Detecting Verbal and Non-verbal Cues in the Communication of Emotions

Committee:
Alex Waibel
Bob Carpenter
Jeff Cohn
Maxine Eskenazi
John Lafferty

Thomas S. Polzin
School of Computer Science
Carnegie Mellon University

November 16, 1999

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Machines are our friends!" These were the words I heard sometimes uttered by an office mate and good friend of mine when he discovered a new feature on his computer, when a program was running smoothly and functioned exactly as it was supposed to do, or when he was amazed by the computer's reliability or power. At that time I was very skeptical about this statement and suspected some psychological problems in my friend's interpersonal skills to be the reason for this extrapolation of friendship to a computer. While I still remain skeptical about this statement, I have to realize that my friend's sentiment is not an unusual one and that similar views are quite common when we interact with computers.

In several experiments Reeves and Nass (1996) show that humans extrapolate their interpersonal interaction patterns to their interaction with computers. Humans are polite to computers, they are flattered by computers, there are good and bad computers, we can form teams with computers, and we have emotional experiences when we interact with them. For instance, in one experiment Reeves and Nass ask people to learn facts about several topics with the aid of a computer. A subsequent computerized test would assess their learning curve. The people are told that at the end of the session they would have to evaluate the computer with which they were working. During the training session the computer only displays text and graphical buttons. Even though the people are told that the computer adjusts the presentation of the topics based on their feedback, every person is presented with the same facts in the same way. In the subsequent test session, people are informed by the computer whether their answer is correct or not. After this test session the computer tells the subjects what it thought about its own performance which in all cases it assessed as a great job. Finally, half of the group is asked to evaluate the computer's performance on the very same computer that they had done the experiment on and which just had praised itself. The rest of the group is asked to evaluate the computer's performance on a different computer. People who have to evaluate the computer's performance on the very same computer give significantly more positive feedback about the computer's performance than the people who judge the computer on a different computer. In other words people were polite to the computer!

If we indeed try to interact with computers very much the same way as we interact with other humans, then this similarity should be reflected in the design of human-computer interfaces to facilitate a more "natural", more human-like interaction. Studies on human computer interaction (HCI) recognize this similarity and also consider emotions to be an important factor in the

communication between humans and machines. However, most investigations in HCI focus on the synthesis of emotional expressions, both visual and acoustic, on the computer side (McCauley et al., 1998; Isbister and Nass, 1998; Ball and Breese, 1998; Olveres et al., 1998; Tosa and Nakatsu, 1996; Nakatsu, 1997; Moriyama, Saito, and Ozwa, 1997; Cassell et al., 1994). We already have cartoon-like characters popping up if we do something wrong or ask for help. Depending on the situation these characters smile at us or frown. Soon these characters will speak to us, have facial expressions, and use gestures to make a point (Cassell et al., 1998; Tosa and Nakatsu, 1996).

But if Reeves and Nass (1996) are right about their thesis then this kind of interface design is dangerously one sided since this design ignores the emoting user. It might even be a little bit confusing when we have to interact with characters emoting heavily while our emotional state is – quite impolitely – ignored. We do need interfaces that not only express emotions but also detect emotions in the user.

The problem of an emoting computer that is unaware of the emotional state of the human user becomes even more evident when we allow the human user to speak to the computer using a speech recognition system in the "front end" of the interface. When a speech recognition interface only pays attention to what is said but ignores how it is said, the interface fails to pick up information that is essential for human-to-human communication. For instance, certain word or syntactic choices might indicate that the speaker is angry or sad. Certain acoustic features might indicate that the speaker is bored or interested. It should be obvious that this kind of information is important for natural interaction with a computer and essential for successful communication. That is, the search for cues that allow the detection of the emotion expressed by a speaker in an utterance becomes an important topic of research.

Emotions can be communicated in various ways by relying both on verbal and non-verbal means. Non-verbal means comprise body gestures, facial expressions, the modifications of prosodic parameters, and changes in the spectral energy distribution. This investigation is confined to information within the speech signal and we show that verbal and non-verbal information within the signal allows an effective decoding of the expressed emotion. In particular, we investigate the role of word choice, spectral energy distribution, and prosody. The focus of this investigation lies on the role of prosodic information and we explore the importance of several prosodic parameters such as speaking rate, pitch, and intensity in signaling the four basic emotions happiness, sadness, anger, or fear (Ortony and Turner, 1990). We show that the combination of verbal and non-verbal information of an utterance allows the automatic detection of the expressed emotion with an accuracy comparable to humans performing the same task.

The obvious way of expressing an emotion by verbal means is to name it explicitly:

(1.1) I am happy!

(1.2) I am angry!

However, speakers tend to be more subtle and implicit and encode emotion, for instance, by a certain word choice:

(1.3) That's terrific!

(1.4) Mind your damn business!

Our corpus studies show that the explicit approach is used very rarely by speakers and we concentrate our investigation on the detection of verbal cues given more implicitly. Modeling verbal cues in the communication of emotions is complicated by the fact that verbal cues are optional. That is, not every emotional utterance necessarily bears verbal cues. There are two reasons why an emotional state would not be expressed by using verbal cues. First, the current utterance does not indicate any particular emotion in its verbal setup because the emotion can be inferred given the preceding discourse or context. Consider the following sentence from the movie Kramer vs. Kramer:

(1.5) Where are you going?

In isolation the verbal information of this sentence does not allow any inference about the emotional state of its speaker. However, given some preceding context such as displayed in the sentence below, we can draw the inference that the speaker is either sad or angry. In this investigation we do not try to model any emotion detection based on inferences relying on the preceding discourse or semantic reasoning.

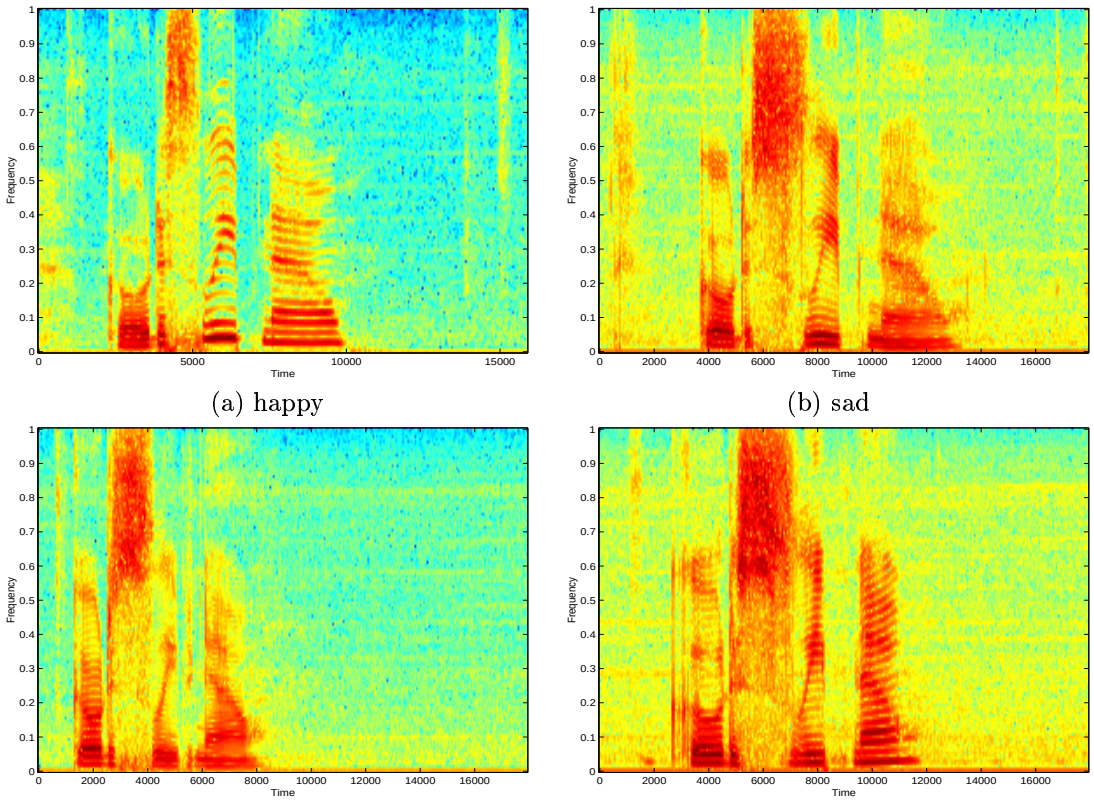
(1.6) I'm leaving you ... and I don't love you anymore ...

A second situation in which no verbal cues are given would be when the speaker relies on acoustic cues to express an emotion. Since we do model non-verbal cues explicitly in this investigation, we compensate for the lack of verbal cues and still be able to detect the expressed emotion.

Emotion specific verbal behavior can be studied at various levels. Questions about how interlocuters manifest and interpret emotions within in a conversation are explored in discourse analysis (Fiehler, 1990b). Other studies focus on other linguistic levels such as the lexical or syntactic level, and investigate how verbal devices are able to encode emotions (Davitz, 1969; Fudge, 1970; Irvine, 1982). While most of these studies are empirical their goal is not to build an automatic system for the recognition of the expressed emotion.

In this investigation we follow the linguistic research and pursue a strictly data driven approach. We do not try to model processes at the discourse or semantic level. Instead we model lexical processes that signal an emotional involvement of the speaker. Some such examples are given in (1.3) and (1.4) above. We model emotion-specific lexical information by computing the probability of a certain word given the previous word and the emotion expressed by a speaker. The idea behind computing these probabilities is that certain combinations are more probable for the expression of certain emotion. Computing the probability of a certain word to given a history of previous words is a technique widely used in speech recognition (language modelling).

In figure 1.1 we show the spectrograms of four utterances of the sentence *"Go to sleep now!"* by the same actress expressing emotions (a) happy, (b) sad, (c) angry, and (d) afraid. The respective spectrograms are quite different indicating that the distribution of energy is employed by the speaker to express a certain emotion. We can also see that the sad and afraid realizations of the sentence resemble each other to some extent in the spectrograms. As it turns out in our experiments, sad and afraid utterances are confused with each other quite frequently both by our automatic classification system and by humans.



(a) happy (b) sad

(c) angry (d) afraid

Figure 1.1: Spectrogram for the sentence *"Go to sleep now!"* spoken by an actress.

One possible way to capture spectral properties of emotional speech is to train emotion-specific speech recognition systems, that is, systems trained on speech samples which encode only a particular emotion. Each of these emotion-specific recognition engines then represents the spectral properties typical for the respective emotion. While this is a valid approach, it is foredoomed if there are not enough available training samples. A recognition system to be trained from scratch requires about 20 hours of data (Finke, 1999). If we want to train systems modeling spectral peculiarities of, for example, three emotions, we would need a total of about 60 hours of emotional speech. No emotion corpora of this size are available. Adaptation, however, requires corpora which can be much smaller than corpora needed for training. We use an already existing recognition system and adapt on emotion-specific speech samples.. We model the spectral information by means of cepstral coefficients to account for the properties of the human auditory system.

The role of prosody within the communication of emotions has been studied extensively in psychology and psycho-linguistics. This kind of research focuses mainly on two questions:

- How do humans express emotions by modifying prosodic parameters of their speech?
- Which prosodic parameters allow the decoding of the emotion expressed by the speaker?

In this investigation we explore the potential of prosodic parameters such as speaking rate, pitch, and intensity, to discriminate among several emotions. For example, in figure 1.2 we display the fundamental frequency of four utterances of the sentence *"Go to sleep now!"*, spoken by the same speaker portraying the emotions (a) happy, (b) sad, (c) angry, and (d) afraid. The

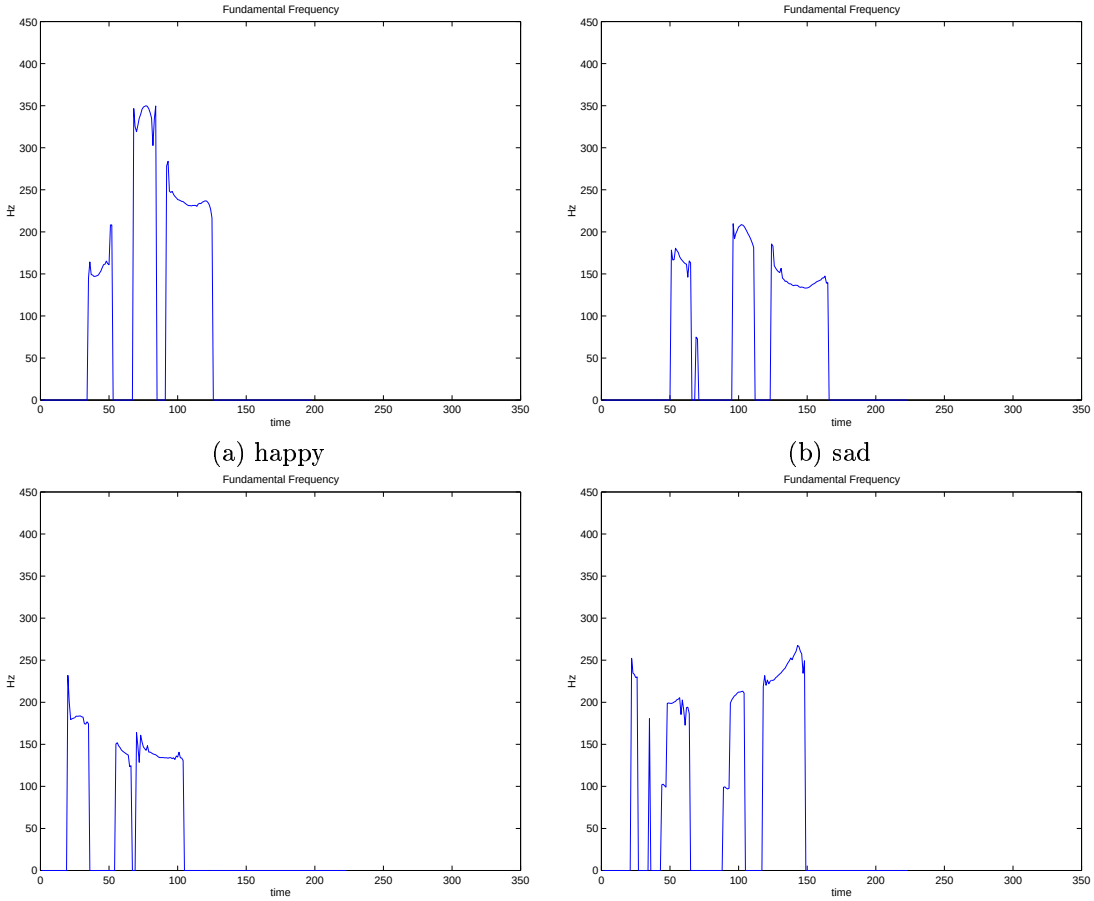


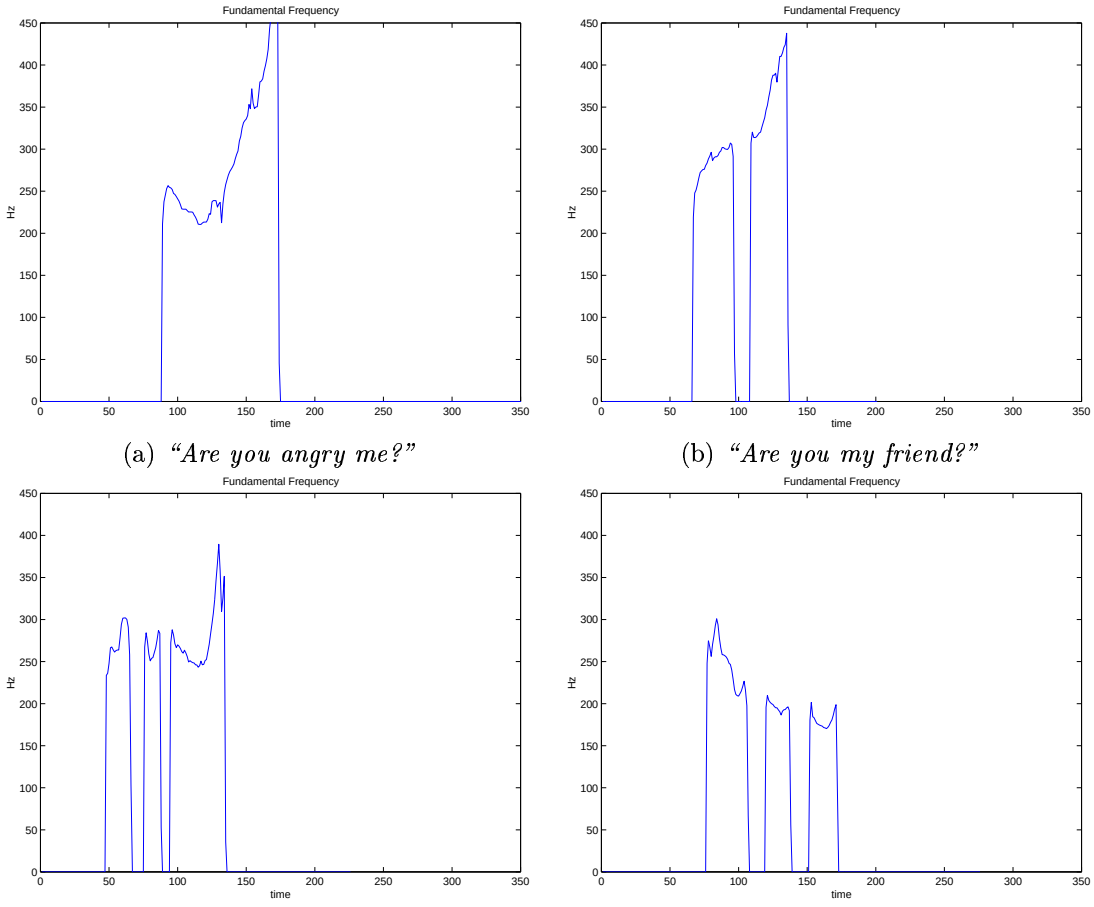
Figure 1.2: Fundamental frequency for the sentence *"Go to sleep now!"* spoken by an actress.

differences are quite striking. Looking at the fundamental frequencies as given in Figure 1.2 above, we can speculate on some prosodic cues. For example, the angry utterance (c) seems to be shorter than (a), (b), (d), indicating an increase in the speaking rate. Moreover, the mean fundamental frequencies of these utterances are quite different. For example, the mean pitch for the sad utterance is substantially lower than the utterance of the same sentence when expressing happiness. Similar tendencies can be shown to be the case for intensity.

Prosodic features are multi-functional. They not only express emotions but also serve a variety of other functions as well, such as word and sentence stress or syntactic segmentation. Other functions of prosody include, for example, the distinction between yes/no questions and statements by using a final rise. We also find that the phonetic content of an utterance has an impact on its prosodic parameters. For instance, each vowel has an inherent fundamental

frequency which can differ substantially among the vowels within a given language. Compare for example, the fundamental frequency of the vowels /i/ and /o/. Differences between two vowels can be as high as 20Hz in their fundamental frequency.

To illustrate the multi-functionality of prosodic features we display in figure 1.3 the fundamental frequency of utterances of the sentences (a) *"Are you angry?"*, (b) *"Are you my friend?"*, (c) *"Are you talking to me?"*, and (d) *"Be my friend Shrimp!"* all uttered by the same actress portraying happiness in all four cases. It is far from clear what renders these utterances in figure 1.3 similar to each other and different from the sad, angry, and afraid utterances in figure 1.2. In this investigation we try to compensate for the multi-functionality of prosodic features by us-



(a) *"Are you angry me?"*

(b) *"Are you my friend?"*

(c) *Are you talking to me?"*

(d) *"Be my friend Shrimp!"*

Figure 1.3: Fundamental frequency all spoken by the same actresst all portraying the emotion happy.

ing several normalization and clustering techniques. Moreover, we investigate not just prosodic features pertaining to a whole utterance but also prosodic features referring to smaller segments such as phones to account, for example, for intrinsic prosodic properties of these segments.

Other studies investigating emotional speech tried to compensate for the multi-functionality

of prosodic features by restricting the corpus to a very small number of sentences – sometimes just one – and by collecting utterances of these sentences through several speakers. Limiting the subject to pronounce sentences from a predefined set also forces the subject to encode his or her emotion using non-verbal (spectral and prosodic) rather than verbal means. Thus, most studies try to compensate for the optionality and variability of prosodic cues by restricting the elicitation conditions.

In this investigation we are more liberal and do not restrict the elicitation conditions as severely as the studies above. This investigation comprises two major experiments. The first experiment is based on a predefined set of sentences uttered in four different emotions. This corpus is used to explore spectral and prosodic properties of emotional speech. The second experiment is based on a corpus comprising sentences from talk shows and movies. Since we do not have control over the productions of utterances in this corpus, we expect actors to use both verbal and non-verbal cues to encode an emotion and, moreover, we expect a large variability of prosodic parameters among the utterances within this corpus. Investigating and combining both verbal and non-verbal information to explore the emotion expressed by a speaker is studied for the first time in this investigation.

