curve` was able to identify `sad` in this way. It would be more methodologically sound to have features that measured voice quality directly, but these are notoriously hard to calculate without hand segmentation (ken,). As a side note, `f0-slope` was not as affected by pitch tracking errors because of the fact that for `sad` they tended to occur at both the start and end of utterances, thus flattening the pitch slope, which was not problematic because this is what we observed for `sad` even without the presence of irregular voicing.

Let us now examine the two clusters of raters we found and discussed previously in Section 5.5. Table 5.13 lists the distinctive features for cluster `c1` of the CU_EPSAT corpus raters. Using the emotion labels for this cluster, we were not able to uniquely identify as many emotions as we were when we used labels generated from all the raters in the entire perception experiment. No distinctions were found between `frustrated`/`anxious`, `interested`/`confident`, or `happy`/`encouraging`. However, each of these emotion pairs are to be quite similar in affect so this lead us to believe that the distinctive features we did find were generally sound. We observed a similar emergent pattern in this distinctive feature set as we did in the previous set; namely, that `db-range` discriminated emotions in terms of activation: `angry` had the highest intensity range, whereas `bored` had the lowest. In the case of cluster `c1`, though, we saw that `f0-curve` was not used to distinguish `sad` and `bored`, instead each had different intensity ranges (`bored` utterances had a smaller intensity range than did `sad` utterances). Also as before, `f0-rising` split the data unequivocally along the valency dimension: all negative emotions were either `L` or `ML` (`bored` had more rising intonation than did `sad`) and all positive emotions were either `MH` (`interested`, `confident`) or `H` (`happy`, `encouraging`, `friendly`).

The feature measuring the slope of the linear regression line over the pitch—`f0-slope`—emerged as a third critical feature with respect to the raters of cluster `c1`. We will describe it here to illuminate the exact role that it played. On average, utterances had negative overall pitch slopes: a z-score value of 0 indicated that an utterance had a pitch slope of -18.98 Hz. Therefore, positive z-score values did not actually guarantee that an utterance had a rising

	bored	sad	frustrated	anxious	angry	interested	confident	happy	encouraging	friendly
db-range	L	ML	MH	MH	H	MH	MH	MH	MH	MH
f0-rising	ML	L	L	L	L	MH	MH	H	H	H
f0-slope	H	H	H	H	H	H	H	L	L	H

Table 5.13: Perceived emotion profiles for cluster `c1` using the minimally distinctive feature set.

pitch slope. Instead, it usually indicated a flat or near-flat pitch slope. Negative z-score values were indicative of steeply falling pitch slopes. We noticed also that the range of quantized `f0-slope` values were somewhat restricted. The labels `happy` and `encouraging` each had `L` values, indicating the presence of an utterance-final pitch fall and high non-utterance-final pitch values that caused the pitch slope to be steep. All other emotions were found to have `H` values for pitch slope; in other words, the regression slope was near flat. Note that this could have resulted from two very different pitch contours. A pitch contour that was generally flat overall would have such a regression line, but one with equal amounts of rising and falling intonation at the beginning and end of the utterance would also result in a flat slope. Thus, on its own, `f0-slope` only painted a partial picture of the role that overall pitch contour played in cueing the perception of emotions. When considered in conjunction with `f0-rising`, though, a sharper picture was rendered.

By abstracting `f0-rising` such that it could take two values—`L` and `H`—then there were four possible `f0-rising/f0-slope` patterns that could be attributed to each utterance: `L/L`, `L/H`, `H/L`, and `H/H`. The emotions observed to have these combinations were the following:

`L/L` : none.

`L/H` : `bored`, `sad`, `frustrated`, `anxious`, `angry`.

`H/L` : `happy`, `encouraging`.

`H/H` : `friendly`, `interested`, `confident`.

The `L/L` combination was not found to be distinctive for any emotion, though we observed that all the negative emotions were `L/H`. In other words, pitch slope did not provide any additional information in discriminating among the negative emotions; a low amount of rising intonation was sufficient. However, the positive emotions displayed two very distinct pitch contours. Though both had a high percentage of rising intonation (`f0-rising` = `H`), some had a steeply falling pitch contour (`f0-slope` = `L`) whereas others had a relatively flat one (`f0-slope` = `H`). Figure 5.8 shows typical pitch tracks for each of the `f0-rising/f0-slope` quantization patterns in the CU_EPSAT corpus. What we noticed was that `L/H` contours had a wide pitch range with long upward sloping pitch to the pitch peaks and a steep fall to the end of the phrase. This contrasted with `H/H` contours that differed only by the fact that

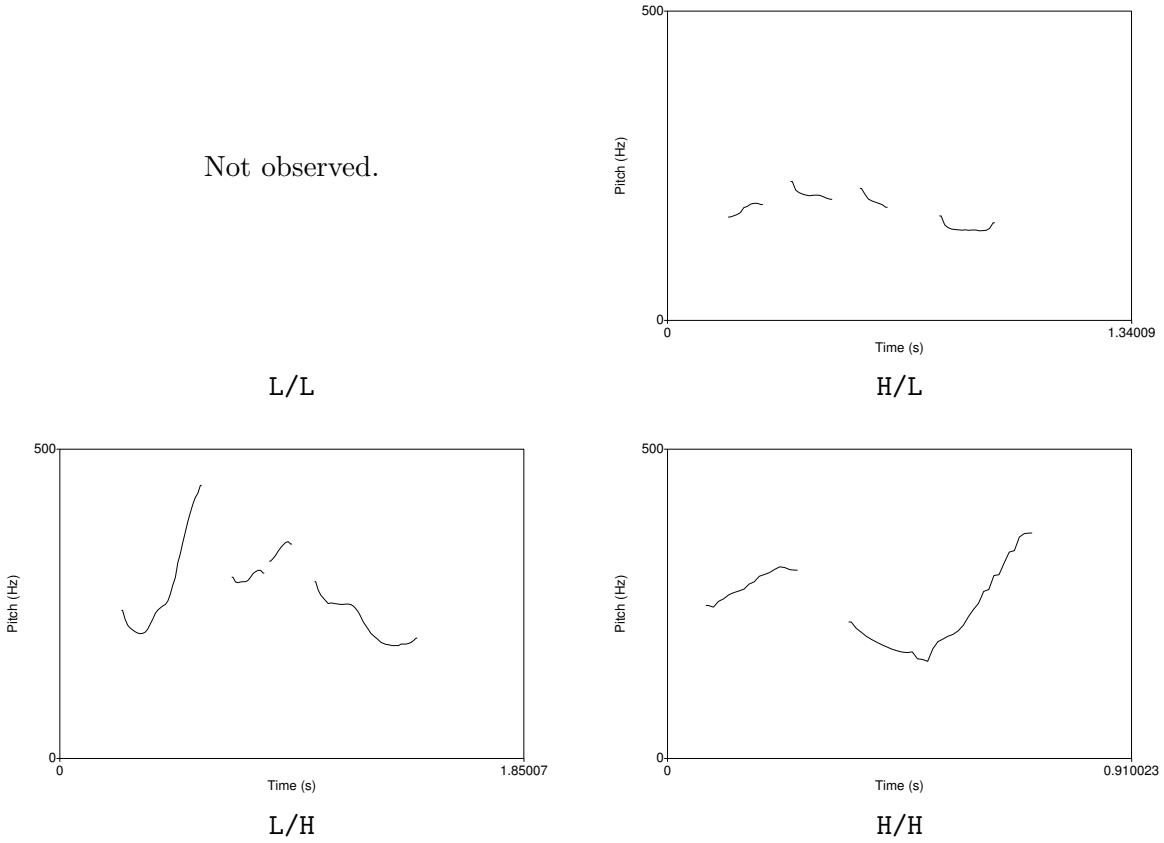


Figure 5.8: Examples of `f0-slope/f0-rising` quantizations.

they had rising instead of falling intonation at the end utterances. This final rise tended to flattened the overall pitch slope. The `H/L` contour also showed a flat pitch slope but for a different reason: the pitch range was quite narrow. It is curious that we did not observe any `L/L` patterns that we might expect of high activation, low valency emotions such as `angry`. We expected, given what we had observed up to this point, that `L/L` contours would look quite similar to `L/H` contours except that they would lack the rising intonation that occurred before each fall.

Analyzing the combination of pitch slope features in this way provided insight into the role of pitch contour in discriminating among discrete emotions and suggested that categorizing pitch contour explicitly might be of use. In fact, each of the contours shown in Figure 5.8 have distinct tone label sequences under the ToBI intonation annotation framework. Roughly speaking `H/L` would be transcribed as `/H* L-L%/`, `L/H` as `/L*+H (!H*)`

	sad	bored	frustrated	anxious	angry	interested	confident	friendly	happy	encouraging
db-range	L	L	M	M	M	M	M	M	M	H
f0-rising	L	ML	ML	ML	ML	MH	MH	MH	H	H

Table 5.14: Perceived emotion profiles for cluster c2 using minimally distinctive feature set.

L-L%/, and H/H as /(H*) L* H-H%/. In Chapter 7 we report on our assessment of the role of categorical intonation units in emotion discrimination.

So that we might compare the aforementioned feature space partitioning with the raters of cluster c2, we conducted the same analyses using the emotion label set provided by that cluster of raters. Table 5.14 lists the derived distinctive feature set along with their values per emotion. In this case, we were less able to uniquely identify each emotion. To the extent that we *could* discriminate, only two features were required: db-range and f0-rising. Once again, f0-rising was found to discriminate on the basis of valency: sad had an f0-rising value of L; bored, frustrated, anxious, and angry all had a value of ML; interested, confident, and friendly were MH; and happy and encouraging were H. Intensity range (db-range) showed the same pattern for sad and bored utterances that we saw in the two previous perceptual label sets (not clustering and cluster c1), but perceived anger (angry) was actually found to be in the mid-level for intensity range, along with all other emotions except sad, bored, and encouraging. Perceived encouragement (encouraging) exhibited the highest intensity range. This difference in the acoustic cues for encouraging and angry emotion suggested that the raters did not attune to the same acoustic cues.

For a clearer picture of the differences between the raters in each cluster, we present an abstract view of the quantized feature values for negative and positive emotions in Table 5.15. What we noticed with cluster c1 is that the two distinctive features operated on the two dimensions of emotionality. Activation was signaled with db-range while valency was signaled with f0-rising. Low percentages of rising intonation were indicative of negative emotions and their ranges of intensity further discriminated among negative emotions. Though positive emotions all lay within the mid range for db-range, they were further dif-

Cluster 1		
	negative	**positive**
db-range	L →H	M
f0-rising	L	M →H

Cluster 2		
	negative	**positive**
db-range	L → M	M →H
f0-rising	L	M →H

Table 5.15: Generalized quantized feature values for perceived emotions for clusters `c1` and `c2`.

ferentiated by the amount of rising intonation, ranging from mid to low. Thus, we observed both dimensions at work orthogonally given the emotion labels of cluster `c1`. This is not the observed pattern for the perception of emotion among the raters of cluster `c2`, though. In fact, both features, though discriminative, operated in cadence. Positive emotions were all found to have more rising intonation and a larger intensity range than negative emotions did.

Table 5.16 on page 68 shows the quantized feature values for all distinctive features using each of our four emotion label sets. We have listed only those emotions that occurred in both the intended and perceived emotion sets. Some prominent similarities across all experiments stand out. First, `f0-mean`, `f0-min`, `f0-rising`, `db-min`, `db-max`, and `db-range` were found to be distinctive in all cases. Furthermore, `bored` was the most consistent across studies. It was always found to have the lowest `f0-mean`, `f0-min`, `db-max`, and `db-range` values; and the highest `db-min` value. The major difference was between the intended and perceived emotion sets. Whereas all of our perception studies found that raters perceived `sad` utterances to have a relatively low percentage of rising pitch, our study of intended emotion found `sad` to have a high percentage of rising pitch. In fact, with the intended emotions, `f0-rising` was only marginally distinctive—used only to identify neutral—whereas it was a highly discriminative feature with the perceived emotions. This was the largest discrepancy among the results. While features that measured global acoustic features were relatively reliable across studies, the features that described the pitch contour were found to be important only in the perceptual cases. This is an important distinction

		f0-mean	f0-min	f0-max	f0-range	f0-voiced	f0-slope	f0-rising	f0-curve	db-mean	db-min	db-max	db-range
happy	EPSAT	MH	ML	-	H	-	-	H	-	-	H	L	M
	CU all	H	H	-	-	-	-	H	L	-	M	M	M
	CU c1	H	M	-	-	-	L	H	L	-	H	MH	MH
	CU c2	MH	H	-	-	L	L	H	H	-	H	H	M
interested	EPSAT	ML	L	-	H	-	-	H	-	-	H	L	ML
	CU all	H	H	-	-	-	-	M	M	-	M	M	M
	CU c1	H	M	-	-	-	H	MH	L	-	H	MH	MH
	CU c2	MH	H	-	-	L	H	MH	H	-	H	H	M
sad	EPSAT	ML	L	-	H	-	-	H	-	-	H	L	L
	CU all	M	L	-	-	-	-	L	H	-	H	L	L
	CU c1	L	L	-	-	-	H	L	H	-	H	ML	ML
	CU c2	ML	M	-	-	L	H	L	H	-	H	M	L
bored	EPSAT	L	L	-	H	-	-	H	-	-	H	L	L
	CU all	L	L	-	-	-	-	L	H	-	H	L	L
	CU c1	L	L	-	-	-	H	ML	M	-	H	L	L
	CU c2	L	L	-	-	L	H	ML	H	-	H	L	L
angry	EPSAT	H	MH	-	H	-	-	H	-	-	L	H	H
	CU all	H	H	-	-	-	-	L	M	-	L	H	H
	CU c1	H	H	-	-	-	H	L	L	-	L	H	H
	CU c2	MH	H	-	-	L	H	ML	H	-	H	H	M
anxious	EPSAT	ML	L	-	H	-	-	H	-	-	H	L	L
	CU all	H	H	-	-	-	-	L	M	-	M	M	M
	CU c1	H	M	-	-	-	H	L	M	-	H	MH	MH
	CU c2	MH	H	-	-	L	H	ML	H	-	H	H	M

Table 5.16: A comparison of the quantized feature values for emotions that are labeled under all four experiment designs.

because the latter features were found to discriminate on the valency dimension, whereas the former discriminated on the activation dimension. To address our early hypothesis—Hypothesis 4.1—we conclude that there were indeed different cues to emotion depending on whether the emotions were labeled based on actor intention or listener perception. In particular, listeners seemed to attune more to pitch contour of the utterances than did the actors.

The main reason for this might be that people find it easiest to control the global acoustic properties of speech rather than the pitch contour. Recall that each of the actors in the EPSAT corpus were asked to produce the utterances several times. Across these utterances it is very likely that they refined the pitch contour until they felt that the correct emotion was conveyed, while they were able to be more consistent with global trends such as intensity range. This would lead to the lower significant differences between the pitch contour features and high significant difference between the global ones. We must acknowledge another factor as well. The data in the EPSAT corpus were much more numerous than in the CU_EPSAT corpus. So the differences we found for the intended emotions could be considered more robust than our findings for the perceived emotions. Though we lowered our threshold for significance in the perception studies, we feel that the logical findings of the perceived studies indicated that these were not simply due to chance. However, the CU_EPSAT corpus was also different from the EPSAT corpus in that utterances were preselected to be those that were considered to be particularly well-conveyed whereas in the EPSAT corpus all the data were included, even those that may not have conveyed the emotions well (even the actors might agree on this). In sum, we feel that the well-motivated creation of the CU_EPSAT corpus makes up for its smaller size. Our findings differed depending on whether intended and perceived emotion labels were considered, an indication of the care that must be taken when labeling one's data. We feel that perceived emotions are most applicable to Artificially Intelligent applications and therefore we place the most importance on the findings we found in those studies.

5.10 Discussion

In this chapter we have explored in depth many aspects of perceived emotion. We described a survey we conducted to elicit perceptual emotion ratings of the CU_EPSAT corpus and the analyses we did to characterize, not only the acoustic cues of emotions but also how our raters systematically judged each emotion. Our experiments were conducted to address several specific hypotheses we had, and we will briefly revisit each of these now.

We found that the acoustic cues of perceived and intended emotions differed (Hypothesis 4.1) and that emotions were highly interrelated (Hypothesis 4.2). We claimed that our monothetic label set comprising intended emotions, by not reflecting such correlation, suffered theoretically because it failed to represent the true nature of emotions. This was confirmed by the fact that classification performance using statistical machine learning was much lower for the monothetic, intended label set than it was for the polythetic, perceived label set, confirming Hypothesis 4.3. This was an important finding with relevance for applications designed to predict spoken emotion, such as Spoken Dialog Systems.

We were able to identify two coherent groups of listeners (Hypothesis 5.1) based on their rating behavior on a substantial subset of utterances (Hypothesis 5.5). The two groups differed with respect to the strength of perceived valency (Hypotheses 5.4 and 5.7). Though there was a clear dominant behavior representing the "standard" listener, we found that a smaller group of listeners systematically differed from the dominant group by rating positive emotions as less positive and negative emotions as more negative. Though the listener groups did not differ statistically with respect to age (Hypothesis 5.3) or speaker sex (Hypothesis 5.6), we did find that the smaller group was more heavily represented by males than was the dominant group (Hypothesis 5.2). In other words, we identified a "standard" listener and a "grumpy man" listener.

Just as differences were found between intended and perceived emotions in terms of acoustic cues, we also found differences between the two listener groups. The "standard" listener seemed to rely on features characterized by pitch contour to distinguish emotions along the valency dimension, whereas these cues did not seem to be used by the "grumpy man" listener (Hypothesis 5.9). Furthermore, the performance of automatic classification of emotions was more successful when the labels of each group were predicted separately

than when the labels were derived from all listeners in aggregate (Hypothesis 5.8). In other words, listener type identification was not only interesting, it was also useful.

On several occasions, we found that emotion dimensions—activation and valency—arose in our analyses. This was somewhat surprising given that we did not structure or label typology in this way. Since we were using discrete emotion labels, we could have found that several emotion label groupings behaved similarity. Instead, we found that activation and valency clearly played a role in identifying the acoustic profiles of emotions and even in rater behavior. This finding supports the idea that emotions can be readily and usefully characterized by such dimensionality, a notion put forth by many scientists. In the next chapter we explore emotion dimensionality directly.

Chapter 6

Dimensions of Perceived Emotion

Heretofore we have explored discrete emotion labels, such as `angry`, `happy`, `sad`, etc. However, it is a long-held contention that discrete emotions labels are colloquial terms used to describe the affective force of an utterance as it exists in a multidimensional space. In this chapter, we explore two of the most commonly proposed dimensions: **activation** and **valency**. The activation dimension is used to describe a state of heightened physiological activity.[1] The valency dimension describes an affective continuum ranging from negative to positive.

The motivating force behind the experiments described in this chapter was to explore whether the suggested findings from the discrete emotion experiments—the role of activation and valency—would hold, were we to explicitly model these dimensions. In particular, we formulated the following hypotheses.

Hypothesis 6.1 (DimAcoustAct) *Global acoustic information is indicative of activation.*

Hypothesis 6.2 (DimAcoustVal) *Pitch contour information is indicative of valency.*

In order to explore these hypotheses, we needed to obtain dimensional ratings for the utterances in the CU_EPSAT corpus. To this end, we designed a computer-based survey similar to the one administered for eliciting ratings of discrete emotion labels. The design of

[1] In the context of emotion, the term **activation** is often synonymous with the term **arousal**.

the survey was as follows. All subjects completed the survey on the same computer. After answering introductory questions about their language background and hearing abilities, subjects were given written instructions describing the procedure. Subjects were asked to rate each utterance—played out loud over headphones—on two scales. The first scale was designed to elicit valency ratings and subjects were asked whether an utterance sounded "negative," "neutral," or "positive." The second scale was designed to elicit activation ratings and possible choices included "calm," "neutral," or "excited."

At the onset of the experiment, subjects were presented with three practice stimuli in fixed order. Then the remaining 44 test stimuli were presented one by one in random order. For each stimulus trial, two boxes with radio-buttons appeared, as depicted in Figure 6.1. The sound file for each trial played repeatedly every two seconds until the subject selected one response for each emotion dimension. Subjects were not allowed to skip any scales or utterances.

Twenty-one (21) native speakers of Standard American English with no reported hearing impairment completed the survey: 10 female and 11 male. All raters were found to be between the ages of 18 and 37 years old: 14 (67%) were between the ages of 18 and 22, 5 (24%) were between 23 and 27 years old, and 1 (5%) was between the ages of 28 and 32.

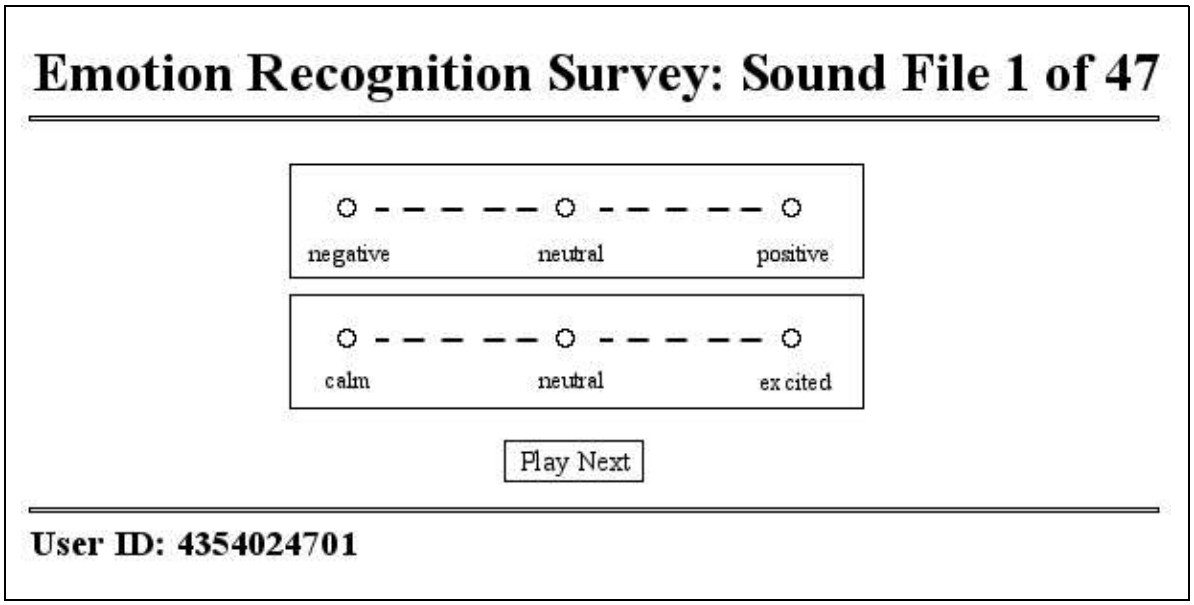


Figure 6.1: Screen-shot of the perceptual experiment for dimensional emotion.

This was a far younger demographic than we had for the discrete emotion survey. In fact, most subjects in this study were Columbia University undergraduates who answered an advertisement and who were paid ten dollars for their participation. None of the subjects who participated in this study were the same as those who participated in the discrete emotion study.

Since the dimensional ratings represented an ordinal scale, we were able to convert the queried terms into integers. The valency terms were converted in the following manner: "negative" → 1, "neutral" → 2, "positive" → 3. Activation ratings were converted in a similar fashion: "calm" → 1, "neutral" → 2, "excited" → 3. Using the majority ordinal rating for each utterance, we correlated each dimension with the acoustic features we explored in earlier experiments. The seven significant correlation coefficients ($p < 0.05$) are shown in Table 6.1.

The `f0-max`, `f0-range`, `f0-voiced`, `f0-slope`, and `db-mean` features did not show significant correlation with any dimension. Of the remaining features, only one—`f0-min`—significantly correlated with both dimensions, and in the same way. As minimum pitch increased, so too did the valency and activation ratings. In other words, negative and calm emotions had low minimum pitch; positive and excited emotions had high minimum pitch.[2] The remaining correlations were quite telling, though, and confirmed our hypotheses. On the one hand, global acoustic measurements—especially those based on intensity—correlated with activation. On the other hand, features that described the shape of the pitch

	f0-mean	f0-min	f0-rising	f0-curve	db-min	db-max	db-range
activation	0.52	0.45			-0.34	0.37	0.55
valency		0.34	0.30	-0.35			

Table 6.1: Significant correlations ($p < 0.05$) between dimensional ratings of the CU_EPSAT corpus and feature values.

[2]This might have been an artifact of the distribution of ratings in our corpus as we had no low activation positive emotions.

contour correlated with the valency dimension. An increase in the percentage of rising pitch in an utterance correlated to an increase in valency (positive emotions showed more rising pitch). As the overall pitch curve became flatter, though, the valency decreased.

We thought it prudent to also calculate the correlation coefficients between the dimensional and discrete ratings of the two sets of subjects of our perceptual surveys. In previous chapters we referred to the activation and valency of our discrete emotion sets based entirely on a general understanding of emotion. It is generally believed, for example, that a `sad` utterance has low valency and low activation, a `happy` utterance has high valency and high activation, and an `angry` utterance has low valency and high activation. By observing the correlations between dimensional and discrete labels, we sought to confirm whether this general understanding of emotion held for the CU_EPSAT corpus. Table 6.2 shows such correlations; all but two of which were found to be significant ($p < 0.05$). Most of the correlations observed were what we would expect. Discrete emotions that are commonly thought to convey negative affect were negatively correlated with the valency ratings. This was true for `angry`, `bored`, `frustrated`, and `sad`. The one exception was `anxious`, which was not found to be significantly correlated with valency ratings, though it is generally thought to

	valency	**activation**
`frustrated`	-0.70	0.33
`angry`	-0.54	0.39
`anxious`	-0.23*	0.50
`sad`	-0.58	-0.47
`bored`	-0.40	-0.77
`encouraging`	0.91	0.34
`friendly`	0.89	0.27**
`happy`	0.88	0.41
`interested`	0.76	0.69
`confident`	0.60	0.34

Table 6.2: Correlation of perceived discrete emotion and perceived dimension ratings. (*$p = 0.14$; **$p = 0.07$; all others: $p < 0.05$)

be negatively valenced. All remaining emotions, as we expected, were positively correlated with valency: `confident`, `encouraging`, `friendly`, `happy`, and `interested`. Furthermore, we found that all emotions that positively correlated with activation also positively correlated with valency as well. The emotions that negatively correlated with valency, though, were divided in relation to activation. The emotions `frustrated`, `angry`, and `anxious` each showed positive correlation with activation ratings, whereas `bored` and `sad` were negatively correlated with activation.

A graphical view of the correlation distribution is shown in Figure 6.2 on page 77. We observed that discrete emotion labels were arranged in a circle in the activation/valency space—as posited theoretically by Cowie (2000) and others—and that those emotions that were found to be the most similar in our discrete emotion survey based on correlation ({`happy, encouraging, friendly`} and {`frustrated, angry`}) were further confirmed to be so in our dimensional survey. We again note the absence of any of our discrete emotion labels showing negative correlation with both activation and valency. These findings serve mainly to corroborate our findings reported earlier for the analysis of discrete emotions but are further referenced in the next chapter when report our exploration of abstract pitch contour and emotion.

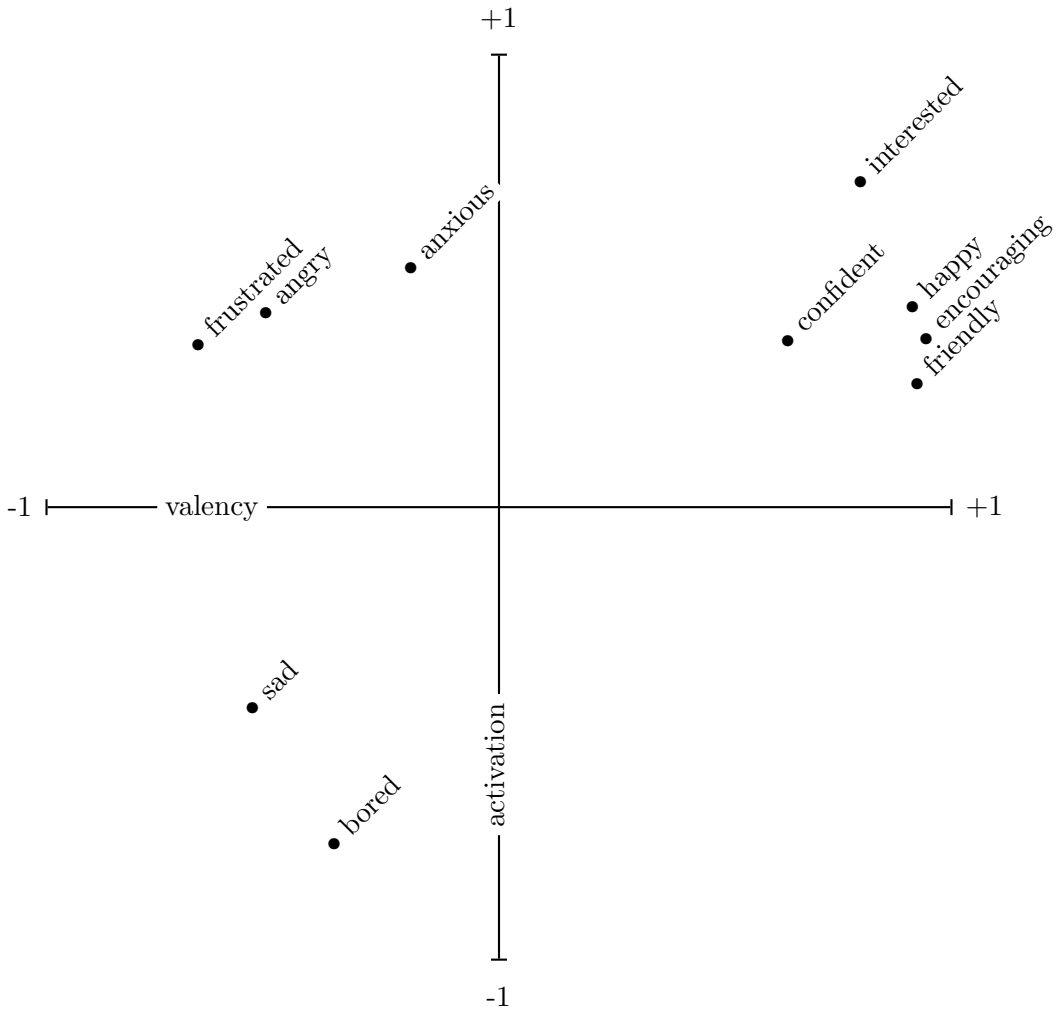


Figure 6.2: Plot of the correlation between ratings for discrete emotion labels and emotion dimensions.