The collection of emotional speech for the corpora constitutes a substantial part of this investigation. Our studies rely on the following corpora:

- Woggles Corpus: We asked 9 female drama students to portray 50 different sentences expressing happiness, sadness, fear, and anger.
- Movie and Talk Shows: We collected several thousand utterances from movies and talk shows. These utterance are transcribed and tagged for the expressed emotion. The tag set comprises ten different emotion labels: neutral, bored, strong joy, weak joy, sad, afraid, irony, angry, disgusted, and suspicious. Unfortunately, only neutral, sad and angry utterances occur frequently enough to allow for a reliable estimation of emotion-specific verbal and non-verbal parameters.
- In order to test how emotions are encoded in languages other than English we also transcribed and tagged Spanish and German movies.
- Spanish Call Home: We tagged 39 dialogues of Spanish spontaneous telephone conversations using the same tag set as for the movies and talk shows above.

To our knowledge these corpora constitute the largest set of corpora used so far to study the impact of the expression on an emotion on an individual's speech. The corpora collected from movies and talk shows also allow to investigate the role of visual information in the communication of emotions using the respective video data. This, however, will have to be part of future work.

Using corpora with several speakers uttering several sentences in various emotions makes the task of finding robust prosodic cues for the detection of the underlying emotion quite challenging. In figure 1.4 we display the fundamental frequency of the sentence *"Go to sleep now!"* from the Woggles Corpus uttered by four different actresses portraying the emotion happy. We would expect the fundamental frequencies to be similar. While it could be argued that utterances (a) and (b) look similar, the case seems to be lost for utterances (c) and (d). Finding prosodic features which capture these changes across speakers and utterances constitutes a major contribution of this investigation.

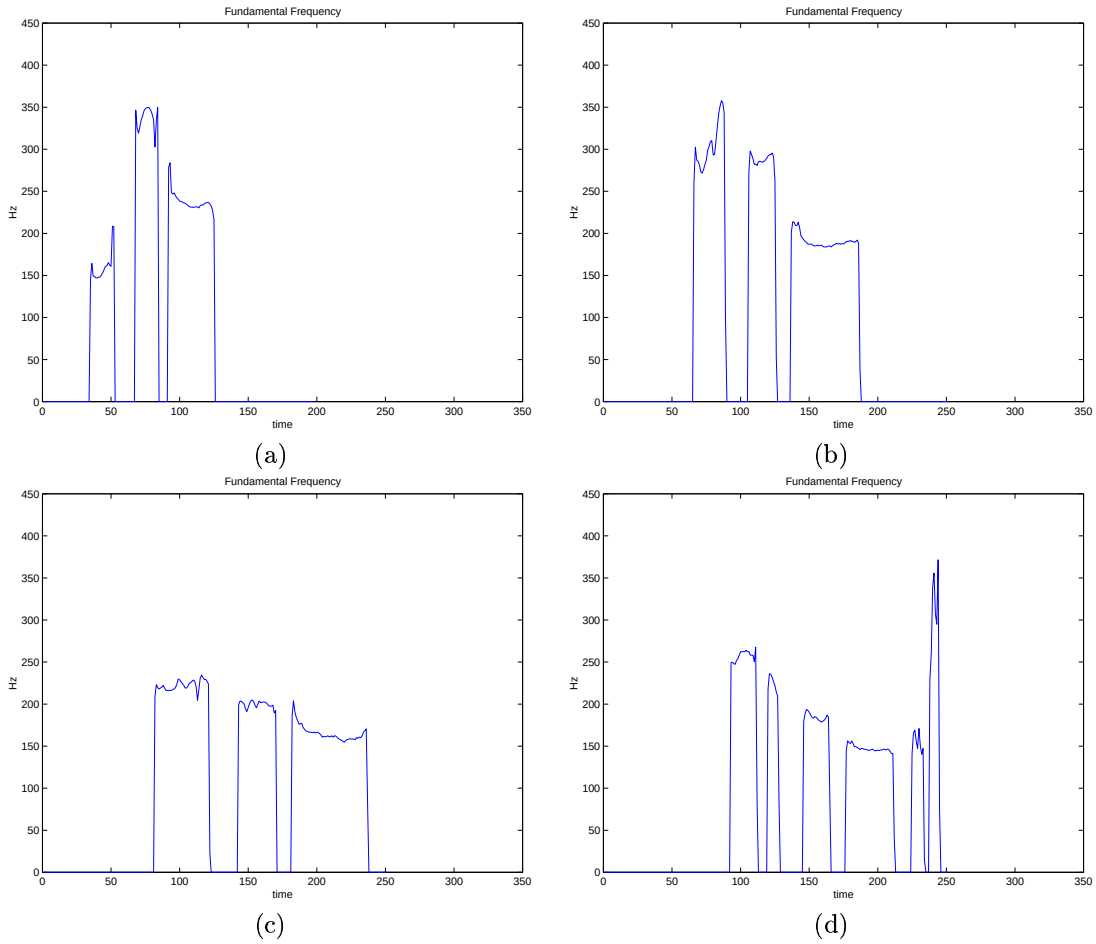


Figure 1.4: Fundamental frequency for the sentence *"Go to sleep now!"* spoken by actresses all portraying the emotion happy.

Research on the expression of emotions can be traced back at least all the way back to Charles Darwin (Darwin, 1998). One of Darwin's central claims was that the expression of emotions is universal both for humans and animals, a claim later to be validated by extensive research lead by Paul Ekman (1994) whose research focused mainly on the facial expression of emotions. Within this investigation we carry out some pilot experiments which suggest that prosodic cues in the expressions of emotions are to some extent universal as well. To substantiate this claim we transcribe and tag Chinese and German movies and try to detect the expressed emotions by using prosodic models developed on English data.

Throughout this investigation we perform several experiments involving human subjects. The purpose behind these experiments is primarily to validate our corpora regarding the emotions expressed and to assess an accuracy baseline to which we can compare the performance of the automatic classification system. For studies focusing in particular on the way humans encode and decode emotions in speech, consult the work of Scherer and Ekman (1984).

We alluded to a possible application of emotion detection earlier on in this chapter. Research on HCI aims to design interfaces which allow a more natural interaction involving emoting artificial agents. So far, however, these emoting agents are designed to operate without regard to the expressed emotion of the human which renders these interfaces even more unnatural. We think that this investigation closes the gap to some extent by demonstrating that the automatic detection of emotion is feasible. In addition, we think that the detection of emotion is not confined to use in HCI but reveals interesting applications in other areas as well.

Verbal and nonverbal information can change or modify the literal meaning of an utterance. Any natural language processing system that relies on a semantic representation, for example speech-to-speech machine translation, has to be aware of the expressed emotion by the speaker to warrant an accurate representation of the input sentence and its felicitous translation. Related to this topic is the detection of irony or sarcasm which reverse the literal meaning of an utterance. Ironic speech employs both verbal and non-verbal cues that allow the listener to decode the irony of the utterance. Having a profound understanding on how emotions are encoded might help to detect irony or sarcasm as well.

As mentioned several times in this introduction, prosody serves several important functions within human communication. By modeling a certain linguistic function of prosody one has to be aware of interferences from other functions that are also implemented by the modification of prosodic parameters. One way to compensate for interferences is to model explicitly several functions of prosody at the same time. To give a simple example, sentence boundaries are often marked by pauses. Pauses, however, are also employed by humans to express sadness. By modeling both functions at the same time, one can hope to improve the overall accuracy for both, boundary and emotion detection. Moreover, while there is an extensive body of research on the synthesis of emotional speech, synthetic speech still sounds unnatural. This study might help to isolate those prosodic features which render the synthesized speech more natural.

One of our experiments in this investigation shows that the accuracy of a speech recognition system depends on the emotion encoded in the input utterances. Modeling emotional speech explicitly within a speech recognition system by using emotion-dependent spectral, prosodic, and language models might improve the overall accuracy of the recognition system.

In the next chapter we describe how emotions are encoded in human communication. The chapter begins with a brief description of the differences between verbal and non-verbal information, followed by a brief introduction into the notion of emotions. We review previous studies investigating the role of verbal and non-verbal cues in the communication of emotions. We pay special attention to the role of prosody within this communication and we also describe how prosodic features such as intensity or speaking rate can be extracted from the speech signal.

In the third chapter we introduce the formal apparatus which we use to model verbal and non-verbal cues. This chapter comprises three sections in which we discuss verbal, spectral, and prosodic cues. In the first section we give a very brief introduction into the language model which we use to model verbal cues. The second and most comprehensive section describes first our approach to model prosodic information. We model these both spectral and prosodic information with the same underlying modeling approach relying on hidden Markov models. We develop a hidden Markov architecture which allows to summarize atomic hidden Markov states into what we call suprasegmental hidden Markov states. The atomic states in this suprasegmental hidden Markov models are used to model spectral (segmental) information. The suprasegmental states, having access to the overall time spent in their constituting states, are used to model prosodic events. The last section describes the adaptation technique which we employ to model emotion-

specific spectral properties.

The fourth chapter contains the experiments of this investigation which we arrange according to the emotional speech corpus studied. The chapter starts with some general statements pertaining to all of the experiments which follow. This includes a description of elicitation techniques to collect emotional corpora and evaluation measures. The Woggles corpus is the starting point to explore spectral and prosodic cues. With the corpus comprising speech segments from movies and talk shows, we also investigate the role of verbal cues and its interaction with spectral and prosodic cues. The prosodic models developed with this corpus are also tested on corpora consisting of segments in different languages. The idea behind this set of experiments is to see whether prosodic cues for the communication of emotions in English extrapolate to other languages. In a final set of experiments we explore the possibilities of detecting emotion cues in telephone speech.

In the final chapter we summarize the results of this investigation and end with a conclusion which points to some possible extensions of the formal apparatus and direction of future research.

For this study we use the JANUS Recognition Toolkit (JRTk) (Zeppenfeld et al., 1997). For this study the JANUS system was extended by several tools for processing and modeling prosodic information. We describe the implementation of the JANUS Prosodic Tool Kit (JPTk) in an appendix. This set of additional JANUS objects allows the extraction of various prosodic information from the signal and the training of different bounded prosodic models, for example, phone, syllable, or world based prosodic models. All these JANUS objects have an Tcl/Tk interface and can be accessed and configured at the Tcl/Tk level without consulting the respective C-code.

Chapter 2

Communicating Emotions

Within human communication, messages can be conveyed by verbal and non-verbal devices. For instance, for the pure exchange of information verbal devices seem to be the most common way whereas for the expression of emotions, non-verbal devices tend to dominate (Knapp and Hall, 1997). In some cases, non-verbal devices might even substitute for the verbal devices completely. For example, a convincing way to express anger at someone might involve just a wiggling finger or the denial to cooperate in the conversation altogether. In other cases, the verbal message is intensified by non-verbal means. A felicitous utterance of 'I love you' should be accompanied by certain facial expressions and additional vocal cues.

Verbal communication is based on the the choice of words and their linear order within an utterance. Non-verbal communication, in contrast, relies on means such as extra speech sounds (hissing, whistling, or laughing), special qualities of the voice (giggling or whiny), or the modification of prosodic parameters (pitch, intensity, or speaking rate). Non-verbal communication is not necessarily confined to the audio channel. It can open additional communication channels by allowing body postures, gestures, or facial expressions to participate in the communication. The visual channel can carry additional independent information or modify the information delivered in parallel within the audio channel.

In this investigation we look at both verbal and non-verbal devices in the communication of emotions and we investigate what specific means speakers employ to express an emotion. We confine the investigation to the audio channel and focus on emotion-specific spectral and prosodic changes and on verbal devices such as word choice.

The following section starts with a definition of emotions, a definition which allows us to distinguish emotions from reflexes, drives, and moods. Following, we describe devices used to communicate emotions. We first look into verbal devices at various linguistic levels such as morphology, semantic, and syntax. We then cover non-verbal devices employed by a speaker to signal an emotion: spectral changes and the modifications of prosodic parameters. The latter, the modification of prosodic parameters, constitutes the focus of this investigation. As the discussion shows, the communication of emotions relies on the modification of parameters which serve other communicative functions as well. In particular this seems to be the case with prosodic parameters which implement several other linguistic functions such as turn taking and syntactic segmentation and disambiguation. In general, there is no reserved emotion-specific verbal or

non-verbal cue which signals a certain emotion. What makes the communication of emotions succeed nevertheless is the combination of cues given by the verbal and non-verbal parts within a message.

2.1 Emotions

This investigation is certainly not the place to go into a detailed discussion of emotions. However, it is necessary for us to describe and explain how we use the word "emotion" in order to avoid further misunderstandings during this investigation.

Following (Lazarus, 1994), we think that an *emotional reaction* comprises three components:

1. Experiences that include a cognitive evaluation of the current situation (appraisal).
2. Physiological reactions at various levels. For example, the sympathetic nervous system is aroused when one feels angry or happy. Along with this arousal the blood pressure and the heart beat increase. In contrast, the antagonistic parasympathetic nervous system is aroused when we feel, for example, sad. The heart beat and the blood pressure decrease. The arousal of both, the sympathetic and the parasympathetic nervous system has specific effects on the speech production process, the details of which are given below.
3. Impulses to act in a certain way. For example, these action can impulses include fleeing, shouting at someone, or attacking. We extend this list of possible action impulses by verbal and non-verbal actions. We consider a particular choice of words or a certain syntactic construction within an individual's utterances to be an action as well. For instance, the utterance of "Shut up!" certainly indicates through the choice of words and the imperative word order an angry speaker. Compare this with an utterance of "Could you be quiet please?"

The importance of these three components of an emotional reaction becomes clearer when we use them to delimit an emotional reaction from, for example, sensorimotor reflexes, such as the patellar or the pupillary reflex, or physiological drives, such as hunger. For a reflex, releasing a specific sensory stimulus automatically triggers a fixed motoric response pattern, not involving any cognitive activity. Thus a reflex utterly misses the cognitive evaluation of the current situational context. In addition, within an emotional reaction, there exists only action tendencies or impulses to act in a certain way; actions that do not have to be carried out necessarily. For a reflex, in contrast, the response action is fixed and performed automatically. Following the reasoning above, we classify distaste as a sensorimotor reflex to offensive substances while we classify disgust as an emotional reaction.

While physiological drives do not dictate the corresponding action as directly as reflexes do, drives still do lack the cognitive component which we consider essential for a reaction to be emotional.

Looking at the other extreme, how can we distinguish between emotional reactions and moods using the definition as given above? The main distinction between emotions and moods is the lack of an evaluation of the current situational context in case of moods. Moods seem to refer to larger segments of our life span and the reason for a particular mood, say melancholy, is not related to a single, current situation but to a number of experiences dating back possibly for quite some time. Also, moods seem to lack specific actions impulses.

We can look at emotions from several perspectives. For this investigation we adapt the perspective of an outside observer of an emoting individual. Thus, we are interested in observables that allow a reliable classification of an emotion expressed by some speaker.

2.1.1 Observing Emotions

How can an outside observer tell someone's emotional state? What are the features in someone's behavior which allow for a reliable classification? We can base our inferences on information from three different domains: the social context, the physiological changes, and actions of the emoting individual.

For this investigation we ignore the social context in which an emotional reaction takes place. That is, we do not consider, for instance, environmental demands or constraints. Thus, we are not modeling inferences of the form: the individual was socially insulted and is therefore angry. Instead we focus on features based on physiological changes and actions as a consequence to an emotional reaction.

When we attempt to determine an individual's emotional state resulting from an emotional reaction, we can consider physiological parameters such as automatic nervous activity, hormonal secretion, brain activity, heart beat, skin resistance, or blood pressure. Most of these parameters are not observable without extensive machinery and are highly inconvenient to the emoting subject, and more appropriate for psychological studies of emotions (Katz, 1997). We do not consider these parameters in our investigation.[1]

Some of the physiological reactions, in particular, the arousal of the sympathetic or parasympathetic nervous system, have an effect on the speech production process. For example, the arousal of the sympathetic nervous system tends to quicken speech and to increase the energy distribution in the high frequency bands. In contrast, an arousal of the parasympathetic nervous system lowers the speaking rate, the fundamental frequency and the overall energy distribution (Scherer, 1986). Spectral and prosodic properties in an individual's voice, consequences of physiological processes from an emotional reaction, are the first domain which we explore for observables (Davis et al., 1996).

The second domain are action impulses of an emoting individual. Remember that the list of possible actions which are responses to an emotional reaction is extended by linguistic actions. We consider a particular choice of words or a certain syntactic construction a verbal action. Thus, we investigate whether a certain word choice or particular syntactic constructions yield information about someone's emotional state.

There are several points we want to mention in this context. First, the arousal of the nervous system is not a reliable sign for an emotional reaction in general. There might be other reasons for its arousal. Second, note that the arousal of, say the sympathetic nervous system, is not specific enough to infer a particular emotional reaction. We mentioned earlier that the sympathetic nervous system is aroused when we are, for example, angry or happy. Third, the observation of certain spectral and prosodic properties in an individual's voice does not necessarily allow the conclusion of a particular expressed emotion. As the discussion in the following sections shows, particular prosodic parameters are modified by several linguistic functions such as segmentation and accentuation. Nor does the absence of spectral and prosodic cues dictate that the speaker is not experiencing a certain emotion. He or she might use verbal or different non-verbal means to encode his or her emotion. The same caveats apply to verbal cues. The presence of certain verbal

[1] Progress in the development of computer hardware will soon facilitate some non-invasive assessments of physiological data without any inconvenience to the human (Picard and Healey, 1997; Scheirer, Fernandez, and Picard, 1999). However, physiological data exhibits a large variance and emotion detection algorithms based on physiological data require specialized modeling techniques (Vyzas and Picard, 1998; Vyzas and Picard, 1999).

material does not necessarily mean that the person is experiencing a certain emotion because there are other reasons for using this material, for instance, certain speaking styles. Nor does the absence of verbal cues necessarily determine that the speaker is not emotionally involved. Thus, verbal or non-verbal cues looked at in isolation are highly unreliable indicators for an emotion. It is the merging of spectral, prosodic, and verbal information which allows the communication of emotions to succeed.

Because of this indetermination of emotional cues we think it necessary to investigate as many verbal and non-verbal cues as possible with the underlying idea that the integration of all these cues allows a reliable detection of an expressed emotion.

2.1.2 Categorical or Dimensional

So far we made two simplifications while discussing emotions. First, we assumed that it is only possible to experience a single emotion at a given time. Second, we treat emotions as categorical.

We certainly do not deny that several emotions can occur at the same time. However, for our investigations we assume that within a speech segment the speaker experiences a single constant emotion. In fact, we try to segment our corpora to ensure that this is the case; see chapter 4 for more details. The main reasons for assuming a single constant emotion in a speech segment are the complications for the tagging, training, testing, and evaluation processes which would arise otherwise.

By assuming emotions to be categorical we do not deny within-category variation. For instance, we consider both annoyance and rage to fall within the anger category, even though they occupy quite different points on the scale ranging from mild anger to intense anger.

There is a completely different way to look at emotions which denies them their categorical status all together. In a dimensional view, emotions are clustered on the basis of their properties along several dimensions (Lazarus, 1994). That is, emotions are thought of as having two or three qualities and each emotion has a characteristic quantitative instantiation of these qualities. For example, Watson and Tellegen (1985) assume two qualities: positive affect and negative affect. Emotions which have a high quantity of positive affect comprise for instance peppy, excited, and elated. Low quantities of negative affect include calm and relaxed, in contrast to nervous or afraid which require a high amount of negative affect.

Given the fact that the following investigation focuses on the very basic emotions happy, sad, afraid, and angry (Ortony and Turner, 1990), the question whether emotions should be treated as categorical or dimensional does not really arise. However, for future studies of emotional speech, involving the whole range of emotional variations, it might be promising to pursue a dimensional approach.

2.2 Verbal Cues

The influence of a speaker's emotional state on his or her linguistic performance has not been a major topic within linguistic research (Fiehler, 1990a). Until recently, linguistic research has been primarily focused on written text – most of the time single sentences – for which the notion of emotion is only marginally relevant. With the emergence of spontaneous language as a central subject of linguistic research, the picture changed somewhat. However, research on verbal cues in the communication of emotions is still very sparse (Fiehler, 1990a; Hübler, 1998). Existing linguistic research on the impact of the speaker's emotional state is mainly concerned with the distinction of neutral and emotional language. Emotion-specific linguistic choices are not discussed in this research. Most studies describe linguistic devices signalling any kind of emotional involvement and are agnostic about the actual underlying emotion. We think that the main reason for not looking into emotion-specific verbal cues is that verbal information is highly ambiguous without extensive context or non-verbal information. The actual emotion expressed depends heavily on the context of the utterance and the way it is delivered, that is, the way non-verbal cues accompany the verbal message. Consider, for instance, an utterance of (2.1). The presence of the adverb *so* indicates some emotional involvement of the speaker.

(2.1) It is so warm!

But depending on the context and the way it is pronounced, (2.1) could either express someone really satisfied and happy with the current temperature or someone really getting upset because of the warm weather.

There is, of course, always the possibility of being very direct about the emotional state you are currently in. That is, after some introspection you directly name the emotional state you are in, for instance:

(2.2) I'm angry!

Following Hübler (1998), we call this way of expressing ones emotional state *explicit*. The other way, the way which concerns us in this investigation, is called *implicit* and refers to expressing the emotional state by verbal cues, such as certain lexemes or certain syntactic constructions.

Affect sounds constitute a borderline case since they usually do not form a lexical unit. Under affect sounds we understand phenomena such as moaning, crying or laughing. We do not review their potential to indicate a particular emotion since we do not explore their impact in this investigation. However, emotions can be expressed by a speaker through interjections which we consider as lexical units. In the following we describe how a speaker can express his or her emotional involvement by making certain choices offered by linguistic devices at various levels, such as the morphological, the lexical and the syntactic levels.[2]

Probably the most obvious way to encode verbally an emotion other than explicitly naming it, is to use certain lexemes. The use of interjections and exclamations indicates, in general, that

[2] We illustrate the following sections with several examples, some of which are drawn from the corpus comprising movies and talk shows. This corpus is described in more detail in Chapter 4.

the speaker is undergoing some emotional experience (Volek, 1987). Some examples are: Alas, Shoot, Darn, Heck.

Other verbal cues make use of sometimes very drastic metaphorizations (Davitz, 1969). One such metphorization – quite common in English – is given in (2.3).

(2.3) Son of a bitch!
(As Good As It Gets)

Other lexemes have certain connotations associated with them (Volek, 1987) and their mentioning within an utterance advertises certain emotions. Some connotation-loaded lexemes are given in (2.4) and (2.5).

(2.4) What are you jabbering/talking about?
(Volek, 1987)

(2.5) You are a spoiled rotten little brat!
(Kramer vs. Kramer)

Sometimes it is enough to intensify certain word meanings by grading adverbs or adjectives to render the whole construction emotional, see (2.6)-(2.9) for some examples.

(2.6) schrecklich heißes Wasser
(terribly hot water)
(Mathesius, 1964)

(2.7) Something terrible has happened! (The Sweet Thereafter)

(2.8) I'm in this fucking city! (The Sweet Thereafter)

(2.9) We ain't got the slightest fucking idea what happened to Mr Blond or Mr Blue!
(Reservoir Dogs)

Certain morphological changes can also signal an emotional involvement of the speaker. These cues seem to be more subtle than the lexical cues mentioned above. Morphological cues indicating emotional involvement include diminutive and augumentative forms. An example is given in (2.10) and (2.11).

(2.10) Mann - Männlein
man - small man
(Volek, 1987)

(2.11) bacio - bacione
kiss - big kiss
(Volek, 1987)

We also find intensification processes at the morphological level achieved by compounding. An example is given in (2.12).

(2.12) zuckersüß
sugar sweet
(Mathesius, 1964)

Intensification seems to be the prevalent tool to render verbal messages emotional. At the syntactic level we find several intensification constructions as well. Two typical syntactic intensification tools, the asyndetic and syndetic repetition, are exemplified in (2.13) - (2.16).

(2.13) She is very very sweet!
(Mathesius, 1964)

(2.14) A long time ago, in a galaxy far far away ...

(2.15) That makes me very very mad that's why I came all the way up here.
(The Sweet Hereafter)

(2.16) She cried and cried!

Other syntactic cues are more complex and involve the movement of a substantial amount of verbal material. One such cue is the subjective word order which we illustrate with (2.17).

(2.17) Where did that bitch disappear to, I would like to know!
(I would like to know where that bitch disappeared to!)
(Volek, 1987)

In addition, in (Hübler, 1998) it was argued that grammatical devices such as the present perfect, the periphrastic *do*, or the get-passive indicate an emotional involvement of the speaker.

In most of the cases the linguistic device indicating a speaker's emotional involvement consists of more language material than the neutral alternative. Using a more linguistically motivated term, it seems that the emotional forms are *marked* with regard to the neutral alternative (Givon, 1991). Assuming that speakers do not tolerate synonyms in language (de Saussure, 1816; Clark, 1990), we have to attribute a special meaning to these choices. That is, the choice of the marked form is interpreted to indicate the speaker's emotional involvement (Hübler, 1998).

As mentioned earlier, linguistic research's recent focus on the communication of emotions through verbal cues is descriptive, not yet formalized. This lack of a formalization prevents a straightforward operationalization of this research. However, the research does suggest that a data driven classification approach is promising. In our investigation we captured emotion-specific verbal cues with a statistical approach which modeled the probability of a certain word given a history of previous words and the emotion expressed by the speaker. The history of previous words is normally confined to one or two to circumvent the sparse data problem and to allow for a reliable estimation. With this constraint on the history, we could not hope to model complex syntactic cues such as (2.13) or (2.17). Lexical and morphological cues, in contrast, do not require a long history of previous words and could be modeled with this kind of probabilistic

model. Finally, note that this probabilistic model is very similar to traditional language models used in speech recognition systems (Jelinek, 1998).

2.3 Spectral Cues

The role of spectral information in the communication of emotions was demonstrated in an experiment by Lieberman and Michaels (1962), in which they resynthesized only pitch and intensity information from the signals of emotional speech segments, thus removing basically all spectral information. While the emotions of the original segments were recognized by human listeners with an accuracy of 85%, this accuracy decreased dramatically to just 47% for the sentences in which the spectral structure was filtered out and only prosodic information was preserved. Thus, information other than prosodic information was also able to signal the emotions originally expressed in the sentences. Voice quality – for instance, its clearness or pleasantness – can carry information about the vocal emotional expression. If changes in the voice quality correlate with the expression of certain emotions than particular distributions of spectral energy might indicate these very emotions.

Scherer et al. (1991) tried to approximate the spectral information responsible for communicating emotions by two numbers: first, the slope of the regression line of the energy decrease from low to high frequencies and second, the percentage of the total energy below a cutoff level of 635Hz. The results of this experiment showed some correlation of these parameters with the expressed emotions. For example, anger seemed to be positively correlated with a very high proportion of high frequency energy. In an additional experiment, Banse and Scherer (1996) compared prosodic and spectral features to predict the emotion expressed in an utterance. Prosodic features outperformed spectral features significantly. Frick (1985) claimed that the spectral bands were narrower in the speech of grieving persons due to a high position of the larynx in the vocal tract. Other studies found spectral changes to correlate with the expression of anxiety (Roessler and Lester, 1976). However, most of the results mentioned above were based on very small data sets and it is far from clear whether the findings transfer to other speech corpora.

Note that the above experiments exploring the systematic variations in vocal quality due to the expression of certain emotions were based on differences in the short-term spectral representation of a speech segment. One of the problems with using the short-term spectrum is that it does not necessarily reflect the properties of the human auditory system. The human auditory system has the best resolution for frequencies under 500Hz (Zwicker and Fastl, 1990). To model the perceptually relevant aspects better, Davis and Mermelstein (1990) proposed to place filters linearly at low frequencies and logarithmically at high frequencies. Furthermore, in several speech recognition experiments Davis and Mermelstein (1990) demonstrated that mel-frequency cepstral coefficients yielded superior performance than other parametric representations of the speech signal. Cepstral coefficients are the result of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale (Davis and Mermelstein, 1990; Schafer and Rabiner, 1990).

The circumstance that cepstral coefficients delivered superior recognition results than other parametric representation of the speech signal motivated J.H.L. Hansen to explore the correlation of these coefficients with the emotions expressed in a speech segment. (Cairns and Hansen, 1994; Hansen, 1992; Hansen and Clements, 1993; Womack and Hansen, 1995; Womack and Hansen, 1996; Hansen and Womack, 1996). Hansen's idea was to detect first the emotional state of the speaker and, second, using information about the underlying emotional state of the speaker, to improve recognition accuracy. For his experiments Hansen relied on the SUSAS database (Speech Under Simulated and Actual Stress). This database had a relative small vocabulary; 35 aircraft words made up over 95% of the database. The words were mono- and multisyllabic and were highly confusable. The database comprised a total of 16,000 speech segments produced by 32

speakers (male and female) covering five domains:

- psychiatric analysis data (fear, depression, anxiety)
- speaking styles (slow, fast, soft, loud, angry, clear, quotation)
- speech produced in a noisy environment (Lombard)
- dual tracking computer task
- subject motion fear tasks (G-force, Lombard effect)

Womack and Hansen (1996) used mel-scale cepstral coefficients as input to their neural-network based classifiers. In addition to these cepstral coefficients, Hansen applied two additional features: auto-correlation and cross-correlation

Classification results across 11 stress conditions ranged from 46% to 79% correctly classified segments depending on whether the respective word was in the training vocabulary or not. Both performance numbers were greater than chance. Using stress dependent hidden Markov model recognizers, Hansen was able to improve the word accuracy when compared to neutral or multi-style trained recognition systems.

In our investigation spectral information was also modeled by means of cepstral coefficients. However, we investigated adaptation techniques to model emotion-specific spectral differences. In addition, for our investigations different corpora were used comprising a substantially larger than Hansen's vocabulary and a greater number of speakers. Note also that Hansen's classifiers were speaker dependent while our classifiers are speaker independent.

2.4 Prosodic Cues

The focus of this investigation lies on the detection of emotion-specific modifications of prosodic parameters. Prosody, however, is not reserved for the communication of emotions but participates in several other linguistic processes as well. The fact that prosody attends to several tasks in parallel might lead to interferences, and makes the interpretation of prosodic observations difficult. At the beginning of this section we give a definition of prosody and related terms followed by a description of some of the linguistic functions of prosody other than the communication of emotions. We conclude this section with a description of research emphasizing the role of prosody in the communication of emotions.

2.4.1 Prosody

Prosody can be thought of as suprasegmental information achieved by the modification of segmental acoustic parameters such as energy, fundamental frequency, and duration. Prosody is, so to speak, "parasitic", i.e. prosodic information is established by modifying acoustic features which are already present in the signal for other reasons.

The essential difference between prosodic (suprasegmental) and acoustic (segmental) information is that prosodic information can only be observed by considering a substantial context. For example, a word accent can only be observed by considering the duration, energy, and fundamental frequency of segments in the context of the respective word (syntagmatic relation). In contrast, acoustic information, such as the roundedness of a vowel, can be observed without reference to neighboring segments (paradigmatic relation) (Lehiste, 1970).

Speech can either be described articulatorially, acoustically, or auditorially, depending on the part of the communication process we are referring to. An articulatory description of speech is mainly concerned with how speech is produced by the speaker. When we are concerned with the perception of speech by a listener we describe speech via auditory processes, i.e. how the listener perceives speech. Finally, acoustic notions refer to aspects of sounds which are independent of the production or perception processes, and which can be measured automatically. It is important not to confuse these notions. For example, *pitch* is an auditory term, that is, the perceived height. It is analogous is to *fundamental frequency* (F_0) in terms of acoustics. In articulatory terms we would say that pitch is increased or decreased by the tension of the vocal cords.

Note that there is not necessarily an isomorphism between these terms. The subjective experience of pitch is primarily based on the fundamental frequency but both duration and energy can modify this experience as well. The frequency resolution of the human ear is best for frequencies under 500 Hz. Within this range humans are able to detect differences as small as 1 Hz. In addition, within this frequency range, the duplication of the frequency is experienced as twice as high: Thus, the difference from 100 to 200Hz is experienced exactly as large as the difference between 200 to 400Hz, an octave in both cases (Zwicker and Fastl, 1990). Thus, it is not surprising that the fundamental frequency of human speech falls within this range. The fundamental frequency typically ranges for men from 60 to 240Hz and for women from 130 to 400Hz. For childrem the fundamental frequency can exceed 500Hz.

Similar points apply to the subjective experience of loudness. The experience of loudness is primarily based on the energy distribution within the signal but both duration and fundamental

frequency have an impact as well (Zwicker and Fastl, 1990).

The importance of prosodic information within the communication process is evident when we look at the difficulty humans experience when they try to understand non-natives trying to speak their mother tongue. Most of the time, it is the distribution of stress and incorrect pitch contours that render this speech unintelligible.(Cowie and Douglas-Cowie, 1996). Another area to observe the importance of prosodic information is automatic speech synthesis. To a large extent, the missing or poorly modeled prosody is the reason why synthetic speech sounds so unnatural (Cahn, 1990; Murray and Arnott, 1993).

We can divide prosodic information into three classes corresponding to the spectral, the energy, and the time dimension:

- timbre, quality, tone pitch,
- intensity,
- rhythm, speaking rate, silence.

As the following sections demonstrate, most of the time prosody relies on the modification of all parameters in these three classes to implement a certain linguistic function. It is therefore important to consider all three dimensions when investigating the role of prosody in human communication.

2.4.2 Multi-Functionality of Prosody

Prosody is employed at various linguistic levels to implement a multitude of communicative functions. The fact that prosody participates and interacts so actively at several linguistic levels makes it difficult to recover a particular function given a particular prosodic observation. For instance, an increased fundamental frequency can implement a yes-no question or indicate an emotional speaker. We review briefly some of the communicative functions of prosody in the following paragraphs to illustrate its multi-functionality.

One major task of prosody is to emphasize certain segments. The saliency of certain segments is used to implement word stress or accents. For example, a given word can belong to more than one part of speech. It is the distribution of the stressed syllable within the word which assigns it a unique part of speech:

(2.18) to im**port** *vs.* the **im**port

(2.19) to in**sult** *vs.* the **in**sult

In other cases, the distribution of the stress distinguishes between homographs:

(2.20) **um**fahren *vs.* um**fah**ren
to run over, to drive round
(Kießling, 1997)

(2.21) **Ten**or *vs.* Te**nor**
(Kießling, 1997)

Other words can be used either as sentential adverbs or as cue phrases that can explicitly indicate discourse structure. An example is the word *incidentally* which, when used as cue phrase, indicates some kind of digression. Hirschberg and Litman (1993) claimed that pitch accent and prosodic phrasing enable a distinction between these two usages.

Stress does not have to be confined to segments within a word. A whole word or phrase can be accentuated to emphasize its importance in the current discourse state or to disambiguate among possible semantic interpretations. Some examples for German are given below:

(2.22) Ich ziehe **Dres**den Frankfurt vor
I prefer Dresden to Frankfurt

(2.23) Ich ziehe Dresden **Frank**furt vor
I prefer Frankfurt to Dresden

(2.24) Dann müßten wir noch einen Termin ausmachen
Then we need another meeting date.
Then we still need a meeting data.
(Kießling, 1997)

In the first two sentences the distribution of the accent which can either be on *Dresden* or *Frankfurt* identifies the accusative objects. In the last example, the interpretation of the sentence depends whether the particle *noch* is accentuated.

Stressed segments appear to be more salient from their surroundings. It seems to be the case that most prosodic parameters can be used to render a segment more salient. In general, the salient segment is characterized and distinguished quantatively from surrounding segments, that is, it is longer, louder, or higher (Kießling, 1997).

Prosody can also be used to segment speech, i.e. which group of syllables form a word (Barry, 1981). Some examples are given in 2.4.2.

(2.25) Staubecken *vs.* Staub ecken
(Kießling, 1997)

(2.26) bewußter leben *vs.* bewußt erleben
(Kießling, 1997)

Prosody also helps to segment and disambiguate at the syntactic level. Syntactic ambiguities arise when a given surface string is assigned more than one valid syntactic structure. Depending on the design and the size of the underlying grammar, the amount of syntactic ambiguity can be quite large. Ambiguity is a difficult problem in every natural language processing system. For spoken language, prosodic cues given in the utterance reduce this ambiguity. The utterance does not simply consist of a homogeneous list of words but of words that are grouped into chunks, prosodic phrases (Wightman et al., 1991). An example for English is given below:

(2.27) There was one bottle under the bridge and another on the park bench.
Andrea moved the bottle under the bridge.
vs. Where did Andrea move the bottle?
Andrea moved the bottle under the bridge. (Wightman et al., 1991).

The prepositional phrase *under the bridge* can either modify the preceding noun phrase *the bottle*, or, as exemplified in the second case, the verb *moved*. Another example for syntactic ambiguity is given below:

(2.28) ja zur not geht's auch am Sammstag
(Kießling, 1997)

Without any prosodic information the above sentence has at least 36 different syntactic analyses with different semantic interpretations. When using prosodic constituency for syntactic disambiguation, the underlying assumption is that non-similar syntactic structures are realized differently at the prosodic level. Syntactic structures and prosodic structures correlate to some degree, i.e. the greater the prosodic break between two words the more the corresponding syntactic constituents are separable; or conversely, the smaller the prosodic break the greater the syntactic cohesion between the corresponding constituents. Note, however, that the correspondence between the prosodic structure and the syntactic structure is not isomorphic. Several prosodic features are used to indicate phrase boundaries: energy, fundamental frequency, durations of preboundary segments, and the duration of of the pause between the phrases. The fundamental frequency typically falls at the end of a major phrase and rises with the beginning of following phrase (Lea, 1990). Another important cue to signal a syntactic or discourse boundary is preboundary lengthening (Wightman et al., 1991) and a silent pause (Lea, 1990).

Pitch contours are also assumed to encode specific pragmatic meanings. A common example is the distinction of statements and Yes-No questions. It is the final drop vs. the final rise of the intonation which distinguishes between the otherwise possible identical utterances (Lehiste, 1970).

The pitch contour can also be used to discriminate among several pragmatic categories in infant-directed speech. Katz, Cohn, and Moore (1996) used prosodic summary features and the pitch contour to discriminate among three pragmatic categories (getting attention, showing approval, and providing comfort).

This incomplete list of prosodic functionality illustrates that prosody interacts with basically all linguistic levels. This fact makes prosody a very important subject of research. However, the following caveats apply to research relying on the observation of prosodic features:

- Within a single utterance several prosodic functions can occur at the same time and are implemented by the modification of the same prosodic parameters. This superimposition of prosodic parameters complicates the attribution of a certain prosodic observation to a particular function.
- Prosody relies on features that are also used in segmental information. Thus, we have to be aware of the influence of segmental information on prosodic parameters. For example, low vowels, such as /a/, have an intrinsic lower fundamental frequency than high vowels, such

as /i/. Moreover, the duration of vowels is not only determined by the overall speaking rate but also by the actual segmental context. For instance, vowels in the context of unvoiced consonants tend to be shorter than in a voiced context.

- Prosody is to some extent optional. Its functionality within an utterance or discourse can be implemented by other linguistic devices.
- Prosodic parameters interact. The realization of a certain prosodic function can be achieved by the modification of different prosodic features. For example, a word accent can be implemented by a either increased intensity, pitch, duration, or a combination of all three.

Besides the linguistic functions listed above, prosody also implements several indexical functions that indicate the age, the sex, regional and social upbringing, and other traits. Other additional functions of prosody comprise the signalling of irony, doubt, or rejection. In the following sections we consider one of these additional functions of prosody: the expression of emotion.

2.4.3 Prosody and Emotions

In order to demonstrate the role of prosody in the communication of emotions, Scherer (1971) asked actors to portray certain emotions in their speech. Subsequently the recorded speech was filtered by a high pass filter to eliminate all spectral energy above 400Hz. This method effectively filtered out the verbal meaning but preserved many of the prosodic features. Listeners were still able to recognize the underlying emotion from a list of possible emotions with a better-than-chance accuracy.

The experiment above is interesting in several ways. First, it raises the question of how to collect emotional speech. Is it valid to use actors, or are there other, more realistic ways to elicit and collect emotional speech? Second, if it is indeed the case that by the modification of prosodic parameters one can communicate emotions as suggested by this very experiment, what prosodic parameters are modified? And, third, are there emotion-specific prosodic parameter settings? In this section, we try to find some preliminary answers to questions number two and three, that is, what are the relevant prosodic parameters in the communication of emotions and are there emotion-specific prosodic parameter settings which allow the detection of particular emotions? The issue of data collection is handled in chapter 4.

The experiment by Scherer (1971) as outlined at beginning of this section relied on the same experimental format found in nearly all studies investigating the role of prosodic parameters in the encoding of emotions: certain prosodic features were removed from the signal by certain filters and subjects were asked to detect the underlying emotion of the thus altered utterance. The most simple of these experiments involves whispered speech. The simple fact that whispered voice also communicates emotions forces us to assume that it is not pitch alone which encodes the underlying emotion of the speaker (Knower, 1941; Tartter and Braun, 1994). Remember that whispered speech does not involve voicing.

Other experiments employed more extensive machinery to control certain prosodic parameters. For example, Lieberman and Michaels (1962) used vocoders to control the parameters of intensity and fundamental frequency. They found that not only fundamental frequency but also the intensity contributed to the encoding of an emotion. Uldall (1960) added a synthetic

pitch contour to four neutral sentences which were originally spoken by humans but were now reproduced through a vocoder. Uldall found that intonation contours carried information about the strength of the emotion, its friendliness, and about the authority relationship between the speaker and the hearer.

In general, it can be observed that the deprivation of prosodic information from an utterance negatively correlates with the accuracy its underlying emotion can be detected. However, the remaining acoustic information in the utterance still allows a better-than-chance recognition of the underlying emotions. All major prosodic features seem to have an impact on the communication of emotions.