Chapter 7

Abstract Pitch Contour

Heretofore our research has concerned intonational aspects associated with emotion—both intended and perceived—of acted speech. The type of information we extracted for this purpose was at a relatively low level and referred broadly to the overall shape of automatically-derived fundamental frequency measurements. Such information was shown to be quite useful for emotion profiling and yet such an approach may fail to capture other types of intonational information that could be of importance for cuing spoken emotions. In this section we address the relationship between abstract intonational units and emotions.

7.1 Annotation

We annotated all tokens in the CU_EPSAT corpus with MAE_ToBI labels. ToBI (**To**nes and **B**reak **I**ndices) is an annotation system designed to encode the underlying phonological representation of a pitch contour as a series of phonemic pitch targets taken from a limited inventory of accent types. We followed the coding practices of MAE_ToBI for the transcription of Mainstream (Standard) American English (Beckman et al., 2005).

The MAE_ToBI tone inventory is limited to two tonic phonemes, specified by height: high (`H`) and low (`L`). Height is not determined by absolute (phonetic) pitch values, but rather perceptual cues and phonological theory. Pitch accents mark the words that are considered to be perceptually prominent and can consist of a simple tone or can be a combination of two simple tones. An asterisk marks the pitch accent tone that aligns with

the stressed (prominent) syllable within the word bearing it. There are six possible pitch accent tones allowable in the ToBI framework: H*, L*, L+H*, L*+H, H+L*, and H*+L.

Tones are also associated with intonational phrasing on two levels: intermediate (moderate juncture) and intonational (strong juncture). An intermediate phrase consists of at least one pitch accent followed by a phrasal tone that extends to the end of the phrase. Phrase accents are demarcated with a dash following the tone (i.e., L- or H-). An intonational phrase may contain one or more intermediate phrases, the last of which is followed by a strong perceptual break and demarcated with a percent sign (i.e., L% or H%). Thus, an intonational phrase boundary is signified by one of the four possible pairings of phrase accents and boundary tones: L-L%, L-H%, H-L%, H-H%. Taken together, a phrase accents plus boundary tone describes what is often referred to as phrase-final intonation. L-L% is used to describe phrase-final falling intonation, a pitch contour that—usually—descends from the last pitch accent to the end of the phrase, to a point at the bottom of the speaker's pitch range. Both L-H% and H-H% are used to describe phrase-final rising intonation, though the former is often considered a **low rise** and the latter a **high rise**. The high phrase accent of H-H% causes **upstep** on the following boundary tone, so that the intonation rises to a very high value in the speaker's pitch range. Finally, H-L% refers to a phrase-final contour characterized as a level **plateau** caused by upstep from the high phrase accent of the final intermediate phrase on the low boundary tone. The pitch height of H-L% is usually in the middle of the speaker's range.

In addition to upstep, ToBI also relies on the notion of **downstep**. Downstep is the reduction of the overall pitch range within an intermediate phrase. Downstep is represented by prefixing a H* pitch accent or H- phrase accent with an exclamation point (e.g., !H* or !H-). While a downstepped high pitch accent (!H*) has a lower pitch value than does a preceding non-downstepped high pitch accent, it still has a higher pitch value than does L*, which is characterized by a pitch excursion towards the bottom of the speaker's pitch range.

Though each token in the CU_EPSAT corpus comprised only one intonational phrase, most phrases contained more than one pitch accent. In the analyses we report on in this

Pitch Accent	Phrase Accent + Boundary Tone				N (%)
	H-H%	H-L%	L-H%	L-L%	
!H*				10 (22.7%)	10 (22.7%)
H*		4 (9.1%)		11 (25.0%)	15 (34.1%)
H+!H*				3 (6.8%)	3 (6.8%)
L*	2 (4.5%)			2 (4.5%)	4 (9.1%)
L*+H				1 (2.3%)	1 (2.3%)
L+H*			1 (2.3%)	10 (22.7%)	11 (25.0%)
N (%)	2 (4.5%)	4 (9.1%)	1 (2.3%)	37 (84.1%)	44 (100%)

Table 7.1: Distribution of ToBI tones in the CU_EPSAT corpus.

chapter, we address only the final pitch accent in each phrase.[1] Table 7.1 lists the distribution of final pitch accents and phrase accents plus boundary tones in the CU_EPSAT corpus. Tone distribution was heavily skewed. The overwhelming majority (84.1%) of contours ended with L-L%. There was more uniform distribution for pitch accents. Roughly one quarter of all pitch accents were either !H* or L+H*, and H* comprised about one third of all pitch accents. Most (70.4%) of all pitch contours observed consisted of three types: /!H* L-L%/ (22.7%), /H* L-L%/ (25.0%), and /L+H* L-L%/ (22.7%).[2]

7.2 Analysis

We ran one-way analyses of variance (ANOVAs) of ToBI labels using the median ratings of each cluster of raters, as described in previous sections, for each emotion and emotion dimension. Due to the low observed frequencies of many tone labels, we ran ANOVAs only when label sets contained members with four or more observed instances. For pitch accents, this set included H*, !H*, L*, and L+H*. The pitch contours with enough observed

[1]The final pitch accent in a phrase is sometimes referred to as the **nuclear pitch accent**.

[2]For our purposes, a pitch contour was considered to be the combination of the last pitch accent, phrase accent, and boundary tone. We enclose pitch contours with slashes for readability.

instances were /H* H-L%/, /H* L-L%/, /!H* L-L%/, and /L+H* L-L%/. We did not run an ANOVA for phrase-final intonation because we observed that only H-L% and L-L% had adequate counts, and an analysis of variance requires there to be at least three labels for statistical significance testing. However, information concerning phrase-final intonation could be inferred to some degree from the ANOVA conducted with overall pitch contours. In all cases, the ANOVA results when using the median ratings of the clustered raters either did not differ from the results when using all raters or were not significant. Therefore, we have restricted our discussion to the median ratings provided by all raters only.

The results of the ANOVAs for several emotions indicated a systematic rating difference given pitch accent type. These emotions were confident ($F(2) = 5.33$, $p = 0.009$), happy ($F(2) = 3.73$, $p = 0.032$), interested ($F(2) = 11.54$, $p = 0.000$), encouraging ($F(2) = 5.39$, $p = 0.008$), friendly ($F(2) = 4.33$,$p = 0.020$), and bored ($F(2) = 4.82$, $p = 0.013$). Table 7.2 lists the mean rating for each pitch accent given each of these emotions. Note that we took the mean of all the ratings assigned to a particular pitch accent and rounded each to the nearest integer. We then converted this integer back to the original label assigned by each subject in our perceptual experiments.[3] We have presented the results here after merging high pitch accents and downstepped high pitch accents to a single class (H*). ANOVAs were conducted under each scenario (merged and non-merged) and the results did not differ, so we have excluded !H* from Table 7.2 for the sake of simplicity.

	Emotion					
Accent	confident	happy	interested	encouraging	friendly	bored
L*	not at all	not at all	a little	not at all	a little	a little
H*	somewhat	a little	a little	not at all	a little	a little
L+H*	somewhat	a little	somewhat	somewhat	somewhat	not at all

Table 7.2: Average rating per nuclear pitch accent for emotions with systematic rating differences.

[3]The labels and their corresponding ordinal values were: "not at all x" (0), "a little x" (1), "somewhat x" (2), "quite x" (3), and "extremely x" (4), where x was the emotion queried.

The results indicated that the mean emotion ratings were relatively weak and ranged from "not at all" to "somewhat."[4] Furthermore, `L*` and `H*` appeared to be quite similar; the presence of either in an utterance indicated that `encouraging` was not at all perceived and that `interested`, `friendly`, and `bored` were perceived a little. The two pitch accents differed in their indication of `confident` and `happy` utterances, though. The presence of `L*` did not convey either whereas an utterance bearing a `H*` pitch accent was perceived to be somewhat `confident` and a little `happy`. `L+H*` tended to elicit stronger ratings from subjects. Utterances were perceived to be somewhat `confident`, `interested`, `encouraging`, or `friendly`, and a little `happy`, when `L+H*` was present. This is contrasted with `bored`, which was not at all perceived in the presence of `L+H*`.

ANOVAs were run for each emotion dimension as well. The perceptual ratings of both valency ($F(2) = 5.39$, $p = 0.008$) and activation ($F(2) = 6.11$, $p = 0.005$) were found to vary systematically in relation to pitch accent. The label associated with the mean ratings of each pitch accent per dimension are shown in Table 7.3. Both `L*` and `H*` were indicative of neutral valency and mild activation. On the other hand, `L+H*` was found to be indicative of an utterance that conveyed positive valency and high activation. This is in line with the ANOVA results mentioned. All the emotions that were a little or somewhat perceived in the presence of `L+H*` were positively valenced and highly activated (e.g., `encouraging`), whereas `bored`—having negative valency and low activation—was not at all conveyed by `L+H*`. The suggestion that `L+H*` is related to emotion in terms of *both* valency and activation

	Dimension	
Accent	valency	activation
`L*`	neutral	mild
`H*`	neutral	mild
`L+H*`	positive	high

Table 7.3: Average rating per nuclear pitch accent for emotion dimensions.

[4]Recall that "somewhat" was the average degree of emotion perception in our survey.

would explain why most of the other emotions were not found to be significant. Utterances that are `angry`, `anxious`, or `frustrated` have negative valency and high activation. We would not expect the ratings for these to pattern systematically with `L+H*` because they are dimensionally ambiguous: their activation suggests that `L+H*` would be appropriate but their valency does not. Though, this would not explain why `sad` was not found to be significant.

We computed ANOVAs for pitch contours as well. As above, we considered `!H*` to be `H*` for simplicity.[5] We observed fewer significant differences than we did when examining only pitch accent. The emotions whose ratings were found to vary systematically with pitch contour were `interested` ($F(2) = 11.35$, $p = 0.000$), `encouraging` ($F(2) = 4.65$, $p = 0.017$), and `bored` ($F(2) = 8.02$, $p = 0.002$); though, both valency ($F(2) = 6.78$, $p = 0.004$) and activation ($F(2) = 3.52$, $p = 0.041$) were found to be significant as well. The mean ratings for these can be found in Table 7.4. `/L+H* L-L%/` was characteristic of an intonational phrase with a rising nuclear pitch accent and a phrase-final fall. This contour was associated with utterances that were somewhat `interested`, a little `encouraging`, and not at all `bored`. It was also neutral in terms of valency and received high ratings for activation. `/H* L-L%/` was actually quite similar, though mean activation rating was "mild" not "high" and mean `interested` rating was "a little" instead of "somewhat." `/H* H-L%/` showed almost the opposite pattern as the other two contours. Whereas it was indicative of utterances that were a little or somewhat `interested` and `encouraging`, `/H* L-L%/` was not at all indicative

	Emotion or Dimension				
Contour	`interested`	`encouraging`	`bored`	valency	activation
`/L+H* L-L%/`	somewhat	a little	not at all	neutral	high
`/H* L-L%/`	a little	a little	not at all	neutral	mild
`/H* H-L%/`	not at all	not at all	a little	negative	mild

Table 7.4: Average rating per pitch contour.

[5] In this case, a few slight differences were observed. `/H* L-L%/` was perceived as "not at all" `encouraging` (instead of "a little") and had high activation (instead of mild); `/!H* L-L%/` remained unchanged.

of these emotions. Instead it was associated—a little—with `bored` utterances. In addition, it was present in utterances with negative valency and mild activation.

These findings appeared to be the effects of phrase-final intonational characteristics. The contours with similar rating distributions both had low phrase accents and low boundary tones (`L-L%`). However, the contour that was found to differ from both of the other contours in terms of rating distribution could be uniquely identified by `H-L%`. In general, `H-L%` seemed to be associated with negative affect, whereas `L-L%` was not. It should be noted, however, that the range of emotions with systematic rating distributions was somewhat restricted and might more accurately be referred to as moods or attitudes rather than full blown emotions. We will revisit this topic in the Chapter 9.

Chapter 8

Non-acted Emotion

Though we have restricted our discussion of emotion to the prosodic cues found in acted speech, there exists a large body of research devoted to studying the cues of emotion in non-acted speech as well (e.g., Huber et al., 2000; Ang et al., 2002; Walker et al., 2002; Batliner et al., 2003; Douglas-Cowie et al., 2003; Devillers & Vidrascu, 2004; Litman & Forbes-Riley, 2004; Lee & Narayanan, 2005; Liscombe et al., 2005a; Liscombe et al., 2005b, inter alia). These studies have primarily focused on the speech of human users of **Spoken Dialog Systems** (SDSs) and have been been geared towards the development of automatic emotion recognition for use in such systems. In this chapter, we report on experiments we conducted to explore the cues of emotions in an SDS domain.

One striking difference between acted and non-acted speech corpora relates to the emotion labels analyzed. In acted corpora, emotions tend to be centered around the canonical **Big Six**: happiness, sadness, fear, surprise, anger, and disgust (Ekman, 1999). However, researchers have found that most of these emotions rarely occur in human-computer interaction, with the notable exception of anger. For these reasons, most studies have focused on the automatic classification of angry or frustrated speech, or simply conflating all negative emotions into one class. These emotions are then defined in relation to non-angry, non-frustrated, or non-negative speech. Another difference between acted and non-acted speech is that the latter exists within the context of a dialog, whereas the former usually does not. Prosodic correlates of emotion in acted domains have been well researched (see Chapter 1 and Table 1.1 on page 8) and the results are largely consistent with what one finds

in the non-acted domain. In this chapter we explore some non-prosodic cues to emotion in a corpus of human-computer dialogs.

8.1 The HMIHY Corpus

"How May I Help You[SM]" (HMIHY), AT&T's natural language human-computer spoken dialog system, enables callers to interact verbally with an automated agent. A caller can ask for an account balance, help with AT&T rates and calling plans, explanation of certain bill charges, or identification of unrecognized numbers on a bill. The task of the automated agent is to understand these requests and to either satisfy the request or, if unable to do so, route callers to the correct department for further trouble-shooting. If the system needs to confirm or clarify a customer's response, the dialog manager asks for more information; if it is still not clear, it routes the caller to a service representative. Speech data from the deployed "How May I Help You[SM]" system has been assembled into a corpus referred to as **HMIHY 0300** (Gorin et al., 1997). Figure 8.1 presents a transcription of an example dialog from the corpus.

In a study by Shafran et al. (2003), 5,147 caller turns sampled from 1,854 HMIHY 0300 calls were annotated with one of seven emotional states: `positive/neutral`, `somewhat frustrated`, `very frustrated`, `somewhat angry`, `very angry`, `somewhat other negative`, `very other negative`. Cohen's κ was calculated on a subset of the data consisting of 627 caller turns to measure the label agreement of two independent human labelers. A score

System:	How may I help you?
Caller:	I need to find out about a number that I don't recognize.
System:	Would you like to look up a number you don't recognize on your bill?
Caller:	Yes I would.
System:	Are you calling from your home phone?
Caller:	Yes I am.
System:	...

Figure 8.1: Sample dialog from the HMIHY 0300 Corpus.

of 0.32 was reported using the full emotion label set, whereas a score of 0.42 was observed when the classes were collapsed to ¬`negative` versus `negative`.

We were primarily interested in studying caller behavior over entire calls; thus, we increased the size of the corpus to 5,690 complete dialogs that collectively contained 20,013 caller turns. Each new caller turn turn was labeled with one of the emotion labels mentioned above. However, this resulted in a highly non-uniform distribution of the emotion labels (73.1% were *positive/neutral*), so we adopted the binary classification scheme instead (¬`negative`, `negative`).

8.2 Extraction of Emotional Cues

Each caller turn was defined by a set of 80 features that were either automatically derived or annotated by hand. The features were grouped into the following four coherent feature sets: prosodic features (PROS), lexical features (LEX), dialog acts (DA), and contextual features (CONTEXT).

8.2.1 Prosodic Features

The first set of features—PROS—included 17 features very similar to those described in the preceding chapters. Due to the relative brevity of speech collected for each caller (an average HMIHY dialog consisted of 3.5 caller turns), we z-score normalized all prosodic feature values by gender instead of by speaker.

The following 10 features were automatically extracted over the entire caller turn using PRAAT: overall energy minimum, maximum, median, and standard deviation, to approximate loudness information; overall fundamental frequency ($f0$) minimum, maximum, median, standard deviation, and mean absolute slope, to approximate pitch contour; and the ratio of voiced frames to total frames, to approximate speaking rate.

The remaining seven prosodic features were semi-automatically extracted. Phones and silence were identified via forced alignment with manual transcriptions of caller turns using a special application of AT&T WATSON, a real-time speech recognizer (Goffin et al., 2005). These features included: $f0$ slope after the final vowel, intended to model turn-final pitch

contour; mean $f0$ and energy over the longest normalized vowel, to approximate pitch accent information; syllables per second, mean vowel length, and percent of turn-internal silence, to approximate speaking rate and hesitation; and local jitter over longest normalized vowel, as a parameter of voice quality. The normalized length of each vowel was conditioned upon durational and allophonic context found in the training corpus.

8.2.2 Lexical Features

The LEX feature set contained features based on the manual transcription of caller utterances. The features themselves were word-level unigrams, bigrams, and trigrams and were encoded in a "bag-of-ngrams" fashion. In addition to lexical items, transcriptions also contained non-speech human noise, such as laughter and sighs.

In the corpus, we noticed that certain words found in the caller transcriptions correlated with emotional state. While these correlations were slight (the highest was less than 0.2), they were very significant ($p < 0.001$). This would seem to indicate that the words people say play a part in their emotional state, although they may not be the only indicators, and there most certainly is not a one-to-one correspondence between a word and the emotional content of an utterance. Some of the more interesting correlations with `negative` caller state were domain-specific words concerning a phone bill (e.g., "dollars," "cents," "call") and those that indicated that the caller wished to be transferred to a human operator (e.g., "person," "human," "speak," "talking," "machine"). Also, the data showed that filled pauses (e.g., "oh") and non-speech human noises (e.g, a sigh) were also correlated with `negative` caller state.

8.2.3 Dialog Act Features

The DA feature set included one feature indicating the dialog act of the current caller turn. Dialog acts are considered to be the function an utterance plays within the context of a dialog and, as such, may be a cue to emotional content. There are different ways to label dialog acts and they range from generic to specific. For this study we used the pre-annotated call-types of the HMIHY 0300 corpus. These were somewhat specific, domain-dependent dialog act tags. Each caller turn was labeled with one or more call-type from a set of

65. A few examples of the most frequent call-types in the corpus were: `Yes`, when the caller confirmed a system-initiated question; `Customer_Rep`, when the caller requested to speak with a customer representative; and `Account_Balance`, when the caller requested information regarding their account balance.[1]

8.2.4 Contextual Features

The CONTEXT feature set was introduced as a way of modeling phenomena at a level that extended beyond the scope of the present caller turn. Caller turns were situated in a larger structure—a dialog–and it therefore seemed natural to use past evidence of caller activity to help inform the emotion classification of the present caller turn. Because the dialogs were relatively short, we decided to use contextual information that extended to the previous two caller turns only. This feature set contained 61 features designed to track how the features described in the aforementioned feature sets compared to those of previous turns and how their values changed over time.

Thirty four (34) features recorded the first order differentials—or rate of change—of the PROS feature set. Half of these recorded the rate of change between the current caller utterance (n) and the previous caller utterance ($n-1$). The other half recorded the rate of change between utterances n and $n-2$. An additional 17 features measured the second order differentials between each feature in the PROS feature set for the current and previous two caller turns.

An additional four features recorded the history of lexical information within the dialog. Two features encoded the (LEX) bag-of-ngrams of the previous two caller turns. Two additional features calculated the Levenshtein edit distance between the transcriptions of caller turns n and $n-1$, as well as n and $n-2$. Edit distance was used as an automatic way to represent caller repetition, a common indicator of misunderstanding on the part of the automated agent and, anecdotally, of `negative` caller state.

Four features were designed to capture dialog act history based on the DA feature set. Two features recorded the dialog acts of caller turns $n-1$ and $n-2$. Additionally, two

[1]For the complete set of dialog act labels in the HMIHY 0300 corpus we refer the reader to Gorin et al. (1997).

features were introduced to record the dialog acts of the system prompts that elicited caller turns n and $n-1$. The HMIHY 0300 system prompts are predetermined by the dialog manager and comprise the following dialog acts: `greeting`, `closing`, `acknowledgment`, `confirmation`, `specification`, `disambiguation`, `informative`, `reprompt`, `apologetic`, `help`.

The final two features of the CONTEXT feature set were the emotional state of the previous two caller turns. For this experiment, we used hand-labeled emotions rather than predicting them.

8.3 Automatic Classification

This section describes machine learning experiments designed to evaluate the usefulness of each feature set in automatically classifying the emotional content conveyed by each caller turn. We applied the machine learning program BOOSTEXTER, a boosting algorithm that forms a classification hypothesis by combining the results of several iterations of weak learner decisions (Schapire & Singer, 2000). For all experiments reported here we ran 2,000 iterations.

The corpus was divided into training and testing sets. The training set contained 15,013 caller turns (75% of the corpus) and the test set was made up of the remaining 5,000 turns. The corpus was split temporally; the caller turns in the training set occurred at dates prior to those in the testing set. In addition, no dialogs were split between training and test sets. The corpus was divided in this way in order to simulate actual system development in which training data is first collected from the field, a system is then constructed using this data, and, finally, performance is evaluated on the newly-deployed system. Table 8.1 shows the classification accuracy when different feature set combinations were used for training the classification model.

We used as a baseline the performance accuracy with a classification model that always assigned the majority class label (¬`negative`). Since 73.1% of all caller turns were ¬*negative*, this was also our baseline for comparing our other classification models. The feature set combinations showed that adding more information increased performance ac-

Feature Sets Used	**Accuracy**	**Relative Improvement over** BASELINE
BASELINE (majority class)	73.1%	0.0%
PROS	75.2%	2.9%
LEX+PROS	76.1%	4.1%
LEX+PROS+DA	77.0%	5.3%
LEX+PROS+DA+CONTEXT	79.0%	8.1%

Table 8.1: Classification accuracy and relative improvement over the baseline of caller emotional state given different feature sets.

curacy. Using prosodic information alone (PROS) was useful, as indicated by the 2.9% relative increase in performance accuracy over the baseline. When lexical information was then added (LEX+PROS), the relative increase in performance accuracy over the baseline was almost doubled (4.1%). Similar results were observed when dialog acts were added—relative performance accuracy increased to 5.3% for LEX+PROS+DA—and when contextual information was added—relative performance accuracy increased to 8.1% for LEX+PROS+DA+CONTEXT.

8.4 Discussion

In this chapter we have shown that automatic emotion classification can be extended to non-acted domains; in particular, Spoken Dialog Systems. Prosodic information was observed to be useful for emotion classification, though higher classification accuracy was observed when non-prosodic information was exploited as well. It has been claimed by some researchers (Cauldwell, 2000, e.g.,) that information signaling emotion in speech is not static, but rather is dynamically interpreted by context. We found support for this, as evidenced by the fact that introducing discoursal information—in the form of dialog acts—improved emotion recognition. This has been noted by other researchers as well (Ang et al., 2002; Batliner et al., 2003; Lee & Narayanan, 2005, e.g.,); though, we found that further contextualizing lexical, pragmatic, and prosodic information—by encoding changes in these features over time—lead to ever greater classification accuracy of `negative` caller utterances.

Chapter 9

Discussion

The first part of the thesis has dealt with paralanguage and the ways in which emotion is conveyed in speech via suprasegmental cues. In particular, we have explored the acoustic nonverbal cues of emotion in a corpus of acted speech. As there has been a large body of research devoted to this topic in the past, we expected certain findings. For example, we hypothesized that measurements of acoustic features would be useful for discriminating among emotions along the activation continuum. This we found. However, we wished to also answer a question that had not been explored as much; namely: are their groups of people who perceive emotion in systematically different ways from other groups of people? And, if so, in what way do they differ and would identifying such groups aid in the automatic classification of discrete emotions using acoustic features?

In the pursuit of these questions, we conducted two perceptual studies to elicit polythetic ratings for discrete emotions and emotion dimensions. Via automatic clustering techniques we identified two groups of raters in the discrete emotion survey whose ratings were similar to other members of their group but different from members of the other group. It has been suggested by a few past studies that people may differ in how they perceive emotion given acoustic information, though these studies have usually restricted the exploration to gender differences (e.g., Fernandez, 2004; Toivanen et al., 2005). Though our observed rater clusters did differ slightly with respect to the gender of the raters, more important seemed to be the pattern of perception itself. We identified a dominant group of raters and a smaller group who rated, systematically, positive emotions as more positive and negative

emotions as less negative than the dominant group. We feel that these results are suggestive of the relationship between personality type and emotion perception, a relationship that has been suggested before (Meyer & Shack, 1989). Though such a relationship was not directly analyzed, we believe this could be a fruitful avenue of future research.

Furthermore, we found that rater clustering was of use in statistical machine learning experiments. In fact, average F-measure increased by 60% when we predicted majority ratings of the two clusters relative to classification of the majority rating using all rates in aggregate. This finding is particularly relevant for applications with the aim of automatically predicting of user emotion, say for Automatic Spoken Dialog Systems.

Another approach we found to increase automatic classification performance was using perceptual polythetic emotion labels rather than intended monothetic emotion labels. We observed that the emotions in the former set were all significantly inter-correlated and that the polythetic labeling scheme captured this notion of non-exclusivity as indicated again by a relative increase in of 80% in average F-measure over the performance observed when the monothetic emotion label set was used. Furthermore, we noted considerable correspondence mismatch between intended and perceived emotion labels.

As was stated at the outset of this section, we found that global acoustic measures, such as mean intensity and pitch, tended to discriminate emotions in terms of activation. This is a finding well supported by past studies. However, our approach of using t-tests to segment emotions into natural classes allowed us to profile emotions by identifying a minimally distinctive feature set that, when considered in combination, maximally partitioned the emotions. In this way, we were able to discover the features with the most impact and also to quantize their gradient feature values in relation to other emotions.

Though acoustic cues of activation are well understood, the acoustic cues of valency are less so. We believe that our research has provided some insight into the role that pitch contour might play in signaling valency. In most of our experiments—though certainly more true of perceived emotion—we found that acoustic measurements of pitch slope were useful in differentiating emotions by valency, especially the percentage of rising pitch slope in an utterance. Negative emotions were found to have less rising slope overall than positive emotions did. Furthermore, by examining several pitch slope predictors, we were able to

discover that, in particular, pitch peaks with long leading slopes were correlated with positive emotions. This finding was confirmed by experiments using abstract tonal morphemes in that `L+H*` pitch accent was found to be associated with positive emotions. As far as we know, this is a novel finding and warrants future investigation, as valency has proved to be the most difficult aspect of emotion to automatically predict, despite the fact that is it arguably the most critical.

In a final set of experiments, we showed that emotion recognition could be extended to non-acted domains by augmenting acoustic information with contextual information found in real-world domains. In the subsequent parts of this thesis, we report on our exploration of two other non-acted domains and describe the role of acoustic cues in signaling other types of cognitive states.

Part III

PRAGMATICS

The term **pragmatics** is used to describe the principles of language use whereby speakers utilize language to achieve goals. Such goals may include changing the state of the physical world (e.g., the closing of a window), or changing beliefs (e.g., convincing another of an idea). As such, pragmatics necessarily describe how language is used in the interaction between (at least) two people: a speaker and a listener. One of the fundamental ways that a speaker attempts to use language to achieve goals is by asking questions of a listener. In so doing, a speaker performs the act of seeking to elicit an answer or action from the listener. In this chapter we describe the form and function of student questions in a corpus of spoken tutorial sessions. We also describe a framework for automatically detecting such information using machine learning techniques. The motivations behind our research were two-fold. Not only were we interested adding to the general understanding of the form and function of questions in spoken Standard American English, but we also wanted to improve Spoken Dialog Systems—specifically, Intelligent Tutoring Systems—so that they might be able to more effectively detect and respond to user questions.

Chapter 10

Previous Research on Questions

Arriving at the definition of a "question" is not a straightforward endeavor. As Geluykens (1988, p. 468) states, "There is no simple correlation between Interrogative form and Question status. Some questions do not have interrogative form (i.e., Queclaratives); conversely, some interrogatives do not function as questions." Echoing this sentiment, Graesser & Person (1994, p. 109) maintain that, with respect to questions, "there is not a simple mapping between the syntactic mood of an utterance and its pragmatic speech act category" and define a question to be an utterance that functions as either an inquiry or an interrogative expression. Ginzburg & Sag (2000, p. 107) define a question to be "the semantic object associated with the attitude of wondering and the speech act of questioning." Bolinger (1957, p. 4) noted long ago that "a Q[uestion] is fundamentally an attitude, which might be called a 'craving'—it is an utterance that 'craves' a verbal or other semiotic (e.g., a nod) response." The ambiguities associated with a comprehensive definition of questions arise from the many different forms and functions that questions can take. In this chapter, we describe some of these issues in detail.

10.1 The syntax of questions

It is well understood that questions in Standard American English can be identified, in part, through syntax. The forms most commonly referred to include yes-no questions,

wh-questions, tag questions, alternative questions, and particle questions.[1] The syntax of English questions often involve inversion of declarative word order with respect to the subject and auxiliary verb. An auxiliary verb is one that is used together with a main verb to add additional syntactic or semantic meaning (e.g., "to be," "to do," "to have"). A **yes-no question**, generally speaking, is formed by such inverted word order, though the actual implementation of inversion can be quite complex. Yes-no questions often elicit a restricted range of responses (e.g., "yes," "no," "I don't know"). A **wh-question** is formed by substituting a wh-word, or interrogative word, in place of the subject or object and moving it to the beginning of the sentence. (This phenomenon is sometimes referred to as **wh-fronting**.) Common English wh-words include "who," "what," "where," "when," "why," and "how." Subject/auxiliary inversion may also be present in a wh-question. A **tag question** is formed by adding a subordinate interrogative clause to the end of a declarative statement. The tag portion consists of an auxiliary verb followed by a pronoun (a simplified version of a yes-no question). An **alternative question** is a question that presents two or more possible answers and presupposes that only one is true. A **particle question** is much more restricted in terms of lexical and syntactic choice; it consists solely of a grammatical particle or function word.