The above experiments show that there is strong evidence that prosody plays an essential role in the communication of emotions. This leads us to the next question, are there specific prosodic parameter settings that correlate with certain emotions?

The studies on emotion-specific prosodic parameter settings do not yield a very coherent picture. This is largely due to the differences in the methods of data collection, the number of emotions and subjects used in these studies. However, the studies do converge on some trends of emotion-specific prosodic parameters which we describe in the following sections. We confine the following description to the emotions studied in our own experiments.

Most studies extracted global prosodic features, i.e. features pertaining to the whole utterance, and tried to correlate these features with the underlying emotion. The shortcoming of these summary features to characterize emotions became evident by a simple experiment: presenting the utterance backwards reduced substantially the recognition accuracy from 89% to 43% (Knower, 1941). The inversion of the utterance left the prosodic summary features untouched while the contours of the fundamental frequency and intensity were changed. Thus, the contours of the fundamental frequency and intensity do carry information relevant for the decoding of emotions. Based on his studies, Scherer (1974) suggested that falling pitch contours correlate with pleasantness. Rising contours, on the other hand, suggest surprise or fear. Note however, that there might be other reasons for a rising or falling pitch contour, see section 2.4.2 above for more details.

Prosodic Features of Happiness

Happiness is often described as having *gentle* contours and some regularity (Davitz, 1964; Fonagy, 1978). The mean, range and variability of both fundamental frequency and intensity increases when compared to neutral speech.

Prosodic Features of Sadness

When compared to neutral speech, sad speech shows a decrease in the mean of the fundamental frequency, and a very low mean intensity (Davitz, 1964). Moreover, the range of changes in the fundamental frequency is very small (Fonagy, 1978). Another important prosodic cue in the communication of sadness is a slow speaking rate, involving extended intra- and inter-utterance pauses (Fairbanks and Hoaglin, 1941; Siegman and Boyle, 1993)

Prosodic Features of Anger

The expression of anger involves an increase in the fundamental frequency, high variability and range. The intensity of angry speech also increases when compared to neutral speech (Frick, 1985).

Prosodic Features of Fear

Due to the difficulty of collecting authentic fearful speech there are only a handful of studies considering this emotion. Fonagy (1978) reported an increased fundamental frequency. Also the range and the variability of the fundamental frequency increase. Intensity also increases compared to neutral speech.

2.4.4 Automatic Detection of Emotions

The research on the automatic detection of emotions using prosodic features is quite limited. We think that the small number of studies is related to limited availability of speech corpora. Not surprisingly, studies which investigate the relation of emotions and prosody and which do not require emotional speech corpora are more frequent. There is, for example, a substantial body of research on the synthesis of emotional speech focusing on prosodic parameters (Mozziconacci and Hermes, 1999; Mozziconacci, 1998; Vroomen, Collier, and Mozzicanacci, 1993; Murray and Arnott, 1996; Murray and Arnott, 1993; Heuft, Portele, and Rauth, 1996; Sato and Morshiama, 1996; Cahn, 1990).

Some studies, however, explicitly used machine learning techniques to automatically classify the emotion expressed by a speaker using prosodic information. Dellaert, Polzin, and Waibel (1996) applied several statistical pattern recognition techniques for the classification of emotional speech. To model the contour of the fundamental frequency, a smoothing spline approximation was used and combined with a majority voting of subspace specialists. The classification accuracy on the Woggles corpus was comparable to human performance. See also section 4.2.

Amir and Ron (1998) used a corpus consisting of utterances of 24 subjects (12 male and 12 female). The subjects were asked to recall a past event which evoked one of the five emotions: happy, angry, sad, afraid, and disgusted. The subjects were asked to talk about that event and to participate emotionally. For the automatic classification of the emotional content, prosodic features were extracted from the speech samples. The interesting aspect of this investigation was that the classification system computed a fuzzy membership index for each emotion. This approach allowed in principle to model intensities of a particular emotion and the superimposition of several emotions.

Preliminary studies by Thymé-Gobbel (1998) on English and Spanish telephone data suggested that prosodic features could also be used to assess the emotions expressed by the speakers participating in telephone conversations. Other studies tried to integrate both visual and prosodic information to detect emotions (Chen et al., 1998).

We think that the exploration of cues signalling a certain emotion has to be based on a

substantial amount of data to prevent idiosyncrasies of speakers or utterances from being considered as reliable cues. For instance, (Tischer, 1993) based his investigation of prosodic features signalling emotions on the sentence:

(2.29) Sag das nochmal. Ich kanns nicht glauben. Was für ein Tag.
Say it again. I can't believe it. What a day.

This sentence was uttered by four speakers in different emotion-provoking story contexts. As mentioned earlier in this chapter, prosody participates in the implementation of several linguistic functions other than the expression of emotions. Investigating the role of prosody in the communication of emotions on such a small data space runs the risk of producing idiosyncratic prosodic features which most probably fail to discriminate the expressed emotion in an utterance of a different sentence by a different person. Thus, our investigation is based on four corpora constituting the largest collection of emotional speech we are aware of.

In addition, we think that emotions are not necessarily expressed solely by acoustic cues but that additional cues can be given by certain word choices or syntactic constructions. Therefore, our approach explored – in contrast to the approaches as sketched above – in parallel spectral, prosodic, and verbal information for cues to communicate emotions. We modeled verbal information with emotion-specific language models. Spectral information was modeled by means of cepstral coefficients and emotion-specific adaptation. The emphasis of this investigation, however, lay on the study of those prosodic features which allowed a robust classification of the expressed emotions across speakers, utterances, and corpora. We explored prosodic features pertaining to the whole utterance, for instance, the mean of the fundamental frequency or the variance of the intensity within an utterance. In addition, we explored prosodic features which referred to smaller segments such as phones. One such features was, for example, speaking rate which we modeled with emotion-specific durations of vowels. We also explored the combination of spectral, prosodic, and verbal information to see whether this combination resulted in an overall improvement of the detection of the expressed emotions in some utterances.

In a possible application, the detection of the emotion expressed in some utterance would probably be combined with a speech recognition module and additional subsequent natural language processing modules. Because of the interdependencies of the expressed emotion with the acoustic and verbal properties of the utterance, we want a tight interaction of the emotion detection process with these modules. For instance, in one experiment we showed that the word accuracy depended on the emotion expressed in an utterance. In addition, expressing a certain emotion also changes the probability of certain words being uttered. Moreover, since prosodic information was a reliable indicator for the emotion expressed in some utterance, prosodic information should be part of this tight interaction as well. In the next chapter we introduce a hidden Markov model architecture which allows the direct integration of prosodic information into the speech recognition module. Thus, in principle, the recognition of speech and the detection of the emotion expressed by spectral, prosodic, and verbal cues become one integrated process by using this architecture.

Chapter 3

Modeling Verbal and Non-Verbal Information

In this chapter, we specify the underlying modeling assumptions to capture emotion-specific verbal and non-verbal information. We describe in the first section our approach to model emotion-specific verbal cues. Intuitively, we computed the probability of certain word combinations depending on the expressed emotion. We used back-off language models to compute these probabilities. The second section describes our approach to model emotion-specific prosodic information. Because of the interdependencies of spectral and prosodic information we modeled these two domains with the same underlying modeling approach relying on hidden Markov models (HMM). We developed a hidden Markov architecture which allows to summarize atomic HMM states into what we call suprasegmental hidden Markov states. The atomic states in this suprasegmental hidden Markov models (SPHMM) were used to model spectral (segmental) information. The suprasegmental states, having access to the overall time spent in their constituting states, were used to model prosodic events. Remember that prosodic events cannot be observed at the segmental level. See section 2.4 for details. Instead of training emotion-specific spectral models from scratch, we adapted existing spectral models to maximize the use of the limited amount of available emotion-specific training data. The description of this adaptation procedure is given in the last section of this chapter.

Note that the modeling assumptions made in the following sections are quite general. In fact, there are no references to emotions in the respective model definitions. In order to capture emotion-specific verbal and non-verbal information, we exposed these models to emotion-specific training data. Thus, we derived models which were only trained, for instance, on data which expressed sadness. For the purpose of testing for the emotion expressed in some utterance, we computed the likelihoods that this utterance was generated by emotion-specific models and took the highest likelihood to be indicative of the emotion expressed.

3.1 Modeling Verbal Information

The description of verbal information within the communication of emotions in section 2.2 shows that the phenomena are distributed among several linguistic levels ranging from morphology to syntax. The adequate modeling of all these phenomena presumably requires a full blown natural language processing system comprising inter alia lexical and syntactical analyses. However, modeling verbal phenomena by relying on such complex natural language tools was beyond the scope of this investigation. Instead we made use of a simpler technique which was proven to be a successful approximation for a variety of verbal phenomena. Language models represent the probability, $P(\mathbf{W})$, that certain words or strings of words, $\mathbf{W} = w_1 w_2 \dots w_N$, occur. Language models have been shown to effectively constrain and guide acoustic hypotheses within speech recognition systems (Jelinek, 1998). Language models are also capable of capturing idiosyncrasies of different text corpora and can be used to discriminate among them. For example, language models can be used to detect topics (Seymore and Rosenfeld, 1997b; Seymore and Rosenfeld, 1997a) or to infer discourse structure (Finke et al., 1998).

We can formulate $P(\mathbf{W})$ using Bayes's rule as

$$P(\mathbf{W}) = \prod_{i}^{N} P(w_i \mid w_1 w_2 \dots w_{i-1}) \tag{3.1}$$

where $P(w_i \mid w_1 w_2 \dots w_{i-1})$ is the probability that word w_i follows words $w_1 w_2 \dots w_{i-1}$. Estimation of $P(w_i \mid w_1 w_2 \dots w_{i-1})$ is impossible even for moderate values of i due to sparse data. Therefore in practice, different word strings $w_1 w_2 \dots w_{i-1}$ are treated as equivalent. A common mapping is to treat all word strings with identical last two words as equivalent:

$$P(\mathbf{W}) = \prod_{i}^{N} P(w_i \mid w_{i-1} w_{i-2}). \tag{3.2}$$

This kind of language model is referred to as a trigram. For the estimation of the probabilities of trigram language models, we count the number of times w_3 follows $w_1 w_2$ in a given training corpus divided by the times the word pair $w_1 w_2$ occurs in the same corpus:

$$P(w_3 \mid w_1 w_2) = \frac{C(w_1 w_2 w_3)}{C(w_1 w_2)}, \tag{3.3}$$

where C is a function counting the occurrences of its argument in the training corpus. Language models based on equation 3.3 face the problem that they assign zero possibility to trigrams which were never encountered within the estimation phase. There are several ways around this problem. For instance, probabilities of bigrams and unigrams can be incorporated into equation 3.3 by linear interpolation. For this investigation we use a different approach and apply a language model which is known as the back-off language model (Katz, 1987). Intuitively we make $P(w_3 \mid w_1 w_2)$ depend on $C(w_1 w_2 w_3)$. If $w_1 w_2 w_3$ occurs in the training set frequently, then the relative frequency is a reasonable estimation. In case there are only a very limited number of occurrences of $w_1 w_2 w_3$ within the training set, we back-off and approximate the trigram probability by the bigram probability of $w_2 w_3$ and so on:

$$\hat{P}(w_3 \mid w_1 w_2) = \begin{cases} \frac{C(w_1 w_2 w_3)}{C(w_1 w_2)} & \text{if } C(w_1 w_2 w_3) \geq K \\ \alpha Q_T(w_3 \mid w_2) & \text{if } 1 \leq C(w_1 w_2 w_3) < K \\ \beta \hat{P}(w_3 \mid w_2) & \text{otherwise,} \end{cases} \tag{3.4}$$

where α and β have to be chosen to normalize $\hat{P}(w_3 \mid w_1 w_2)$, K is a threshold, and $Q_T(w_3 \mid w_1 w_2)$ is a Good-Turing-type function. Note that equation 3.4 constitutes a recursion with the call of $\hat{P}(w_3 \mid w_2)$. For a detailed discussion of back-off language models consult (Jelinek, 1998). For our experiments, we used the language model toolkit CLAUSI developed by Klaus Ries (Ries, 1997). Note that back-off language models relying on a Good-Turing-type function do not require a development set to estimate the parameters α and β.

There are no references to emotions in the above formulas. In our experiments we made language models also to depend on the expressed emotion by training them on respective subsets of the training corpora:

$$\hat{P}(w_3 \mid w_1 w_2, \textit{expressed emotion}). \tag{3.5}$$

Thus, if a certain phrase occurs in one of these sets more often than in the remaining sets, other things being equal, the respective language model assigns a higher probability to this phrase than the language models trained on the remaining sets. In order to detect the expressed emotion in some utterance we computed the probabilities that the utterance was produced by each of the emotion-specific language models and then took the highest probability to indicate the expressed emotion.

3.2 Modeling Prosodic Information

We mentioned in section 2.4.1 the interdependence of prosodic (suprasegmental) and spectral (segmental) information. The influence of segmental on prosodic parameters can be divided into two classes. The first class consists of cases in which the influence is directly related to the underlying phone. For instance, low vowels have an intrinsic lower fundamental frequency than high vowels (Beckman, 1986; Lehiste, 1970). The second class comprises cases of coarticulation. For instance, vowels in a context of non-voiced consonants tend to be shorter than in a voiced context (Kießling, 1997). Several linguistic functions of prosody have an impact on spectral information (Campbell, 1995). Consider for example, the realization of an accented word which can be achieved by increasing the energy and the fundamental frequency. Remember also that prosodic information is implemented by the modification of segmental parameters.

Because of these interdependencies of prosodic and spectral information we developed an HMM architecture which models both spectral and prosodic information. However, in order to integrate prosodic information into a hidden Markov model, we had to overcome two inherent problems of the traditional HMM architecture (Ostendorf, Digalakis, and Kimball, 1997):

1. State durations are implicitly modeled by a geometric distribution.

2. Features are confined to be based on frame-based observations.

Features relying on observations confined, for instance, to 10ms frames, do not yield information about prosodic events. Thus we cannot observe prosodic events at the segmental level. In order to model prosodic events we have to be able to observe the behavior of prosodic parameters over several frames spanning for example a phone, syllable, or word. In addition, if we want to observe the behavior of prosodic parameters over the duration of, say, a phone we have to make sure that phone durations are assessable and modeled accurately.

In the following sections we show in more detail why hidden Markov models show these two weaknesses. We specify a new suprasegmental hidden Markov model (SPHMM) architecture which tackles the problematic duration modeling and the frame based extraction of features in traditional hidden Markov models. This new architecture permits the summarization of several atomic states within a hidden Markov model into what we call a suprasegmental state. These suprasegmental states allow the consideration of the observation sequence spanned by their constituting atomic states, thus overcoming the frame based extraction of features in atomic hidden Markov models. In addition, because suprasegmental states know how many time steps were spent by their constituent states, duration can be modeled by non-geometric parametric distributions.

The next section introduces the basic architecture of HMMs and its underlying procedures and describes the modifications to this architecture which have been proposed to address its deficiencies as mentioned above. The final sections explain the extensions to HMMs which are required to incorporate suprasegmental states in the overall architecture and processing.

3.2.1 Hidden Markov Models

A specification of a hidden Markov model requires the number of states, N, and three probability distributions A, B, and π. A hidden Markov model is characterized by:[1]

1. The number of states, N, where individual states are labelled as $1, 2, \ldots, N$. An individual state at time t is denoted by q_t. The notation q_t^i is used as an abbreviation for $q_t = i$ with $1 \leq i \leq N$.
2. $A = a_{ij}$ is a state-transition probability distribution where $a_{ij} = P(q_{t+1}^j | q_t^i)$, $1 \leq i, j \leq N$.
3. $\pi = \pi_i$ is the initial state distribution where $\pi_i = P(q_1^i)$ for $1 \leq i \leq N$.
4. An observation sequence is referred to by $O = (o_1 o_2 \ldots o_T)$ where T is the number of observations in the sequence.
5. $B = b_j(o)$ is the observation probability distribution where $b_j(o) = P(o_t \mid q_t^j)$, $1 \leq t \leq T$, defines the symbol distribution in state j for $j = 1, 2, \ldots, N$.

$$b_j(o) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid C_j \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(o-\mu_j)' C_j^{-1} (o-\mu_j)} \tag{3.6}$$

where

o is the n-dimensional observation vector,
μ_j is the n-dimensional mean vector, and
C_j is the $n \times n$ covariance matrix.

6. A state sequence is referred to by $q = (q_1 q_2 \ldots q_T)$. T is the number of observations in the sequence.

Following Rabiner and Juang (1993), we use $\lambda = (A, B, \pi)$ to refer to the complete parameter set of a hidden Markov model. In order to apply an HMM as specified above we have to find a solution for the following problems:

- Given a sequence O and a model λ we want to compute the probability of the observation sequence given the model:

$$P(O \mid \lambda).$$

- Given an observation sequence O and model λ we want to find the most likely state sequence $q_1 q_2 \ldots q_T$.
- How can we maximize $P(O \mid \lambda)$ by adjusting the model parameters λ?

We compute $P(O \mid \lambda)$ with the forward/backward procedures (Rabiner and Juang, 1993). Because of the similarity of theses two procedures we confine ourselves to the forward procedure in

[1] The description of the theory of hidden Markov models in this section closely follows Rabiner and Juang (1993).

the next section. We use the Viterbi algorithm in order to compute the most likely state sequence given an observation sequence O and model parameters λ. We describe the Viterbi algorithm after the specification of the forward procedure. Note that the maximization of $P(O \mid \lambda)$ by adjusting the model parameters λ is mainly based on the forward/backward and the Viterbi procedures. Consult (Rabiner and Juang, 1993) for details. Note also that the forward/backward procedures and the Viterbi algorithm have to be adjusted to accommodate prosodic observations in an SPHMM.

The Forward Procedure

The forward procedure is used to efficiently compute the probability of $P(O \mid \lambda)$. We define

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t^i \mid \lambda), \tag{3.7}$$

i.e. the probability of observing the partial sequence $o_1 o_2 \ldots o_t$ and being in state i at time t given the model λ:

Initialization:

$$\alpha_1(i) = \pi_i \cdot b_i(o_1), \qquad 1 \leq i \leq N. \tag{3.8}$$

Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) \cdot a_{ij}\right] \cdot b_j(o_{t+1}), \qquad \begin{array}{l} 1 \leq t \leq T-1, \\ 1 \leq j \leq N. \end{array} \tag{3.9}$$

Termination:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{3.10}$$

The Viterbi Algorithm

The optimal state sequence $q = (q_1 q_2 \ldots q_T)$ for the observation sequence $O = (o_1 o_2 \ldots o_T)$ given a model λ can be computed by the Viterbi algorithm. We define the quantity

$$\delta_t(i) = \max_{q_1, q_2, \ldots q_{t-1}} P(q_1 q_2, \ldots q_{t-1}, q_t^i, o_1 o_2 \ldots o_t \mid \lambda) \tag{3.11}$$

as the highest probability along a single path which ends in state i at time t accounting for the first t observations. We transform equation 3.11 into 3.12 using induction:

$$\delta_{t+1}(j) = [\max_{1 \leq i \leq N} \delta_t(i) \cdot a_{ij}] \cdot b_j(o_{t+1}). \tag{3.12}$$

Keeping equation 3.12 in mind, the Viterbi algorithm can be specified with the following four steps where the array $\psi_t(j)$ is used to store the argument maximizing equation 3.12 at time t. In the last step of the Viterbi algorithm, the array $\psi_t(j)$ is used to recover the state sequence.

Initialization:

$$\delta_1(i) = \pi_i \cdot b_i(o_1), 1 \leq i \leq N. \tag{3.13}$$
$$\psi_1(i) = 0, 1 \leq i \leq N. \tag{3.14}$$

Recursion:

$$\delta_t(j) = [\max_{1 \leq i \leq N} \delta_{t-1}(i) \cdot a_{ij}] \cdot b_j(o_t), \qquad 2 \leq t \leq T, \quad 1 \leq j \leq N. \tag{3.15}$$
$$\psi_t(j) = [\arg \max_{1 \leq i \leq N} \delta_{t-1}(i) \cdot a_{ij}], \qquad 2 \leq t \leq T, \quad 1 \leq j \leq N. \tag{3.16}$$

Termination:

$$P^* = [\max_{1 \leq i \leq N} \delta_T(i)]. \tag{3.17}$$
$$q_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)]. \tag{3.18}$$

Recovering the state sequence:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \ldots, 1. \tag{3.19}$$

In order to be applicable to SPHMMs, these specifications of the forward procedure and the Viterbi algorithm have to be modified. Section 3.2.2 describes these modifications. Since the modeling of duration is essential for the SPHMM, we describe in the next section alternative approaches to model duration within hidden Markov model architectures. The development of the modeling assumptions of SPHMMs is based on these approaches.

Duration Modeling

Before we describe our approach of modeling durations in SPHMMs, let us have a look at how duration is modeled within conventional HMMs. In a fully connected hidden Markov model duration is modeled by self-transitions. It is interesting to look at the inherent duration probability density $P(d_i)$ associated with some state i with a self-transition coefficient a_{ii}:

$$P(d_i) = (a_{ii})^{d-1}(1 - a_{ii}) \tag{3.20}$$

which gives us the probability that the hidden Markov model will stay in state i for d consecutive times. That is to say, duration is modeled by an exponential function (geometric distribution), seemingly a poor estimate (Rabiner and Juang, 1993). A straightforward extension to this model is to deviate from a fully connected transition matrix and specify topologies in which certain states are obligatory (Jelinek, 1976; Bakis, 1976). Other approaches rely on the idea of imposing a fixed lower or upper bound on the duration of some segment. For instance, Gupta et al. (1992) implemented a minimum duration constraint for phones. Only state sequences through a phone which meet the minimum duration constraint of the respective phone are valid.

Levinson (1986) and Rabiner and Juang (1993) set the self-transition coefficients to zero and model durations explicitly as a duration density. Let us have a brief look at the definition of

the forward probabilities in this framework because it bears resemblance to our definition of SPHMMs in the next section. Let $P(d \mid q^i)$ be the probability of staying in state q^i for d time steps and let D be the maximal number of time steps you can spend in a state. The forward variable $\alpha_t(i)$ is defined as follows:

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, \text{the stay in } q^i \text{ ends at } t \mid \lambda). \tag{3.21}$$

The inductive definition of $\alpha_t(j)$ is as follows:

$$\alpha_t(j) = \sum_{i=1}^{N} \sum_{d=1}^{D} \alpha_{t-d}(i) \cdot a_{ij} \cdot P(d \mid j) \cdot \prod_{s=t-d+1}^{t} b_j(o_s), \tag{3.22}$$

The additional $\sum_{d=1}^{D}$ sums over all possible durations of q^j. Note that for a given duration d we have to consider the $\alpha(i)$ which was computed d time steps before the current time t, i.e. $\alpha(i)_{t-d}$. The final product, $\prod_{s=t-d+1}^{t} b_j(o_s)$, computes the probability to observe the last d observations in state q^j. Note the explicit state duration modeling achieved by $P(d \mid j)$, i.e., the probability of staying in state j for d consecutive time steps.

In order to incorporate explicit state duration densities can improve performance significantly (Rabiner and Juang, 1993). At the same time, the computational effort increases quadratically with $D^2/2$ and the required storage increases by a factor of D. In a simple semi Markov model we have to compute D additional parameters for each state. In order to decrease this large number of parameters Juang and Rabiner (1985; Russel and Moore (1985) introduced parametric state duration densities.

3.2.2 A Suprasegmental Hidden Markov Model

Keep in mind that the modifications to the hidden Markov model architecture as discussed in the previous section are still confined to extensions at the state level. That is, durations and feature extractions are still modeled at the hidden Markov state level which models segmental, not prosodic (suprasegmental) events. It is common practice to construct phone models with several HMM states to account for idiosyncratic acoustic properties at the beginning, the middle, and the end of a phone (Zhan et al., 1997; Zeppenfeld et al., 1997). Once the final state of a phone model is reached, we cannot access information from its initial state because of the Markov assumption. In particular, by leaving the final state we do not know how much time was spent in the phone model of which it was a part. However, having access to the duration of larger segments, such as phones, syllables, or words, is essential for the extraction and integration of prosodic observations into an HMM architecture. In the following sections we develop a suprasegmental hidden Markov model (SPHMM) architecture which allows to summarize atomic states into what we call suprasegmental states. These suprasegmental states have access to the overall time spent in their constituting states and, thus, allow the observation of prosodic events.

The following specification of an SPHMM is confined to one additional level of suprasegmental states in order to keep the notational overhead to a minimum. However, the architecture and the corresponding procedures allow multiple levels of suprasegmental states at the same time. Thus, it is possible to model prosodic properties of phones, syllables, and words in parallel with this architecture. The basic idea of a suprasegmental hidden Markov model is given in Figure 3.1 where the states $\overline{q}^1$, $\overline{q}^2$, and $\overline{q}^3$ constitute a suprasegmental state $\overline{\overline{q}}^1$ where $\overline{q}^1$ is the unique initial and $\overline{q}^3$ the unique final state. Let Q_1 denote the set of all atomic hidden Markov states.

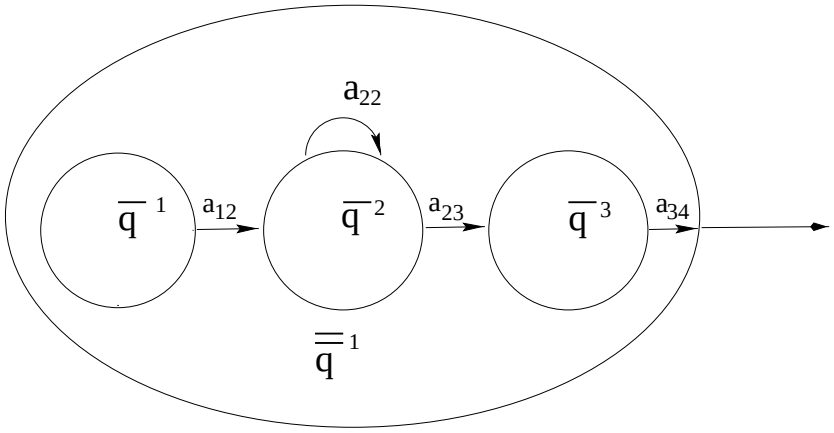


Figure 3.1: Simple Suprasegmental Hidden Markov Model (SPHMM). The suprasegmental state $\overline{\overline{q}}^1$ comprises the atomic hidden Markov states $\overline{q}^1$, $\overline{q}^2$, and $\overline{q}^3$ where $\overline{q}^1$ is the unique initial and $\overline{q}^3$ the unique final state.

We require that the set of all suprasegmental states, Q_2, forms a partition of Q_1. We use N_2 to refer to the number of suprasegmental states and use double bars to mark these states as suprasegmental states:

$$Q_2 = \{\overline{\overline{q}}^1, \overline{\overline{q}}^2 \ldots \overline{\overline{q}}^{N_2}\}. \tag{3.23}$$

$$Q_1 = \overline{\overline{q}}^1 \cup \overline{\overline{q}}^2 \ldots \cup \overline{\overline{q}}^{N_2}. \tag{3.24}$$

$$\forall \overline{\overline{q}}^k, \overline{\overline{q}}^{k'} \in Q_2 \text{ iff } \overline{\overline{q}}^k \cap \overline{\overline{q}}^{k'} = \emptyset \text{ then } k \neq k' \qquad (3.25)$$
$$1 \leq k, k' \leq N_2.$$

We also require that each suprasegmental state has a unique beginning and end state which we indicate by the superscripts s and e. In addition, we require that the last atomic state within a suprasegmental state has no self-transition. We use D_2 to represent the maximal number of time steps you can stay in a suprasegmental state.

The probabilistic model is very similar to the semi HMM architecture as described above with equations 3.21 and 3.22. The main difference is that the explicit duration probability is not defined for atomic first level states, as in the semi model, but for suprasegmental states. Similar to a semi Markov models, the segmentation is uniquely specified by a sequence of segment durations $D_1^S = \{d_1, \ldots, d_S\}$. Note that

$$\sum_{n=1}^{S} d_n = T. \qquad (3.26)$$

Given a state sequence of atomic HMM states $\overline{q}_1 \ldots \overline{q}_T$ and a corresponding suprasegmental state sequence $\overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S$, the probability of observing $o_1 \ldots o_n$ can be defined as follows where $\sum_{D_1^S}$ quantifies over all possible segmentations:

$$\begin{aligned} & P(o_1 \ldots o_T \mid \overline{q}_1 \ldots \overline{q}_T, \overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S) \\ &= \sum_{D_1^S} P(o_1 \ldots o_T, d_1 \ldots d_S \mid \overline{q}_1 \ldots \overline{q}_T, \overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S) \\ &= \sum_{D_1^S} P(o_1 \ldots o_T \mid d_1 \ldots d_S, \overline{q}_1 \ldots \overline{q}_T, \overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S) \\ &\quad \cdot P(d_1 \ldots d_S \mid \overline{q}_1 \ldots \overline{q}_T, \overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S), \end{aligned} \qquad (3.27)$$

where

$$\begin{aligned} P(o_1 \ldots o_T \mid d_1 \ldots d_S, \overline{q}_1 \ldots \overline{q}_T, \overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S) \quad &= \\ & \prod_{i=1}^{S} P(o_{t_i+1} \ldots o_{t_i+d_i} \mid d_i, \overline{q}^s_{t_i+1} \ldots \overline{q}^e_{t_i+d_i}, \overline{\overline{q}}_i), \\ & \text{where } t_i = \sum_{j=1}^{i-1} d_j \text{ if } i > 1 \text{ and } d_1 \text{ if } i = 1, \\ & \text{and } \overline{q}^s_{t_i+1}, \ldots, \text{ and } \overline{q}^e_{t_i+d_i} \in \overline{\overline{q}_i}, \end{aligned} \qquad (3.28)$$

and where

$$P(d_1 \ldots d_S \mid \overline{q}_1 \ldots \overline{q}_T, \overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S) \quad = \quad \prod_{i=1}^{S} P(d_i \mid \overline{\overline{q}}_i). \qquad (3.29)$$

Thus, equation 3.28 defines the probability of observing $o_1 \ldots o_T$ given the state sequences $\overline{q}_1 \ldots \overline{q}_T$ and $\overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S$ in which we spent d_1 time steps in $\overline{\overline{q}}_1$ and d_2 time steps in $\overline{\overline{q}}_2$ and so on. The constraint on t_i ensures that all observations are accounted for in the right order by the respective states. With $\overline{q}^s_{t_i+1}, \ldots,$ and $\overline{q}^e_{t_i+d_i} \in \overline{\overline{q}}^i$ we make sure that we stay the correct number

of time steps in the suprasegmental state $\overline{\overline{q}}^i$ which we had entered at time $t_i + 1$ with the unique start state $\overline{q}^s$ and left after d_i time steps with the unique end state $\overline{q}^e$. With equation 3.29 we define the probability of a segmentation $d_1 \ldots d_S$ given state sequences $\overline{q}_1 \ldots \overline{q}_T$ and $\overline{\overline{q}}_1 \ldots \overline{\overline{q}}_S$.

In the following we explain the forward variable α which we defined earlier for traditional hidden Markov models, see equation 3.7, and for semi Markov models, see equation 3.21. The idea behind the following specification of the forward variable for SPHMMs is to specify the probability of leaving some suprasegmental state at time t in terms of the probabilities of leaving its corresponding unique end state at time t. We, therefore, define two different forward variables, $\overline{\alpha}_t$ and $\overline{\overline{\alpha}}_t$, where the bars indicate whether α is defined for atomic or suprasegmental hidden Markov states. The inductive definition of $\overline{\overline{\alpha}}$ is outlined below where D_2 indicates the maximal number of time steps possible in a suprasegmental state:

$$\overline{\overline{\alpha}}_t(\overline{\overline{q}}^j) = \sum_{i=1}^{N_2} \sum_{d=1}^{D_2} \overline{\overline{\alpha}}_{t-d}(\overline{\overline{q}}^i) \cdot a_{\overline{q}^{e'}_{t-d}, \overline{q}^s_{t-d+1}} \cdot P(d \mid \overline{\overline{q}}^j_t) \cdot \overline{\alpha}_t^{j,t-d+1,t}(\overline{q}^e), \tag{3.30}$$

where
$\overline{q}^{e'}$ is the unique end state of the suprasegmental state $\overline{\overline{q}}^i$,
and $\overline{q}^s$ and $\overline{q}^e$ are the unique start and end states
of the suprasegmental state $\overline{\overline{q}}^j$ and $1 \leq t \leq T$.
The definition of $\overline{\alpha}$ is given in equation 3.31 below.

Thus, the probability of observing t events $o_1 \ldots o_t$ and the stay ends in the suprasegmental state $\overline{\overline{q}}^j$ comprises four main parts:

1. the probability of observing $t - d$ events by leaving suprasegmental state $\overline{\overline{q}}^i$,
2. the transition probability from the unique final state $\overline{q}^{e'}_{t-d}$ of suprasegmental state $\overline{\overline{q}}^i$ to the unique initial state $\overline{q}^s_{t-d+1}$ of suprasegmental state $\overline{\overline{q}}^j$,
3. the probability of staying in $\overline{\overline{q}}^j$ for d time steps, and
4. the probability of observing the last d events with the state sequences $\overline{q}^s_{t-d+1} \ldots \overline{q}^e_t$ and $\overline{\overline{q}}^j_{t-d+1} \ldots \overline{\overline{q}}^j_t$.

We define $\overline{\alpha}_t^{k,t_s,t_e}(i)$ as:

$$\overline{\alpha}_t^{k,t_s,t_e}(\overline{q}^e) = P(o_{t_s} \ldots o_{t_e} \mid \overline{q}^s_{t_s} \ldots \overline{q}^e_{t_e}), \tag{3.31}$$

i.e. the probability of observing $o_{t_s-d} \ldots o_{t_e}$ in the suprasegmental state $\overline{\overline{q}}^k$ starting in the start state $\overline{q}^s$ at time t_s and ending in the end state $\overline{q}^e$ at time t_e where $\overline{q}^s$ and $\overline{q}^e$ are the unique initial and final states of suprasegmental state $\overline{\overline{q}}^k$. Note that only these initial and final states are fixed and the only constraint applying to the states $\overline{q}_{t_s+1} \ldots \overline{q}_{t_e-1}$ is that they have to belong to suprasegmental state $\overline{\overline{q}}^k$. The computation of this probability is very similar to the computation of the genuine forward variable as given in equations 3.8-3.10, except that the observation starts at time t_s and ends at time t_e. Moreover, only states of the suprasegmental state $\overline{\overline{q}}^k$ can be used to account for the observation sequence.

The inductive definition of $\overline{\alpha}^{k,t_s,t_e}(\overline{q}^i)$ is as follows:

Initialization:

$$\overline{\alpha}_{t_s}^{k,t_s,t_e}(i) = \begin{cases} \pi_i \cdot b_i(o_{t_s}) & \text{iff } t_s = 1, \\ b_i(o_{t_s}) & \text{otherwise,} \end{cases} \quad \forall \overline{q}^i \in \overline{\overline{q}}^k. \tag{3.32}$$

Induction:

$$\overline{\alpha}_{t+1}^{k,t_s,t_e}(j) = \left[\sum_{\overline{q}^i \in \overline{\overline{q}}^k} \overline{\alpha}_t^{k,t_s,t_e}(i) \cdot a_{ij} \right] \cdot b_j(o_{t+1}), \quad t_s \leq t+1 \leq t_e, \quad \overline{q}^j \in \overline{\overline{q}}^k. \tag{3.33}$$

Termination:

$$P(o_{t_s} \ldots o_{t_e} \mid \lambda) = \sum_{\overline{q}^i \in \overline{\overline{q}}^k} \overline{\alpha}_{t_e}^{k,t_s,t_e}(i). \tag{3.34}$$

Unfortunately, the complexity involved in computing $\overline{\overline{\alpha}}_t(\overline{\overline{q}}^j)$ is immense. First, note the additional sum over D_2 in equation 3.30. Second, the forward variable $\overline{\alpha}_t^{k,t_s,t_e}(\cdot)$ is strictly local and has to be computed again for every d in equation 3.30.

Let us assume the special case in where we have an one-to-one mapping between atomic and suprasegmental states. In particular, we require $\overline{q}^i = \overline{\overline{q}}^i$ for $1 \leq i \leq N$. That is, $N_2 = N$. In addition, we require that the maximal number of time steps in a conventional HMM state and a suprasegmental state are indentcial: $D_2 = D$. The definition of the forward variable, $\overline{\overline{\alpha}}_t(\overline{\overline{q}}^j)$ reduces to equation 3.22 as defined for a semi Markov model:

(3.35)

$$\begin{aligned}
\overline{\overline{\alpha}}_t(\overline{\overline{q}}^j) &= \sum_{i=1}^{N_2} \sum_{d=1}^{D_2} \overline{\overline{\alpha}}_{t-d}(\overline{\overline{q}}^i) \cdot a_{\overline{q}^{e'}\overline{q}^s} \cdot P(d \mid \overline{\overline{q}}_t^j) \cdot P(o_{t-d} \ldots o_t, \overline{q}_{t-d}^s \ldots \overline{q}_t^e, \overline{\overline{q}}_{t-d}^j \ldots \overline{\overline{q}}_t^j) \\
&= \sum_{i=1}^{N_2} \sum_{d=1}^{D_2} \overline{\overline{\alpha}}_{t-d}(\overline{\overline{q}}^i) \cdot a_{\overline{q}^{e'}\overline{q}^s} \cdot P(d \mid \overline{q}_t^j) \cdot P(o_{t-d} \ldots o_t, \overline{q}_{t-d}^s \ldots \overline{q}_t^e) \\
&\quad \text{because } \overline{q}^i = \overline{\overline{q}}^i \text{ and } \overline{q}^j = \overline{\overline{q}}^j \\
&= \sum_{i=1}^{N_2} \sum_{d=1}^{D_2} \overline{\overline{\alpha}}_{t-d}(\overline{\overline{q}}^i) \cdot a_{\overline{q}^{e'}\overline{q}^s} \cdot P(d \mid \overline{q}_t^j) \cdot \prod_{s=t-d+1}^{t} b_j(o_s) \\
&\quad \text{because } \overline{q}_{t-d}^s = \overline{q}_t^e = \overline{\overline{q}}_t^j = \overline{q}_t^j, \\
&= \sum_{i=1}^{N} \sum_{d=1}^{D} \overline{\overline{\alpha}}_{t-d}(\overline{\overline{q}}^i) \cdot a_{ij} \cdot P(d \mid \overline{q}_t^j) \cdot \prod_{s=t-d+1}^{t} b_j(o_s) \\
&\quad \text{because } N_2 = N,\ D_2 = D,\ \overline{q}^{e'} = \overline{q}^i, \text{ and } \overline{q}^s = \overline{q}^j.
\end{aligned}$$

The increase in complexity and storage requirements for a SPHMM for the implementation of the forward procedure as mentioned above transfers to the Viterbi algorithm. In the next section, we specify an approximation of the Viterbi which is computationally more reasonable.

The Approximation of the Viterbi Algorithm

In our approximation of the Viterbi algorithm we include the duration probability distribution for suprasegmental states, $P(d \mid \overline{\overline{q}}_t)$ into the overall score computation when we leave a suprasegmental state. Note that $P(d \mid \overline{\overline{q}}_t)$ is a posterior probability and its integration into the Viterbi algorithm does not guarantee that the most likely state sequence $\overline{q}_1 \ldots \overline{q}_T$ is found. As mentioned in the specification of the Viterbi in section 3.2.1 the algorithm computes the best preceding state for each time step t and state. Because $P(d \mid \overline{\overline{q}}_t)$ is posterior, however, a previously optimal state might turn out to be suboptimal. In order to gurantee an optimal path we would have to reconsider all possible paths starting at $t-d$ in the unique start state of $\overline{\overline{q}}_t$ and end at time t in the unique final state of $\overline{\overline{q}}_t$. Obviously, depending on the maximal segment size, the computational cost can be immense. For the purposes of this investigation, we approximated the Viterbi algorithm by storing not just the previous best state but by storing the r-best previous states of some state $\overline{q}^i$ at time t. We will use R to refer to the size of this ordered stack. We require three arrays:

1. $\psi_{t,r}(i)$ denotes the r-best previous state of state i at time t.
2. $\delta_{t,r}(i)$ denotes the r-best previous probability of having arrived in state i at time t.
3. $\tau_{t,r}(i)$ denotes the number of time steps spent in the suprasegmental state of state i at time t depending on the r-best previous states of state i where $1 \leq \tau_{t,r}(i) \leq D_2$ for $1 \leq t \leq T, 1 \leq r \leq R$ and $1 \leq i \leq N$,

where the index r indicates the position on the stack.