It is widely known that questions can take another form as well. Some utterances, even though they do not differ in syntax from declarative statements, still function as questions and are clear to most listeners that a response on their part is in order. Such questions are referred to as **declarative questions**. Cues other than syntax are necessary for the identification of such questions and we will discuss previous investigations to such cues in Section 10.3.

10.2 The intonation of questions

Syntax is not the only sort of information associated with the signaling of questions in English. Intonation, or pitch contour, has often been cited as playing a role as well. How-

[1]Though terminology differs with respect to question form, the terminology we have adopted is that which has been most commonly used in the research literature.

ever, there is far less of a consensus on the intonational structures associated with different question forms, though it is generally believed that final rising at the end of an utterance or phrase is in some ways indicative of questions. For example, Pierrehumbert & Hirschberg (1990, p. 277) state that, "A typical interrogative contour is represented with L* H-H%."[2] However, not all question forms are considered to exhibit rising intonation; for example, Pierrehumbert & Hirschberg also state that the canonical intonational form of wh-questions is that of the phrase-falling intonation typically associated with declaratives (H* L-L%, p. 284). Bartels (1997) offers an excellent overview of the earlier descriptive works of Bolinger (1957), Schubiger (1958), and Rando (1980). Bartels contends that different sentence types have corresponding canonical, or prototypical, intonational form. On page 33 she asserts, "Syntactically declarative sentences uttered as 'statements,' or 'assertions,' are prototypically associated with a final fall...." This is the same intonation found on alternative questions (p. 84) and wh-questions (p. 169) as well. Falling intonation (H* L-L%) is contrasted with either the low rise (L* H-H%) or high rise (H* H-H%) typical of yes-no questions (p. 123).

Though the cited work has described intonational patterns as "prototypical" or "natural" or "canonical" for questions that take different forms, no author has claimed that an intonational contour exists that *always* maps to a particular question form, nor even that intonational contours exist that are reserved *exclusively* for signaling questions in general. Indeed, as Bartels made explicit, "[T]here is no direct correlation in questions between syntactic subtype and intonation, and further, that in empirical descriptions, there is little use for the notion of a uniform 'question intonation'" (p. 8). Such empirical studies include Fries (1964), who claimed that most yes-no question in a corpus of American English television panel games in fact did *not* have rising intonation and Geluykens (1988) who claimed a similar finding in a corpus of spoken British English conversations. Nevertheless, such studies have also found that rising intonation is present on the types of questions we would expect them to be—including yes-no questions—at a rate above that found on sentence forms we would not expect them on (e.g., declaratives).

[2]Most discussion of intonational form in this section uses ToBI notion, described in full in Section 7.1, even though not all the studies cited here necessarily used such notation when reported originally.

The consensus is that the pitch contour shape at the end of a phrase is indicative of some semantic or pragmatic meaning, but that this meaning is not directly tied to the question-status of an utterance. Rather, phrase-final rising intonation is thought to indicate either non-commitment (Gunlogson, 2001; Steedman, 2003), non-assertiveness (Bartels, 1997), relevance testing (Gussenhoven, 1983), uncertainty (Stenström, 1984; Šafářová, 2005), or forward reference (Pierrehumbert & Hirschberg, 1990) on the part of the speaker. Phrase-final falling intonation is thought to signal the opposite. Though each of these meanings is distinct, they are all similar to the extent that one could associate such meaning with the function of many types of questions and could explain why rising intonation is both often thought to be indicative of questions and more likely to be found on questions than non-questions. However, since questions do not always function as one of the aforementioned meanings, it would also explain why not all questions, in all contexts, exhibit rising intonation. This notion has lead to further investigation of the different types of functions that questions can serve, which we will address in Section 10.4. First, we would like to discuss declarative questions, because the issues that pertain to them are intrinsically related to syntactic and intonational form.

10.3 Declarative questions

Declarative questions are of interest precisely because they are those questions that show no difference, syntactically, from proper declarative statements. Thus, non-syntactic cues must be present in order for listeners to determine that they have the pragmatic force of a question. It has been theorized that the intonation of declarative questions might be of use and that their intonation is similar in range to what might be expected of yes-no questions (e.g., Bartels, 1997, p. 227). Though if, as we've seen, yes-no questions do not always exhibit rising intonation, then the question remains: how it is that declarative questions are conveyed? A few researchers have examined this issue directly and we will discuss their findings presently.

Geluykens (1987) is possibly the most skeptical of the phrase-final rising intonation of-

ten attributed to declarative questions in British English.[3] In fact, he stated that "[r]ising intonation is irrelevant for the recognition of a declarative utterance as a queclarative, provided pragmatic factors contribute to the utterance's question-status" (p. 492). Motivation for this claim was based on perceptual experiments in which subjects were asked to judge whether spoken utterances functioned as questions or not. In particular, he examined the role that personal pronouns played in signaling whether an utterance was "question-prone" or "statement-prone." His claim was that utterances that contain second person pronouns (e.g., "you") are question-prone (as in, "**you** feel ill") because they address the cognitive state of the listener, whereas utterances that contain first person pronouns (e.g., "I") are statement-prone (as in, "**I** feel ill"). In other words, when the content of a speaker's utterance refers to the listener, as indicated by the use of a second person pronoun, then the meaning is related to the cognitive state of the listener, which is presumably unknown to the speaker and is therefore likely being queried by the speaker. This is contrasted with an utterance that contains a first person pronoun because the content refers to the cognitive state of the speaker, which is known to the speaker, and thus not likely to serve as a question. Geluykens hand-constructed five declarative sentences of the form PRONOUN+VERB+PHRASE, as in, "I/you like the apples a lot." These sentences were then recorded by one speaker using phrase-final falling intonation. Analogous sentences with phrase-final rising intonation were created through resynthesis. Irrespective of intonation, he found that statements containing "you" were more often perceived as questions than were statements containing "I," thus justifying his hypothesis that lexico-pragmatic cues are more crucial for the identification of declarative questions than is intonation.

Similarly for British English, Stenström (1984) found only 23% of declarative questions exhibited phrase-final rising intonation in a corpus of transcribed conversations, yet 77% were found to contain q-markers. Q-markers were defined as lexical phrases indicative of questioning, including tentative expressions (e.g., "I think"), modal verbs (e.g., "might"), and particles (e.g., "so").

Beun (1990) found lexico-pragmatic cues to be important for declarative questions in Dutch, though he considered his conclusion to be far less severe than Geluyken's: "[A]t

[3]Geluykens referred to declarative questions as "queclaratives."

least for Dutch spoken utterances, rising intonation is an important cue for questioning and cannot easily be overrruled by pragmatic cues" (p. 52). Beun examined a corpus of Dutch telephone dialogs and found 20% of all questions to be declarative in form and that 48% of all such questions exhibited phrase-final rising intonation. However, he also found that the presence of phrase-initial particles (e.g., *en* "and," *dus* "so," and *oh* "oh")—which can be thought of a discourse cues—increased dramatically the likelihood that a sentence of declarative form functioned as a question.

Another major study concerning cues to declarative questions was conducted by Šafářová & Swerts (2004) for spoken American English that built on the research of the studies previously mentioned. They identified 93 declarative questions in a corpus of spontaneously spoken dialogs and asked human judges to determine whether each utterance served to elicit a response from the listener, under two conditions: (1) with transcripts and no audio and (2) with transcription and audio. They coded the lexico-pragmatic cues put forth by both Geluykens (1987) and Beun (1990) and found that when no audio was present judges relied most heavily on whether a second person pronoun was present or not. Phrase-initial particles and first person pronouns were found to be less useful. For the second study, they transcribed the intonation of the declarative questions and coded them as having question-intonation as suggested under four paradigms: (1) L* H-H% (Pierrehumbert & Hirschberg, 1990); (2) L* H-H%, L* H-L%, H* H-H%, or H* H-L% (Bartels, 1997); (3) H* H-H%, L* H-H%, L* H-L%, or L* L-H% (Gunlogson, 2001); (4) H% (Steedman, 2003). They found that coding question-intonation as suggested by Gunlogson correlated best with a subject's ability to classify declarative questions, though they observed significant correlation using all intonation coding schemes. They also found that the the ability of the subjects to classify declarative questions did not significantly differ between the two studies, indicating that access to prosodic information, at least of the utterance itself, is redundant to some extent with other factors, such as lexico-pragmatic cues.

A comparative analysis of previous studies of declarative questions reveals that contextual information in the form of lexico-pragmatic cues is indeed important. However, most studies also found that rising intonation is present in such types of questions as well. It is reasonable to assume that the findings for declarative questions extend to questions that

take different syntactic forms as well.

10.4 The function of questions

Identifying the pragmatic function of a question is a less straightforward endeavor than classifying its syntactic form. Often, form and function are conflated. For example yes-no questions are defined, in part, because they anticipate either "yes" or "no" (or "I don't know") as an answer. This differs from the expected response of wh-questions that seek information that corresponds to the wh-word in the question. An alternative question, as we've already stated, expects the answer to be a choice among options explicitly specified in the question.

However, it has long been noted that the anticipated answer to a question is not fully governed by the syntax of the question. As just one example, Hirschberg (1984) pointed out that there are other ways to felicitously answer a yes-no question than by responding with "yes," "no," or "I don't know." Specifically, yes-no questions can seek to elicit indirect responses. For example, the expected response to the indirect question "Do you have the time?" is not "yes" or "no" but rather the time (e.g., "5:30"). In fact, any question can function as an indirect speech act, and thus one cannot rely on form to signify function.

Of the seminal studies referenced so far addressing question form, most, if not all, have alluded to some of the functions that questions of a particular form might serve. However, one of the first comprehensive descriptions of question function and how it relates to question form was put forth by Stenström (1984). She proposed eight distinct question functions: `Q:acknowledge`, `Q:confirm`, `Q:clarify`, `Q:repeat`, `Q:identify`, `Q:polar`, `Q:action`, `Q:offer`, `Q:permit`, and `Q:react`. A `Q:acknowledge` question takes the form of a tag question (or, possibly, a declarative question) and invites the listener to accept what has been suggested by the speaker. A `Q:confirm` question asks for confirmation of what was proposed in the question and can be realized also as a tag or declarative question. A question that invites the listener to clarify something uttered previously in the discourse is a `Q:clarify` question and is realized only as a wh-question. A `Q:repeat` question is similar but more restricted, in that it functions as a request for the listener to repeat

part or all of the previous utterance. `Q:repeat` questions are realized as particle questions (e.g., "Pardon?"). Both `Q:identify` and `Q:polar` seek information. `Q:polar` questions seek information in the form of a yes-no answer and are thus realized as yes-no questions or alternative questions. `Q:identify` questions are realized as wh-questions or alternative questions and seek information in the form of an adequate substitute for the wh-word or alternative choices put forth by the question. `Q:action`, `Q:offer`, and `Q:permit` are indirect versions of `Q:identify`, `Q:polar`, and `Q:confirm`, respectively, and therefore expect the same forms that their direct counterparts have. The final question function, `Q:react`, is realized only as a tag question and is used to show the speaker's surprise in relation to the listener's actions.

Tsui (1992) agreed with many of the functional categories put forth by Stenström but divorced form from function entirely, claiming that a particular function could, in fact, be realized with any question form. Tsui proposed six question functions: `Elicit:inform`, `Elicit:confirm`, `Elicit:agree`, `Elicit:commit`, `Elicit:repeat`, and `Elicit:clarify`. `Elicit:inform` invites the addressee to supply a piece of information (e.g., "What time will you be finished?"). `Elicit:confirm` invites the addressee to confirm the speaker's assumption (e.g., "Is that you Henry?"). `Elicit:agree` invites the addressee to agree with the speaker's assumption that the expressed proposition is self-evidently true (e.g., "Lovely day isn't it?" (when it is sunny out)). `Elicit:commit` elicits more that just a verbal response from the addressee, it also elicits commitment of some kind (e.g., "Can I talk to you?"). `Elicit:repeat` asks the addressee to repeat what was just said in the discourse (e.g., "What did you say?" or "Pardon?"). Finally, `Elicit:clarify` seeks clarification of a topic referred to earlier in the discourse (e.g., "Did you say you wanted to go to the movies?").

There is clear overlap between the function categories put forth by both Stenström and Tsui, though some differences as well. For our own research, described in subsequent chapters, we chose to adopt a subset of question function labels that could be mapped to both taxonomies. We chose to treat question function as orthogonal to question form, as proposed by Tsui, while allowing for the possibility that certain question forms and functions might be more highly correlated than others.

10.5 Student questions in the tutoring domain

A learning environment in which a tutor and a student work one-on-one has been shown to engender more student learning than does classroom instruction alone. Graesser et al. (2001) reviewed several tutoring studies and reported that the learning gain of students who utilized tutoring and classroom instruction was between 0.4 and 2.3 standard deviations above the learning gain of students who participated in only classroom instruction.[4] There is no lack of competing theories to explain why this is the case. For example, Merrill et al. (1992) suggested that human tutors are effective because they use the technique of "guided learning by doing" that allows students to explore on their own, but prevents them from going down the wrong path. Others have suggested that social interaction may play a role in that students may be too embarrassed in a social setting of their classmates to express their misunderstandings, whereas this is not the case in tutoring interactions. One of the most intriguing reasons was offered in an earlier study by Graesser and has generally been accepted by the educational community. Graesser & Person (1994) provided a review of research studies, including a few conducted by the main authors, that examined student question frequency in different educational settings. They found that a typical student in a classroom environment asks about 0.1 questions per hour, whereas in a one-on-one tutoring session that same student will ask an average of 26.5 questions per hour. In other words, student questions in tutoring are more than 200 times more frequent than in classroom instruction. Furthermore, Graesser & Person, in their own studies, found a significant positive correlation between examination scores and the rate of deep-reasoning student questions and a negative correlation between total questions asked and examination scores.[5] Taken together, these findings suggest that the rate and function of student questions may in some way be related to the increased learning gain observed with one-on-one tutoring.

An Intelligent Tutoring System (ITS) is an educational software program designed to

[4] **Learning gain** is calculated as the difference between scores on a test taken before instruction (pre-test) and one taken after instruction (post-test). One standard deviation represents approximately one letter grade.

[5] It should be noted that examination score cannot be considered the same as learning gain because it does not take into account the learning attributed to being tutored, only overall student proficiency.

tutor students using Artificial Intelligence. Such systems generally serve as domain experts and guide students through problem-solving questions by offering step-by-step feedback on solution accuracy and context-specific advice. The pedagogical approach of ITSs is usually that of the Socratic method: system-initiated questions help guide the student to solving a larger problem. Though early ITSs employed no Natural Language Processing capabilities, many today do so in a text-only format. Far fewer utilize spoken discourse as the means of interaction, but they are becoming more common and will be addressed shortly. The potential advantages of ITSs include increasing access to tutoring among students who otherwise would not be able to receive it, either due to lack of available tutors or to the prohibitive cost of human tutors.

State-of-the-art ITSs have been shown to be effective at fostering student learning. Graesser et al. (2001) noted that the best ones increase student learning gain by 0.3 to 1.0 standard deviations over classroom instruction alone. Note that while promising, such learning gains are still well below those observed when tutoring is conducted with human (rather than automated) tutors. Most solutions offered as ways to close this gap are those that involve improving the Natural Language capabilities of ITSs. Aist et al. (2002) showed, through Wizard-of-Oz experiments, that enabling an ITS to provide feedback as a human tutor does increase student persistence and willingness to continue.[6] The authors noted that this type of student behavior improves the likelihood that learning gain would increase as well. Pon-Barry et al. (2006) examined the role that hedges, response latency, and filled pauses played in signaling a student's "feeling-of-knowing" (Fox, 1993) in the SCoT ITS. They found that long student response latency correlated with a lower feeling-of-knowing on the part of the student and suggested that human tutors use such information in formulating their next move.

Both of the aforementioned studies highlight some of the advantages of developing ITSs that behave more like human tutors. A notable exception to this view was put forth by Anderson et al. (1995) who claimed, after working on cognitive tutors for ten years, that the

[6]A Wizard-of-Oz experiment is one in which the automated agent is actually controlled by a human operator, a fact unknown to the user. Such experiments are usually used as control experiments or gold standards so that performance of true automated system use can be compared with optimal system use.

goal should *not* be to emulate human tutors, but rather to develop tutoring tools. We believe that there is room for such tools in the education software domain, but that Intelligent Tutoring Systems with Natural Language capabilities are also essential. In the following chapters we describe work we conducted on the analysis of student question behavior in a corpus of human-human tutoring dialogs. Our ultimate goal was to improve the question handling capabilities of ITSs such that they respond appropriately to both the function and form of student questions, as a human tutor would.

10.6 Automatic question identification

There have been innumerable theoretical studies on question behavior in the past. There have also been quite a few corpus-based approaches, several of them mentioned above. However, there have been relatively few studies that have conducted machine learning experiments designed to automatically detect and classify spoken questions. Most machine learning experiments that have examined questions have done so in a larger framework of general dialog act prediction. Since such tag sets tend to be relatively large, the performance of question-related tags in isolation is rarely discussed in such studies.[7] Furthermore, most dialog act prediction studies conducted to-date have looked only at information available from text (e.g., words, syntax, dialog act history) and not speech (e.g., prosody and voice quality). Therefore, we choose to focus our discussion of previous work on findings of the few studies that have examined the automatic classification of questions in spoken English.[8]

Shriberg et al. (1998) conducted machine learning experiments to automatically classify utterances as either questions or statements in over one thousand spoken conversations in the SWITCHBOARD corpus of telephone conversations.[9] From each utterance they extracted lexical information (language models of true transcriptions and ASR output) as well

[7]For example, Surendran & Levow (2006) used 13 MapTask tags, Reithinger & Klesen (1997) used 18 VERBMOBIL tags, and Stolcke et al. (2000) report findings on the full SWBD-DAMSL set (42 tags).

[8]There have been similar studies in other languages as well; for example, Liu et al. (2006) conducted machine learning experiments for spoken questions in Mandarin.

[9]This was actually a subtask; the full experiment examined the classification of seven dialog acts: statement, question, backchannel, incomplete, agreement, appreciation, other.

as automatic prosodic information (duration, pausing, fundamental frequency ($f0$), energy, and speaking rate measures). These cues were then used as features for machine learning experiments using binary decision trees. The data was down-sampled such that questions and statements were equally likely (50% baseline performance accuracy). Performance accuracy using prosodic features alone was observed to be 76.0%. Pitch information was found to be informative (consistent with previous theoretical discussion), but so were pausological and durational features. Accuracy using only language models trained on hand-transcriptions yielded a significantly higher performance accuracy of 85.9%. However, when both lexical and prosodic information were combined, a slightly higher overall performance accuracy of 87.6% was observed. The authors contrasted these results with the accuracy when predicted lexical information (ASR transcriptions) was used and found lower prediction rates: 75.4% when predicted words were used in isolation (about the same as using automatic prosodic information) and 79.8% when used in conjunction with prosodic information. The conclusion reached was that prosodic and lexical information were often redundant with respect to the binary classification of questions vs. statements but that together they were both considered useful for the automatic classification of spoken questions.

As a second experiment, Shriberg et al. (1998) ran a four-way classification task exploring question form: `statement`, `y/n-question`, `wh-question`, `declarative question`. Performance accuracy using only prosodic information was 47%.[10] Here, pitch information was found to be the most important of the prosodic features, by a substantial margin. The authors note that the learned decision tree confirmed long-held beliefs that declarative and yes-no questions have phrase-final rising intonation whereas statements and wh-questions have phrase-final falling intonation. Furthermore, the authors contend that conflating disparate question forms into one category obscures some of the intonational cues of some forms of questions.

Surendran & Levow (2006) report results of machine learning experiments for questions as well.[11] They used the HCRC MapTask corpus, comprising 64 dialogs and 14,810 utter-

[10] Baseline performance, due to down-sampling, was 25%. No performance accuracy was reported using lexical information.

[11] Again, this was in the context of a larger 13-way dialog act classification task.

ances. They, too, extracted both lexical information (in the form of unigrams, bigrams, and trigrams) and prosodic information (similar to Shriberg et al.) from each utterance. For each dialog act they reported F-measures for prosodic, lexical, and combined feature information. There were four question types, a mixture of form and function. Agreement-checking questions (`align`) showed an F-measure of 0.22 using acoustic features, 0.43 using lexical features, and 0.60 using the combined feature set. Information-checking questions (`check`) had an F-measure of 0.40 using acoustic features, 0.45 using lexical features, and 0.57 using the combined feature set. The performance of the classification of yes-no questions resulted in F-measures of 0.16 (acoustic), 0.64 (lexical), and 0.63 (combined). The final question category was a catch-all for all other questions (`query-w`). F-measures for this type were 0.05 (acoustic), 0.63 (lexical), and 0.68 (combined). In summary, their findings suggested that acoustic-prosodic features helped most for information-seeking and agreement-checking questions; less so for all other questions. For all question types, lexical information alone was more predictive than acoustic information alone, though the combination of the features resulted in further improvement.

Though the two aforementioned studies provide useful information about the type of performance and the most useful features one might expect to see in the automatic detection of questions in spoken English, the tags themselves don't discriminate between the form and function of questions—something we were very interested in. In the (reduced) SWBD-DAMSL set used by Shriberg et al. (1998) only the form of questions were tagged. Alternatively, while Surendran & Levow (2006) tagged some questions in the the HCRC MapTask corpus by form (e.g., `query-yn`), others were tagged according to function (e.g., `align` and `check`). In Chapter 16 we report results on the automatic classification of both form and function, independently. We felt that such an approach was necessary because both types of information may help guide an automated tutor to the correct response. Question form often dictates the form that the appropriate response should take and, similarly, the function of the question guides the content of expected answer. In other words, our goal was to detect student questions in such a way that automated tutor answers may be supplied that are felicitous both in *how* and *what* is said.

Chapter 11

The HH-ITSPOKE Corpus

Our corpus was a set of tutoring dialogs obtained from the University of Pittsburgh's Learning Research & Development Center. This corpus was collected for the development of ITSPOKE, an Intelligent Tutoring Spoken Dialog System designed to teach principles of qualitative physics (Litman & Silliman, 2004). ITSPOKE is a speech-enabled version of Why2-Atlas, a text-based tutoring system (VanLehn et al., 2002). Several corpora have been collected related to ITSPOKE development, including dialogs between human students and an automated agent. The corpus we analyzed, though, contained dialogs between human students and a human tutor and we refer to it as HH-ITSPOKE.

The HH-ITSPOKE corpus comprises tutorial sessions between 17 undergraduate students (7 female, 10 male; all native American English speakers) and a (male) professional tutor. An excerpt of a dialog from the corpus is shown in Figure 11.1 on page 112; disfluencies have been eliminated and punctuation added for readability.

The recording procedure for each session proceeded as follows. One student and the tutor were seated in the same room but were separated by a partition such that they could not see each other. They interacted via microphones and a graphical user interface (GUI), as shown in Figure 11.2 on page 112. Each student was first asked to type an essay in response to one of five qualitative physics questions, shown in the box in the upper right hand of the GUI (e.g., "Suppose a man is in a free-falling elevator and is holding his keys motionless in front of his face ...."). Each student then typed his or her response into the text box at the bottom right of the GUI (e.g., "The keys will hit the floor of the elevator because of the

force of gravity.") after which point the tutor evaluated the student essay and proceeded to tutor the student verbally until he determined that the student had successfully mastered the material. It was an iterative process whereby the student was asked to retype their essay until the tutor was satisfied that it successfully answered the question. A transcript of the conversation appeared in a text box at the bottom left of the GUI.

The tutor and each student were recorded using separate microphones and each channel was manually transcribed and segmented into turns. Each dialog contained, on average, 53 student turns. Each student turn averaged 2.5 seconds and 5 words in length. The total number of student turns in the corpus was 7,460.

... 17.4 minutes into dialog 71-61-1 ...

TUTOR:	What does the acceleration mean?
STUDENT:	That the object is moving through space?
TUTOR:	No. Acceleration means that object's velocity is changing
STUDENT:	What?
TUTOR:	Object's velocity is changing.
STUDENT:	Uh-huh, and then once you release it the velocity remains constant.

Figure 11.1: A transcribed excerpt from the HH-ITSPOKE corpus.

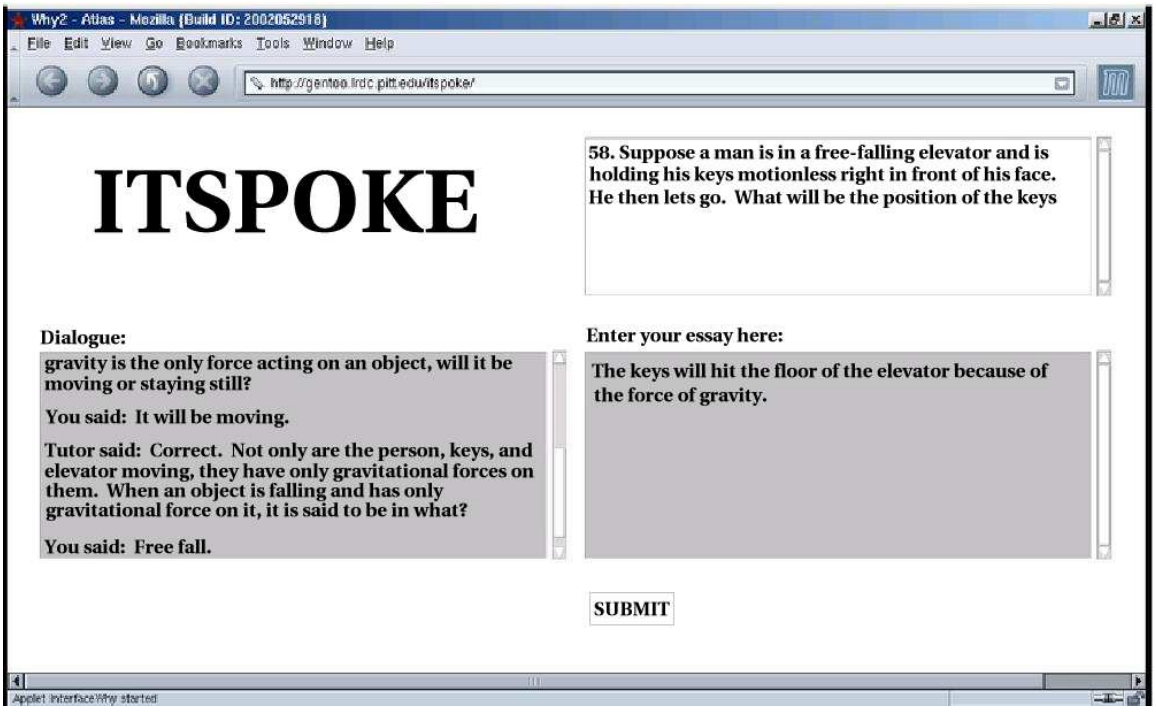


Figure 11.2: Screenshot of the ITSPOKE GUI during a tutoring session.

Chapter 12

Question Annotation

Student questions were identified in the HH-ITSPOKE corpus by one annotator using the construct of Bolinger (1957): student questions were considered utterances judged as seeking a substantive verbal response from the tutor. One thousand and thirty (1,030) questions were identified, and the beginning and end of each question were recorded. The rate of student questions per hour, averaged across all students, was calculated to be 25.2 (SD = 13.0) and student questions were found to comprise 13.3% of total student speaking time. Such a high rate of time spent asking questions is consistent with other findings in one-on-one human tutoring (e.g., Graesser & Person, 1994). In additional to identifying student questions, we labeled the form and function of each question independently. The labels are discussed below and were first introduced in Liscombe et al. (2006).

12.1 Question type

Each question in the HH-ITSPOKE corpus was coded according to its syntactic **form** with one of the labels described in Table 12.1 on page 115. We chose the most commonly accepted question forms: yes-no questions (`ynQ`), wh-questions (`whQ`), yes-no tag questions (`tagQ`), alternative questions (`altQ`), particle questions (`partQ`), and declarative questions (`dQ`). All questions forms have already been described in full in the literature review, though a few clarifications with respect to coding are in order. Questions in the `partQ` class were those that consisted solely of a grammatical particle or function word, though we found

such questions to be restricted to "pardon," "huh," and "hm." Though `altQ`s are generally defined as questions that present two or more possible presupposed answers, we coded such questions using the rather broad heuristic of any question that contained the word "or." Finally, in our coding scheme, a non-clausal fragment was considered `dQ`—as was a wh-question with no wh-fronting, since the surface word order resembled a declarative (e.g., "A what?").[1]

The coding of question **function** was done following the proposed label sets of both Stenström (1984) and Tsui (1992). Due to sparsity in our data, we conflated their larger sets to a smaller set containing four function categories (mapping is shown in Table 12.2 on page 115), with special care taken to maintain distinctions of importance for ITSs. Table 12.3 on page 115 presents examples of each of the function labels that were used to code all the questions in our data. A clarification-seeking question (`clarQ`) was identified as a question asked by the student that sought clarification of something the tutor said, something about the essay, or something about the computer interface. These were questions that could naturally be followed by, "is that what you just said/intended/meant?." They were not considered to be questions concerning the student's own conceptual understanding; such a question was coded as a confirmation-seeking/check question (`chkQ`) because it was one in which the student sought to check or confirm his or her present understanding with that of the tutor's. A question in which the student was asking for new, previously undiscussed information from the tutor was labeled as an information-seeking question (`infoQ`). Any question that did not adhere to one of the aforementioned functions was labeled as other (`othQ`). Such questions included rhetorical and self-addressing questions, as shown as an example in Table 12.3.

[1] Wh-questions with no wh-fronting are often referred to as "in-situ questions."