We represent transitions from a state $\overline{q}_t^i$ to a state $\overline{q}_{t+1}^j$ by the function $a_{ij}(d,t)$. This function reduces to a_{ij} in the case where both states $\overline{q}_t^i$ and $\overline{q}_{t+1}^j$ belong to the same suprasegmental state. The function includes the duration probability distribution for a suprasegmental state if the respective transition leaves a suprasegmental state. We require these *complex* transitions to obey standard stochastic constraints:

$$a_{ij}(d,t) \geq 0 \qquad (3.36)$$
$$1 \leq i,j \leq N,\ 1 \leq d \leq D_2, \text{ and } 1 \leq t \leq T.$$

$$\sum_{j=1}^{N} \sum_{d=1}^{D_2} a_{ij}(d,t) = 1 \qquad (3.37)$$
$$1 \leq i \leq N \text{ and } 1 \leq t \leq T.$$

The approximation of the Viterbi algorithm using ordered stacks of size R is as follows:

Initialization:

$$\hat{\delta}_{1,r}(i) = \pi_i \cdot b_i(o_1), \qquad (3.38)$$
$$\psi_{1,r}(i) = 0, \qquad (3.39)$$
$$\tau_{1,r}(i) = 1, \qquad (3.40)$$
$$\text{where } 1 \leq i \leq N \text{ and } 1 \leq r \leq R.$$

Recursion:

$$\hat{\delta}_{t,r'}(j) = \max_{1 \le i \le N} \max_{1 \le r \le R} \hat{\delta}_{t-1,r}(i) \cdot a_{ij}(\tau_{t-1,r}(i), t) \cdot b_j(o_t), \tag{3.41}$$

$$\text{where } a_{ij}(d,t) = \begin{cases} a_{ij} & \text{iff } \exists \overline{\overline{q}}^k \in Q_2: \\ & \overline{q}^i \text{ and } \overline{q}^j \in \overline{\overline{q}}^k, \\ a_{ij} \cdot P(d \mid \overline{\overline{q}}^k) & \exists \overline{\overline{q}}^k \in Q_2: \\ & \overline{q}^j \in \overline{\overline{q}}^k, \text{ otherwise.} \end{cases}$$

$$2 \le t \le T,\ 1 \le r, r' \le R,\ 1 \le j \le N, \text{ and } 1 \le k \le N_2.$$

Thus, if we stay by traversing from an atomic state $\overline{q}^i_{t-1}$ to an atomic state $\overline{q}^j_t$ in the suprasegmental state $\overline{\overline{q}}^k$, the function reduces to the atomic transition probability a_{ij}. On the other hand, if we leave a suprasegmental state, the transition probablity is multiplied by the duration probability of ramaining in the suprasegmental state $\overline{\overline{q}}^k$ for d time steps: $P(d \mid \overline{\overline{q}}^k)$.

We informally specify the updating of the arrays $\hat{\delta}_{t,r}(j)$, $\psi_{t,r}(j)$ and $\tau_{t,r}(j)$ with the procedure given below:

For each time step t and each state j we keep an ordered stack of size R where we store:

(a) the probability $\hat{\delta}_{t,r}(j)$,

(b) the time spent so far in the respective suprasegmental state, $\tau_{t,r}(j)$, and

(c) the predecessor $\overline{q}^i_{t-1}$ of $\overline{q}^j_t$.

The stack will be ordered according to the above probability where the top of the stack, indexed with 1, stores the highest probability. As we approach the bottom of this stack the probability decreases. The index r denotes the respective stack index.

For all states i, $1 \leq i \leq N$ **and** r, $1 \leq r \leq R$
begin

compute
$\hat{\delta}_t(j)$ = $\hat{\delta}_{t-1,r}(i) \cdot a_{ij}(\tau_{t-1,r}(i), t)$.

update
We start at the top of the stack and search for a $\hat{\delta}_{t,r'}(j)$ which is smaller than the $\hat{\delta}_t(j)$ as computed above.
In case we find one:
$\hat{\delta}_{t,r'}(j)$ = $\hat{\delta}_t(j)$
$\psi_{t,r'}(j) = i$
$\tau_{t,r'}(j)$ = 1 **if** we leave a suprasegmental state and
$\tau_{t,r'}(j) = \tau_{t-1,r}(i) + 1$ **otherwise**.

end

Termination:

$$P^* = [\max_{1 \leq i \leq N} \hat{\delta}_{T,1}(i)]. \tag{3.42}$$

$$q^*_T = \arg \max_{1 \leq i \leq N} [\hat{\delta}_{T,1}(i)]. \tag{3.43}$$

Recovering the state sequence:

In order to recover the state sequence correctly we need an additional variable which keeps track of the number of time steps required to leave the suprasegmental state corresponding to current best atomic state q^*_t. We use d_t to refer to this number. Recall that recovering the state sequence starts at time T and moves backwards to $t = 1$. Thus, for example,

$d_t = 5$ means that you leave the current suprasegmental state by moving 5 time steps backwards. The variable d_t is specified as follows:

$$d_t = \begin{cases} \tau_{T,1}(q_T^*), & \text{iff } t = T, \\ \tau_{t+1,1}(q_{t+1}^*), & \text{iff } d_{t+1} = 1, \\ d_{t+1} - 1, & \text{otherwise,} \end{cases} \quad (3.44)$$
$$1 \leq t \leq T.$$

The first case handles the situation where we want to know how many time steps we stay in the suprasegmental state which ends at T. The second case describes the situation in which we leave a suprasegmental state, i.e. d_{t+1} equals 1. In this case we choose the duration of the suprasegmental state corresponding to the best previous atomic state of the current best atomic state. The last case covers the situation in which we stay within a suprasegmental state and the duration correspondingly decreases by one.

Using d_t we can start recovering the best state sequence. The first case covers the situation in which we try to find the best atomic HMM state at time T. If $d_{t+1} = 1$ we leave the current suprasegmental state and we choose the best predecessor as stored in the array ψ. Remember that the array ψ was ordered and the index 1 refers to the best predecessor. The third case corresponds to a situation in which we stay in a suprasegmental state. We have to find the predecessor that agrees with the time steps yet to be spent in the current suprasegmental state. These three cases are formalized below:

$$\overline{q}_t^* = \begin{cases} q_T^*, & \text{iff } t = T, \\ \psi_{t+1,1}(\overline{q}_{t+1}^*), & \text{iff } d_{t+1} = 1, \\ \psi_{t+1,r}(\overline{q}_{t+1}^*), & \text{where } r \text{ is the smallest } r \text{ such there is an } r' \text{ such that} \\ & \tau_{t,r'}(\overline{q}_t^*) = d_{t+1} - 1, \text{ otherwise,} \end{cases} \quad (3.45)$$
$$t = T, T-1, T-2, \ldots, 1, \text{ and } 1 \leq r, r' \leq R.$$

For this approximation of the Viterbi algorithm the memory increase is confined to the two additional arrays $\hat{\delta}_{t,r}(j)$ and $\tau_{t,r}(j)$. The size of these arrays depend upon the stack size, R, total time steps, T, and the number of atomic states, N. Additional memory requirements arise to store parameters of suprasegmental models. The requirements depend on the number of suprasegmental models, N_2. Additional computations are needed in three places:

1. the inclusion of suprasegmental duration probabilities,
2. the maintaining of the stacks, and
3. the additional maximization over R in equation 3.41.

The last two additional computational increases depend upon the stack size, R. In our experiments, we used this approximation of the Viterbi algorithm to train prosodic models.

3.2.3 Context Sensitive Prosodic Models

In order to reliably estimate the parameters of prosodic models, we have to be concerned with two interdependent issues:

- Models based on prosodic information are highly sensitive to their surrounding context. For example, the intensity, intonation, and duration of the same vowel depends whether a vowel is a nucleus within a stressed syllable, whether it is word final, or utterance final (Kießling, 1997). Thus, in the best of all worlds, we would like to have different models, one for each of these situations, to estimate their parameters as detailed as possible.
- We have to make sure that each model occurs in the training corpus with a frequency that allows a reliable parameter estimation.

Thus, there is a tradeoff. If we design prosodic models too specifically, the estimation of these models might be based on an insufficient number of training samples, and if we design the models too generally, the parameter estimation procedure, although based on a sufficient number of training samples, might be unreliable due to context effects. We use a clustering algorithm (Breiman et al., 1984) to handle this tradeoff. Based on binary questions about the prosodic segment and its context, the algorithm builds a regression tree in which a leaf node represents a model for which we insure:

1. The parameter estimation process is based on sufficient samples.
2. The training instances of a leaf node are prosodically similar with each other and contrast prosodically with the instances in the daughter node.

Note that we do not a priori stipulate specific context effects. By allowing various binary questions about the context, the clustering algorithm finds those questions first which maximize the prosodic difference of the instances of two daughter nodes and the similarity of the instances within a node. The technical detail are given below.

Following Kannan, Ostendorf, and Rohlicek (1994), we evaluate a likelihood ratio for each allowable partition of the training data by a set of predefined binary questions. The null hypothesis is that the observations were generated from one distribution which corresponds to the maximum likelihood estimate of the parent node. The alternative hypothesis is that the observations were generated by two different distributions represented by the maximum likelihood estimates of the daughter nodes. We define the likelihood ratio, λ, as the ratio of the observations being generated from one distribution and the likelihood of the observations being generated by two different distributions. For normal distributions, we can express λ as a product of the quantities λ_{MEAN} and λ_{COC} as defined below:

$$\lambda_{MEAN} = \left(1 + \frac{\eta_l \eta_r}{\eta^2}(\hat{\mu}_l - \hat{\mu}_r)^t W^{-1}(\hat{\mu}_l - \hat{\mu}_r)\right)^{\frac{-\eta}{2}} \tag{3.46}$$

$$\lambda_{COV} \left(\frac{\mid \sigma_l \mid^{\alpha} \mid \hat{\Sigma} u_r \mid^{1-\alpha}}{\mid W \mid}\right)^{\frac{\eta}{2}} \tag{3.47}$$

where η_l and η_r are the number of observations in the left and right daughter node, $\eta = \eta_l + \eta_r$, $\hat{\mu}_l$ and $\hat{\mu}_r$ are the sample means of the left and right daughter nodes, $\hat{\Sigma}_l$ and $\hat{\Sigma}_r$ are the sample covariances for the left and right child nodes, $\alpha = \frac{\eta_l}{\eta}$, and W is the frequency weighted tied covariance $W = \frac{\eta_l}{\eta}\hat{\Sigma}_l + \frac{\eta_r}{\eta}\hat{\Sigma}_r$. If we want prosodic models to share a common covariance matrix we grow the tree minimizing $-\log \lambda_{COV}$. In case we want to cluster distribution means, we minimize $-\log \lambda_{MEAN}$. If we want to cluster both distribution means and covariance, we have to minimize $-(\log \lambda_{COV} + \log \lambda_{MEAN})$. We compute $\log \lambda_{COV} - \log \lambda_{MEAN}$ for every binary question within a set of allowable questions and build the tree top-down choosing those questions which minimize the above difference and guarantee that the subsequent estimation of model parameters is based on sufficient training data. We will use this algorithm to cluster prosodic models. The actual question sets and resulting trees are given in the corresponding experiment in Chapter 4.

3.3 Modeling Spectral Information

In principle, it is possible to train entire emotion-specific recognition systems. By comparing the scores of these systems, we would find that system which most likely produced the emotional utterance. However, training a recognition system requires a substantial amount of training data. Since the corpora in our experiments are significantly smaller than the minimum corpus size required to train a recognition system, we pursued a different approach. In order to model emotion-specific spectral information, we started with an existing recognition system and performed emotion-specific adaptations. In our investigation, we used an adaptation technique which is frequently used to adapt a speech recognition system to novel speakers (Gales, 1996; Legetter and Woodland, 1994). This approach adapts only the mean parameters of acoustic models to novel speech data by a set of linear transformations which are found using the maximum likelihood training algorithm.

If the adaptation data is very limited, adaptation techniques face the problem that many acoustic models are not present at all in the data. In order to find linear transformation for these models as well, we first pool together acoustic models which behave similarly in acoustic space of the original training set. We assume that these pooled models behave also acoustically similar in the novel data and, thus, can be linearly transformed in the same way. This way, we can ensure that we find adaptation transformations for all models, even for models which do not occur in the adaptation data.

Having pooled models which behave alike in acoustic space according to some distance measure, we have to find the linear transformations for the parameters of these tied models in the next step. To be more detailed, we reestimate the distribution means, $\hat{\mu}_j$, of some state j using a linear transformation of the original means, μ_j. Thus, the probability density function as specified in equation 3.6 for a state j observing a vector o of dimension n becomes

$$b_j(o) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid C_j \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(o-\hat{\mu}_j)' C_j^{-1} (o-\hat{\mu}_j)} \tag{3.48}$$

where C_j is the corresponding covariance matrix, $\hat{\mu}_j$ is the linear transformation of μ_j by the transformation matrix W_j, and n is the dimension of the observation vector o. As shown by (Legetter and Woodland, 1994) for maximum likelihood linear regression, the matrix W_j can be estimated using the Baum-Welch reestimation procedure as outlined in section 3.2.1.

To summarize, our approach to model emotion-specific spectral information comprises the following steps:

1. We start with a fully trained speech recognition system and pool models according to their acoustic similarity using k-means clustering.
2. We collect the necessary statistics to compute the transformation matrix, W, based on an alignment of the novel speech data using the non-adapted recognition system.
3. We compute the transformation matrices making sure that for each model pool j, the estimation of the matrix W_j is based on a sufficient number of training tokens within the adaptation data.
4. We repeat steps 2 and 3 until a satisfactory result is achieved using the adapted models for a realignment.

In order to model emotion-specific spectral information, we estimate emotion-specific transformation matrices, W, for all model pools by adaptation on data which consistently expressed a certain emotion. Thus, we had, for instance, acoustic model pools which were adapted only on utterances expressing sadness. In order to test for the emotion expressed in a given utterance, we computed the probabilities that the utterance was generated by the emotion-specific acoustic models and took the highest probability to be indicative of the expressed emotion.

Chapter 4

Experiments

In this chapter, we describe the experiments carried out within this investigation. The experiments are divided with respect to the corpus studied. Thus, we have four major sections, in each of which we explore a particular speech corpus. In the first section, we report experiments carried out with the Woggles corpus, a corpus consisting of 50 sentences portrayed by drama students in happy, sad, afraid, and angry variations. This corpus explored, in particular, emotion-specific prosodic and spectral cues. The second corpus, comprised of segments from movies and talk shows, studied the combination of spectral, prosodic, and verbal information. With the third corpus, we conducted some pilot experiments to see whether prosodic models developed on English data extrapolated onto other languages. We tried to detect the underlying emotion of Spanish and German movie segments using prosodic models estimated on the English movie corpus. The last corpus of Spanish spontaneous telephone conversations investigated whether acoustic information could be used to detect emotions in a natural telephone conversation.

Before we go to the individual experiments, we will describe several notions pertaining to all experiments.

4.1 Experimental Set-Up

In this section we describe the training and testing procedures we used to model spectral, prosodic, and verbal information. During the following experiments we frequently carried out small pilot experiments involving human subjects. The tools for these pilots are introduced in the following sections as well. We start with a discussion of elicitation techniques to collect emotional speech.

4.1.1 Elicitation of Emotional Speech

Since it is our ultimate goal to classify the emotions expressed in natural and spontaneous speech, a corpus providing this kind of data would be suited best for our experiments. Unfortunately,

this kind of data is very difficult to procur for several reasons. Privacy issues make it difficult, for instance, to use doctor-patient dialogues. Additional problems arise if we are concerned about the recording quality (Greasley et al., 1995). Some studies, however, were based on natural occurring emotional speech data (Utsuki and Okamura, 1976; Sulc, 1977). Both studies used the radio transmissions of pilots experiencing dangerous situations. Because of the relative extreme nature of the pilots' situations, we have to questions whether the results transfer to more mundane situations. In addition, the emotional categories were quite limited and basically confined to fear, stress, or neutral.

A different way of collecting emotional speech data is to induce a certain emotion in a subject and subsequently collect his or her utterances. One such approach to induce an emotion is to embed the utterance to be collected in a story or to play emotion provoking music or movies before recording the subject's utterances. This technique was used in several studies (Scherer et al., 1985; Scherer and Oshinsky, 1977; Scherer et al., 1991; Katz, 1997) and offers the advantages of a recording situation where the collection of utterances is controllable.

The most common way of collecting emotional speech data is to ask subjects to simulate emotional utterances. The subjects, usually actors, are asked to recreate an emotional state and to utter the sentence to be collected. This technique offers two major advantages. First, the recording situation and the sentence to be collected are now controllable. Second, this technique allows the collection of a large number of utterances in a reasonable amount of time. This last point is important, in particular, if we want to train probabilistic models which require a large number of training tokens for a reliable parameter estimation. However, there are also problems with this kind of data collection. For example, actors might portray only stereotyped cues and might fail to reproduce other more subtle cues. Investigations based on induced and simulated emotional speech also face the question of whether or not their results can be compared with natural emotional speech. That is, is the induced or simulated emotion really the actual emotion we are seeking to elicit? To answer this question at least partially we can test whether a group of control subjects can identify the emotion as they listen to the respective recorded utterances. All our corpora in this investigation had to undergo this kind of "quality control". Studies based on this data category are found, for example, in Walbott and Scherer (1986), Tischer (1993), Bezooijen (1984), and Scherer, Ladd, and Silverman (1984).

Another very interesting approach for collecting emotional speech data is to embed the elicitation and collection directly into a system prototype. The user is told that the system is sensitive to his or her emotional state and the appropriate reaction of the system can be achieved by a Wizard of Oz set-up. A step in this direction of data collection was done by Johnstone (1996) and Riseberg et al. (1997). Healey, Seger, and Picard (1999) used small wearable computers which recorded information about the human subject while he or she was interacting naturally with the environment.

Three of our corpora comprise speech segments produced by actors. Another corpus consists of natural Spanish spontaneous telephone conversations. In our investigation, these four corpora were subject to the following kind of experiments:

- We carried out performance experiments with human subjects to control the quality of the expression of emotions and to asses an upper performance bound.
- We conducted speech recognition experiments on some of these corpora.
- We investigated emotion-specific spectral information.

- We explored emotion-specific prosodic information.
- We studied emotion-specific verbal information.

Each of these tasks is described in more detail in the following sections.

4.1.2 Assessing Human Performance

In order to assess the quality of a corpus consisting of emotional speech segments, experiments were performed with human subjects who first listened to a speech segment and then were asked to judge its expressed emotion. A screen snapshot of one such interface is displayed in figure 4.1. The subject was allowed to listen to the segment of speech as many times as wished by using the

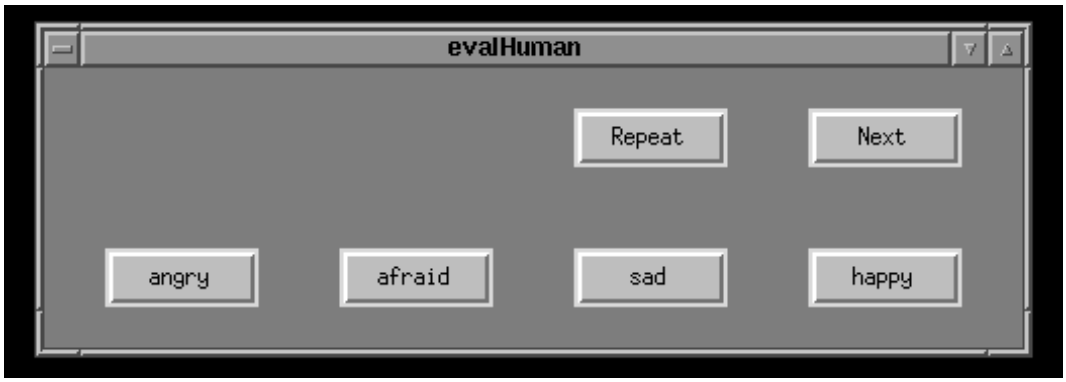


Figure 4.1: Interface for Experiments with Human Subjects.

repeat-button. When an emotion was chosen by clicking on the respective button with the mouse cursor, the button changed its color to black. To proceed to the next segment the subject had to press the *next*-button. We used Sennheiser headsets and the audio capabilities of sun work stations for these experiments.

4.1.3 Spectral Information

We modeled spectral information by means of cepstral coefficients to account for the properties of the auditory system. See sections 2.3 and 2.4 for details. If not indicated otherwise, the speech samples in our corpora were sampled with 16kHz. From the short time spectral analysis we derived a 16ms wide power spectrum that was calculated every 10ms. For the extraction of the speech feature a 30 dimensional melscale filterbank was used and we derived 16 cepstral coefficients from it. We also added the first and second order derivative of these coefficients. In addition, we considered log power and its first and second derivative. Thus, we have a total of 51 features which we reduced to 32 coefficients by linear discriminative analysis. In the experiments investigating emotion-specific spectral information we did not evaluate individual cepstral coefficients. Since the emphasis of this investigation lay on the exploration of prosodic cues, we assessed the potential of the whole set of cepstral coefficients to discriminate among

the emotions. For the experiments involving English corpora we used triphone acoustic models comprising a total of 9358 states each with 70 mixture components per state. For our experiments involving the Spanish telephone corpus, we had 2100 states each with 16 mixture components per state.

To model emotion-specific spectral information we adapted acoustic models on a representation of the speech signal as described above. Consult section 3.3 for the details of the adaptation technique. For the classification of the expressed emotion in some speech segment we used the following procedure:

1. Depending on the experiment set-up, we used either the transcribed utterance or the utterance as recognized by a recognition system.
2. Using the same recognition system as in step 1, we computed for all emotions the probability that the utterance from step 1 was produced by the respective emotion-specific acoustic models:
$$P(speech\ signal \mid sentence, acoustic\ models_i), \tag{4.1}$$
where i was an element from the set of emotions, E, we wanted to discriminate.
3. We compared the probabilities as computed in step 2 and took the highest probability to be indicative of the underlying emotion, that is to say, we maximized equation 4.1:
$$expressed\ emotion\ = \arg \max_{i \in E} P(speech\ signal \mid sentence, acoustic\ models_i), \tag{4.2}$$
where E denoted the set of emotions we wanted to discriminate.

4.1.4 Prosodic Information

One of the main aims of this investigation was to find robust prosodic features which detect the expressed emotion across speakers and utterances. In this section we briefly describe how we extracted prosodic information from the speech signal and how we trained and tested emotion-specific prosodic models.

As mentioned above, the auditory experience of loudness or pitch can not be reduced solely to energy or fundamental frequency, respectively. Both experiences are based on an intricate interplay of energy, pitch, and duration (Zwicker and Feldkeller, 1967; Zwicker and Fastl, 1990). In our experiments, we modeled these interdependencies by considering information from all three dimensions in parallel. Thus, we operationalized loudness, for instance, by some energy measure and assumed that dependencies with fundamental frequency and duration were taken care of by modeling this information in parallel. Thus, a prosodic model observes a vector of prosodic events comprising information about the fundamental frequency, intensity, and duration.[1]

Intensity

The computation of intensity consists of two parts: a transformation function, $T()$, of the digitized speech signal, s, and a window function, w, to model some context (Rabiner and Schafer, 1978).

[1] This approach seems to be quite common, see (Kießling, 1997; Kompe, 1996) for similar approaches.

We use N to indicate the window size and n and m to refer to some frame within the speech signal. Thus, the energy at some time, m, is given by equation 4.3.

$$E_m = \sum_{n=0}^{N-1} T(s_n) w_n \tag{4.3}$$

For our experiments we will use a Hamming window which is given by equation 4.4.

$$w_n^H = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{4.4}$$

Other possible windows are the rectangle window and the Hann window, given in equations 4.5 and 4.6.

$$w_n^R = 1 \text{ where } n \in [0; N-1] \tag{4.5}$$

$$w_n^H = 0.50 - 0.50 \cos\left(\frac{2\pi n}{N-1}\right) \tag{4.6}$$

For the transformation function we have several options which are given in equations 4.7, 4.8, and 4.9, respectively.

$$T_s(x) = x^2 \tag{4.7}$$

$$T_{rms}(x) = \sqrt{x^2} \tag{4.8}$$

$$T_{abs}(x) = \mid x \mid \tag{4.9}$$

For our experiments, we will use equation 4.7 as the transformation function in combination function with a Hamming window as the context function.

Fundamental Frequency

Extracting the fundamental frequency from a speech signal turns out to be quite complex because of certain properties of the speech signal (Hess, 1983). For instance, depending on the articulation of different sounds, the spectral content of the signal changes constantly. In addition, the glottal impulses do not always have the same amplitude and the signal is, therefore, amplitude modulated. In order to compensate for these properties of speech, pitch tracking algorithms rely most of the time on short time analysis windows. Such a window, however, can contain several pitch periods of the fundamental frequency and, in addition, can comprise voiced and unvoiced regions (Medan, Yair, and Chazan, 1991). All these circumstances make the exact determination of the fundamental frequency difficult and we have to be aware that the estimation of the fundamental frequency of a given utterance might be errornous.

For our experiments we used a pitch tracker developed by the Cambridge University Engineering Department. The pitch tracker is based on an algorithm as described in (Medan, Yair, and Chazan, 1991) which relies on two passes through the signal. In the first pass a set of possible pitch locations is computed for every 10ms using two adjacent non-overlapping windows and cross-correlation. In the second step we compute the overall best pitch locations using dynamic programming. In order to validate the results of our experiments investigating the fundamental frequency we used a second pitch tracker which was developed by the Entropic Research Lab based on an algorithm by (Secreset and Doddington, 1983). We did not find any significant differences in our results using this pitch tracker.

Speaking Rate

The discussion in chapter 2 showed that speaking can signal an emotional involvement of the speaker. We will operationalize speaking rate by segment durations. That is, the longer the average duration of the segments within an utterance, the slower the speaking rate. The duration of segments is also an important prosodic parameter which participates in, for example, the implementation of accents and prosodic boundaries (preboundary lengthening), (Wightman et al., 1991). In order to observe a salient lengthening of a segment, that is, a slow speaking rate, we have to know about the segments normal duration. Following Wightman et al. (1991), we can compensate for the inherent durational differences of a segment, i.e., a phone, by a normalization step as given in equation 4.10 where d_i refers to the actual duration of the ith segment labelled as p, and $\mu_{p,i}$ and $\sigma_{p,i}$ are the mean and standard deviation of the duration of phone p.

$$\hat{d}_i = \frac{d_i - \mu_{p,i}}{\sigma_{p,i}} \tag{4.10}$$

Crystal and House (1988) showed that speaking rate has a strictly linear influence on the duration of segments. Hence, we can model the impact of the overall speaking rate by a scaling factor, α. Thus, equation 4.10 becomes:

$$\hat{d}_i = \frac{d_i - \mu_{p,i}\alpha}{\sigma_{p,i}\alpha} \tag{4.11}$$

This leaves us with the estimation of the scaling factor α. We can approximate α using segments of speech with a constant speaking rate:

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\mu_{p,i}} \tag{4.12}$$

where N is the number of segments in the speech segment used for the estimation of α. For this investigation, we were interested in emotion-specific variations of the speaking rate, that is, emotion-specific variations of the scaling factor α. We modeled emotion-specific speaking rate variations by estimating the average duration of segments in emotion-specific subsets of the respective training corpora using the suprasegmental hidden Markov model introduced in section 3.2.2. Thus, we computed, for example, the average duration of the phone /a/ in utterances which expressed sadness.

For a hierarchical definition of speaking rate, see also (Chung and Seneff, 1997). Other approaches approximated the speaking rate directly from the signal by measuring the length or the frequency of voiced segments in a speech sample (Amir and Ron, 1998; Dellaert, Polzin, and Waibel, 1996; Thymé-Gobbel, 1998).

Postprocessing of Prosodic Base Features

Before we could estimate emotion-specific models, we had to account for flaws in the extraction of prosodic base features. For instance, postprocessing the output of the pitch tracker included:

- removal of unvoiced regions for global features such as mean pitch,
- median or mean smoothing,

- normalization with respect to a speaker's baseline.

Pitch normalization, turns out to be particularly problematic if one tries to normalize based on very limited data. For some of our experiments, the speaker's identity was known and we disposed of a substantial number of speech tokens to normalize with respect to his or her baseline. Normalization based on this procedure yields, in general, better classification results. If the speaker's identity is not known or there is not a sufficient number of speech tokens, normalization became more difficult. Ladd (1983) proposed that a speaker's pitch baseline could be approximated based on the fundamental frequency of the last syllable, provided the sentence is not a question or request. Scherer and Bergmann (1984) suggested to approximate a speaker's baseline by the average of the lowest 5% of the fundamental frequency values within an utterance segment.

For intensity, similar caveats apply. In addition to differences among actors, we also had to account for intensity differences among movies due to different recording settings. In general, we used the minimum and maximum intensity values in voiced regions for normalization. Normalizing intensity of a given speech segment in a movie by the overall average can be flawed when the distribution of emotional segments is not uniform.

In general, normalization techniques have to be applied very carefully, since the purpose is to normalize with regard to variations among individuals or movies while preserving differences due to the expression of an emotion. Disregarding these problems can result in deletion of prosodic information which is essential for the detection of cues used by the speaker to express a certain emotion.

Derived Prosodic Features

We will illustrate the extraction of prosodic features using figure 4.2 which shows an idealized fundamental frequency (red lines) of some utterance. We can divide the prosodic features into three major groups:

1. **Global features** refer to features pertaining to the whole utterance, for instance mean pitch – referred to as line 1 in figure 4.2 – or standard deviation (2).

2. **Local features** refer to segments smaller than the utterance. These segments can refer either to words, syllables, phones, or some other segment. Within these segments, we are able to compute features such as mean or variance of the fundamental frequency or intensity. In addition, we can take advantage of the segment size and compute features such as the slope of the fundamental frequency over this segment (3 and 4) using, for instance, regression methods for curve fitting. We can compute additional local features by moving a small window over the fundamental frequency and compute the slope of the contour in this window. Using the slope information we can compute the following features:

   (a) the numbers of changes from a positive to a negative slope normalized by the total number of windows,

   (b) the number of positive slopes divided by the total number of windows, and

   (c) the number of negative slopes divided by the total number of windows.

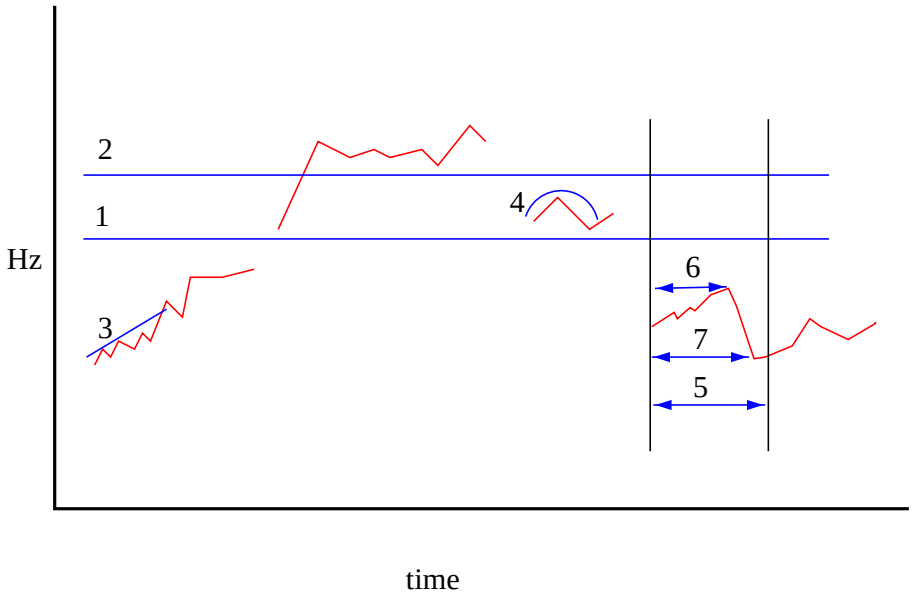


Figure 4.2: Illustration of Prosodic Features. See text for an explanation.

Additional features involve the goodness of the fit of the regression line to the actual contour, χ^2:

(a) the sum of all χ^2 of all positive slopes normalized by the number of positive slopes,

(b) the sum of all χ^2 of all negative slopes normalized by the number of negative slopes,

(c) the sum of all χ^2 of all slopes normalized by the number of windows.

3. **Durational features** allow us to compute the speaking rate (5). In our investigation, this computation is based on the duration of phones. Knowing the duration of a segment allows also the computation of vehemence features, that is, the duration up to the maximum (6) or the minimum (7).

Using the prosodic features as described above we can train emotion-specific prosodic models with Gaussian mixtures and the Viterbi algorithm of the suprasegmental hidden Markov model as described in section 3.2.2. In order to test the accuracy of prosodic models on the classification of emotional speech segments, we used the following procedure:

1. Depending on the experiment set-up we used either the transcribed utterance or the utterance as recognized by a recognition system.

2. Using the same recognition system as in step 1, we computed for all emotions the probability that the utterance from step 1 was produced by the respective emotion-specific prosodic models.

$$P(speech\ signal \mid sentence, prosodic\ models_i), \tag{4.13}$$

where i was an element of the set of emotions we want to discriminate.

3. We compared the probabilities as computed in step 2 and took the highest probability to be indicative of the underlying emotion, that is, we maximized equation 4.13:

$$expressed\ emotion = \arg\max_{i \in E} P(speech\ signal \mid sentence, prosodic\ models_i), \tag{4.14}$$

where E denoted the set of emotions we want to discriminate.

In this investigation we conducted various experiments to evaluate prosodic features with regard to their potential to predict the expressed emotion in an utterance. In the following we report only results of those experiments which proved that the respective prosodic feature allowed a reasonable classification accuracy. Several of the prosodic features suggested by previous research turned out not to yield reliable cues. For instance, Tischer (1993) proposed vehemence features, that is, the time from the beginning of a segment until the minimum or maximum of the fundamental frequency or the intensity. We illustrated these features in Figure 4.2 by (6) and (7). We were not able to duplicate his findings in our experiments. We computed these features on phone segments and on voiced segments but the respective classification accuracies did not significantly exceed chance level. Note that Tischer found these features by studying only one utterance pronounced by four speakers. An additional set of features which did not yield reliable classification results were intonation based features. We conducted several experiments in which we investigated whether certain contours of the fundamental frequency systematically signal a certain emotion. We modeled contours on phone segments and on voiced segments but neither approach lead to an reasonable accuracy. The failure of our experiments to demonstrate that these prosodic features are reliable indicators for the expressed emotions does, of course, not imply that these features can not be used to signal an emotion in general; see, for example, (Davitz, 1964; Fonagy, 1978; Katz, Cohn, and Moore, 1996). We think that there are several reasons for these features not to become reliable indicators in our experiments. Note that our corpora were substantially larger than the corpus used, for instance, by Tischer (1993). If these features are to some extend optional and are, as a consequence, not employed in every utterance expressing a certain emotion, then a probabilistic approach has difficulties to reliably estimate the corresponding parameters. In addition, these features are much more susceptible to interferences arising from other communicative functions also implemented by the modification of prosodic parameters. See section 2.4.2 for details. Thus, in order to model these prosodic features in future research an integrated approach is needed that models several communicative functions of prosody.

4.1.5 Verbal Information

We used emotion-specific back-off language models to detect the expressed emotion in an utterance. We trained these language models on emotion-specific data to model verbal information. Consult section for 3.1 more details. In order to test the accuracy of these models to detect the expressed emotion in an utterance, we used the following procedure:

1. For each emotion we computed the probability that the words in the utterance were produced by the respective language model.

$$P(w_1 w_2 \ldots w_N \mid language\ model_i), \tag{4.15}$$

where i was an element in the set of emotions we wanted to discriminate and N denoted the number of words in the utterance.

2. We compared the probabilities as computed in the previous step and took the highest probability to be indicative of the underlying emotion, that is, we maximized equation 4.15:

$$expressed\ emotion = \arg\max_{i \in E} P((w_1 w_2 \ldots w_N \mid language\ model_i), \tag{4.16}$$

where E denoted the set of emotions we wanted to discriminate.

4.1.6 Combining Prosodic, Spectral, and Verbal Information

In order to use concurrently prosodic, spectral, and verbal information as a mean to detect the emotion expressed in an utterance, we combined linearly their corresponding individual probabilities:

$$\begin{aligned} expressed\ emotion = \arg\max_{i \in E} & \\ & \lambda_1 P(w_1 w_2 \ldots w_N \mid language - model_i) \\ & \lambda_2 P(speech\ signal \mid sentence, prosodic\ models_i) \\ & \lambda_3 P(speech\ signal \mid sentence, acoustic\ models_i), \end{aligned} \tag{4.17}$$

where E denoted the set of emotions we want to discriminate, and N the number of words in the input sentence. The interpolation weights, λ, were determined empirically on some independent development set.

4.1.7 Displaying Results

We use two ways of reporting and displaying results of our experiments: confusion matrices which give information about the confusability among the emotions to be classified and a measure which combines recall and precision to give a condensed form of the accuracy of a classification.

Within a confusion matrix, we display truth in the columns which, therefore, add up to 100%. Thus, a cell in column x and row Y indicates the percentage of how many of the segments signalling emotion x are classified as Y. For illustration, we display a confusion matrix in Table 4.1 below. For instance, 75.8% percent of the happy segments are classified as HAPPY, or 25.0%

Table 4.1: Example of a confusion matrix.

	happy	sad	afraid	angry
HAPPY	75.8	3.2	11.1	6.1
SAD	7.1	65.1	24.4	8.0
AFRAID	5.7	25.0	58.5	5.8
ANGRY	11.4	6.7	6.0	80.1

of the sad segments are classified as AFRAID. Note that columns add up to 100%.

A more condensed form for reporting results of experiments is the f1-score which combines precision and recall. Under *precision* we understand the ratio of the number of segments classified correctly as i and the number of segments in the corpus classified as i, regardless whether correctly or not. We define *recall* as the ratio of the number of segments classified correctly as i and the total number of segments in the respective corpus belonging to class i. The corresponding formulas are given below:

$$\begin{aligned} precission_i &= C_i/T_i \\ recall_i &= C_i/I_i \end{aligned} \tag{4.18}$$

where

C_i *is* the number of segments in the corpus classified correctly as i,

T_i the number of segments in the corpus classified as i, regardless whether correctly or not, and

I_i the actual number of segments in the corpus belonging to class i.

Combining precision and recall of some class i, we get the corresponding f1-score, defined as

$$f1_i = \frac{2 * precission_i * recall_i}{precission_i + recall_i} \tag{4.19}$$

The closer the precision, the recall, and the corresponding f1-score is to 1, the more accurate the classification of the respective classification system.

4.2 The Woggles Corpus

The first corpus in our investigation, hence forth the Woggles Corpus, consisted of 50 different sentences portrayed by 9 female drama students expressing either happiness, sadness, fear, or anger. That is, the same material was available in all four different emotional realizations. By having the same sentence pronounced in four different emotional variations we prevented the students from using verbal cues and forced them to rely on spectral and prosodic cues only. The Woggles Corpus was, therefore, used to explore the discriminative power of prosodic and spectral features.

4.2.1 The Corpus

The sentences in the corpus comprised questions, statements, and orders. The sentence length varied from 2 to 12 words; the mean sentence length was 5.8 words. The corpus consisted of 291 word tokens (87 word types).

We asked drama and linguistic students to express the emotion given in square brackets at the beginning of the sentence on a computer screen. An example is given in Figure 4.3. The subjects

[HAPPY] I want to be friends.
[HAPPY] I want to see you go through the chute again.
[HAPPY] I want to sleep.
[HAPPY] I would love to play follow the leader.
[HAPPY] I'm angry because you won't play.

Figure 4.3: Subjects were asked to express the emotion as indicated by the label in the square brackets.

were asked to portray five sentences in the same emotion, followed by the same five sentences to be pronounced in the remaining emotions. Thus, we had a maximum of 200 sentences for a given speaker. We used the utterances of 5 speakers for training and testing. We reserved the utterances of two speakers as a development set. The utterances of the last two speakers were reserved to test for speaker independence of spectral and prosodic models.

SennHeiser HMD 410 or SennHeiser HMD 414 microphones were used for all recordings. We used a gradient box (Gradient Desklab Model 14) with a sampling rate of 16 kHz for recoding. During the recording sessions, attention was paid on the energy of the utterance in order to ensure that it lay within a certain range. In case the energy distribution of an utterance exceeded or fell below this range, the subject was asked to repeat the utterance and to speak louder or softer. While this recording schema guaranteed a smooth signal, it also excluded very dynamical utterances. As we see in the following experiments based on this corpus, intensity as a consequence failed to be become a reliable indicator for emotions. Finally, all utterances were transcribed by trained students.

4.2.2 Assessing Human Performance

In order to control the quality of the expression of emotions in the Woggles corpus, we carried out an experiment in which we asked humans to classify utterances as either happy, sad, afraid, or angry. For this experiment, we asked eleven subjects, both men and women, to listen and classify 240 utterances drawn from the Woggles test corpus. The experiment included a brief training period in which the subjects could familiarize themselves with the data by listening to 40 utterances. During that training session the respective emotion-button became dark, thus informing the subject about the underlying emotion of the current utterance. For more details about the experimental design see section 4.1.2 above.

The confusion matrix based on the performance of eleven subjects is given in Table 4.2. The emotions in lower case indicate the actual emotion as portrayed by the actors and, thus, the percentages in the columns add up to 100%. As apparent from the confusion matrix above, subjects

Table 4.2: Confusion matrix for Human Subjects. Overall about 69% of the utterances were classified correctly.

	happy	sad	afraid	angry
HAPPY	**75.8**	3.2	11.1	6.1
SAD	7.1	**65.1**	24.4	8.0
AFRAID	5.7	25.0	**58.5**	5.8
ANGRY	11.4	6.7	6.0	**80.1**

experienced problems with distinguishing sad and afraid utterances. Most of the confusion took place between this pair of emotions. For instance, 25.0% of the sad sentences were classified as afraid and 24.4% of the afraid sentences were classified as sad. Looking at the f1-scores, given in Table 4.3, we see that the confusion between sad and afraid utterances was reflected in the respective f1-scores. Sad had an f1-sore of 0.64, afraid of 0.60. Compare these f1-scores with the f1-score of happy (0.77) and angry (0.78). The overall f1-score was 0.69.

Table 4.3: Precision, recall, and f1-scores for human subjects.

	happy	sad	afraid	angry
precision	0.79	0.62	0.62	0.77
recall	0.76	0.65	0.58	0.8
f1	0.77	0.64	0.60	0.78

The percentage of correctly classified sentences for each of the four actors is given below in Table 4.4. The percentage of correctly classified sentences ranged from 65.3% to 75.0%. That is, the sentences of all speakers were detected by the eleven subjects with a comparable consistency and well above chance level (25%).