Label	Form	Example
`ynQ`	yes-no question	"Is it a vector?"
`whQ`	wh-question	"What is a vector?"
`tagQ`	tag question	"It's a vector, isn't it?"
`altQ`	alternative question	"Is it a vector or a scalar?"
`partQ`	particle question	"Huh?"
`dQ`	declarative question	"It's a vector?" or "A vector?"

Table 12.1: Question form labels with examples.

Stenström's labels	Tsui's label(s)		Our Label
{acknowledge, confirm}	{confirm}	→	confirmation-seeking/check
{clarify, repeat}	{clarify, repeat}	→	clarification-seeking
{identify, polar}	{inform}	→	information-seeking
{action, offer, permit, react}	{agree, commit}	→	other

Table 12.2: Stenström's and Tsui's question function labels mapped to ours.

Label	Function	Example
`chkQ`	confirmation-seeking/check question	"Is that right?"
`clarQ`	clarification-seeking question	"What do you mean?"
`infoQ`	information-seeking question	"Is the initial velocity the same?"
`othQ`	other	"Why did I do that?"

Table 12.3: Question function labels with examples.

12.2 Question-bearing turns

Each of the turns containing a question was coded as a **question-bearing turn** (QBT). Though most QBTs (89%) bore only one question, 10% were observed to contain two questions, and the remaining 1% bore three questions. Thus, there were fewer QBTs (918) than there were questions (1030) in the HH-ITSPOKE corpus. Furthermore, we note that 70% of QBTs consisted entirely of the question itself. Of the remaining QBTs—those that contained speech other than a question—63% bore questions that ended the turn. In other words, most (89%) of all QBTs were observed to bear questions that occurred at the end the turn.

Chapter 13

Question Form and Function Distribution

We turn now to the distribution of the question form and function labels. Table 13.1 lists the counts of all form and function labels as well as their intersection. Each cell contains the raw count of the form and function label pairs along with their overall percentage over all 1,030 questions. For example, the first cell states that 416 questions were labeled as both `dQ` and `chkQ` and that this label pair comprised 40.4% of all label pairs. The last column and last row (with the header "N") in the table list the overall frequency of the form and function labels, respectively. For example, 577 out of the 1,030 questions (56.0%) were labeled as `chkQ`. We will first discuss these terminal cells and then turn to the intersection of the two label sets (the internal cells). The column with the header "`?`" represents questions whose function could not be determined.

With respect to the syntactic form of student questions, almost half (54.0%) were found to be declarative (`dQ`). This figure is much higher than has been reported for other corpora. For example, Beun (1990) found 20% of all questions in a Dutch corpus of travel-planning dialogs to be of declarative form and Jurasfky et al. (1997) found 12% in the SWITCHBOARD telephone dialog corpus. The second most frequent question form was `ynQ` at 23.9%, while yes-no questions in tag form (`tagQ`) made up another 7.4% of all questions. Wh-questions (`whQ`) constituted 10.1% of all questions, alternative questions (`altQ`) comprised 3.0%, and

	Function					
Form	chkQ	clarQ	infoQ	othQ	?	N
dQ	416 (40.4%)	126 (12.3%)	2 (0.2%)	7 (0.7%)	6 (0.6%)	557 (54.0%)
ynQ	81 (7.9%)	99 (9.6%)	33 (3.2%)	10 (1.0%)	23 (2.2%)	246 (23.9%)
whQ		68 (6.6%)	28 (2.7%)	2 (0.2%)	6 (0.6%)	104 (10.1%)
tagQ	67 (6.5%)	7 (0.7%)			2 (0.2%)	76 (7.4%)
altQ	13 (1.3%)	12 (1.2%)	2 (0.2%)		4 (0.4%)	31 (3.0%)
partQ		16 (1.6%)				16 (1.6%)
N	577 (56.0%)	328 (31.9%)	65 (6.3%)	19 (1.8%)	41 (4.0%)	1030 (100%)

Table 13.1: Syntactic form and discourse function of all questions.

particle questions (partQ)—such as "Huh?"—were the least frequent question form at 1.6%.

Concerning the pragmatic function of student questions, most (56.0%) sought a tutor response that validated the students' understanding (chkQ). The second most frequent question function (31.9%) was that which sought clarification of something the tutor had said (clarQ). Far fewer student questions (6.3%) sought new information (infoQ), and other types of questions (othQ) made up only 1.8% of all student questions. There were 41 questions (4.0%) whose function could not be determined by the annotator.

Turning now to <form, function> label pairs, we observed that most (40.4%) student questions were declarative in form and confirmation-seeking/check in function: <dQ, chkQ>. The second most frequent pairing, at 12.3%, were questions that were declarative in form and clarification-seeking in function: <dQ, clarQ>. The remaining pairs made up relatively small fractions. We ran a χ^2 test of independence between the two label sets, excluding the column of unknown question function, and found a systematic relationship between question form and function ($\chi^2 = 370.95$, df = 15, $p < 0.001$). The correlation strength of the two sets was found to be 0.37, as calculated using Cramer's phi (ϕ_c).

	Function					
Form	chkQ	clarQ	infoQ	othQ	?	N
dQ	75	23	0	1	1	557
ynQ	33	40	13	4	9	246
whQ	0	65	27	2	6	104
tagQ	88	9	0	0	3	76
altQ	42	39	7	0	13	31
partQ	0	100	0	0	0	16

Table 13.2: Question <form, function> counts normalized by total form count.

To more fully understand the nature of the relationship between question form and function, we observed the frequency of each question label normalized by type. Table 13.2 shows the relative counts of the label pairs when normalized by question form. We see that dQs were more than three times as likely to be chkQs than they were to be clarQs (75 compared with 23). The function of other dQs were marginal. Yes-no questions (ynQs) had a more uniform distribution with respect to question function, though most were found to function as either clarQ (40) or chkQ (33) and these functions were more than two and a half times as likely as any other function. Wh-questions were primarily clarQ (65) and were half as often infoQ (27). The function of tagQ was overwhelmingly chkQ (88); it was almost ten times more likely any other function. It is interesting that tag and yes-no questions, though most similar in form, did not seem to serve the same functions. Rather, the function of altQs was most similar to the function of ynQs in that their functions were primarily chkQ or clarQ, at a rate very similar to ynQs. A particle question (partQ) functioned exclusively as a clarification-seeking question (clarQ).

Table 13.3 lists the relative frequency of question function with respect to form, when normalized by the overall function label counts (the row totals). We observed that chkQs took the form of dQs more than 5 to 1 over the next most frequent form, ynQ (72:14). There were three primary function of clarQs: dQ (38), ynQ (30), and whQ (21). Questions that functioned as information-seeking (infoQ) were overwhelmingly either of the form ynQ (51) or whQ (43). Half (53) of all other types of discernible question functions (othQ) were

	Function				
Form	chkQ	clarQ	infoQ	othQ	?
dQ	72	38	3	37	15
ynQ	14	30	51	53	56
whQ	0	21	43	11	15
tagQ	12	2	0	0	5
altQ	2	4	3	0	10
partQ	0	5	0	0	0
N	577	328	65	19	41

Table 13.3: Question <form, function> counts normalized by total function count.

produced in the form of a yes-no question (ynQ) and outnumbered the next most frequent question form—dQ—by nearly one and a half times (53/37). A question whose function could not be determined was primarily ynQ.

In examining the relationship between the form and function of student questions in the HH-ITSPOKE corpus, we observed a systematic relationship between the two types of information. Through analysis, we found the most significant coupling to exist between questions of declarative form and confirmation-seeking/check function: <dQ, chkQ>. Other observations indicated that most wh-questions were clarification-seeking, most yes-no tag questions were confirmation-seeking/check, and all particle questions were clarification-seeking. Not included in the statistical analysis, but observed from normalized frequency counts, we also observed that most questions of ambiguous function took the form of a yes-no question. Clearly, question form and function—at least insofar as could be concluded by our data—were not truly orthogonal. However, it would be unwise, we think, to exclude the independent label sets because though a relationship was found, the strength was relatively low (0.34)—or at least was not overwhelming—indicating that there was still some amount of independence between the form and function of questions. Furthermore, we felt that it was important for an Intelligent Tutoring System to be able to distinguish between the two because identifying a question by it's form does not always indicate the expected type of response. For example, yes-no and alternative questions were found to be equally indicative

of questions that either sought confirmation or clarification from the tutor. These entail different response strategies on the part of the tutor to ensure felicitous dialog flow and, presumably, facilitation of student learning.

Chapter 14

Extraction of Question Cues

Motivated by previous research on dialog act classification, and by the theoretical claims put forth in relation to questions cues, we extracted several features from each student turn in the HH-ITSPOKE corpus. Descriptions of these features follow.

The vast majority of the features we examined as potential indicators of questions were automatically-extracted acoustic-prosodic features, including features associated with pitch, loudness, and rhythm.[1] Each prosodic feature was z-score normalized by the speaker's mean and standard deviation for all feature values. All acoustic-prosodic features are listed in Table 14.1 on page 123.

We used fundamental frequency ($f0$) measurements to approximate overall **pitch** behavior. Features encapsulating pitch statistics—minimum (`f0-min`), maximum (`f0-max`), mean (`f0-mean`), and standard deviation (`f0-stdv`)—were calculated on all $f0$ information, excluding the top and bottom 2%, to eliminate outliers. Global pitch shape was approximated by calculating the slope of the all-points regression line over the entire turn (`f0-rslope`). In addition, we isolated turn-final intonation shape by smoothing and interpolating the $f0$ using built-in Praat algorithms and then isolating the last 200 milliseconds of the student turn, over which we calculated the following $f0$ features: minimum (`f0-end-min`), maximum (`f0-end-max`), mean (`f0-end-mean`), standard deviation (`f0-end-stdv`), difference between the first and last $f0$ points (`f0-end-range`), slope of all-points regression line

[1] Acoustic processing was done in Praat, a program for speech analysis and synthesis (Boersma, 2001).

	Feature	Description
Pitch	f0-min	minimum $f0$
	f0-max	maximum $f0$
	f0-mean	mean $f0$
	f0-stdv	standard deviation of $f0$
	f0-rslope	slope of regression line through all $f0$ points
	f0-end-min	minimum $f0$ in last 200 ms
	f0-end-max	maximum $f0$ in last 200 ms
	f0-end-mean	mean $f0$ in last 200 ms
	f0-end-stdv	standard deviation of $f0$ in last 200 ms
	f0-end-range	difference of first and last $f0$ points of last 200 ms
	f0-end-rslope	slope of regression line through all $f0$ points in last 200 ms
	f0-end-rising	percent rising $f0$
Loudness	db-min	minimum intensity
	db-max	maximum intensity
	db-mean	mean intensity
	db-stdv	standard deviation of intensity
	db-end-mean	mean intensity of last 200 ms
	db-end-diff	difference of mean intensity of of last 200 ms and entire turn
Rhythm	pause-count	number of pauses
	pause-dur-mean	mean length of all pauses
	pause-dur-cum	cumulative pause duration
	phn-rate	phonation rate (speech duration - pause duration)/total time)
	spk-rate	speaking rate of non-pause regions (voiced frames/all frames)

Table 14.1: Acoustic-prosodic features extracted from each turn in the HH-ITSPOKE corpus.

(`f0-end-rslope`), and the percent of rising slope between all consecutive time-point pairs (`f0-end-rising`).

To examine the role of **loudness** we extracted the minimum (`db-min`), maximum (`db-max`), mean (`db-mean`), and standard deviation (`db-stdv`) of signal intensity, measured in decibels, over the entire student turn. In addition, we calculated the mean intensity over the last 200 milliseconds of each student turn (`db-end-mean`), as well as the difference between the mean in the final region and the mean over the entire student turn (`db-end-diff`).

Rhythmic features were designed to capture pausing and speaking rate behavior.[2] We adapted a procedure to automatically identify pauses in student turns.[3] The procedure isolated spans of silence 200 milliseconds or longer in length by using background noise estimation for each dialog, defined as the 75th quantile of intensity measurements over all non-student turns in that dialog. In an earlier study we showed that this semi-automatic process reliably identified inter-turn pauses in the HH-ITSPOKE corpus (Liscombe et al., 2005a). We found there to be 1.62 pauses per student turn and the mean length of pauses to be 1.59 seconds. Pausing behavior in each student turn was represented as the number of pauses (`pause-count`), the mean length of all pauses (`pause-dur-mean`), the cumulative pause duration (`pause-dur-cum`), and the percentage of time that pausing occupies relative to the entire student turn (`phn-rate`), often referred to as **phonation rate**. Speaking rate (`spk-rate`) was calculated by counting the number of voiced frames in the turn, normalized by the total number of frames in non-pause regions of the turn.

We also encoded both syntactic and lexical information as features, which are listed in Table 14.2 on page 126. Our representation of **lexical** information consisted of manually-transcribed word unigrams (`lex-uni`) and bigrams (`lex-bi`) uttered in each student turn. In addition to words with semantic content, we also included filled pauses, such as "um" and "uh." To capture **syntactic** information we applied a part-of-speech (POS) tagger (Ratnaparkhi, 1996) trained on the SWITCHBOARD corpus to the lexical transcriptions of student turns. Syntactic features consisted of automatically predicted POS unigrams

[2]Note that the term **rhythm** here is defined differently than it is in Part IV, following the conventions of each discipline.

[3]The original pause detection program can be found at http://www.helsinki.fi/~lennes/praat-scripts

(pos-uni) and bigrams (pos-bi).

The remaining features were designed to capture knowledge about the student not present in either the aural or linguistic channels and are referred to as the **student and task dependent** feature set, also summarized in Table 14.2. Included in this feature set were: the score the student received on a physics test taken before the tutoring session (pre-test), the gender of the student, the hand-labeled correctness of the student turn (correct), and the tutor dialog act (ptda) immediately preceding the student turn (also hand-labeled). Tutor dialog acts were labeled according to pragmatic function by Forbes-Riley et al. (2005) and could be a tutor turn that sought to elicit a certain type of student answer (short-answer-question, long-answer-question, deep-answer-question), provided feedback to the student (positive-feedback, negative-feedback), reformulated a previously-introduced idea (recap, restatement, expansion), or gave the student a hint. The remaining tutor dialog acts included when the tutor gave the student the answer (bottom-out) or a directive.

	Feature	Description
Lexis	lex-uni	lexical unigrams
	lex-bi	lexical bigrams
Syntax	pos-uni	part-of-speech unigrams
	pos-bi	part-of-speech bigrams
Student and Task	pre-test	student pre-test score
	gender	student gender
	correct	turn correctness label: { fully partially none not-applicable }
	ptda	previous tutor dialog act: { short-answer-question long-answer-question deep-answer-question positive-feedback negative-feedback restatement recap request bottom-out hint expansion directive }

Table 14.2: Non-acoustic-prosodic features assigned to each student turn in the HH-ITSPOKE corpus

Chapter 15

Learning Gain

We mentioned previously that research has indicated that students learn more when tutored one-on-one than they do through classroom instruction, and that one of the differences between the two environments is that students ask questions at a dramatically higher rate in the former environment than in the latter. It has been suggested that certain types of questions might correlate with student learning (e.g., Graesser & Person, 1994). To examine this suggestion further, we correlated student learning gain with several of features mentioned in the previous section. Before and after each tutoring session, the 14 students in the HH-ITSPOKE corpus were tested on their knowledge of physics. Learning gain was measured, simply, as the difference between the two 100-point test scores. Learning gain in our corpus ranged from 15.0% to 47.5%, with a mean of 31.0% and a standard deviation of 8.0%.

We calculated the Pearson Product Moment correlation coefficient (r) between student learning gain and several features calculated over all question-bearing turns, including counts of overall questions asked, each question form and function, and each previous tutor dialog act. We did the same for the z-score normalized continuous acoustic features of QBTs as well (those listed in Table 14.1 on page 123). In total, 46 independent correlations were calculated. The overwhelming majority proved non-significant ($p > 0.10$), though five were found to be somewhat significant ($p \leq 0.07$) and relatively strong (average absolute correlation strength was 0.51). Table 15.1 lists the most significant correlations found.

The fewer yes-no tag questions (`tagQ`) a student asked (e.g., "um the plane right"),

Label or Feature	r	p
tagQ	-0.52	0.06
directive	-0.50	0.07
db-end-mean	0.53	0.05
f0-mean	0.50	0.07
f0-min	0.51	0.06

Table 15.1: Labels and QBT features correlated with student learning gain.

the more they seemed to learn. Since we found that most (88.2%) tag questions were confirmation-seeking in function (see page 119), perhaps this might explain the observed correlation; as learning increases then, possibly, the need to confirm answers decreases. Those students who were less likely to respond with a QBT following explicit directives from the tutor (e.g., "use newton's second law of motion") also tended to learn more. Again, questions following such directives may be indicative of confusion or non-understanding, even when key concepts are made explicit by the tutor. Presumably, the those students with such fundamental non-understanding would show lower learning gains.

The few acoustic measures observed to be the most significantly correlated with student learning gains were the loudness of the last 200 milliseconds of the student turn (db-end-mean), mean pitch (f0-mean), and minimum pitch (f0-min). Students who learned most were those that exhibited higher values for each of these features, normalized by student. It remains unclear why students who were louder at the end of QBTs and exhibited higher pitch values over the entirety of QBTs should learn more, but it is intriguing nonetheless and warrants future investigation.

Chapter 16

Automatic Classification of QBTs

Understanding the nature of spoken questions is of both theoretic and practical importance. By identifying the form and function that student questions can take in the HH-ITSPOKE corpus, for example, we might be able to then use that knowledge to improve the ITSPOKE system by enabling it to behave more as a human tutor does. In our view, this is essential for increasing the benefits offered by Intelligent Tutoring Systems and for closing the learning gap between human tutoring and automated tutoring. In this section we describe machine learning experiments we conducted and report on both the observed prediction performance as well as the most salient features.

We conducted three sets of classification experiments: (1) question-bearing turns (QBTs) vs. non-question-bearing turns (¬QBTs), (2) QBT form, and (3) QBT function. For each of the three tasks, we conducted a set of nine classification experiments to evaluate the usefulness of each feature set (as described in Chapter 14), as well as to examine the predictive power of all feature sets combined. An additional experiment was conducted using all prosodic features calculated over only the last 200 milliseconds of each student turn. Each classification experiment used the WEKA machine learning environment (Witten et al., 1999). While we experimented with several machine learning algorithms, including decision trees, rule induction, and support vector machines, we present results here using the decision tree learner C4.5 boosted with the meta-learning algorithm ADABOOST (Freund

& Schapire, 1999), which provided the best results.[1] Performance for each experiment was averaged after running 5-fold cross-validation.

We present results in two formats. For comparing the performance of feature sets we report performance as overall classification **accuracy** of the labels.[2] There are well-known arguments against using accuracy as a performance metric. One of the drawbacks is that it does not generally describe the classification performance of each individual label. Secondly, it's usefulness depends greatly on the distribution of the data. If the data is highly biased towards one label, then high prediction accuracy can be obtained simply by always choosing the most frequent label as the prediction, without using learning of any kind. As an example, the most frequent question form in our corpus was `dQ` and it represented 54% of all labels. Without learning any rules and by classifying all QBTs as `dQ` we would observe a classification accuracy of 54%, though this would actually indicate the uselessness of our machine learning approach.

To combat the drawbacks of accuracy, **F-measure** is often considered to be a more useful performance metric because it reports performance as a balance between precision and recall.[3] However, F-measure must be reported for each label and calculated against all other labels. Thus, it can be quite difficult to paint an overall picture of performance when several experiments are compared. Acknowledging the advantage and drawbacks of each approach, we decided to use both metrics, but for different purposes. When comparing the performance of different feature sets on the same task we list classification accuracy so that we might easily compare the overall performance of each set by examining the accuracy of each set. However, we report the performance of the best-performing feature set combination in terms of F-measure so that we provide a realistic expectation of the performance for each class label if our models were to be used to predict unseen data.

Another technique we employed was a method described by Hall (1998) in which an optimal feature subset is selected by finding the minimal feature set that predicts the labels

[1] The procedure is implemented as the J4.8 function in the WEKA machine learning software package.

[2] Accuracy is defined as the count of correctly predicted tokens divided by the count of all tokens.

[3] We used the standard method of F-measure calculation: $(2 \cdot \text{precision} \cdot \text{recall})/(\text{precision} + \text{recall})$.

most accurately while also reducing the redundancy among features.[4] In other words, the features in the derived optimal subset for a given task are the features that correlate most strongly with the labels and most weakly among each other. This technique was employed to derive the most useful features for classification.

16.1 QBT vs. ¬QBT

In our corpus of tutorial dialogs most student turns did not contain questions. Excluding student turns that functioned only to maintain discourse flow, such as back-channels (e.g., "uh huh"), non-question-bearing turns (¬QBTs) outnumbered question-bearing-turns (QBTs) nearly 2.5 to 1. In order to learn meaningful cues to question-bearing turns and to avoid a machine learning solution that favored ¬QBTs *a priori*, we randomly down-sampled ¬QBTs from each student to match the number of QBTs for that student. Thus, we explored how well our features could automatically discriminate between 918 QBTs and 918 ¬QBTs; a majority class baseline of 50%.

Table 16.1 reports the performance accuracy of each feature set in isolation. Here we observed that the least predictive feature sets were rhythmic (52.6%), and student and task dependent (56.1%). The most predictive feature set comprised all prosodic information (74.5%), though it appeared that the most significant contributor to this set was pitch information (72.6%). The performance accuracies of the remaining feature sets, including the lexical and syntactic ones, fell somewhere in between. QBT precision, recall, and F-measure using the feature set with highest accuracy (all features) were all 0.80, indicating that our ability to discriminate between turns containing student questions and those that did not was robust.

Using the feature subset evaluation metric mentioned above, 15 features were considered to be the most informative and least redundant. In decreasing order of importance these were: `f0-end-rslope`, `f0-end-range`, `f0-end-rising`, `f0-rslope`, `f0-max`, `db-end-mean`, `pos-bi` (PRP+VB), `ptda`, `lex-uni` (yes), `db-max`, `lex-bi` ("# I"), `pos-bi`

[4]The procedure is implemented as the CfsSubsetEval function in the Weka machine learning software package.

Feature Set	Accuracy
none (majority class baseline)	50.0%
prosody: rhythmic	52.6%
student and task dependent	56.1%
prosody: loudness	61.8%
syntax	65.3%
lexis	67.2%
prosody: last 200 ms	70.3%
prosody: pitch	72.6%
prosody: all	74.5%
all feature sets combined	79.7%

Table 16.1: Performance accuracy of each feature set in classifying QBTs vs. ¬QBTs.

(UH+PRP), `db-end-diff`, `lex-uni` ("acceleration"), and `pos-bi` (IN+INNS). Quite dramatically, the features encoding phrase-final pitch shape were the most important. Other important features were those that encoded pitch information of the entire turn as well the loudness of the end of the turn. Table 16.2 on page 134 lists the acoustic-prosodic features in this subset and reports the mean value of each calculated separately over all QBTs and ¬QBTs. The values listed are for descriptive purposes and so are shown in raw, rather than z-score normalized, format. Though it is well known that different speakers have different pitch ranges and that pitch information should therefore be speaker-normalized in order to compare across speakers, we computed mean feature values aggregated across all speakers, for whom we had an equal number of QBTs and ¬QBTs. Thus, in terms of describing average behavior, using raw values in this context is appropriate and more illuminating than using normalized values. We observed unequivocally that QBTs, on average, exhibited phrase-final rising intonation, whereas ¬QBTs exhibited phrase-final falling intonation. For example, `f0-end-rising` indicated that 70% of all consecutive time points in the last 200 milliseconds of QBTs were rising, whereas less than half were rising in ¬QBTs. Both `f0-end-rslope` and `f0-end-range` showed this as well, as the average slope was positive in QBTs and negative in ¬QBTs. This finding can be generalized to the pitch slope over

the entire turn (`f0-rslope`) as well. Furthermore, the average maximum pitch (`f0-max`) of QBTs was over 100 Hz higher than for ¬QBTs. The remaining three features measured loudness and, though less dramatic in their differences, suggested that QBTs tended to be louder than ¬QBTs.

Let us now turn to the most informative lexical and syntactic features. Table 16.3 on page 134 lists the number of ngram instances observed in question-bearing and non-question bearing turns. The raw counts have been normalized by their row totals (N) in order to generalize the distribution trends. We noticed immediately that some of the suggested lexico-pragmatic cues to questions were supported by our findings. In particular, the word "yes" was highly biased towards ¬QBTs, as was use of the 1st person pronoun at the start of an utterance.[5] The remaining word—"acceleration"—is clearly domain-dependent and is something that most likely would not be indicative of questions in general, though we found it, somewhat inexplicably, to be so in the HH-ITSPOKE corpus.

The most informative syntactic ngrams were also somewhat unexpected. These included a personal pronoun followed by a verb (PRP+VB), an interjection followed by a personal pronoun (UH+PRP), and a preposition or coordinating conjunction followed by a noun phrase (IN+NNS). The fact that the first two bigrams encoded information about personal pronouns was not wholly unexpected considering previous research suggesting that "I" is associated with statements and "you" with questions. However, these features actually did not make such discriminations between pronouns, and yet they were still considered to be highly informative. Perhaps the syntactic contextualization of the other parts of speech in the bigrams played a role here. It is not immediately obvious why a preposition followed by a noun phrase would be more indicative of a question-bearing turn, either. It should be noted that the features that were included in the most informative subset were done so in combination with the other features in the subset, so we cannot arrive at a truly exhaustive description of the phenomenon by examining the features in isolation.[6] However, it does seem clear that information encoding both syntax and lexis was useful

[5]The token "#" was used to designate a turn boundary.

[6]Examination of the induced decision trees could help, but in all cases ours were far too complex to extrapolate generalizations from.

Feature	QBT	¬QBT
f0-end-rslope	1.22	-0.19
f0-end-range	20.36	-3.49
f0-end-rising	0.70	0.43
f0-rslope	0.26	-0.27
f0-max	213.53	193.47
db-end-mean	51.59	48.16
db-max	64.24	63.22
db-end-diff	7.01	5.94

Table 16.2: Mean non-normalized acoustic-prosodic feature values for QBTs and ¬QBTs.

Ngram	QBT	¬QBT	N
"yes"	6	94	71
"# I"	17	83	77
"acceleration"	69	31	149
prp+vb	94	6	107
uh+prp	30	70	185
in+nns	81	19	26

Table 16.3: Normalized lexical and syntactic ngram counts for QBTs and ¬QBTs.

for the automatic classification of QBTs and that the findings were largely consistent with previous research. As a final note, it was surprising that wh-pronouns were not among the most useful syntactic ngram features; neither were modal verbs nor hedges, which are often cited as being indicative of questions. Their absence suggests that the importance placed on these types of information with respect to identifying questions may be overestimated, at least in our domain.

Before we turn to the form and function classification experiments, the role of the previous tutor dialog should be addressed since it was also found to be one of the best indicators of whether a student turn contained a question or not. Table 16.4 shows the normalized tutor dialog act label counts per student turn type. Students responded with a QBT two-thirds more often when the previous tutor turn functioned as a `long-answer-question` and two-thirds *less* often after `positive-feedback` or a `restatement`. Students responded with a QBT only a quarter of the time when the tutor gave them the solution to the answer

Tutor dialog act	**QBT**	**¬QBT**	**N**
`bottom-out`	25	**75**	56
`deep-answer-question`	56	44	313
`expansion`	46	54	128
`hint`	46	54	157
`long-answer-question`	**67**	33	82
`negative-feedback`	46	54	24
`positive-feedback`	36	**64**	77
`recap`	13	**87**	31
`directive`	58	42	57
`restatement`	34	66	107
`social-coordination`	47	53	163
`short-answer-question`	55	45	619

Table 16.4: Normalized counts of previous tutor dialog act (`ptda`) for QBTs and ¬QBTs.

(`bottom-out`) and only 15% of the time when the tutor provided a `recap` of previously discussed concepts. The rest of tutor dialog acts elicited student QBTs and ¬QBTs equally as often. These findings are relatively intuitive and support the role of discourse contextualization to aid in dialog act classification, including turns that function as questions.

16.2 QBT form

A 6-way classification design was established to automatically classify the form of questions present in QBTs. In other words, the task was to predict question form given that we knew the student turn contained a question. The best performance accuracy was observed when all features were used together. Precision, recall, and F-measures for each form label are shown in Table 16.5. All but alternative questions (`altQ`) exhibited quite high F-measures; classification of `partQ` was the most robust (F = 0.93). Average F-measure was 0.71.

Table 16.6 displays the performance accuracy for the nine feature sets described in Chapter 14. We found that only lexical and syntactic features were useful in predicting QBT form. Furthermore, syntactic information did not improve prediction accuracy when used alongside lexical unigrams and bigrams.

The best feature subset was found to include 22 features and, apart from the previous tutor dialog act, all were either lexical or syntactic ngrams. Table 16.7 on page 138 lists the most informative ngrams in order of decreasing informativeness along with normalized counts per question form.[7] Though a bit convoluted, it nevertheless manages to show general tendencies that speak to what we already know about question form. For example, student turns that ended in "right" indicated that 90% of the time the question was in tag form (e.g., "um the plane **right**"). Several other bigrams jump out as well. For example, wh-pronouns, such as "what," and wh-adverbs, such as "why," at the beginning of a turn (#+WP and #+WRB, respectively) were found to be most often present in wh-questions. Student turns starting with the word "huh" co-occurred with particle questions in all cases. We also observed a few biases that support the findings of previous studies. For example,

[7]Normalized ngram counts with more than 50% co-occurrence with a question form have been highlighted in Table 16.7 for clarity.

Form	Precision	Recall	F-measure
altQ	0.29	0.18	0.22
dQ	0.84	0.92	0.88
part	1.00	0.88	0.93
whQ	0.83	0.75	0.79
ynQ	0.73	0.65	0.69
tagQ	0.78	0.66	0.72

Table 16.5: Classification precision, recall, and F-measures of QBT form.

Feature Set	Accuracy
prosody: last 200 ms	44.3%
prosody: rhythmic	46.7%
prosody: loudness	46.9%
prosody: all	51.4%
prosody: pitch	51.5%
student and task	53.7%
none (majority class baseline: dQ)	54.9%
syntax	71.0%
lexis	79.8%
all feature sets combined	79.8%

Table 16.6: Performance accuracy of each feature set in predicting QBT form.

Ngram	altQ	dQ	partQ	whQ	ynQ	tagQ	N
"right #"	0	2	0	0	7	**91**	44
PRP+VB	5	3	1	25	**64**	2	99
PRP	4	38	0	10	38	10	442
UH+#	1	30	20	1	7	41	74
WP	2	20	1	**50**	23	5	106
MD+PRP	10	0	2	8	**79**	2	52
"or"	38	28	0	8	23	4	53
#+WP	3	3	0	**87**	7	0	30
UH	2	**50**	4	6	22	15	375
VB	4	37	0	11	36	11	313
"# huh"	0	0	**100**	0	0	0	9
WP+VBD	5	5	0	**90**	0	0	20
"you"	3	28	0	15	48	6	142
#+MD	6	3	0	3	**88**	0	32
CC	12	**52**	0	5	21	10	167
NN+UH	0	33	0	8	5	**54**	39
#+WRB	0	20	0	**75**	5	0	20
"# the"	1	**90**	0	3	6	0	68
"# is"	0	10	0	0	**90**	0	20
"how's"	3	26	0	44	26	0	34
"isn't"	0	10	0	5	**62**	24	21

Table 16.7: Normalized lexical and syntactic ngram counts for QBT form.

modal verbs and personal pronouns were considered to be important. In fact, a QBT containing a modal verb followed by a personal pronoun (MD+PRP) was found to be a yes-no question 79% of the time, (e.g., "**would**/MD **it**/PRP keep going"). A second person pronoun anywhere in the utterance was also considered to be informative, but not overwhelmingly so on its own. Almost half time the question was considered to be `ynQ` (e.g., "did **you** receive it") and a quarter of the time it occurred in dQ form (e.g., "**you** have greater constant acceleration"). A few other lexico-syntactic words were included in the most informative feature set as well, including "or," "how's" and "is/isn't."

There were a few rather unsuspected findings concerning the syntactic ngrams associated with different question forms. In particular, interjections (UH) arose quite often in association with both `dQ`s and `tagQ`s. When associated with the latter they appeared in the tag fragment in the form of "really" and "huh." With respect to `dQ`s, though, they were quite frequently interjections in the form of filled pauses (e.g, "**uh**/UH vertical zero" and "**oh**/UH the magnitude and direction"). It was also surprising to find that conjunctions (CC) and the base form of verbs (VB) should be useful, since one would assume those to be common to all types of utterances equally. Instead, we found that both were much more frequent in `dQ`s and `ynQ`s than they were in any other question form. Their presence might indicate that these two question forms are less likely to be fragments (and more likely to be complex sentences) than are the other forms of questions.

The previous tutor dialog act, though one of the most informative features in combination with the others we have looked at so far, did not seem to be particularly discriminating on its own (see Table 16.8). In fact, given most tutor dialog acts, students tended to respond overwhelmingly with questions in declarative form or—to a much lesser degree—in yes-no question form. There were a few notable exceptions, though. When the tutor provided the solution to the student (`bottom-out`) and the student then responded with a question, this question was equally likely to be `ynQ` or `dQ`. The same can be said for question forms following a tutor `directive`. When the tutor recapped an earlier conversation (`recap`), if the student replied with a question then it was most likely to be a wh-question. Furthermore, yes-no questions were the most likely question form after tutor `social-coordination` turns.

Tutor dialog act	altQ	dQ	partQ	whQ	ynQ	tagQ	N
bottom-out	7	43	0	7	**43**	0	14
deep-answer-question	3	68	1	8	16	4	176
expansion	0	53	0	15	27	5	59
hint	3	48	1	13	21	14	71
long-answer-question	5	65	4	5	15	5	55
negative-feedback	0	55	0	18	27	0	11
positive-feedback	4	64	0	7	25	0	28
recap	0	25	0	**50**	25	0	4
directive	0	36	6	18	**36**	3	33
restatement	3	50	3	6	25	14	36
social-coordination	4	25	6	14	**44**	6	77
short-answer-question	2	59	1	9	18	10	343

Table 16.8: Normalized counts of previous tutor dialog act (ptda) for QBT form labels.

16.3 QBT function

A 4-way classification task was designed to predict question function independent of question form. Similar to the aforementioned experiments for question form, we considered only those student turns we knew to contain questions. We excluded all QBTs that contained questions whose functions were ambiguous or indeterminate. This left us with data comprising 885 tokens with the following distribution: 521 chkQ (58.9%), 289 clarQ (32.7%), 56 infoQ (6.3%), and 19 othQ (2.1%).

As in all previous experiments, the best prediction accuracy was obtained when considering all features in combination. Precision, recall, and F-measures for each QBT function label are shown in Table 16.9 on page 142. As we can see, performance for chkQs was quite robust (F = 0.83), though it was less so for the other labels. The label with the lowest F-measure was othQ (F = 0.15). Mean F-measure was 0.51, 39% below the average

F-measure we observed for question form. Together, these results indicated that given our features, classifying question function was more difficult than classifying question form.

Listed in Table 16.10 are the performance accuracies of the different features sets. The most striking similarity here with respect to the results of QBT form classification was that lexical items were the most important predictor in both experiments. In fact, even though prosodic, syntactic, and student and task dependent features showed improvement over the baseline performance, they did not increase performance when combined with lexical features.

The number of features in the most informative feature subset was substantially larger than the number for question form classification, another indication of the increased difficulty in finding reliable cues to question function. In total, 39 features were found to be the most informative and least redundant. Not only was the subset large, but it was also diverse. Features from each feature set were represented: 19 lexical ngrams, 8 syntactic ngrams, 10 acoustic features, and 1 task dependent feature. In fact, the four most informative features were all from different features sets: previous tutor dialog act (`ptda`), wh-pronoun (WP), "what," and cumulative pause duration (`pause-dur-cum`). The important syntactic features recorded whether a wh-pronoun occurred at the beginning or end of a turn (presumably as an indication of wh-fronting), whether the turn contained a interjection, and more general syntactic categories such as nouns and adjectives. The most useful lexical features encoded both lexico-syntactic information (e.g., "# what," "what #," "do," "be," "how's that," "did") and lexico-pragmatic information (e.g., "you," "I," "repeat," "right #"), consistent with previous findings. A few domain-specific words arose as well, including "force" and "acceleration." Finally, all types of prosodic information were among the most informative acoustic-prosodic features; pausing, loudness, pitch range, and pitch slope were all included.

Given that classification of question form was more robust than question function, we decided to run an additional machine learning experiment to predict question function assuming we knew the question form. Surprisingly, F-measures did not increase substantially compared with those reported above when form was not considered as a feature. F-measures for `chkQ`, `clarQ`, and `othQ` increased by 2 percentage points to 0.85, 0.69, and 0.17, respec-

Function	Precision	Recall	F-measure
chkQ	0.79	0.87	0.83
clarQ	0.68	0.66	0.67
infoQ	0.70	0.25	0.37
othQ	0.29	0.11	0.15

Table 16.9: Classification precision, recall, and F-measures of QBT function.

Feature Set	Accuracy
prosody: last 200 ms	54.6%
prosody: loudness	54.9%
prosody: rhythmic	57.0%
prosody: pitch	58.1%
none (majority class baseline: chkQ)	58.9%
prosody: all	60.3%
student and task	61.7%
syntax	70.1%
lexis	75.5%
all feature sets combined	75.5%

Table 16.10: Classification accuracy of each feature set in predicting QBT function.

tively. Only the classification of `infoQ` was notable: an increase in F-measure of 22%, from 0.37 to 0.45.

Though we did not observe a dramatic improvement in question function classification by including question form as a feature, the size of the most informative feature subset was dramatically reduced. Fourteen (14) features, including question form, were among the most informative and least redundant. In order of decreasing importance, these were: question form, previous tutor dialog act (`ptda`), WP, `pause-dur-cum`, "how," "be," "mass," `db-end-mean`, "did you," "no #," "on," "what's the," NN+NN, and "should I."

The mean feature values per question function for the two acoustic features are shown in Table 16.11. It was very interesting to find that QBTs containing pauses greater than a second in length were indicative of `chkQ` and `infoQ` functions, whereas a total pause time of less than a second was associated with `clarQ` and `othQ` functions. Furthermore, the total time spent pausing in `chkQ`s was almost double what it was in `infoQ`s. It was not entirely clear why this would be, but it should be noted that this feature was not normalized for turn length so we cannot be sure whether there was something specifically salient about pauses or whether the real indicator was, in fact, turn length. Regardless, either would be an interesting cue to question function warranting further exploration. Also of potential future interest is why it appears that questions seeking confirmation (`chkQ`) were less loud near the end of a turn than were questions of other functionality.

Ngram	**chkQ**	**clarQ**	**infoQ**	**othQ**
`pause-dur-cum` (s)	2.58	0.66	1.35	0.71
`db-end-mean` (db)	50.08	53.61	53.13	53.95

Table 16.11: Mean non-normalized acoustic-prosodic feature values for QBT function.

The usefulness of most of the informative ngrams with respect to question function, listed in Table 16.12, have already been addressed, though they have not be attributed to specific question functions. QBTs containing clarification-seeking questions (`clarQ`) were those that most often contained wh-pronouns (e.g., "I'm sorry **what**/WP did you say") and the bigram "should I" (e.g., "**should I** answer this"). The bigram "what's the" most often occurred with information-seeking questions (e.g., "**what's the** definition of displacement"). Confirmation-seeking/check questions (`chkQ`) were associated with several lexical ngrams, some domain-dependent (e.g., "mass"), some lexico-syntactic (e.g., "be," "how"), and some a bit more interesting. For example, ending a turn with the word "no" can be considered a way for a student to explicitly ask the tutor to confirm his or her current understanding (e.g., "um I guess I am right or **no**"). Another interesting cue to `chkQ`s was the presence of compound noun phrases (NN+NN), which might be an indication that students tended to need verification of their understanding of complex ideas.

Ngram	**chkQ**	**clarQ**	**infoQ**	**othQ**	**N**
WP	8	**75**	16	1	99
"how"	36	18	45	0	33
"be"	**89**	8	3	0	76
"mass"	**92**	7	2	0	61
"did you"	7	47	47	0	15
"no #"	**71**	0	0	29	14
"on"	**86**	12	2	0	66
"what's the"	0	20	**80**	0	5
NN+NN	**78**	13	7	3	76
"should I"	0	**83**	0	17	6

Table 16.12: Normalized lexical and syntactic ngram counts per QBT function.

As was the case in all of our other classification experiments, the previous tutor dialog act was considered to be a useful feature for classification of question function. Table 16.13 lists the normalized dialog act counts per function label. We can see that a student QBT immediately preceded by any type of tutor question was most likely to be confirmation-seeking (which is also what we found for declarative questions). The same can also be said following tutor `restatement`s and `positive-feedback`. The remaining tutor dialog act counts were either spread evenly among the function labels or were too rare to generalize from.

Tutor dialog act	**chkQ**	**clarQ**	**infoQ**	**othQ**	**N**
`bottom-out`	38	46	8	8	13
`deep-answer-question`	**73**	26	1	1	168
`expansion`	48	30	16	6	50
`hint`	44	42	13	2	62
`long-answer-question`	**83**	17	0	0	53
`negative-feedback`	22	22	0	**56**	9
`positive-feedback`	**79**	17	4	0	24
`recap`	25	**50**	25	0	4
`directive`	44	44	11	0	27
`restatement`	**62**	24	9	6	34
`social-coordination`	14	47	33	6	70
`short-answer-question`	**70**	28	2	0	333

Table 16.13: Normalized counts of previous tutor dialog act (`ptda`) for QBT function labels.

Chapter 17

Discussion

The original motivation for examining student questions in the HH-ITSPOKE corpus was rather narrow. We wanted to automatically predict student questions, identifying both their form and function, so that we might use such predictions to improve the naturalness and relevance of automated tutor discourse in the ITSPOKE Intelligent Tutoring System. Along the way, we made several interesting discoveries that transgressed simple reporting of automatic classification performance. We found that these findings could be extrapolated to the structure of Spoken American English much more generally.

In relation to the types of questions found in naturally occurring speech, we found that their form and function were not one and the same, though they were not wholly unrelated either. It is true that all question forms were observed to serve (virtually) every function, though it is also true that some syntactic forms significantly favored some pragmatic functions. Clarification-seeking questions were found to be predominately declarative in form, though not infrequently were also manifested as yes-no and tag questions. Wh-, alternate, and particle questions very rarely served that function. These form/function correspondences were different from what we found for those questions that sought clarification from the tutor. They were approximately equally as likely to be declarative, yes-no, or wh-questions, but were seldom tag, alternative, or particle questions. Questions of information-seeking function displayed yet another pattern. Such questions were predominately yes-no and wh-questions and rarely took any other form. The fact that no consistent grouping of question forms held for multiple functions is a testament to the relative independence of the

two dimensions. However, knowledge about the most frequent associations between form and function, possibly dependent somewhat on the discoursal domain, likely informs felicitous discourse among participants and such knowledge can be of potential use to designers of ITSs.

Another indication of the influence of domain on question type distribution was the inordinately high frequency of a few question types in our corpus. Over half of all student questions were found to be declarative in form and over half were also found to be clarification-seeking in function. The intersection between the two accounted for 40% of all students questions. These are rates much higher than have been observed elsewhere and suggest that there is something about tutorial discourse that encourages this behavior. Again, knowledge of this phenomenon can help construct better ITSs.

Beun (2000) offers some insight into why declarative form might be used for interrogative purposes (p. 10):

> The use of a declarative form as a syntactic device for questioning is positively correlated with the certainty of the speaker's belief about the truth value of the propositional content of the question.... Hence, speakers often use a declarative for questioning if they have a weak belief about the content..... [Declarative questions] are often used when the information is literally provided in the dialogue.... So, if speakers do not express this evidence, they give the impression that relevant parts of the discourse were not well understood.... Note that also in normal circumstances it is considered inattentive to repeat the same question in an interrogative form if the question was already answered.... In this respect, the declarative question seems to fulfill the control function of a acknowledgment.

Many of these points speak directly to the role of the student in tutorial domain. In most cases the propositional content has been made explicit by the formal statement of a problem which is to be solved. Furthermore, hints and feedback are offered along the way by the tutor, such that the student can rarely claim that the necessary knowledge for solving the problem has not been mentioned in, or is inferable from, the previous discourse. The main task then becomes to apply a previously mentioned principle to the current problem. It would impress upon the tutor that the student had not been attentive were he or she to use

an interrogative form in relation to a previously discussed concept. In other words, both face-saving and politeness protocol might explain the high frequency of declarative questions in our corpus. Furthermore, if declarative questions seek an acknowledgment of a weakly held idea, as put forth by Beun, then this would also support the high rate of confirmation-seeking/check questions, which can be considered to function as acknowledgment-seeking.

A final note concerning declarative questions relates to phrase-final rising intonation. Beun has claimed that as certainty increases, so too does the use of declarative questions. However, this would seem at odds with the commonly held notion that phrase-final rising intonation is associated with uncertainty (e.g., Stenström, 1984; Šafářová, 2005), since most questions in our corpus—including those in declarative form—were found to have phrase-final rising intonation. As it so happens, we labeled the student turns in our corpus for perceived uncertainty for another study (Liscombe et al., 2005a) and were able to address this issue without having to conjecture. We found that declarative questions were twice as likely to convey certainty as would be expected given the distribution.[1] The only other question types perceived to convey certainty at rates higher than expected were tag questions and clarification-seeking/check questions, both of which have been shown to co-occur at high rates with declarative questions. We can say with a relatively high degree of confidence that rising intonation is not indicative of uncertainty, at least for our corpus, though it is very likely indicative of some other cognitive state or discoursal function (cf. Gunlogson, 2001; Steedman, 2003; Bartels, 1997; Gussenhoven, 1983; Pierrehumbert & Hirschberg, 1990).

Phrase-final rising intonation was found to be highly indicative of whether a question-bearing turn contained a question or not, much more so than either lexical, syntactic, or even pragmatic information. This finding seems also to be at odds with some of the previous research mentioned at the outset of this chapter and suggests that such studies have tended to dismiss the role of intonation as an indicator of question-status too readily. Though we found intonation to be critical for identifying the presence of a question, it was considered virtually useless at discriminating among the different forms and functions that questions can take. Not because phrase-final rising intonation was not present, but precisely

[1] Expected frequencies were computed using the χ^2 statistic: $\chi^2 = 102.7$, df $= 10$, p ≤ 0.001.

because it was present on virtually all questions. In such cases, lexical cues, especially words that encoded syntactic and pragmatic meaning, were the most critical for both tasks. These findings are not strictly at odds with previous studies, but show how analyzing several facets of observed question behavior in a corpus of spontaneously occurring speech can offer insight into the roles of different channels of communication. We should also reiterate that the tutor dialog act immediately preceding a student question was found to be highly informative in all tasks, thus underscoring the importance of discourse context, often discussed as a critical component of question identification.

A final note is order with respect to the comparisons that can be made between the findings of our study and those of previous empirical studies. Several of the studies cited have not been conducted on Standard American English (e.g., Geluykens, 1987; Beun, 1990) so findings that differ are not necessarily contradictory; they might simply be language specific. Also, few studies have conducted classification experiments that separated question form and function, so comparisons with them are similarly limited. The only direct comparison we can make is with the experiment conducted by Shriberg et al. (1998) on the automatic classification of questions vs. statements. Whereas they found that lexical information was more important than prosodic information in this task, we found the opposite to be true. It is not entirely clear why this would be, but it is most likely attributable to different discoursal domains, which would further highlight the importance of understanding the expected behavior of participants in different domains.

We were also able to compare our findings with other studies of student questions in the tutoring domain. The high rate of student questions per hour that we observed is consistent with previous reported measures. Our findings with respect to learning gain offer more than confirmation of previous studies. Whereas Graesser & Person (1994) showed that question type significantly correlated with examination score, the findings are weakened by the fact that no pre-test was administered, so high examination scores cannot be assured to correlate with the tutoring process. Nevertheless, our findings supported theirs, in part, in that one question form—yes-no tag—negatively correlated with learning gain. Furthermore, we observed that tutor directives followed by a student question-bearing turn also negatively correlated with learning gain. This finding is provocative in that it suggests that the

structure of tutoring discourse may play a part in learning and that an ITS might benefit from continuously updating anticipated learning gain given the current state of the dialog, and dynamically adapting in ways that might increase learning gain. It was also found, somewhat surprisingly, that several acoustic-prosodic features positively correlated with student learning gain. Such information could be monitored as well and used to guide the discourse of ITSs in similar ways.

For practical reasons, we chose to classify units of speech at the level of student turns instead of isolating student questions from such turns. We might have expected, therefore, that our classification performance would suffer because not all content in a question-bearing turn can be relied on to be part of a question. In such cases we extracted extraneous—and possibly contradictory—information from the utterance. Even so, there is a significant advantage to using turn units over question units. It is a difficult task in and of itself to automatically detect question regions, and errors in so doing could cause question classification accuracy to suffer. Based on our statistical and manual analyses of the questions by students, it is safe to say that when questions were asked in our corpus, they were the primary function of student turns. Most question-bearing turns contained only one question and most consisted entirely of the question itself. Furthermore, due to the nature of questions—that they seek a response—most questions occurred at the end of question-bearing turns, presumably because students were waiting for a response from the tutor. From this evidence, we conclude that when questions were asked, they operated as the primary function of a student turn, so there was actually little conflicting information in the turn. The relatively high performance accuracy we observed in our classification experiments further supported this conclusion, and gives us confidence that implementation of automatic detection of question form and function is an achievable goal in the ITS framework.

Part IV

PROFICIENCY

Chapter 18

Non-native Speaking Proficiency

There are thousands of human languages spoken in the world today, an indication that the world in which we live is still a multilingual one. A person's **native language** is learned very early on in life and is used to communicate with others in his or her speech community. However, one is capable, through practice, to learn other languages later in life. Such a language is usually referred to as a **second language**. In the study of language acquisition, a person's native language is referred to as **L1** and a second language is referred to as **L2**. It is an interesting and well-established fact that an L2 speaker is rarely perceived to have native-like proficiency by L1 speakers of that language. Indeed, the later in life that one begins to learn a new language, the harder it becomes to sound like a native speaker of that language. It is an equally intriguing fact that people are inherently capable of instantly assessing the proficiency of a speaker of their own L1.

Though people of different L1 speech communities have always been in contact with one another, modern technology has enabled virtually every member of a speech community to easily be in contact with a member of a different speech community. It should come as no surprise then, that research using modern technology includes studies that aim to assess and teach L2 proficiency automatically using computer-based methods.

Many of today's academic and business institutions require English as the primary means of communication. These same institutions often judge the proficiency of applicants by requiring standardized testing designed to assess command of the English language. Though traditionally devised to measure listening, reading, and writing skills, many language pro-

ficiency exams now offer speaking portions as well. Computer-based automated scoring of speech has several potential advantages. It promises to reduce the cost of language assessment, has the potential to yield more consistent assessment scoring, and could also be used as an essential component of computer-assisted English language-learning software, thus providing more people with access to quality instruction.

Chapter 19

Communicative Competence

Being a proficient speaker of a language demands more than proper grammar, vocabulary, and pronunciation. Though these are the three aspects of language proficiency that are usually taught in language courses, it is not uncommon for a person who has mastered these skills to nevertheless be perceived as a non-native speaker. This is because language use comprises more that syntactic and segmental properties. Discoursal, pragmatic, and sociolinguistic behavior are all critical components as well. In other words, how a language is used to accomplish different types of goals is an essential part of proficient language use.

A global view of language proficiency that includes all aspects of language use is referred to as **communicative competence** and was first put forth by Hymes (1971) and was later refined by Canale & Swain (1980) and Bachman (1990). This is the paradigm under which we are operating when we refer to spoken language proficiency in this chapter. The three dimensions of communicative competence are enumerated in Table 19.1. The first dimension, **language use**, includes aspects of language proficiency that have traditionally been emphasized when teaching an L2 and it includes grammar and vocabulary proficiency. The second dimension, **topic development**, can be thought of as the way in which a language structures discourse. The third dimension, **delivery**, includes segmental (**pronunciation**) as well as suprasegmental or prosodic behavior (**intonation**, **rhythm**, and **fluency**). We will be focusing on the delivery dimension of communicative competence in this chapter.

Pronunciation assessment concerns proficiency at the level of the phoneme. The goal of pronunciation assessment is to evaluate the spectral qualities of intended phonemes in

I. Language Use			
A.	*Vocabulary*	*B.*	*Grammar*
	Diversity		Range
	Sophistication		Complexity
	Precision		

II.	**Topic Development**
	Coherence
	Idea progression
	Content relevance

III.	**Delivery**
	Pronunciation
	Fluency
	Rhythm
	Intonation

Table 19.1: Dimensions of communicative competence for speaking proficiency.

order to determine how similar these qualities are to what a native speaker would produce. Most studies on automatic delivery assessment have focused on this sub-dimension of communicative competence, as we will show in the next section.

The term **fluency** can be defined in (at least) two ways. In its colloquial usage, it refers to global oral proficiency in all areas of speech production. As a technical term, however, it is often used by the L2 language assessment community to refer to temporal qualities and smoothness of flow in speech production—including phenomena such as speaking rate, segment duration, and pausological behavior. Filled pauses and hesitations are often included in this sub-dimension as well. Automatic assessment of the fluency proficiency of L2 learners has been studied to some degree in recent years and in this chapter we describe experiments we conducted in this area so that we might compare our findings with those studies. The remaining two sub-dimensions of delivery proficiency—**rhythm** and **intonation**—have been less studied with respect to automatic scoring techniques. These two sub-dimensions are the focus of the experiments presented in this chapter.

Rhythm—more specifically, rhythmic stress—describes the relative emphasis given to syllables in an utterance. English is known to be a rhythmically stress-timed language;

that is, stressed syllables appear at a roughly constant rate. Syllable stress is conveyed in English through a variety of means: pitch, loudness, duration, and spectral information. Rhythmic stress is not equivalent to lexical stress. In English, lexical stress refers to the fact that each word has underlying stress on at least one syllable in a word. Where the stress lies is unpredictable and must be learned. Clearly, lexical stress is an important part of communicative competence but is more appropriately grouped with pronunciation. This is not to say that there is no relation between lexical and rhythmic stress. In spontaneous speech, many syllables that bear underlying stress become destressed; however, the converse is not generally true: underlying unstressed syllables rarely become stressed.

Intonation describes how pitch varies in speech. Intonational languages, such as English, use pitch variation to signal several types of meaning change. In such languages, intonation can be used syntactically; for example, to distinguish questions, which may exhibit phrase-final rising intonation, from statements, which tend to have phrase-final falling intonation. It can also be used to structure discourse flow by signaling to the listener when to expect a change in topic or turn-taking. Intonation can even be used to distinguish between paralinguistic meaning such as emotion (a topic addressed in depth in Part I). Clearly, intonation is an important aspect of communicative competence and yet it is rarely explicitly taught to language learners and is, in fact, one of the last skills L2 speakers acquire. This should not indicate that it is not important for an L2 language learner to neglect this area of language proficiency, though. As Gass & Selinker (1994, p. 184) state, "Miscommunication resulting from native speaker perceptions of relatively proficient non-native speakers as opposed to learners with low level comprehension and productive skills is more often the more serious in terms of interpersonal relations because the source of the difficulty is attributed to a defect in a person (or a culture)." In other words, though the literal content of the message may be correctly communicated, low proficiency in suprasegmental delivery may yield negative perceptions of the speaker.

Chapter 20

Previous Research

Most studies on automatic delivery assessment have focused on pronunciation, with much success (Bernstein et al., 1990; Neumeyer et al., 1996; Franco et al., 1997; Cucchiarini et al., 1997, inter alia). In fact, there are even a few automatic proficiency assessment products on the market today, including EduSpeak (Franco et al., 2000) and PhonePass (Bernstein et al., 1990). With respect to the other sub-dimensions of delivery proficiency, however, not nearly as much work has been done. Table 20.1 presents a comparison of notable recent studies on the automatic assessment of L2 fluency, rhythm, and intonation.

All studies but one (Toivanen, 2003) have investigated frequency assessment, and most studies have looked at the delivery of spoken English (notable exceptions are Neumeyer et al., 1996; Cucchiarini et al., 2002; Morgan et al., 2004). Additionally, most studies have used data from one L1 language community, though these communities have ranged from study to study and have included Japanese, Finnish, Hungarian, and Swedish. Another important variable to consider is the type of speech data used for analysis. Overwhelmingly, researchers have chosen to use read speech for the reason that it is highly constrained and can be automatically segmented into words easily. Two studies (Kormos & Dénes, 2004; Xi et al., 2006) examined only spontaneous speech while two others (Bernstein et al., 2000; Cucchiarini et al., 2002) have looked at both read and spontaneous speech. Hincks (2005) used data from oral presentations wherein the speakers knew ahead of time what they planned to say, though they did not read from scripts.

There is a consensus among the studies that have explored fluency. It has been found

Study	Delivery			L1	L2	Speech
	flu	**rhy**	**int**			
Neumeyer et al. (1996)	✓			English	French	read
Franco et al. (2000)	✓			Japanese	English	read
Bernstein et al. (2000)	✓			various	English	read/spontaneous
Teixeira et al. (2001)	✓	✓	✓	Japanese	English	read
Cucchiarini et al. (2002)	✓			various	Dutch	read/spontaneous
Herry & Hirst (2002)	✓		✓	French	English	read
Toivanen (2003)			✓	Finnish	English	read
Kormos & Dénes (2004)	✓	✓		Hungarian	English	spontaneous
Morgan et al. (2004)	✓			unreported	Mandarin	read
Hincks (2005)	✓			Swedish	English	planned
Xi et al. (2006)	✓			various	English	spontaneous
Present study	✓	✓	✓	various	English	spontaneous

Table 20.1: Previous research on the automatic assessment of three sub-dimensions of the delivery proficiency of L2 speech: fluency (**flu**), rhythm (**rhy**), and intonation (**int**).

that proficient speakers have a faster speaking rate, longer stretches of speech between pauses (run length), and shorter pause lengths than do non-proficient speakers, regardless of L1 and L2. The reason proposed for these findings is usually described in terms of cognitive load. Non-proficient speakers have a higher cognitive load when accessing vocabulary and grammar rules than do proficient speakers and this increase in load is manifested as speech that is slower and that has longer and more pauses.

Most studies that have examined the automatic assessment of delivery proficiency have first extracted features from the speech signal and then have correlated these values with human assessment scores. Cucchiarini et al. (2002) found significant correlations between human raters and speaking rate (r = 0.6), run length (r = 0.5), pause length (r = -0.5), and pause rate (r = -0.3). Neumeyer et al. (1996) reported a correlation coefficient of 0.4

for speaking rate; Teixeira et al. (2001) observed a correlation coefficient of 0.4 for speaking rate and 0.5 for pause length; and Hincks (2005) reported a correlation coefficient of 0.4 for speaking rate and 0.5 for run length. Kormos & Dénes (2004) observed the highest correlation of fluency features and human assessment scores: 0.8 for speaking rate, -0.6 for pause length, and 0.9 for run length. It is difficult to compare absolute correlation values across studies because the factors between one study and the next can vary greatly by sample size, feature extraction method, L1 and L2, and the nature of human assessment scores. Nevertheless, it is clear that relatively simple fluency measurements can be extracted from the speech of non-native speakers that significantly correlate with human assessment scores.

The number of studies exploring reliable rhythm and intonation features that correlate well with human assessment scores have been few, and no consensus with respect to these sub-dimensions has yet emerged. Of the two, rhythm has been shown to be the most important. Kormos & Dénes (2004) found a correlation coefficient of 0.9 between the number of stressed words per minute and human assessment scores for Hungarian L1 speakers of English L2. Teixeira et al. (2001) approximated syllable stress by identifying the vowel of longest duration, the vowel with underlying lexical stress, and the vowel with the highest pitch. Using these stressed vowels as reference points, the author computed rhythm features by recording the stressed vowel length, the time between a stressed vowel and the previous and following intra-word vowels, and the time between the consecutive stressed vowels. No results were reported for the rhythm features alone, but they were shown not to increase correlation when combined with fluency features.

Teixeira et al. (2001) also examined intonation proficiency by extracting fundamental frequency ($f0$) measurements from their speech data as an approximation of global pitch behavior. Several measurements were calculated, including maximum $f0$, $f0$ slope, and $f0$ variation. The pitch feature set was found to have the weakest correlation with human raters but was found, when combined with the fluency features, to increase correlation by 7.7% over using fluency features alone.[1]

[1]Correlation using pitch features was 0.23, when using fluency features was 0.32, and when using both in combination was 0.34.

Herry & Hirst (2002) also approximated intonation by using a direct pitch modeling approach, but instead of using absolute values they compared $f0$ statistics of non-native speakers with productions of two native English speakers. The features included $f0$ range, $f0$ standard deviation, and $f0$ slope variation. They found that the intonation features were not as critical, as were the fluency features on all assessment tests except one that measured the "repetition quality" of read speech, in which case they found $f0$ range to be most useful.[2]

Toivanen (2003) took a different approach to the analysis of intonation proficiency of non-native speakers by adopting the "British school" of abstract intonation analysis. Tonic syllables were identified by their pitch and loudness peaks and then labeled as one of the following perceptual tones: fall, rise-fall, rise, fall-rise, fall-plus-rise, or level. The distribution of these tones were compared between two corpora of read British English: one produced by L1 speakers of British English and the other by L1 speakers of Finnish. The author found that the Finnish speakers produced far fewer rising tones, and far more falling and level tones than did the British speakers. It was also found that Finnish speakers produced rising intonation with yes/no sentence types and falling intonation with declarative sentence types far more often than did British speakers, indicating the importance of intonation in the pragmatic nature of communicative competence. Toivanen (2005) described work on further annotation of the data in terms of word-level stress, voice quality, and ToBI intonation labels, though no results have been reported on this to-date.

The approach we have taken to the automatic assessment of fluency, rhythm, and intonation proficiency builds on the fluency work of Xi et al. (2006) and is most similar to the rhythm approach of Kormos & Dénes (2004) and the intonation analysis of Toivanen (2003). Our work benefits from the fact that our corpus of non-native speech was unconstrained and spontaneously generated and that the speakers came from different speech communities. In this way our findings aimed to generalize across language backgrounds and to emphasize the communicative competence of non-native speakers in real-life speaking scenarios. The following sections detail our corpus, annotation scheme, human rater correlation, and prosodic event detection techniques.

[2]The authors reported no correlation tests for these results.

Chapter 21

The DTAST Corpus

The Test of English as a Foreign Language™ Internet-based Test (TOEFL iBT), developed and administered by the Educational Testing Service (ETS), measures the ability of non-native speakers of English to use and understand English as it is spoken, written, and heard in college and university settings. Today, there are more than 6,000 institutions in 110 countries that use TOEFL scores. The assessment test is designed for non-native English speakers at the 11th-grade level or above and is most frequently used to assess English proficiency before beginning academic work at an institution of higher education.

The TOEFL Academic Speaking Test (TAST) evaluates English speaking proficiency at intermediate to advanced proficiency levels. TAST presents integrated speaking tasks that simulate academic settings, just like the speaking portion of TOEFL iBT. TAST consists of six tasks. The first two tasks are independent tasks that require test-takers to talk about familiar topics. The other four tasks are integrated tasks that ask test-takers to listen to a short talk or to read a passage, and to then respond verbally to questions regarding these talks or passages. In total, the test takes about 20 minutes to complete and the data is recorded over the telephone. The audio data is recorded digitally at an 8 kHz sampling rate with 8 bit resolution. TAST is not an official test but a subset of the TOEFL iBT's speaking portion, which is used to evaluate test-takers' ability to communicate orally in English and prepare them for the TOEFL iBT and, more generally, for course work conducted in English.

Xi et al. (2006) described the TAST 2003/180 corpus, a set of responses from 180 test-takers of TAST collected in 2003. Each response was scored holistically by two trained

human raters on a scale from 1 (poor proficiency) to 4 (excellent proficiency). Additionally, responses from a subset of the subjects (80) were rated separately for delivery and language use proficiency. The scoring of different dimensions of communicative competence was undertaken to separate different aspects of proficiency. For example, a test-taker might have great delivery skills but poor language use skills, resulting in a mediocre holistic score but assessment ratings that differ on these two dimensions.

The corpus we used for the analyses reported in subsequent chapters was a further subset of the 80 test-takers of TAST 2003/180 corpus that were rated separately on delivery and language use dimensions. We refer to this corpus as DTAST because we were most interested in assessing the spoken delivery proficiency of the test-takers. DTAST comprised 87 responses distributed among 61 test-takers. We used only one task response from the majority of the speakers in our corpus (67.2%), though for some speakers we used two (9.8%) or three (23.0%) task responses. The self-reported native language (L1) of the test-takers varied considerably. The native languages of the test-takers were distributed among 12 distinct language families, though the most frequent L1s by far were Chinese dialects, which constituted 36% of all L1 speakers in DTAST.

Each task response in the corpus was rated by a trained human judge using the rubrics discussed above. For this analyses, we used only the delivery scores because doing so allowed us to study fluency, intonation, and rhythm proficiency while normalizing for language use and topic development dimensions of communicative competence. A subset of the responses (55) were rated by two raters. On this subset, we measured the rate of inter-rater agreement over chance using the quadratically-weighted Kappa statistic (κ_{qw}, defined in Section 5.5). Inter-rater κ_{qw} was found to be 0.72 and the correlation between human ratings was also 0.72. The mean ratings of both raters were 2.29 and 2.92; the standard deviations of both rater were 0.96 and 0.90. In all experiments we conducted that relied on human raters, we used only the ratings from the person who scored all of the DTAST corpus.

Each response in DTAST was hand-segmented into words and syllables. In total, there were found to be 9,013 words and 14,560 syllables. The shortest response was 19.5 seconds and the longest was 60 seconds—the maximum time allowed for an answer in TAST. Average response length was 48.0 seconds with a standard deviation of 10.7 seconds.

Chapter 22

Annotation

Intonation and rhythm were hand-labeled as sequences of abstract prosodic events aligned to syllables under the Tones and Break Indices (ToBI) framework (Beckman et al., 2005). ToBI is a system for transcribing intonation patterns and other aspects of the prosody of English utterances, though it has been adapted for other languages as well. Refer to Chapter 7 for a more thorough description of ToBI phrase accents and boundary tones.

22.1 Intonation

Since intonation is phonological in nature it is by definition a language-specific phenomenon. In other words, there are well-governed rules for the distribution of prosodic events that native speakers adhere to. Knowledge of these rules are utilized by the human labelers under the ToBI framework. In this sense, it is particularly difficult to label non-native speech with an abstract labeling system designed for native productions.

However, it has been shown that inter-labeler agreement is high for intonational phrase boundary tones and we hoped it would be the same when labeling non-native data.[1] Full intonation phrase boundaries were marked when strong perceptual prosodic juncture was detected.[2] The syllable immediately preceding each prosodic juncture was assigned one of

[1] Yoon et al. (2004) found inter-rater κ of phrase accent + boundary tones to range from 0.7 to 0.8.

[2] A "strong perceptual prosodic juncture" corresponds to a break index of value 4 in the ToBI framework.

the following phrase accent + boundary tone labels: `L-L%`, `L-H%`, `H-L%`, `H-H%`.[3]

Labeling of boundary tones was accomplished through quasi-consensus labeling by two labelers. One labeler annotated each of the files in the corpus. A second, more experienced annotator then checked each annotation and indicated where there was disagreement. The two labelers then conferred and reached a mutually agreed-upon label. We observed a κ value of 0.77 between the annotations before consensus, indicating that disagreements were acceptably rare. It should be noted that a much stricter test would be to measure κ between annotations that were truly independently labeled.

22.2 Rhythm

Syllable stress is not incorporated into the standard set of labels under the ToBI framework. This is due largely to the fact that it can be predicted in English based on the underlying lexical stress of each word and whether it bears a nuclear pitch accent or not (which is labeled in the ToBI framework). However this assumes, again, that speakers adhere to the rhythmic stress rules of English. With non-native speakers we could not assume this, so explicit labeling of syllable prominence was necessary.

Two levels of syllable prominence were labeled independently of phrase accents and boundary tones. Primary stress was marked with a `1` and secondary stress (less prominent than primary stress) was marked with a `2`. Unstressed syllables were left unmarked. In addition to listening to the speech, labelers also had access to visualization of the fundamental frequency, intensity, and waveform of the speech, as they did when labeling phrase accents and boundary tones.

One annotator labeled all of the stressed syllables in the corpus. A second annotated a subset of these (5 test-taker responses totaling 590 syllables). We found κ on the subset of the data labeled by two annotators to be 0.53 for full rhythmic stress and 0.71 when primary and secondary stress were conflated to a single category `stressed`. It has been noted that labeling of secondary (and tertiary) stress in spontaneous speech can be quite difficult and unreliable. Some linguists claim it may not even exist in spoken English (e.g.,

[3]Refer to Section 7.1 for a description of the ToBI annotation scheme.

	Label	Distribution	
stress	¬stress	11099	(76.2%)
	stress	3451	(23.8%)
tone	¬tone	12976	(89.1%)
	L-L%	878	(6.0%)
	L-H%	377	(2.6%)
	H-L%	243	(1.7%)
	H-H%	86	(0.6%)

Table 22.1: Distribution of stress (top) and phrase accents + boundary tones (bottom) in the DTAST corpus.

Ladefoged, 1993). For these reasons, we decided to adopt a binary annotation scheme for stress; a syllable was considered to be be either stressed (stress) or unstressed (¬stress).

Table 22.1 shows the label distribution in our corpus for stress and tone at the syllable-level. Most (76.2%) syllables bore no stress and most (89.1%) did not end at a phrase boundary. The most common phrase accent and boundary tone combination was L-L% (6.0%) and L-H% was observed nearly half as often (2.6%) as this tone. The remaining intonation labels—H-L% and H-H%—were observed less often (1.7% and 0.6%, respectively).

Chapter 23

Automatic Scoring Metrics

In this chapter we describe the features that were extracted based on the intonation and rhythm annotation. In order to determine whether these features were useful for scoring the English delivery skills of non-native speakers of DTAST, we used the Pearson product-moment correlation coefficient (r) to measure the linear relationship between the features and the delivery scores provided by the human raters.

23.1 Fluency

The fluency metrics calculated for the TAST 2003/180 corpus (Xi et al., 2006) were recalculated for the DTAST corpus and are described in Table 23.1 on page 167. A **silence** was defined to be any non-speech segment greater than 150 milliseconds, a **long pause** was identified as a silence greater than 500 milliseconds, and a **chunk** was any segment of speech without internal long pauses. Significant correlations between fluency measurements and human ratings are shown in Table 23.2 on page 167. All features except `longpstddv`, `silpsec`, and `dpsec` were found to significantly correlate with human ratings. These findings were consistent with other studies of the fluency of non-native speech: proficiency ratings increase as speaking rate (`wpsec`, `wdpchk`, `secpchk`) increases and both the rate and duration of silences and long pauses (`silpwd`, `silmean`, `longmn`) decrease. We also found that proficient speakers tended to be more consistent in their length of silences (`silstddv`) than less proficient speakers (though this was not found to hold for long pauses).

Variable	Description
silmean	Mean silence duration.
silstddv	Standard deviation of silence duration.
silpsec	Silences per second.
longpmn	Mean long pause duration.
longpstddv	Standard deviation of long pause duration.
silpwd	Number of silences per word.
wpsec	Words per second.
wdpchk	Words per chunk.
secpchk	Seconds per chunk.
dpsec	Disfluencies per second.

Table 23.1: Descriptions of the fluency features computed for DTAST, as proposed in Xi et al. (2006).

Variable	r	p
silpwd	-0.54	0.000
wpsec	0.51	0.000
wdpchk	0.48	0.000
silmean	-0.40	0.008
secpchk	0.33	0.000
longpmn	-0.25	0.008
silstddv	-0.25	0.009

Table 23.2: Significant correlations between delivery scores and fluency measurements.

23.2 Intonation

Intonation features were extracted in a way that quantized the distribution of prosodic boundary tones observed in the DTAST corpus. Table 23.3 on page 169 describes the intonation features under consideration. Recording the present or absence of phrase accent + boundary tone labels (`tonedist`) was the same as recording the distribution of intonational phrases, which we hypothesized to be an important delivery skill; however, this metric may be somewhat redundant with the fluency metrics that take note of pausing information, considering that most intonational phrase boundaries occur before a pause. However, we also hypothesized that metrics that accounted for the distributional properties of specific tones would also be a significant indicator of delivery proficiency.

Table 23.4 on page 169 shows significant correlations ($p < 0.05$) between human delivery scores and intonation features. We observed that the strength of the correlations were on par with the correlations observed for the fluency features; the average absolute correlation strength with human scores using the fluency features was 0.39, while it was 0.35 when using the intonation features. We also note that the features that described phrase-final rising intonation—`L-H%` and `H-H%`—made up most (62.5%) of the significant correlations. The message here appears to be that judicious use of rising intonation correlated with higher human proficiency scores. More specifically, the greater the distance between `L-H%` and `H-H%` tones (`LHdist` and `HHdist`, respectively) the higher the delivery score. Conversely, the lower the rate of `L-H%` and `H-H%` tones per syllable (`LHrate` and `HHrate`, respectively) the higher the score. It was also found that if there were too many `H-H%` tones relative to other tones (`HH2tone`) then this indicated a low delivery proficiency score. One possible explanation for the observation that frequent phrase-final rising intonation correlated with low delivery proficiency could be pragmatic in nature. Given that student answers to TAST questions should be affirmative and assertive, and that phrase-final rising intonation is thought to convey non-commitment and/or non-assertiveness in English (cf. Chapter 10), using rising intonation too often in this context might convey poor proficiency. These findings motivate future exploration of features that measure the relationship of phrase accent and boundary tone to syntactic or pragmatic sentence type, as has also been suggested by Toivanen (2003).

Another pattern emerged from the observed significant correlations as well: the seg-

Variable	Description
LLdist	Average number of syllables between successive L-L% tones.
LHdist	Average number of syllables between successive L-H% tones.
HLdist	Average number of syllables between successive H-L% tones.
HHdist	Average number of syllables between successive H-H% tones.
tonedist	Average number of syllables between any successive tones.
LLrate	Number of L-L% tones per syllable.
LHrate	Number of L-H% tones per syllable.
HLrate	Number of H-L% tones per syllable.
HHrate	Number of H-H% tones per syllable.
tonerate	Number of any and all tones per syllable.
LL2tone	Number of L-L% tones / number of all tones.
LH2tone	Number of L-H% tones / number of all tones.
HL2tone	Number of H-L% tones / number of all tones.
HH2tone	Number of H-H% tones / number of all tones.

Table 23.3: Descriptions of the intonation features computed for DTAST.

Variable	r	p
HHdist	0.53	0.025
LHdist	0.47	0.000
tonerate	-0.46	0.000
tonedist	0.38	0.000
HHrate	-0.29	0.002
HH2tone	-0.24	0.012
LHrate	-0.22	0.024
LLrate	-0.22	0.024

Table 23.4: Significant correlations between delivery scores and intonation features.

mentation of a stream of speech into too many intonational phrases (`tonerate`) that were too close together (`tonedist`) indicated lower proficiency scores. Due to the predominance of `L-L%` tones to other tone types in the corpus, we can say that the `LLrate` feature was indicative of this pattern as well, rather than of something inherent to the `L-L%` tone type in particular. In other words, the frequent display of strong prosodic juncture was indicative of low proficiency in spoken Standard Academic English.

23.3 Rhythm

Rhythm features, described in Table 23.5, were extracted in a way very similar to the intonation features. The two rhythm features recorded the distance between successive stressed syllables (`stressdist`) and the rate of stressed syllables (`stressrate`) in each task response. Table 23.6 shows significant correlations between human delivery scores and the rhythm features. The observed strength of the correlations here was much lower than that observed for the fluency and intonation features, though we can remark on a pattern similar to one already discussed. To some extent, proficient speakers were distinguished from non-proficient speakers by the fact that the former stressed syllables less frequently than the latter did. This pattern is analogous to the pattern found for phrase accent + boundary tone distribution, as discussed in Section 23.2.

Variable	**Description**
stressdist	Average number of syllables between successive stressed syllables.
stressrate	Number of stressed syllables / number of syllables.

Table 23.5: Descriptions of the rhythm features computed for DTAST.

Variable	**r**	**p**
stressrate	-0.26	0.007
stressdist	0.23	0.017

Table 23.6: Significant correlations between delivery scores and rhythm features.

23.4 Redundancy

Above, we established that there were distribution fluency, intonation, and rhythm metrics that significantly correlated with human-assigned delivery proficiency scores of the spoken answers of test-takers in the DTAST corpus. In this section we describe the redundancy of each feature set, as indicated by the significant pairwise inter-correlation of all features, in order to identify the strongest and least redundant features. We hypothesized that many features would encode redundant information.

Table 23.7 shows all significant inter-correlations ($p < 0.001$) within the fluency feature set. Of the 45 possible pairwise tests, 15 (33.3%) were found to significantly correlate. In other words, the features in the fluency feature set were redundant to some extent and, in fact, the average absolute correlation strength was found to be quite high ($M = 0.74$, $SD = 0.18$). Among the strongest correlations were those that measured information about silences and long pauses (e.g., the correlation coefficient for `silstddv` and `longpstddv` was 0.97). This implied here that it was not necessary for fluency assessment to distinguish between silences of different lengths. Another redundancy seemed to exist between mea-

	silmean	**silstddv**	**silpsec**	**longpmn**	**longpstddv**	**silpwd**	**wpsec**	**wdpchk**	**secpchk**
silstddv	0.85								
silpsec									
longpmn	0.81	0.96							
longpstddv	0.71	0.97		0.95					
silpwd	0.33		0.70						
wpsec			-0.54			-0.47			
wdpchk						-0.80	0.66		
secpchk			-0.73			-0.78		0.89	
dpsec									

Table 23.7: Significant ($p < 0.001$) correlations of fluency features.

surements based on seconds and those based on linguistic units, such as words (e.g., the correlation coefficient for seconds per chunk (`secpchk`) and words per chunk (`wdpchk`) was 0.89). Interestingly, the rate of disfluencies (`dpsec`) was not found to correlate with any of the other fluency features, indicating that it modeled fluency information that was truly independent of the other fluency features.

It was not surprising that many of the fluency features were found to be somewhat redundant, and it is not presented here as a way to be controversial. If the cost of calculating the features is minimal, which is true for the fluency features, then there is no harm in having redundant features. The analysis is presented as a way to compare subsequent examination of the redundancy of the intonation and rhythm feature sets, which have been far less studied.

We calculated correlations for all pairs of intonation features and noticed first and foremost that there was near perfect correlation between the features that measured specific phrase accent + boundary tone labels per syllable and the features that measured the ratio of specific phrase accent + boundary tone combination to all boundary tones: (`LLrate`, `LL2tone`) = 0.83, (`LHrate`, `LH2tone`) = 0.89, (`HLrate`, `HL2tone`) = 0.94, (`HHrate`, `HH2tone`) = 0.95. This can be explained by the distribution of the phrase accents and boundary tones observed in DTAST. Since the overwhelming majority of phrase accent + boundary tone labels were `L-L%`, both sets of features effectively measured the frequencies of phrase accent + boundary tone labels in a response. Thus, we have excluded `LL2tone`, `LH2tone`, `HL2tone`, `HH2tone` from subsequent discussion.

Of the remaining intonation feature pairs, shown in Table 23.8, we note that there were the same number of significant correlations as there were with the fluency features (15/45 = 33.3%), though the mean correlation (M = 0.55, SD = 0.18) was quite a bit lower than they were for the fluency features (M = 0.74, SD = 0.18). Taken together, we can state that the intonation feature set had at least some internal degree of redundancy as did the commonly used fluency feature set, though it was possibly more diverse in terms of representing the intonation sub-dimension.

	LLdist	LHdist	HLdist	HHdist	tonedist	LLrate	LHrate	HLrate	HHrate
LHdist									
HLdist									
HHdist			-0.97						
tonedist	0.52	0.43							
LLrate	-0.72				-0.67				
LHrate		-0.57			-0.44				
HLrate						-0.44	-0.33		
HHrate	0.33								
tonerate	-0.46	-0.44			-0.95	0.63	0.51		

Table 23.8: Significant ($p < 0.001$) correlations of intonation features.

In Table 23.8 we also note two very strong correlations: (`HHdist`, `HLdist`) = -0.97 and (`tonerate`, `tonedist`) = -0.95. The latter makes sense if we consider both the observed boundary tone rate and average distance between tones to be relatively stable. The former correlation is more perplexing, though. It is not entirely clear why the distance between `H-H%` and `H-L%` labels would correlate so strongly, though it was likely due to the sparsity of the tones in the corpus. We required there to be at least two of the same tone type in a response in order to calculate these measures and in all likelihood this happened rather infrequently for these tones, so this potentially skewed our calculations of these measures. This reasoning is further supported by the fact that we did not see similarly strong correlations for `L-L%` or `L-H%`, of which we had more instances.

The remaining correlations tended to mirror the correlation we saw between `tonerate` and `tonedist` to a much lesser degree, but for the same reason. For example, the correlation between `LLrate` and `LLdist` (-0.72) indicated that `L-L%` labels were distributed at regular intervals in the data. However, the fact that the correlation strength was weaker than the correlation strength between `tonerate` and `tonedist` implies that recording the

distributional properties of individual boundary tones was justified because their behavior was less predictable and, hopefully, more indicative of speaker proficiency.

The rhythm features (`stressrate` and `stressdist`) were highly correlated with one another (r = -0.97) indicating that they were redundant. However, when examining the extent to which each feature set was redundant with one another, we found that no rhythm feature significantly correlated with any other feature in the fluency or intonation feature sets. In other words, the rhythm features can be considered to have modeled delivery information independently of both fluency and intonation.

There were 140 correlations we might have observed between the fluency measurements and the intonation measurements (10 fluency features · 14 intonation features). We observed only 16 significant correlations, though, which are shown in Table 23.9. The fact that the overwhelming majority of correlations were *not* significant suggested that there was some general degree of independence between the fluency measurements and those we developed to measure intonation. This notwithstanding, speaking rate—as defined as the number of words per second (`wpsec`)—correlated with five of the intonation features. Furthermore, information about general boundary tone distribution (`tonedist` and `tonerate`) correlated

	dpsec	longpmn	silpwd	silstddv	wdpchk	wpsec
HHdist	0.85		-0.73			
HHrate				0.34		
HL2tone					0.36	0.49
HLrate						0.39
LHrate						-0.36
LLrate			0.31			
tonedist	0.39		-0.45			0.40
tonerate	-0.38	0.37	0.52	0.36		-0.40

Table 23.9: Significant correlations ($p < 0.001$) between intonation and fluency measurements.

with five fluency measurements as well.[1] Taken together, and noting that speaking rate is known to be one of the best correlates of fluency, this would seem to imply, as we hypothesized, that some intonation information need not be separately modeled from fluency when correlating with human raters. A closer look at the correlations of specific phrase accent + boundary tone labels seems to indicate otherwise, though. We note that `wpsec` negatively correlated with `tonerate` and `LHrate`, but positively correlated with `HLrate`. Forgoing intonation measurements might run the risk of discarding important distributional aspects of boundary tones. Also, most correlations, though significant, were relatively weak (from 0.3 to 0.4) considering that here we were using correlation to determine whether feature sets were redundant or not. Clearly, intonation and fluency features are redundant *to some extent*, but not so dramatically that we can say that they model the same information.

The strongest correlations, by far, concerned the average number of syllables between successive high rising boundary tones (`HHdist`). It is very intriguing to note that the distance between `H-H%` boundary tones positively correlated with the rate of filled pauses (`dpsec`) and negatively correlated with the rate of silences (`silpwd`). However, upon closer examination, we discovered that there were only 12 task responses that had more than one `H-H%` label in the DTAST corpus; in other words, in most of our data `HHdist` was actually undefined. With such a small sample it would be unwise to conjecture about the relationships concerning `H-H%` except to keep it in mind as a future avenue of exploration.

To summarize, many of the metrics we used to evaluate speaking proficiency encoded redundant information. Foremost among such redundancy was discrimination between the mean distance between phrase accent + boundary tone labels and the rate of these labels. Also redundant was recording the relationship between different phrase accent + boundary tone labels to one another. However, there were few significant correlations found between the fluency, intonation, and rhythm features sets, strengthening our belief that each feature set was representative of complementary aspects of delivery proficiency.

[1] Since `tonerate` and `tonedist` were highly inter-correlated (r = -0.95) the number of correlations here is less dramatic than would appear.

Chapter 24

Automatic Estimation of Human Scores

As has been shown by other studies for fluency, we demonstrated in preceding chapters that measurements of intonation and rhythm correlated with human ratings of the delivery of non-native spoken English. Furthermore, analysis indicated that the feature sets postulated as reflecting the intonation and rhythm sub-dimensions of delivery proficiency were independent from one another as well as from the commonly-proposed fluency metrics. Though these findings may shed light onto the sometimes intangible human rating process, the larger goal of our research was to estimate with high accuracy and reliability the human rating scores themselves so that the assessment of the communicative competence of non-native English speakers can be done automatically. This section describes experiments conducted to assess the degree to which human rating scores could be automatically assigned.

We chose to use the support vector regression method (Smola & Schoelkopf, 1998) as implemented in the WEKA software package because we wanted to classify human ratings as a continuous variable ranging from 1 to 4. We used a quadratic kernel and ran 5-fold cross validation. We report average performance on the five sets of held-out data in terms of both correlation coefficient (r) and quadratically-weighted Kappa (κ_{qw}) after rounding the predictions. Table 24.1 lists the performance of support vector regression using all combinations of feature sets. Our gold standard was the agreement of human raters (r = 0.72, $\kappa_{qw} = 0.72$).

Task	r	κ_{qw}
human agreement	0.72	0.72
fluency+intonation+rhythm	0.69	0.74
fluency+intonation	0.69	0.67
fluency+rhythm	0.70	0.67
intonation+rhythm	0.65	0.64
fluency	0.60	0.51
intonation	0.50	0.52
rhythm	0.26	0.20

Table 24.1: Performance measured as correlation and κ_{qw} of estimated and human delivery ratings.

We noticed somewhat different performance depending on whether correlation or κ_{qw} was used as the measurement of agreement between the automatically estimated ratings and those assigned by the human rater, though a global trend did emerge. When each feature set was considered in isolation, rhythm performed much worse (r = 0.26, $\kappa_{qw} = 0.20$) that both fluency and intonation. When considering κ_{qw}, fluency and intonation feature sets were comparable to one another ($\kappa_{qw} = 0.51$ and $\kappa_{qw} = 0.52$, respectively). However, when considering correlation, performance of the fluency feature set was substantially better than than performance of the intonation feature set (r = 0.60 and r = 0.50, respectively).

Both performance measures showed an increase in human rating estimation as feature sets were combined, though the exact nature of these improvements differed slightly. For all pair-wise feature set combinations, the sets containing the fluency features performed the best, and equally as well for both correlation (fluency+intonation = 0.69, fluency+rhythm = 0.70) and κ_{qw} (fluency+intonation = 0.67, fluency+rhythm = 0.67). The performances of the paired fluency+intonation and fluency+rhythm feature sets were substantially higher than intonation+rhythm (r = 0.65, $\kappa_{qw} = 0.64$) which, in turn, was higher than when using each feature set in isolation. Finally, The combination of all feature sets (fluency+intonation+rhythm) was shown to perform the best when measured by κ_{qw} (0.74),

though was not shown to be better than other feature sets when measured by correlation.

Despite the observed differences between correlation and κ_{qw} performance, the findings supported the following general assessment. We observed estimation of human delivery ratings at or above what we had observed for human agreement on the task. The fact that this was corroborated by two performance measures was reassuring because it signified that reaching human performance levels on the assessment of the delivery dimension of spoken English is indeed obtainable. Furthermore, we observed that the addition of either intonation or rhythm features increased performance over the use of fluency features alone. This supported our earlier hypothesis that using intonation and rhythm metrics is an important (in fact, essential) part of spoken language assessment. It remains to be seen how critical the rhythm features are. On their own, they were quite poor predictors of human ratings, though in combination with the other features they were considered useful. These findings warrant further investigation into the role of rhythm in spoken language proficiency.

A final caveat is in order. Unlike classification tasks in some other domains, when machine learning is used in the context of language assessment, care must be taken to ensure that the automatically-learned models reflect a pedagogical rubric, in this case the construct of communicative competence shown in Table 19.1. For example, before our models could be used to assess the delivery proficiency of non-native English speakers using TAST or TOEFL-iBT, we would need to verify that no single feature dominated the prediction algorithm, and that features were not used to drive scores in a direction opposite of what we would expect. Despite this cautionary message, we are optimistic that intonational and rhythmic assessment are critical for this task, and we feel that our experimental results have supported this claim. Future research would be to validate the models given these considerations.

Chapter 25

Automatic Detection of Prosodic Events

As the motivating force behind our research was to automate proficiency scoring of non-native English speakers, we must ultimately automate every part of the process. In the previous chapter we demonstrated how human ratings might be estimated using a machine learning approach utilizing fluency, intonation, and rhythm features. However, these features themselves must be automatically derived from the speech stream. Work on automatically calculating fluency features is quite robust. Most of the information necessary for these features can be obtained from state-of-the-art continuous speech recognizers (Franco et al., 2000; Cucchiarini et al., 2002; Xi et al., 2006; Zechner & Bejar, 2006, inter alia). For the automatic detection of prosodic boundary tones and stressed syllables we can also utilize speech recognition technology to segment and transcribe the data. However, one must go beyond the capabilities of this technology to fully model prosodic events.

The general framework we chose to use for automatic detection of prosodic events was the binary decision tree, a type of machine classifier especially good at predicting categorical data, such as boundary tones and stress.[1] Decision trees allow for non-linear learning and the ability to easily understand the types of decisions made in the learning process. For the experiments we describe here, some of our features relied on hand-labeled information;

[1]We used the Weka J4.8 implementation of C4.5 decision tress.

specifically, orthographic transcription and, as well as word and syllable alignments. Thus, our prosodic prediction results should be taken as upper bounds as a fully implemented automatic proficiency scorer must use predictions for this information as well.

25.1 Feature extraction

From each DTAST response, features were extracted based on several known cues to prosodic events. These features included acoustic, segmental, lexical, and syntactic information and are shown in Table 25.1 on page 181. The basic unit of analysis was the syllable, but many features comprised information about the word containing the syllable in question. For most features, information from surrounding syllables was taken into account as well, extending up to five syllables in the future and the past.

The largest group of features encapsulated acoustic information automatically extracted from the speech signal. As a known correlate of pitch, fundamental frequency (f0) was extracted using the speech analysis software PRAAT (Boersma, 2001) and the minimum (`f0-min-`σ), mean (`f0-mean-`σ), maximum (`f0-max-`σ), standard deviation (`f0-stdv-`σ), and range (`f0-range-`σ) were calculated for each syllable. Additionally, two approximations of pitch contour shape were calculated using linear regression of the pitch values against time. The first calculated a straight line through the data and the feature value recorded was the slope of this line (`f0-rslope-`σ). The second calculated a second-degree polynomial through the data and the feature value recorded was the coefficient of the quadratic term (`f0-curve-`σ). Finally, the ratio of voiced frames to unvoiced frames was calculated as an approximation of speaking rate (`f0-voiced-`σ).[2]

A second set of acoustic features were computed based on the intensity (measured in decibels) of the sound waveform; again, as estimated by PRAAT. This is known to be a correlate of perceptual loudness of speech. The statistics recorded for intensity were the mean (`db-mean-`σ), minimum (`db-min-`σ), and maximum (`db-max-`σ) over each syllable.

Segmental information of each syllable and word containing that syllable was also recorded. For syllables, this was the the phonemic transcription of the vocalic nucleus,

[2]Whether a frame was voiced or not was determined by the pitch detection algorithm in PRAAT.

	Feature	Description
Acoustic	`f0-min-`σ	minimum f0 of syllable
	`f0-max-`σ	maximum f0 of syllable
	`f0-mean-`σ	mean f0 of syllable
	`f0-stdv-`σ	standard deviation of f0 of syllable
	`f0-range-`σ	range of f0 of syllable
	`f0-rslope-`σ	slope of regression line through all f0 points
	`f0-curve-`σ	coefficient of quadratic regression curve of f0 against time
	`f0-voiced-`σ	percent of voiced frames in syllable
	`db-min-`σ	minimum intensity of syllable
	`db-max-`σ	maximum intensity of syllable
	`db-mean-`σ	mean intensity of syllable
Segmental	`trans-`σ	transcription of syllable nucleus
	`ms-`σ	duration of syllable
	`ms-w`	duration of word bearing syllable
	`ms-`σ`2w`	ratio of syllable duration to word duration
	`count-`σ`2w`	number of syllables in word
	`loc-`σ`2w`	location of syllable in word: {`initial, medial, final, entire`}
	`pos-w`	part of speech of word bearing syllable
Prosodic	`pstress`	stress label of previous syllables
	`ptone`	phrase accent + boundary tone of previous syllables
	`span-stress`	number of syllables between current σ and last with stress
	`span-tone`	number of syllables between current σ and last with tone

Table 25.1: Features extracted from each syllable in the DTAST corpus.

including lexical stress, as obtained from the CMU Pronouncing Dictionary based on the hand-labeled word transcriptions of each task response (`trans-`σ).[3] Note that orthographic transcriptions included demarcation of silences and filled pauses. Both the syllable and word durations, in milliseconds, were noted as well (`ms-`σ and `ms-w`, respectively). Other features in this set recorded the relationship between the syllable and word that bore the syllable: the number of syllables in the word (`count-`σ`2w`), the ratio of the duration of the syllable to the duration of the word (`ms-`σ`2w`), and the position of the syllable in the word (`loc-`σ`2w`).

The part of speech of each word, as automatically determined by a part of speech tagger (Ratnaparkhi, 1996) trained on Switchboard transcripts, was recorded to encapsulate syntactic information. Assigned to each syllable was the predicted part of speech for the word containing that syllable (`pos-w`).

Included as additional features were the corresponding values of all aforementioned features for of each of the five preceding and following syllables. For example, `db-mean-`σ_1 recorded the intensity of the syllable following the syllable under consideration. In this way, we hoped to contextualize information of the current syllable by those surrounding it.

A final set of features recorded information about stress and tone labels occurring previous to the syllable in question. In particular, these features measured the number of syllables and words since the observations of the last stress and tone boundaries (`span-stress` and `span-tone`). Also, the stress and phrase accent + boundary tone labels (as well as simply the presence or absence of each) were recorded for each of the previous five syllables (`pstress` and `ptone`). For example, `pstress`$_{-1}$ recorded whether the previous syllable bore stress or not.

Given the features enumerated in Table 25.1 and the features derived from exploiting a 5-syllable contextual window, there were 200 features associated with each syllable. Real-valued features were further contextualized by z-score normalizing per task response.

[3]The CMU Pronuncing Dictionary can be accessed here: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

25.2 Performance

We trained separate classifiers for stress and tone detection. Performance of each classifier is reported as the average of 5-fold cross-validation and is shown in Table 25.2. Expectedly, classification of the absence of stress (¬`stress`) and/or tone (¬`tone`) was quite reliable, in part because of the high frequency of these labels. Though not as good, the detection of stressed syllables (`stress`) was quite robust (F = .72). Tone detection did not perform as well. `L-L%` was the only boundary tone that was predicted reasonably well (F = 0.51). `L-H%` prediction had an F-measure of only 0.21 and both `H-L%` and `H-H%` were essentially not modeled with any degree of accuracy.

An examination of the induced decision trees provided some insight into the usefulness of our features. For stress detection, all features types were represented but varied with respect to which of the surrounding syllables were most useful. Intensity features of the current and surrounding syllables were heavily used, as was information about past stress and tone labels. Somewhat surprisingly, pitch information of surrounding syllables was never used. The duration of current and surrounding syllables was deemed important, as were the transcriptions of the current (but not surrounding) words and syllables. Finally, the part of speech of the current and following words were used often, as were the number of syllables in the current and five preceding words.

Label	Precision	Recall	F-measure
¬`stress`	0.91	0.92	0.91
`stress`	0.73	0.71	0.72
¬`tone`	0.95	0.99	0.97
`L-L%`	0.54	0.49	0.51
`L-H%`	0.33	0.16	0.21
`H-L%`	0.08	0.01	0.02
`H-H%`	0.06	0.01	0.02

Table 25.2: Precision, recall, and F-measure for each prosodic label in the corpus.

The features used for boundary tone detection were much more restricted. Little contextual information was used beyond the tone and stress labels of preceding syllables. In a more limited capacity, other useful features were the intensity and pitch measurements of the syllable.

Despite the fact that we did not achieve perfect performance when automatically detecting prosodic events, we still considered it important to observe whether the automatic intonation and rhythm scoring metrics of Chapter 23, based on the prosodic event predictions that were made, correlated with human delivery scores. Table 25.3 lists the significant correlations of predicted rhythm and intonation features with human delivery ratings. Also shown (reprinted from Tables 23.4 and 23.6 from Chapter 23) are significant correlations using hand-labeled intonation and rhythm and intonation features, for comparison.

It is somewhat surprising to have found that the strength of the observed correlations did

		Predicted		**Hand-labeled**	
	Feature	**r**	**p**	**r**	**p**
Intonation	tonedist	0.34	0.000	0.38	0.000
	tonerate	-0.31	0.001	-0.46	0.000
	HH2tone	-0.29	0.002	-0.24	0.012
	HHrate	-0.29	0.002	-0.29	0.002
	LLrate	-0.20	0.034	-0.22	0.024
	LL2tone	0.27	0.005	—	—
	LLdist	0.23	0.016	—	—
	HHdist	—	—	0.53	0.025
	LHdist	—	—	0.47	0.000
	LHrate	—	—	-0.22	0.024
Stress	stressrate	-0.19	0.050	-0.26	0.007
	stressdist	0.20	0.038	0.23	0.017

Table 25.3: Comparison of delivery correlations between predicted and hand-labeled intonation/rhythm correlations.

not differ as much as might be expected given the imperfect prediction of prosodic events. The mean absolute correlation strength using predicted prosody was 0.26 (SD = 0.05) and was 0.33 (SD = 0.11) when considering hand-label prosody. Though, as the standard deviations indicate, the correlations for predicted prosody were all relatively the same, whereas when using the hand-labeled prosody, there was more variation. In other words, it seems that we lost some of the stronger correlations when predicting prosodic events. We will examine the possible impact of this shortly.

Though there was less variation with respect to the correlations based on predicted prosodic events, the relative correlation strength was nevertheless maintained. The `tonedist` and `tonerate` features correlated most strongly with the human ratings; `stressdist` and `stressdist` showed the weakest correlations. It should also be noted that the significance factor (p) for the correlations using predicted prosody were slightly higher, on average, than they were for the correlations using hand-labeled prosody. This indicates that the correlations themselves were slightly less significant.

In the hand-labeled scenario, correlations involving boundary tones other than `L-L%` were found to correlate with human ratings, whereas this was largely not the case with the prediction scenario. In fact, `HHrate` and `HH2tone`, based on predicted prosody, were the only metrics that referred to specific tone labels and that were found to have a significant correlation coefficient. Additionally, `L-L%` boundary tones were represented more often in the prediction scenario than they were in the hand-labeled scenario. These two findings were most likely due to the fact that boundary tones other than `L-L%` were not reliably predicted in the prosodic event classification task. This was probably the most egregious shortcoming of our prosodic event detection, considering our earlier suggestion that non-`L-L%` boundary tones might be an essential aspect of proficiency assessment of non-native English speakers.

In keeping with our previous experiments using hand-labeled prosodic events, we conducted machine learning experiments to automatically predict human delivery ratings given predicted prosodic events. We ran 5-fold cross-validation using support vector regression for all combinations of feature sets. The results of these experiments can be seen in Table 25.4. Also shown are the results of the same experiments using delivery features derived from

	Predicted		Hand-labeled	
Task	**r**	κ_{qw}	**r**	κ_{qw}
human agreement	0.72	0.72	0.72	0.72
fluency+intonation+rhythm	0.69	0.74	0.66	0.66
intonation+rhythm	0.65	0.64	0.48	0.46
fluency	0.60	0.51	0.60	0.51
intonation	0.50	0.52	0.28	0.30
rhythm	0.26	0.20	0.09	0.09

Table 25.4: Performance of predicted and human delivery ratings using predicted and hand-labeled prosody labels.

hand-labeled prosody, for comparison (reprinted from Table 24.1). Fluency feature values were the same across scenarios because we used hand-labeled word segmentation for all experiments.

Human score estimation using intonation and rhythm features (in isolation) in the predicted scenario suffered greatly compared with the same features in the hand-labeled scenario. Score estimation, as measured by κ_{qw} between estimated and actual human ratings, dropped from 0.20 to 0.09 when using the rhythm features and from 0.52 to 0.30 when using intonation features. Combining the two feature sets (intonation+fluency) increased κ_{qw} to 0.46, but this was still well below the 0.64 value we saw when using hand-labeled prosodic events. Reassuringly, we note that score estimation when using all features combined (fluency+intonation+rhythm) was better than when using only the fluency features. In other words, though predicting prosodic events degraded our ability to predict human delivery ratings, intonation and rhythm features still contributed to the improvement of human score estimation to a level almost comparable to the agreement we observed between the two human raters on the task.

Chapter 26

Discussion

Intonation, rhythm, fluency, and pronunciation are believed to be independent factors in the delivery dimension as set forth by the construct of communicative competence. While there has been substantial empirical research concerning fluency and pronunciation of non-native English speakers within the assessment community, little research has been conducted with respect to intonation and rhythm. Our work has provided valuable insight into these dimensions by validating their usefulness and independence in the automatic scoring of delivery proficiency. In the DTAST corpus, we showed that these phenomena, for the most part, modeled distinct information from that encoded by well-known fluency correlates, and yet correlated equally strongly with human scores. Furthermore, though fluency measurements were somewhat better at automatically estimating human scores, using a combination of all delivery features together approached the agreement of human raters on the task. What remains, and is planned for future work, is to validate that automatically-learned assessment models use features in a way consistent with the human scoring rubric.

It was also shown that prediction of prosodic events in non-native speech can be achieved to some extent using some automatically derived information, such as pitch and energy. Furthermore, it was reassuring to observe that significant correlations of delivery features derived from these predicted prosodic events still significantly correlated with human ratings in a similar (albeit weaker) manner. However, low prediction accuracy of some prosodic events, especially boundary tones, indicated that there is still more work to be done in this area. A decreased ability to predict human rating scores also supported this position. One

solution may be to model boundary tones independently or use graphical models that find the best sequence of tones given acoustic and segmental information. Also, with respect to segmentation, our feature set relied heavily on hand-labeled segmentation and transcription. It remains future work to evaluate our method using speech recognition technology that automatically predicts this information.

We should strive in particular to predict phrase-final rising intonation more accurately, as it appeared that distributional features quantifying these tones dominated the correlations with human raters. We conjecture that these correlations are due to the inadvertent signaling of non-commitment or non-assertiveness by test-takers through the use of rising intonation at inappropriate times. One of our future goals is to encode such information into the delivery features by recording the rate of tones given pragmatic and syntactic sentence types. We also believe that a similar technique might yield more informative rhythmic features as well by recording the rate of different parts of speech that are stressed (e.g., function words should not generally be stressed).

There are two distinctly different directions one could take with this research, each motivated by a different goal. The first direction would be to continue the line of research we have set forth in this paper. Though our focus was rating the communicative competence of non-native speakers of English, we were also very interested in modeling behavior that could be used for diagnostic purposes in language-learning software. The delivery features we have described here could be of use to a foreign language learner by informing them of general tendencies that lead to low proficiency perception. For example, students could be informed that they are pausing too often or stressing too many words. However, our approach of modeling prosodic events actually allows us to be even more specific by, ideally, pinpointing areas of poor proficiency. For example, a student might be told that the use of a particular boundary tone on a specific utterance is not what a proficient speaker might do. Of course, before reaching this point one would need to ensure reliable prediction accuracy of prosodic events, and this remains a major challenge.

The second direction would forgo diagnosis and strive only to increase scoring accuracy. There may be a more direct way of measuring intonation and rhythm that does not rely on segmentation or transcription of any kind, thus decreasing potential sources of noise.

Instead of modeling discrete prosodic events, one might directly model the intensity and fundamental frequency of speech in a fashion similar to acoustic modeling for pronunciation evaluation. Though the output of such models might be more predictive of human ratings, they would most likely be of little use to the language-learner.

Part V

CONCLUSION

In this thesis we explored the role of prosody in conveying three different yet related types of speaker state: emotion, question-status, and non-native language proficiency. We addressed the automatic classification of each type of speaker state using statistical machine learning approaches, as our ultimate goal was to enable artificially intelligent applications to monitor user state and respond appropriately. The most applicable application for the work we have conducted are those related directly to the aforementioned topics; namely, Spoken Dialog Systems, Intelligent Tutoring Systems, and Computer Assisted Language Learning. However, in light of the fact that humans react to computers socially (Reeves & Nass, 1998), virtually any application might benefit from knowing more about the mental state of its user.

In the first part of this thesis, we explored affective speaker state by characterizing the prosodic cues present in emotional speech. Discrete emotions were found to be most successfully characterized when they were defined using a perceptual, polythetic labeling typology. As per past studies, global acoustic characteristics were found to be characteristic of the activation level of emotion. Intonational features describing pitch contour shape were found to further discriminate emotion by differentiating positive negative emotions. A procedure was described for clustering groups of listeners according to perceptual emotion ratings that fostered further understanding of the relationship between prosodic cues and emotion perception. In particular, it was found that some raters systematically judged positive emotions to be more positive and negative emotions to be more negative than did other raters. Furthermore, automatic prediction of the ratings of each rater cluster performed better than when the raters were not clustered.

The role of prosody in signaling the form and function of questions was explored in the second part of the thesis. Student questions in a corpus of one-on-one tutorial dialogs were found to be signaled primarily by phrase-final rising intonation. Moreover, over half of all student questions were found to be syntactically identical to declarative statements. It could therefore be inferred that intonation was of use for differentiating the pragmatic force of sentence type. Lexico-pragmatic and lexico-syntactic features were found to be of particular importance for further differentiating the form and function of student questions.

In the final part of the thesis we explored the role of prosody in communicating the spo-

ken language proficiency of non-native English speakers. Intonational features, including syllable prominence, pitch accent, and boundary tones were found to correlate with human assessment scores to the same degree that more traditional fluency metrics have been shown to do. Additionally, it was found that these different aspects of speaking proficiency operated relatively independent of one another and that the combination of all three correlated best with human proficiency scores.

The common thread tying these explorations together has been the assumption that prosody is informative in signaling temporary internal (metacognitive) states of speakers. We have addressed three such states and have shown that automatic identification is possible using shared techniques. It is our belief that by modeling internal speaker state, human-computer interaction can be improved. Furthermore, we believe that this is a promising future direction of Spoken Dialog Systems. It is our hope that one day we arrive at a unified model of the acoustic-prosodic cues to a generalized metacognitive state, and that we can design automated agents that respond appropriately.

Part VI

APPENDICES

Appendix A

CU_EPSAT Corpus

The CU_EPSAT corpus is a subset of 44 utterances selected from the EPSAT corpus (Liberman et al., 2002). The neutral utterances plus the three tokens selected as the control for the web surveys were the following:

new label	**old label**	**actor I.D.**	**start ms**	**transcript**
neutral	neutral	MM	774.20	two thousand three
neutral	neutral	GG	238.99	two thousand nine
neutral	neutral	CC	42.26	two thousand one
neutral	neutral	CL	69.39	two thousand ten
control	boredom	MF	2994.15	five hundred one
control	elation	JG	1336.72	three thousand ten
control	hot anger	MK	1113.17	fifteen hundred

The positve emotion utterances were:

new label	**old label**	**actor I.D.**	**start ms**	**transcript**
confident	pride	MM	3010.07	ten thousand ten
confident	pride	GG	2130.85	five thousand one
confident	pride	CC	2586.16	two hundred eight
confident	pride	CL	1524.62	nine thousand four
encouraging	elation	MM	2212.71	eight hundred ten
encouraging	elation	GG	1529.11	march seventeenth
encouraging	happy	CC	1943.35	september first
encouraging	happy	CL	910.42	eight hundred three
friendly	happy	MM	2353.51	three hundred nine
friendly	happy	GG	1672.49	four hundred nine
friendly	happy	CC	1940.11	may twentieth
friendly	happy	CL	899.59	march thirtieth
happy	elation	MM	2227.76	nineteen hundred
happy	happy	GG	1691.56	six thousand five
happy	happy	CC	1995.58	six thousand five
happy	happy	CL	915.05	five thousand eight
interested	interest	MM	2468.49	four thousand six
interested	interest	GG	1775.43	june thirtieth
interested	interest	CC	2126.93	one thousand ten
interested	interest	CL	1020.42	one hundred four

The negative emotion utterances were:

new label	**old label**	**actor I.D.**	**start ms**	**transcript**
angry	hot anger	MM	1505.31	november third
angry	hot anger	GG	973.14	ten thousand nine
angry	cold anger	CC	1442.35	three thousand four
angry	hot anger	CL	424.91	september fourth
anxious	anxiety	MM	3273.15	may twenty third
anxious	anxiety	GG	805.95	august sixteenth
anxious	anxiety	CC	961.15	six thousand twelve
anxious	anxiety	CL	339.40	five thousand three
bored	boredom	MM	2614.58	april fifteenth
bored	boredom	GG	1890.46	april fifteenth
bored	boredom	CC	2295.60	two hundred one
bored	boredom	CL	1085.60	august thirteenth
frustrated	contempt	MM	3167.00	six hundred three
frustrated	contempt	GG	2235.51	may eleventh
frustrated	contempt	CC	2769.96	october third
frustrated	cold anger	CL	524.77	nine hundred nine
sad	sadness	MM	2093.41	six thousand ten
sad	sadness	GG	1421.40	six thousand ten
sad	sadness	CC	1710.61	five hundred five
sad	sadness	CL	729.86	eight thousand four

Appendix B

Mean Feature Values Per Emotion

This appendix lists the quantized z-scored feature values given several labeling schemes. The first labeling scheme is the orthogonal intended emotion labels of the full EPSAT corpus. The remaining three use the perceived emotion labels of the CU_EPSAT corpus: the first uses the majority label given all the raters and the second two use the majority label given the raters in one of two clusters.

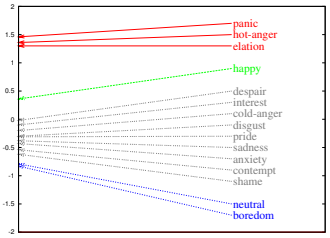
B.1 EPSAT intended

The mean z-scored feature value for each EPSAT emotion is listed below:

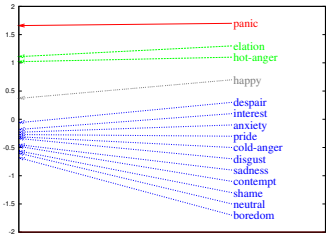
	db-max	db-mean	db-min	db-range	f0-max	f0-mean	f0-min	f0-rising	f0-range	f0-curve	f0-slope	f0-voiced
anxiety	-0.69	-0.43	0.14	-0.64	0.02	-0.43	-0.23	-0.28	0.09	0.18	0.21	-0.23
boredom	-0.42	0.03	0.51	-0.72	-0.21	-0.83	-0.68	0.40	0.03	0.14	0.12	-0.08
cold-anger	0.16	-0.23	-0.22	0.30	-0.27	-0.20	-0.31	-0.13	-0.11	-0.11	0.12	-0.31
contempt	-0.29	-0.39	-0.07	-0.17	-0.25	-0.54	-0.48	-0.37	-0.08	0.23	0.00	-0.13
despair	-0.12	0.09	0.29	-0.31	0.15	-0.02	-0.06	-0.21	0.13	0.21	-0.07	-0.03
disgust	0.39	0.34	0.10	0.23	-0.04	-0.30	-0.34	0.05	0.08	-0.06	-0.15	-0.18
elation	0.73	0.62	-0.44	0.81	0.27	1.30	1.11	0.15	-0.14	-0.28	-0.02	0.60
happy	-0.09	-0.34	-0.53	0.35	-0.08	0.36	0.37	0.15	-0.19	-0.32	-0.16	0.26
hot-anger	0.81	-0.08	-1.32	1.69	0.40	1.36	1.02	0.01	0.07	-0.36	-0.37	-0.02
interest	-0.17	0.08	0.30	-0.35	-0.01	-0.10	-0.18	0.45	0.07	0.10	0.23	-0.02
neutral	-0.27	0.30	0.64	-0.72	-0.11	-0.79	-0.60	-0.63	0.07	0.11	-0.05	0.45
panic	0.80	0.41	-0.19	0.78	0.29	1.46	1.66	-0.17	-0.34	-0.52	-0.57	0.19
pride	-0.05	-0.19	-0.05	-0.04	-0.11	-0.31	-0.27	0.09	-0.01	0.17	0.03	0.04
sadness	-0.54	-0.14	0.31	-0.63	0.16	-0.38	-0.45	0.07	0.31	0.31	0.29	-0.06
shame	-0.12	0.17	0.65	-0.60	-0.17	-0.62	-0.56	0.12	0.05	0.14	0.28	-0.27

The following plots show where each emotion lies in z-score space given its mean feature value. Natural classes of emotions are identified by grouping statistically different emotions based on the results of t-test at a significance of $p < 0.01$.

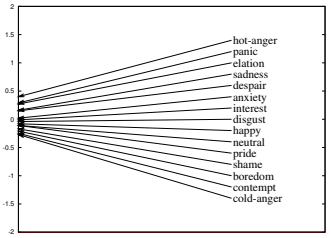
f0-mean



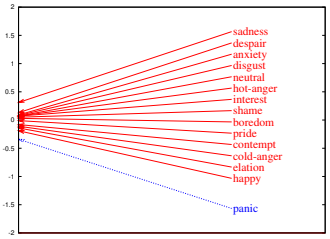
f0-min



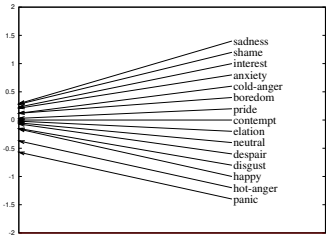
f0-max



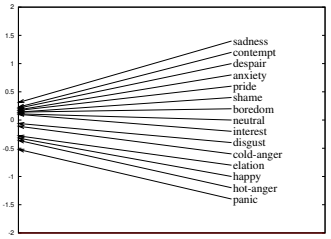
f0-range



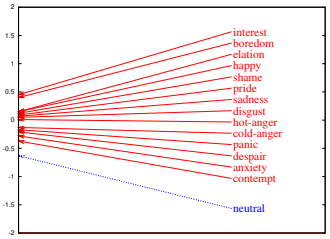
f0-slope



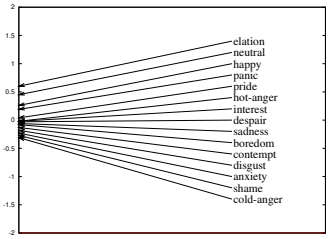
f0-curve



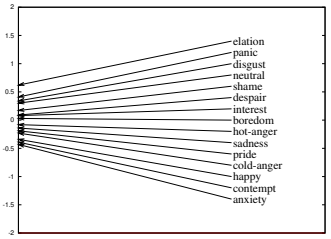
f0-rising



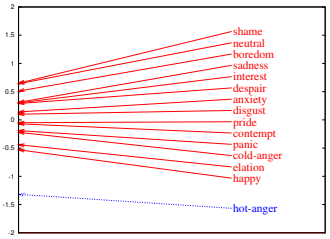
f0-voiced



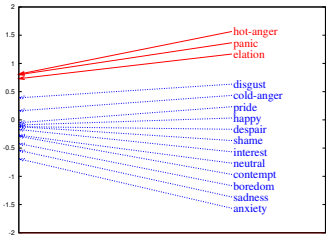
db-mean



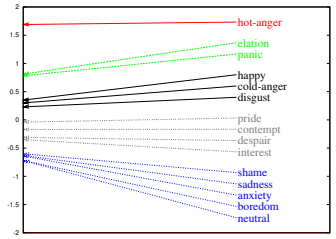
db-min



db-max



db-range



B.2 CU_EPSAT perceived: All raters

The mean z-scored feature value for each CU_EPSAT emotion is listed below.

	db-max	db-mean	db-min	db-range	f0-max	f0-mean	f0-min	f0-rising	f0-range	f0-shape	f0-slope	f0-voiced	db-tilt	nuc-tilt
happy	0.17	-0.14	-0.35	0.37	-0.10	0.16	0.30	0.51	-0.15	-0.54	-0.59	0.05	-0.02	-0.18
sad	-0.43	-0.10	0.34	-0.60	0.24	-0.62	-0.51	-0.43	0.32	0.50	0.37	-0.01	-0.14	-0.19
frustrated	0.38	0.00	0.06	0.22	0.32	-0.13	-0.16	-0.53	0.38	0.02	0.23	-0.10	0.22	0.11
confident	0.23	-0.02	-0.15	0.29	-0.18	-0.01	0.12	0.18	-0.16	-0.43	-0.14	0.04	0.17	0.13
friendly	0.00	-0.06	-0.28	0.24	-0.05	0.22	0.31	0.50	-0.10	-0.19	-0.13	0.15	-0.13	-0.25
interested	0.17	0.12	-0.03	0.14	-0.00	0.28	0.27	0.23	-0.07	-0.12	-0.08	0.08	-0.07	-0.08
angry	0.65	-0.03	-0.63	1.00	0.12	0.42	0.48	-0.49	0.07	-0.03	0.06	0.18	0.75	0.71
encouraging	0.26	-0.16	-0.29	0.40	-0.00	0.25	0.18	0.61	-0.06	-0.53	-0.33	0.16	0.03	-0.40
anxious	0.14	0.40	0.16	0.05	0.20	0.36	0.34	-0.39	0.12	0.28	0.16	-0.18	-0.80	-0.53
bored	-0.39	-0.12	0.52	-0.69	-0.09	-0.86	-0.61	-0.19	0.06	0.10	0.20	0.09	0.00	0.28

B.3 CU_EPSAT perceived: Cluster 1

The mean z-scored feature value for each CU_EPSAT emotion is listed below.

	db-max	db-mean	db-min	db-range	f0-max	f0-mean	f0-min	f0-rising	f0-range	f0-curve	f0-slope	f0-voiced	db-tilt	mfc-tilt
happy	0.03	-0.09	-0.07	0.03	0.01	0.03	0.12	0.59	-0.01	-0.42	-0.40	0.25	0.13	-0.15
sad	-0.29	-0.12	0.40	-0.63	0.26	-0.54	-0.53	-0.42	0.33	0.59	0.50	-0.16	-0.19	-0.24
frustrated	0.01	-0.20	0.09	-0.10	0.27	-0.11	-0.18	-0.45	0.31	0.24	0.37	-0.12	0.24	0.16
confident	0.09	-0.07	-0.11	0.19	-0.16	0.07	0.12	0.04	-0.17	-0.24	-0.07	0.01	0.06	0.09
friendly	0.04	-0.04	-0.16	0.14	-0.05	0.21	0.17	0.49	-0.10	-0.24	-0.17	0.10	-0.16	-0.19
interested	0.20	0.09	-0.17	0.26	-0.15	0.19	0.19	0.10	-0.18	-0.17	-0.09	0.13	0.01	-0.01
angry	0.65	-0.03	-0.63	1.00	0.12	0.42	0.48	-0.49	0.07	-0.03	0.06	0.18	0.75	0.71
encouraging	0.03	-0.09	-0.07	0.03	0.01	0.03	0.12	0.59	-0.01	-0.42	-0.40	0.25	0.13	-0.15
anxious	-0.02	0.22	0.26	-0.14	0.34	0.06	-0.01	-0.56	0.33	0.37	0.38	-0.17	-0.85	-0.54
bored	-0.56	-0.22	0.51	-0.79	-0.01	-0.68	-0.61	-0.20	0.13	0.21	0.14	-0.03	0.06	0.31

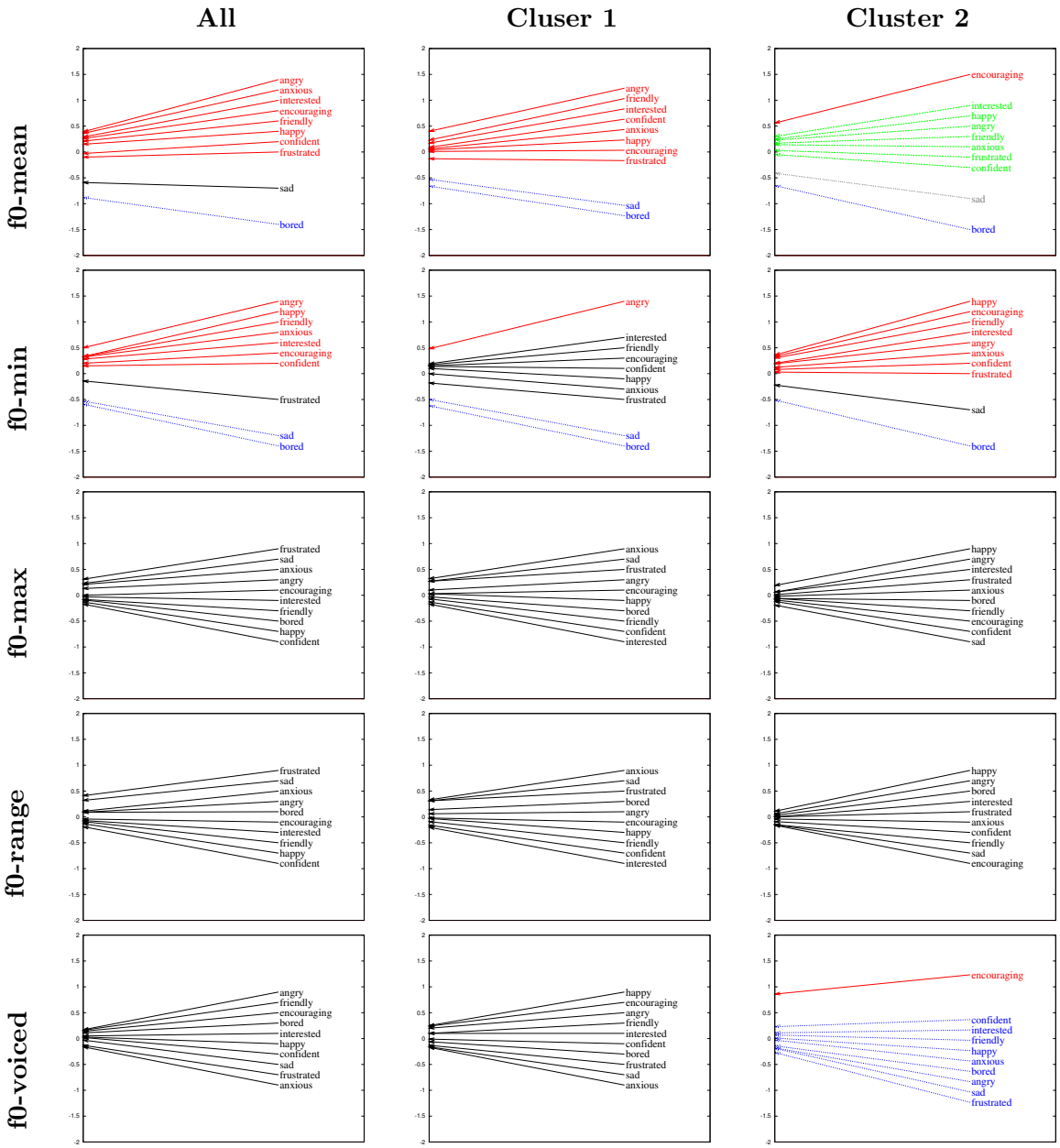
B.4 CU_EPSAT perceived: Cluster 2

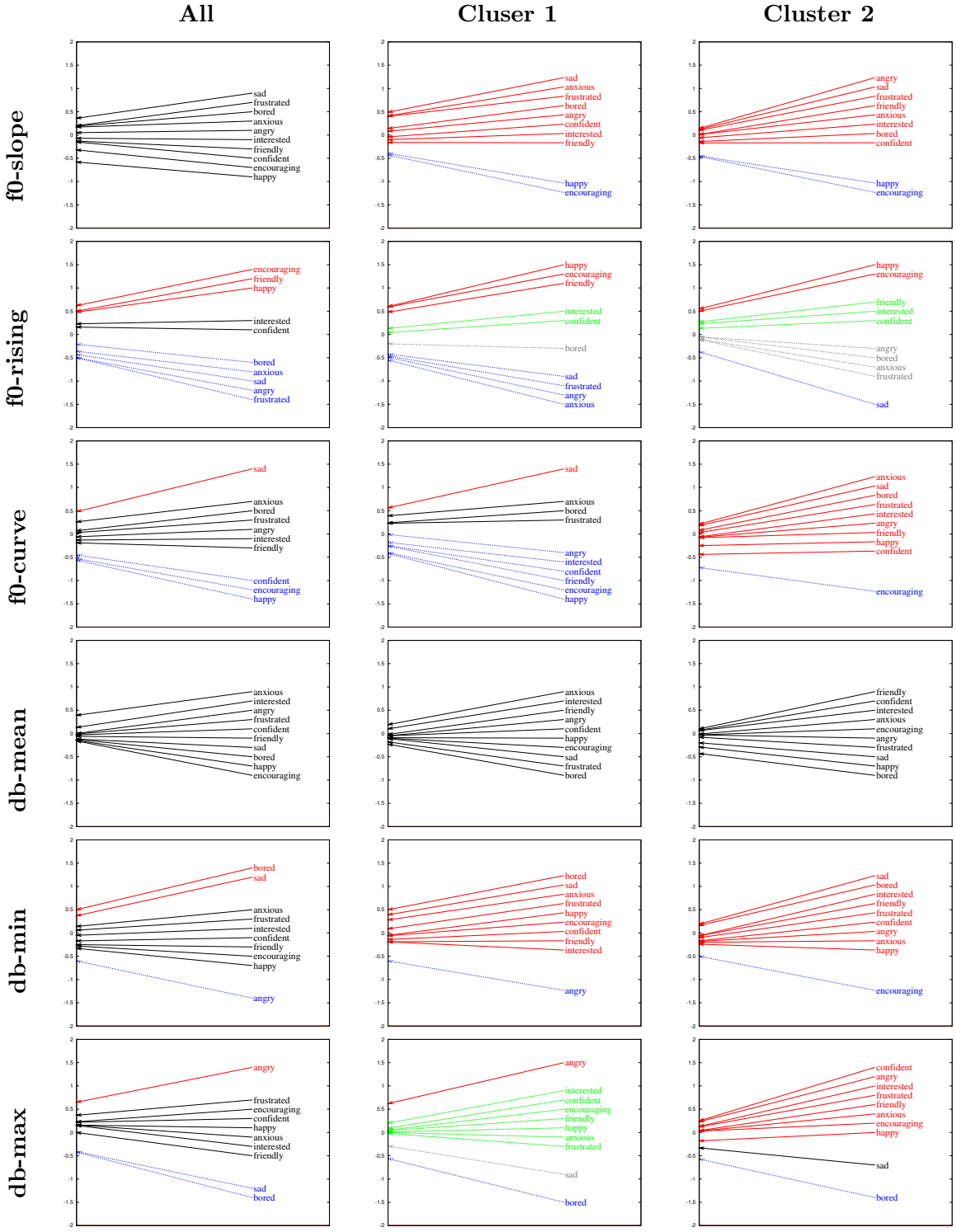
The mean z-scored feature value for each CU_EPSAT emotion is listed below.

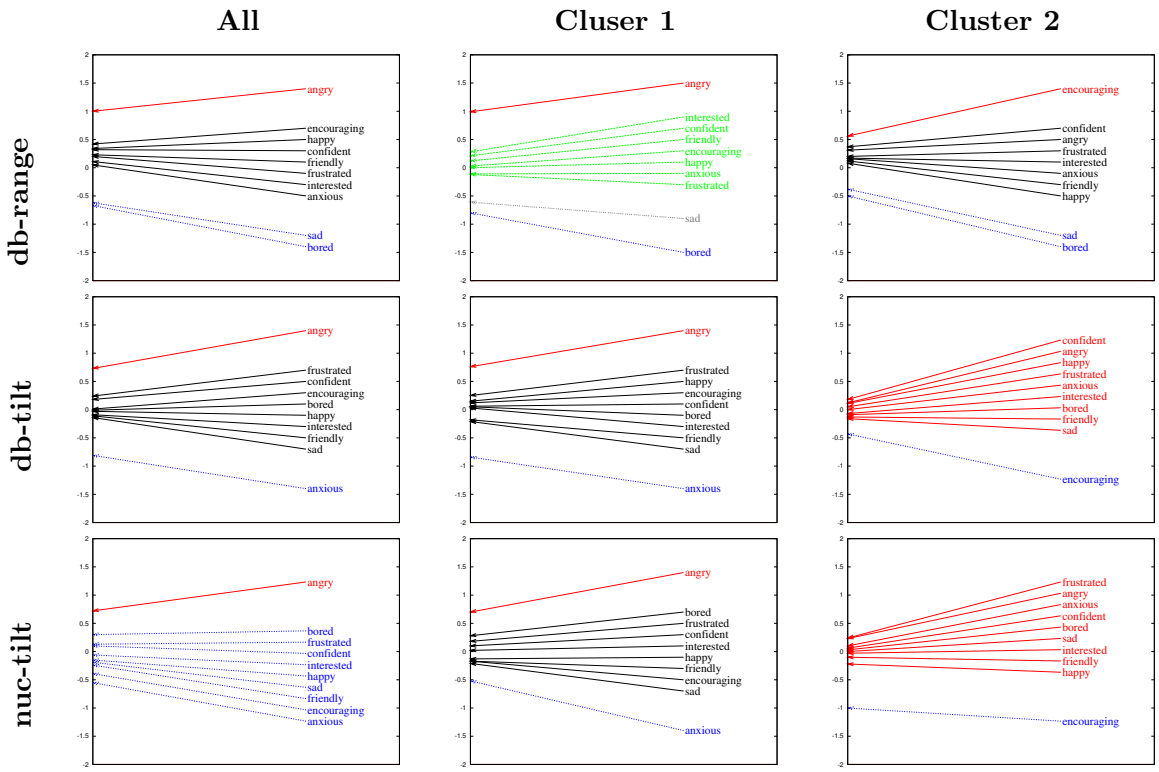
	db-max	db-mean	db-min	db-range	f0-max	f0-mean	f0-min	f0-rising	f0-range	f0-curve	f0-slope	f0-voiced	db-tilt	mic-tilt
happy	-0.20	-0.31	-0.27	0.10	0.16	0.28	0.38	0.57	0.11	-0.28	-0.44	0.04	0.09	-0.22
sad	-0.34	-0.22	0.17	-0.36	-0.18	-0.41	-0.24	-0.35	-0.14	0.17	0.11	-0.20	-0.16	-0.02
frustrated	0.14	-0.09	-0.10	0.19	0.00	0.03	0.00	-0.09	-0.00	0.01	0.13	-0.29	0.04	0.21
confident	0.26	0.09	-0.16	0.35	-0.10	-0.03	0.09	0.12	-0.09	-0.45	-0.20	0.23	0.16	0.07
friendly	0.07	0.09	-0.08	0.14	-0.09	0.15	0.30	0.25	-0.16	-0.08	-0.01	0.06	-0.14	-0.11
interested	0.13	0.09	-0.09	0.16	0.05	0.31	0.23	0.23	0.00	-0.06	-0.06	0.10	-0.09	-0.05
angry	0.24	-0.02	-0.16	0.32	0.07	0.23	0.22	-0.07	0.04	-0.06	0.17	-0.21	0.09	0.21
anxious	0.06	-0.02	-0.17	0.17	-0.04	0.12	0.10	-0.09	-0.06	0.18	0.02	-0.03	0.00	0.10
encouraging	0.00	-0.01	-0.53	0.53	-0.11	0.55	0.31	0.51	-0.15	-0.70	-0.47	0.89	-0.45	-1.03
bored	-0.59	-0.43	0.18	-0.53	-0.09	-0.63	-0.49	-0.03	0.01	0.07	-0.15	-0.16	-0.08	0.04

B.5 CU_EPSAT perceived: Plots

The following plots show where the mean values per emotion per cluster lie in z-score space. The first column uses all raters while the second two columns use only the perceived labels from the raters in cluster 1 and cluster 2, respectively.




All
Cluser 1
Cluster 2
f0-slope
f0-rising
f0-curve
db-mean
db-min
db-max


All
Cluser 1
Cluster 2
db-range
db-tilt
nuc-tilt
angry
encouraging
happy
confident
friendly
frustrated
interested
anxious
sad
bored
angry
interested
confident
friendly
encouraging
happy
anxious
frustrated
sad
bored
encouraging
confident
angry
frustrated
interested
anxious
friendly
happy
sad
bored
angry
frustrated
confident
encouraging
bored
happy
interested
friendly
sad
anxious
angry
frustrated
happy
encouraging
confident
bored
interested
friendly
sad
anxious
confident
angry
happy
frustrated
anxious
interested
bored
friendly
sad
encouraging
angry
bored
frustrated
confident
interested
happy
sad
friendly
encouraging
anxious
angry
bored
frustrated
confident
interested
happy
friendly
encouraging
sad
anxious
frustrated
angry
anxious
confident
bored
sad
interested
friendly
happy
encouraging

Part VII

BIBLIOGRAPHY

Bibliography

Aist, G., Kort, B., Reilly, R., Mostow, J. & Picard, R. (2002). Experimentally augmenting an Intelligent Tutoring System with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proceedings of the Intelligent Tutoring Systems Conference (ITS2002)*. Biarritz, France.

Anderson, J. R., Corbett, A. T., Koedinger, K. R. & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4(2)*, 167–207.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E. & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing*. Denver, Colorado, USA.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford U. Press.

Banse, R. & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70(3)*, 614–636.

Bänziger, T. & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46(3-4)*, 252–267.

Bartels, C. (1997). *Towards a Compositional Interpretation of English Statement and Question Intonation*. PhD thesis, University of Massachusetts Amherst.

Batliner, A., Fischer, K., Huber, R., Spilker, J. & Nöth, E. (2003). How to find trouble in communication. *Speech Communication*, *40(1-2)*, 117–143.

Beckman, M. E., Hirschberg, J. & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S. A. Jun (Ed.), *Prosodic Typology – The Phonology of Intonation and Phrasing*. Oxford, OUP.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D. & Weintraub, M. (1990). Automatic evaluation and training in english pronunciation. In *Proceedings of ICSLP*. Kobe, Japan.

Bernstein, J., DeJong, J., Pisoni, D. & Townshend, B. (2000). Two experiments in automatic scoring of spoken language proficiency. In *Proceedings of InSTIL 2000*. Dundee, Scotland.

Beun, R. J. (1990). The recognition of Dutch declarative questions. *Journal of Pragmatics*, *(14)*, 39–56.

Beun, R. J. (2000). Context and form: Declarative or interrogative, that is the question. In W. Black & H. C. Bunt (Eds.), *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam: John Benjamins.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, Volume 17 (pp. 97–110).

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5(9/10)*, 341–345.

Bolinger, D. L. (1957). Interrogative structures of American English. In *Publication of the American Dialect Society*, number 28. Univ of Alabama Press.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1(1)*, 1–47.

Cauldwell, R. T. (2000). Where did the anger go? The role of context in interpreting emotion in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*. Belfast, Northern Ireland.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Cornelius, R. R. (1996). *The Science of Emotion*. New Jersey: Prentice Hall.

Cowie, R. (2000). Describing the emotional states expressed in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*. Belfast, Northern Ireland.

Cucchiarini, C., Strik, H. & Boves, L. (1997). Using speech recognition technology to assess foreign speakers pronunciation of Dutch. In *Proceedings of New Sounds* (pp. 61–68). Klagenfurt, Austria.

Cucchiarini, C., Strik, H. & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111(6)*, 2862–2873.

Davitz, J. R. (1964). Auditory correlates of vocal expression of emotional feeling. In J. R. Davitz (Ed.), *The Communication of Emotional Meaning* (pp. 101–112). New York: McGraw-Hill.

Devillers, L. & Vidrascu, L. (2004). Reliability of lexical and prosodic cues in two real-life spoken dialog corpora. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal.

Douglas-Cowie, E., Cowie, R. & Schröder, M. (2003). The description of naturally occurring emotional speech. In *Proceedings of the 15th International Conference of Phonetic Sciences* (pp. 2877–2880). Barcelona, Spain.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & T. Power (Eds.), *The Handbook of Cognition and Emotion* (pp. 45–60). Sussex, U.K.: John Wiley & Sons, Ltd.

Fairbanks, G. & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotions. *Speech Monographs*, *6*, 87–194.

Fernandez, R. (2004). *A Computational Model for the Automatic Recognition of Affect in Speech*. PhD thesis, Massachusetts Institute of Technology.

Forbes-Riley, K., Litman, D., Huettner, A. & Ward, A. (2005). Dialogue-learning correlations in spoken dialogue tutoring. In *Proceedings 12th International Conference on Artificial Intelligence in Education (AIED 2005)*. Amsterdam.

Fox, B. (1993). *Human Tutorial Dialogue*. New Jersey: Lawrence Erlbaum.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R. & Cesari, F. (2000). The SRI Eduspeak system: Recognition and pronunciation scoring for language learning. In *Proceedings of InSTIL 2000*. Dundee, Scotland.

Franco, H., Neumeyer, L., Kim, Y. & Ronen, O. (1997). Automatic pronunciation scoring for language instruction. In *Proceedings of ICASSP*. Los Alamitos, CA, USA.

Freund, Y. & Schapire, R. E. (1999). A short introduction to boosting. *Journal of the Japanese Society for Artificial Intelligence (JSAI), 14(5)*, 771–780.

Fries, C. C. (1964). On the intonation of yes-no questions in English. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott & J. L. M. Trim (Eds.), *In Honour of Daniel Jones* (pp. 242–254). Longmans, London.

Gass, S. & Selinker, L. (1994). *Second Language Acquisition: An Introductory Course*. University of Turku: Publications of the Department of Phonetics.

Geluykens, R. (1987). Intonation and speech act type: An experimental approach to rising intonation in queclaratives. *Journal of Pragmatics*, *11*, 483–494.

Geluykens, R. (1988). On the myth of rising intonation in polar questions. *Journal of Pragmatics*, *12(4)*, 467–485.

Ginzburg, J. & Sag, I. A. (2000). *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. Stanford, CA, USA: CSLI Publications.

Gobl, C. & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*, 189–212.

Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tür, D., Ljolje, A., Parthasarathy, S., Riccardi, M. R. G. & Saraclar, M. (2005). The AT&T WATSON speech recognizer. In *Proceedings of IEEE ICASSP-2005*. Philadelphia, PA, USA.

Gorin, A. L., Riccardi, G. & Wright, J. H. (1997). How may I help you? *Speech Communication*, *23*, 113–127.

Graesser, A. C. & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31(1)*, 104–137.

Graesser, A. C., VanLehn, K., Rose, C. P., Jordan, P. W. & Harter, D. (2001). Intelligent Tutoring Systems with conversational dialogue. *AI Magazine*, *22(4)*, 39–52.

Gunlogson, C. (2001). *True to Form: Rising and Falling Declaratives as Questions in English*. PhD thesis, University of California, Santa Cruz.

Gussenhoven, C. (1983). Focus, mode and the nucleus. *Journal of Linguistics*, *19*, 377–417.

Hall, M. A. (1998). *Correlation-Based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato.

't Hart, J., Collier, R. & Cohen, A. (1990). *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambrigde: Cambridge Univ. Press.

Herry, N. & Hirst, D. (2002). Subjective and objective evaluation of the prosody of English spoken by French speakers: The contribution of computer assisted learning. In *Proceedings of Speech Prosody*. Aix-en-Provence, France.

Hincks, R. (2005). *Computer Support for Learners of Spoken English*. PhD thesis, Speech and Music Communication, The Royal Institute of Technology in Sweden (KTH).

Hirschberg, J. (1984). Toward a redefinition of yes/no questions. In *Proceedings of the 22nd Annual Meeting of the ACL*. Stanford University, CA, USA.

Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V. & Niemann, H. (2000). Recognition of emotion in a realistic dialog scenario. In *Proceedings of ICSLP* (pp. 665–668). Beijing, China.

Hymes, D. H. (1971). *On Communicative Competence*. Philadelphia: University of Pennsylvania Press.

Johnstone, T. & Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proceedings of the XIV International Congress of Phonetic Sciences*.

Johnstone, T. & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland (Eds.), *The Handbook of Emotions* (2nd Ed.) (pp. 226–235). New York: Guilford.

Jurafsky, D., Shriberg, L. & Biasca, D. (1997). *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual* (13 Ed.).

Juslin, P. N. & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129(5)*.

Kienast, M. & Sendlmeier, W. F. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*. Belfast, Northern Ireland.

Kleinginna Jr., P. R. & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5(4)*, 345–379.

Kormos, J. & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32(2)*, 145–164.

Ladd, D., Silverman, K. A., Tolkmitt, F., Bergmann, G. & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect. *Journal of the Acoustical Society of America*, *78(2)*, 435–444.

Ladefoged, P. (1993). *A Course in Phonetics* (Third edition Ed.). Harcourt Brace Jovanovich, New York.

Laukka, P., Juslin, P. N. & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, *19(5)*, 633–653.

Lee, C. M. & Narayanan, S. (2005). Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13(2)*.

Liberman, M., Davis, K., Grossman, M., Martey, N. & Bell, J. (2002). Emotional prosody speech and transcripts. Linguistic Data Consortium, Philadelphia.

Liscombe, J., Hirschberg, J. & Venditti, J. (2005a). Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech*. Lisbon, Portugal.

Liscombe, J., Riccardi, G. & Hakkani-Tür, D. (2005b). Using context to improve emotion detection in spoken dialogue systems. In *Proceedings of Interspeech*. Lisbon, Portugal.

Liscombe, J., Venditti, J. & Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Eurospeech*. Geneva, Switzerland.

Liscombe, J., Venditti, J. & Hirschberg, J. (2006). Detecting question-bearing turns in spoken tutorial dialogues. In *Proceedings of Interspeech*. Pittsburgh, PA, USA.

Litman, D. & Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Barcelona, Spain.

Litman, D. & Silliman, S. (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In *Proceedings of the 4th Meeting of HLT/NAACL* (Companion Proceedings). Boston, MA.

Liu, F., Surendran, D. & Xu, Y. (2006). Classification of statement and question intonation in Mandarin. In *Proceedings of Speech Prosody*. Dresden, Germany.

Merrill, D. C., Reiser, B. J., Ranney, M. & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and Intelligent Tutoring Systems. *Journal of the Learning Sciences, 2(3)*, 277–305.

Meyer, G. J. & Shack, J. R. (1989). The structural convergence of mood and personality: Evidence for old and new 'directions'. *Journal of Personality and Social Psychology, 57(4)*, 691–706.

Morgan, J., Gallagher, M., Molholt, G., Holland, M. & LaRocca, S. (2004). Towards the automatic evaluation of fluency. In *The 12th Annual Army Research Laboratory/United States Military Academy Technical Symposium*. West Point, New York, USA.

Mozziconacci, S. & Hermes, D. J. (1999). Role of intonation patterns in conveying emotion in speech. In *Proceedings of ICPhS*. San Francisco, CA, USA.

Neumeyer, L., Franco, H., Weintraub, M. & Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. In *Proceedings of ICSLP* (pp. 1457–1460). Philadelphia, PA, USA.

Nöth, W. (1990). *Handbook of Semiotics*. Indiana University Press, Bloomington & Indianapolis.

O'Connor, J. D. & Arnold, G. F. (1973). *Intonation of Colloquial English : A Practical Handbook*. London: Longmans.

Oudeyer, P. (2002). The synthesis of cartoon emotional speech. In *Proceedings of Speech Prosody* (pp. 551–554). Aix-en-Provence, France.

Pierrehumbert, J. B. & Hirschberg, J. (1990). The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan & M. E. Pollack (Eds.), *Intentions in Communication* (pp. 271–311). MIT Press.

Pollermann, B. Z. (2002). A place for prosody in a unified model of cognition and emotion. In *Proceedings of Speech Prosody* (pp. 17–22). Aix-en-Provence, France.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B. & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education (IJAIED)*, *16*, 171–194.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann Publishers.

Rando, E. (1980). Intonation in discourse. In L. Wuagh & C. H. van Schooneveld (Eds.), *The Melody of Language* (pp. 243–278). Baltimore: University Park Press.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 133–142). Somerset, New Jersey.

Reeves, B. & Nass, C. (1998). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York, New York, USA: Cambridge University Press.

Reithinger, N. & Klesen, M. (1997). Dialogue act classification using language models. In *Proceedings of Eurospeech* (pp. 2235–2238). Rhodes, Greece.

Schapire, R. E. & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning, 39(2/3)*, 135–168.

Scherer, K. (2000). A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. In *Proceedings of ICSLP*. Beijing, China.

Scherer, K. R., Ladd, D. R. & Silverman, K. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America, 76(5)*, 1346–1356.

Schröder, M. (2001). Emotional speech synthesis: A review. In *Proceedings of Eurospeech* (pp. 561–564). Aalborg, Denmark.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M. & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proceedings of Eurospeech* (pp. 87–90). Aalborg, Denmark.

Schubiger, M. (1958). *English Intonation: Its Form and Function*. Niemayer Verlag.

Shafran, I., Riley, M., & Mohri, M. (2003). Voice signatures. In *Proceedings of The 8th IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*. St. Thomas, U.S. Virgin Islands.

Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M. & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech, 41(3-4)*, 439–487.

Smola, A. J. & Schoelkopf, B. (1998). A tutorial on support vector regression. In *In NeuroCOLT2 Technical Report Series*.

Steedman, M. (2003). Information-structural semantics for English intonation. In *LSA Summer Institute Workshop on Topic and Focus*. Santa Barbara, CA, USA.

Stenström, A. B. (1984). *Questions and Responses in English Conversation*. CWK Gleerup, Malmo.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C. & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26(3)*, 339–373.

Surendran, D. & Levow, G. A. (2006). Dialog act tagging with support vector machines and hidden markov models. In *Proceedings of Interspeech/ICSLP*.

Tato, R., Santos, R., Kompe, R. & Pardo, J. M. (2002). Space improves emotion recognition. In *Proceedings of ICSLP* (pp. 2029–2032). Denver, Colorado, USA.

Teixeira, C., Franco, H., Shriberg, E., Precoda, K. & Sonmez, K. (2001). Evaluation of speaker's degree of nativeness using text-independent prosodic features. In *Proceedings of the Workshop on Multilingual Speech and Language Processing*. Aalborg, Denmark.

Toivanen, J. (2003). Tone choice in the English intonation of proficient non-native speakers. In *Proceedings of Fonetik* (pp. 165–168). Umeå/Lövånger, Sweden.

Toivanen, J. (2005). Pitch dynamism of English produced by proficient non-native speakers: Preliminary results of a corpus-based analysis of second language speech. In *Proceedings of FONETIK*. Stockholm, Sweden.

Toivanen, J., Väyrynen, E. & Seppänen, T. (2005). Gender differences in the ability to recognize emotional content from speech. In *Proceedings of Fonetik* (pp. 119–122). Gothenburg, Sweden.

Tsui, A. (1992). A functional description of questions. In M. Coulthard (Ed.), *Advances in Spoken Discourse Analysis*. Routledge.

Turk, O., Schröder, M., Bozkurt, B. & Arslan, L. (2005). Voice quality interpolation for emotional text-to-speech synthesis. In *Proceedings of Interspeech* (pp. 797–800). Lisbon, Portugal.

Uldall, E. (1964). Dimensions of meaning in intonation. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott & J. L. M. Trim (Eds.), *In Honour of Daniel Jones* (pp. 271–279). Longmans, London.

VanLehn, K., Jordan, P. & Rose, C. P. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of the Intelligent Tutoring Systems Conference*. Biarritz, France.

Väyrynen, E. (2005). Automatic emotion recognition from speech. Master's thesis, Department of Electrical and Information Engineering, University of Oulu.

Šafářová, M. (2005). The semantics of rising intonation in interrogatives and declaratives. In *Proceedings of Sinn und Bedeutung (SuB9)*.

Šafářová, M. & Swerts, M. (2004). On recognition of declarative questions in English. In *Proceedings of Speech Prosody*. Nara, Japan.

Walker, M. A., Langkilde-Geary, I., Hastie, H. W., Wright, J. & Gorin, A. (2002). Automatically training a problematic dialogue predictor for a spoken dialogue system. *Journal of Artificial Intelligence Research*, *16*, 293–319.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In *Proceedings of the ISCA Workshop on Speech and Emotion*. Belfast, Northern Ireland.

Wichmann, A. (2002). Attitudinal intonation and the inferential process. In *Proceedings of Speech Prosody* (pp. 11–15). Aix-en-Provence, France.

Williams, C. & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, *52(4)*.

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. In *ICONIP/ANZIIS/ANNES'99 International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems* (pp. 192–196). Dunedin, New Zealand.

Xi, X., Zechner, K. & Bejar, I. (2006). Extracting meaningful speech features to support diagnostic feedback: An ECD approach to automated scoring. In *Proceedings of NCME*. San Francisco, CA, USA.

Yoon, T. J., Chavarría, S., Cole, J. & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceeding of the International Conference on Spoken Language Processing – Interspeech*. Jeju, Korea.

Zechner, K. & Bejar, I. (2006). Towards automatic scoring of non-native spontaneous speech. In *Proceedings of the Human Language Technology Conference of the NAACL* (pp. 216–223). New York City, USA: Association for Computational Linguistics.

Zetterholm, E. (1999). Emotional speech focusing on voice quality. In *Proceedings of FONETIK: The Swedish Phonetics Conference* (pp. 145–148). Gothemburg, Sweden.