On a side note, actor C was able to portray all of the afraid sentences in a way that none of the subjects confused it with an angry portrayal. The performance of the eleven subjects was consistently better than chance and ranged from 65% to 89% correctly classified segments. The performance of the female listeners tended to be better than the performance of the male listeners.

Table 4.4: Speaker-specific Human Discrimination Performance

	A	B	C	D
Correct	65.3	70.8	71.3	75.0

However, the number of subjects is too small to make any claims about gender differences in the perception of emotions (Bonebright, Thompson, and Leger, 1996). With this experiment we validated the Woggles corpus. Subjects were able to detect the emotions expressed by the actors consistently and well above chance level. This was the case for all four actors in the test set.

In the following experiments, we explored the possiblity of achieving a comparable accuracy by building classification systems relying on spectral and prosodic information. First, however, we investigated whether the expression of a particular emotion in someone's speech has an impact on the accuracy with which an utterance can be recognized automatically. That is, does the expression of certain emotions correlate with particular speech recognition accuracies?

4.2.3 Recognizing Emotional Utterances

For the following experiments, we used a speech recognition system which was developed independently on a different corpus of English spontaneous speech and achieved an accuracy of about 86% (Lavie et al., 1997). Before carrying out emotion-specific word recognition tests, we adapted the recognition system to the Woggles corpus using adaptation of acoustic models. See section 3.3 for a detailed description of this adaptation technique. Note that the adaptation was carried out on the whole training set, that is, on all utterances regardless of the expressed emotion.

For testing, however, we divided the test corpus into emotion-specific subsets to see whether there are emotion-specific differences in word accuracy.[2]

As it turned out – the results are given in Table 4.5 – depending on the emotional variation in the test sentences, the differences in word accuracy were quite large. Angry sentences achieved the highest recognition accuracy with 76% while happy sentences scored about 10% worse. Sad and afraid sentences were recognized with a word accuracy of about 70%. The average word

Table 4.5: Word accuracy depending on the underlying emotion of the test utterances. The average word accuracy was about 70%.

	happy	sad	afraid	angry
word accuracy	63.3	70.3	68.8	76.2

accuracy was 69.7%. For this experiment we used a vocabulary with about 1000 words and a language model with a perplexity of 46.8. Note that the sentences were the same for all four emotions. Thus, differences in word accuracies were based solely on acoustic differences.

This experiment demonstrated the necessity to model acoustic differences in emotional speech in future speech recognition systems. Even though we will not pursue the speech recognition problem any further, the following experiments point to major acoustic differences of emotional speech and thus might suggest where to improve acoustic modeling in recognition systems.

[2] We use the standard definition of *word accuracy* (wa) given in 4.20 to evaluate the performance of the recognition system.

$$wa = \frac{refs - subst - del - ins}{ref} \tag{4.20}$$

where

$refs$ is the number of reference words,
$subst$ is the number of substitutions,
del is the number of deletions, and
ins is the number of insertions.

4.2.4 Emotion-Specific Spectral Information

In order to model emotion-specific spectral information, we adapted spectral models on emotion-specific subsets of the training corpus. Consult section 4.1.3 for the adaptation procedure. Using acoustic models from the recognition system as described in section 4.2.3, we adapted spectral models on emotion-specific subsets of the corpus. That is, we had spectral models which were adapted only to happy sentences, models which were adapted only to angry sentences and so on. Using these emotion-specific adapted spectral models, we tried to recover the expressed emotion of an utterances in the test corpus by applying the algorithm as described in the beginning of this chapter in section 4.1.3. Note that we used the transcribed text for both training and testing.

Table 4.6: Confusion matrix. Overall, about 68.8% of the segments were correctly classified.

	happy	sad	afraid	angry
HAPPY	68.3	5.0	11.7	8.3
SAD	13.3	70.0	33.3	8.3
AFRAID	11.7	23.3	55.0	1.7
ANGRY	6.7	16.7	0.0	81.7

In Table 4.6 we show the confusion matrix of the outcome of this experiment. Most of the confusion took place between sad and afraid utterances. For instance, 23.3% of the sad utterances were classified as afraid and a third of the afraid utterances were classified as sad. We encountered a similar confusion when we asked human subjects to perform this task, see Table 4.2 for details. Similar to the results in the experiment with the human subjects, happy and angry sentences could be discriminated from the remaining emotions quite accurately. Note that none of the afraid segments was classified as angry.

Table 4.7: Precision, recall, and f1-scores for spectral information.

	happy	sad	afraid	angry
precision	0.73	0.56	0.60	0.91
recall	0.68	0.70	0.55	0.82
f1	0.71	0.62	0.57	0.86

As a result of the confusion between sad and afraid sentences, the corresponding f1-scores for these two emotions, given in Table 4.7, were only around 0.6, whereas the f1-score for happy and angry utterances were 0.71 and 0.86, respectively. This accuracy was comparable to the accuracy of human listeners as assessed in section 4.2.2.

Using emotion-specific spectral models for recovering the underlying emotion of an utterance was surprisingly robust against a decrease in word accuracy. Instead of using the transcription as the input for the emotion detection, we used in a second experiment the actual recognized words as the input. We decreased the word accuracy to about 52% by decreasing the weight of the language model in the speech recognition step. The subsequent emotion detection accuracy only dropped to about 65%. Apparently, the remaining islands of correct recognized speech were still sufficient for a reasonable discrimination among the emotions. In addition, note that the

recognized sentences still bore some phonetic similarity with the actual sentence. An example of an actual sentence uttered and the corresponding recognized sentence is given in 4.21 and 4.22 below.

(4.21) what do you want

(4.22) would you want

Finally, note that all four emotion classes were affected by the degradation in the word accuracy.

4.2.5 Emotion-Specific Prosodic Information

In the following sections we evaluated particular prosodic features with regard to their potential to discriminate among the four emotions represented in the Woggles corpus. Moreover, for each feature, we also computed the relative order of the emotion-specific values to each other. Later on, we compared the relative order of emotion-specific values across different corpora to explore whether the respective features were consistent.

Fundamental Frequency (Utterance Mean and Variance)

In the first experiment, we computed the global mean and variance of the fundamental frequency within a given utterance segment. We removed spikes to compensate for errors in the pitch detection process before mean smoothing. In addition, we performed a speaker-dependent normalization. We then trained emotion-specific prosodic models based on these two features. We used the test procedure as described in section 4.1.4 to classify the utterances in the test corpus. The corresponding confusion matrix is given in Table 4.8. Angry segments were classified best since 83% of the angry segments were detected correctly. Most of the confusion took place between sad and afraid segments. For instance, a third of the sad segments were classified as afraid and 21.7% of the afraid segments were classified as sad. Overall, 55.8% of the segments were classified correctly.

Table 4.8: Confusion matrix based on the mean and variance of the fundamental frequency. Overall 55.8% of the utterances were classified correctly.

	happy	sad	afraid	angry
HAPPY	41.7	5.0	15.0	5.0
SAD	20.0	46.7	21.7	10.0
AFRAID	20.0	33.3	51.7	1.6
ANGRY	18.3	15.0	11.6	83.0

If we look at the corresponding f1-scores, i.e. the combination of precision and recall, we see that all emotions except angry had an f1-score of about 0.5. The f1-score for angry was 0.73. Precision, recall, and f1-scores are given in Table 4.9 below.

Table 4.9: Precision, recall, and f1-scores for mean and variance of the fundamental frequency.

	happy	sad	afraid	angry
precision	0.63	0.47	0.48	0.65
recall	0.42	0.47	0.52	0.83
f1	0.50	0.47	0.50	0.73

The relative order of the emotion-specific mean and variance values of the fundamental frequency are given in Table 4.10. The values for angry occupied the most salient positions in this

Table 4.10: Relative order of mean and variance of F_0

	happy	sad	afraid	angry
happy		>	<	>
sad	<		<	>
afraid	>	>		>
angry	<	<	<	

(a) F_0 mean

	happy	sad	afraid	angry
happy		>	>	>
sad	<		<	>
afraid	<	>		>
angry	<	<	<	

(b) F_0 variance

table since their respective values were lower for both mean and variance than the values of the remaining emotions. These salient positions of the values for angry explained the high accuracy with which angry utterances could be detected. Remember, the f1-score was 0.73. The positions of the other emotions were less pronounced and their detection accuracy was as a consequence lower.

Fundamental Frequency (Jitter))

With the following experiment we tried to capture some of the dynamics of the fundamental frequency. We moved a window over the fundamental frequency and computed the slope of the pitch contour in the corresponding segments. Using this slope information we computed the following two features:[3]

1. the number of changes from a positive to a negative slope (or vice versa) normalized by the total number of windows and
2. the sum of all χ^2 of all slopes normalized by the total number of windows.

Note that for the computation for these features the fundamental frequency was not smoothed. We trained and tested these two features using the procedures as described in section 4.1.4. The confusion matrix for a classification system relying on these two features is given in Table 4.11. We found the substantial confusion between sad and afraid segments. For instance, 41.7% of the sad segments were classified as afraid and 18.3% of the afraid segments as sad. Another large source of confusion was that 58.3% of the happy segments were classified as angry. Only 6.7% of the happy segments were correctly classified. Jitter information seemed to help mainly the discrimination of afraid and angry segments, since 61.7% of the afraid and 65.0% of the angry segments could be detected.

Table 4.11: Confusion matrix based on jitter information. Overall 40.4% of the utterances were classified correctly.

	happy	sad	afraid	angry
HAPPY	6.7	3.3	6.6	0.0
SAD	20.0	28.3	18.3	31.7
AFRAID	15.0	41.7	61.7	3.3
ANGRY	58.3	26.7	13.4	65.0

We found the assumption that these two jitter features mainly helped to discriminate afraid and angry segments confirmed when we looked at the corresponding f1-scores given in Table 4.12. Only the f1-scores for afraid and angry segments lay with 0.56 and 0.49 significantly above

Table 4.12: Precision, recall, and f1-scores for jitter features.

	happy	sad	afraid	angry
precision	0.40	0.29	0.51	0.40
recall	0.07	0.28	0.62	0.65
f1	0.11	0.29	0.56	0.49

chance level. The f1-score for happy segments was 0.11, the one for sad segments 0.29. Overall, about 40% of the segments were classified correctly.

[3]See the section 4.1.4 at the beginning of this chapter for a more detailed description of these features.

We found the reasons that these two jitter features detected afraid and angry segments better than happy and sad segments when we looked at the relative positions of the respective emotions-specific values given in Table 4.13. For both features it was the case that the values for angry

Table 4.13: Relative order of the two jitter features.

	happy	sad	afraid	angry
happy		<	<	>
sad	>		<	>
afraid	>	>		>
angry	<	<	<	

(a) normalized number of changes

	happy	sad	afraid	angry
happy		<	<	>
sad	>		<	>
afraid	>	>		>
angry	<	<	<	

(b) normalized χ^2

were smaller than of any other emotion. For the afraid values the opposite was true, the values were larger than for any other emotion. Thus, afraid and angry values occupied the extreme positions while the values of happy and sad lay in between.

Intensity (Mean and Variance)

In the following experiment we explored intensity mean and variance within an utterance. As mentioned in section 4.2.1, the recording procedure prevented the collection of very dynamical utterances since the energy of all utterances had to lie within a certain range. As a consequence, intensity failed to become a reliable indicator for emotional speech, contrary to previous research (Scherer, Ladd, and Silverman, 1984; Frick, 1985; Katz, 1997). For this experiment, intensity was normalized with respect to the speaker and the computation of the two intensity features, mean and variance, was based only on voiced segments. The confusion matrix of the resulting classification system based on these two intensity features mean and variance is given in Table 4.14. Overall, only about 33% of the segments were classified correctly. The most surprising numbers in the matrix are that 58% of the angry segments were classified as afraid and that only five percent of the angry segments were correctly classified. We also found substantial confusion between sad and afraid segments. For instance, 50% of the sad segments were classified as afraid. Moreover, 55% of the happy segments were misclassified as afraid. In general, more than half of the segments in the test corpus was classified as afraid.

Table 4.14: Confusion matrix based on intensity. Overall only about a third of the utterances were classified correctly.

	happy	sad	afraid	angry
happy	26.7	3.3	6.7	13.4
sad	16.7	46.7	38.3	23.3
afraid	55.0	50.0	53.3	58.3
angry	1.6	0.0	1.7	5.0

The strong tendency of the system to classify segments as afraid led to low f1-scores which are given in Table 4.15. Only sad segments seemed to profit from intensity information. Sad segments could be detected with an f1-score of 0.41, while happy and afraid had an f1-score of only 0.36 and 0.34, respectively. Angry segments had a very low f1-score of 0.09.

Table 4.15: Precision, recall, and f1-scores for mean and variance of the intensity.

	happy	sad	afraid	angry
precision	0.53	0.37	0.25	0.60
recall	0.27	0.47	0.53	0.05
f1	0.36	0.41	0.34	0.09

Looking at the relative positions of the mean and variance values in Table 4.16, we see that the mean intensity for angry was larger than for any other emotion. Moreover, the intensity mean for sad was smaller than the values of the remaining emotions. Finally, the mean intensity for happy was bigger than the mean intensity of afraid. Surprisingly, the internal relations between the emotions for intensity variance were identical to the relations for intensity variance. That is, the variance for angry was larger and the variance for sad was smaller than for the remaining emotions. These extreme positions of the sad and angry values in these tables explained the

respective values in the confusion matrix above. Note that no sad segment was mistaken as angry.

Table 4.16: Relative order of intensity mean and variance.

	happy	sad	afraid	angry
happy		>	>	<
sad	<		<	<
afraid	<	>		<
angry	>	>	>	

(a) Intensity mean

	happy	sad	afraid	angry
happy		>	>	<
sad	<		<	<
afraid	<	>		<
angry	>	>	>	

(b) Intensity variance

Overall, classification based on intensity information yielded a very low accuracy of about 33%. Remember, for example, that classification based on information about the utterance mean and variance of the fundamental frequency resulted in an accuracy of 55.8%. In subsequent experiments, involving other corpora, we demonstrated that this low accuracy of these two intensity features was not necessarily the case since intensity information could become a reliable indicator for emotions. As mentioned earlier, we attributed the failure of intensity to discriminate among the respective emotions to the recording conditions of this corpus's data collection.

Intensity (Tremor)

The previous experiment showed that intensity mean and variance failed to become a reliable indicator for the expressed emotion of a speech segment. We attributed this failure to particular recording conditions which forced the drama student's intensity to fall within a certain range, thus, preventing them from producing very soft or very loud speech. This constraint, however, should not have had any effect on tremor features. While these features are still based on intensity, they only measure small perturbations in the overall intensity contour. These small perturbations should still be present in the signal regardless of the constraint on the recording conditions. To test this hypothesis, we computed the following two features on the logarithms of the intensity in voiced segments within an utterance by moving a window over the voiced regions of the intensity contour:

1. the number of changes from a positive to a negative slope (or vice versa) normalized by the total number of windows and

2. the sum of all χ^2 of all regression slopes normalized by the number sum windows.

We trained emotion-specific prosodic models using these two features and used the test procedure as described in section 4.1.4. The confusion matrix of this experiment is given in Table 4.17. The system had the tendency to classify segments as either afraid or angry. 38.3% of the happy and 46.7% of the sad segments were misclassified as afraid and 23.3% of the happy and 25.0% of the sad segments as angry. As a consequence, only 31.7% of the happy and 10.0% of the sad segments were classified correctly. Tremor features helped mainly to detect afraid and angry segments since 70.0% of the afraid and 70.0% of the angry segments were classified correctly.

Table 4.17: Confusion matrix based on tremor information. Overall 45.4% of the utterances were classified correctly.

	happy	sad	afraid	angry
happy	31.7	18.3	15.0	15.0
sad	6.7	10.0	1.7	5.0
afraid	38.3	46.7	70.0	10.0
angry	23.3	25.0	13.33	70.0

As a consequence of the high confusion of happy and sad with afraid and angry, only the f1-scores for afraid and angry lay significantly above chance level. The f1-scores are given in Table 4.18. For afraid and angry the f1-score were 0.53 and 0.60, respectively, whereas happy achieved only an f1-score of 0.32. Sad was most difficult to detect, its f1-score was only 0.16. Overall, 45.4% of the segments were classified correctly.

Looking at the relative emotion-specific positions of these two features, given in Table 4.19, we see that the value of the first tremor feature was largest for afraid, followed by angry. Sad had the smallest value. The value for angry was the smallest for the second feature and the highest for afraid. These extreme positions of the values for angry and afraid explained their high f1-scores.

Table 4.18: Precision, recall, and f1-scores for tremor features.

	happy	sad	afraid	angry
precision	0.40	0.43	0.42	0.53
recall	0.32	0.10	0.70	0.70
f1	0.32	0.16	0.53	0.60

Table 4.19: Relative order of the two tremor features.

	happy	sad	afraid	angry
happy		>	<	>
sad	<		<	<
afraid	>	>		>
angry	>	>	<	

(a) normalized number of changes

	happy	sad	afraid	angry
happy		>	<	>
sad	<		<	>
afraid	>	>		>
angry	<	<	<	

(b) normalized χ^2

This experiment proved our hypothesis that these two tremor features were not affected by the particular recording conditions. Tremor features turned out to be quite reliable features for the detection of afraid and angry segments. The overall accuracy was well above chance level and well above the accuracy we achieved with intensity mean and variance in the previous experiment.

Speaking Rate (Phone Duration)

The suprasegmental hidden Markov model which we introduced in section 3.2.2 allowed us to model speaking rate by state occupancy. That is, we used suprasegmental states to model phones and we recorded the durations spent in these suprasegmental states. In the following experiment, we computed the durations of vowels to model emotion-specific speaking rates.

The confusion matrix of a system using this single duration feature is given in Table 4.20. We found the typical confusion between afraid and sad segments: 18.3% of the sad segments were classified as afraid and 13.3% of the afraid segments as sad. Moreover, substantial confusion took place between afraid and happy, 41.7% of the afraid segments were classified as happy, and 23.3% of the happy segments were classified as afraid. Despite all this confusion, about 42% of the segments in the test set were classified correctly.

Table 4.20: Confusion matrix based on phone duration. Overall, about 42% of the utterances were correctly classified.

	happy	sad	afraid	angry
HAPPY	46.7	18.3	41.7	28.3
SAD	15.0	45.0	13.3	5.0
AFRAID	23.3	18.3	28.3	20.0
ANGRY	15.0	18.4	16.7	46.7

In Table 4.21 we give the corresponding f1-scores. The most difficult emotion to detect was afraid with an f1-score of 0.3. Sad segments could be detected best with an f1-score of 0.5. Happy and angry achieved an f1-score of 0.4 and 0.47, respectively.

Table 4.21: Precision, recall, and f1-scores for speaking rate.

	happy	sad	afraid	angry
precision	0.35	0.57	0.31	0.48
recall	0.47	0.45	0.28	0.47
f1	0.40	0.50	0.30	0.47

It is no so evident to display the relative order of emotion-specific phone durations since not all phones behaved consistently within a given emotion class. The relative order of emotion-specific vowel durations is given in Table 4.22. The additional number indicates the percentage of vowels for which the respective relation was true. The set of vowels comprised in total 17 vowels. As can be seen from the table, vowels in sad speech were consistently longer than in any other emotion. Following our operationalization of speaking rate as the inverse of phone duration, this means that the speaking rate of sad utterances was slower than the speaking rate of the remaining emotions. The speaking rate of happy speech was highest but close to the rate with which angry was produced. The speaking rate of an afraid speaker was lower than the rate of of an angry speaker. Also, note that all vowels in angry speech were longer than in sad speech, a situation reflected in the confusion matrix in which only 5% of the angry sentences were classified as sad.

Table 4.22: Relative order of phone durations for vowels. The additional number indicates the percentage of vowels for which the respective relation holds true for the respective emotion pair. For instance, 88.2 % of the sad vowel models had a larger mean duration than the respective happy models.

	happy	sad	afraid	angry
happy		<	<	<
sad	> (88.2%)		>	>
afraid	> (64.7%)	< (88.2%)		>
angry	> (52.8%)	< (100.0%)	< (82.3%)	

Context Sensitive Prosodic Phone Models

In the previous experiments we explored prosodic features pertaining to the whole utterance segment. We ignored the phonetic context from which these prosodic features arose. As mentioned in section 2.4.2, the phonetic context, however, has an impact on the prosodic appearance of the overall utterance. For instance, low vowels have a lower intrinsic fundamental frequency than high vowels. With the following experiments we explored whether emotion-specific information at the phonemic level led to an overall improvement in the detection of emotions. Similar to the modeling of speaking rate, we used prosodic phone models to model fundamental frequency features. See section 3.2.2 for a detailed description of the underlying modeling assumptions of the suprasegmental hidden Markov model.

In the first experiment, we trained emotion-specific context independent phone models relying on mean and variance information of the fundamental frequency. The corresponding confusion matrix is given in Table 4.23. The resulting system had the tendency to classify segments as afraid since half of the happy and half of the sad segments were missclassified as afraid.

Table 4.23: Confusion matrix based on context independent phone models relying on mean and variance information of the fundamental frequency. Overall, about 46% of the utterances were correctly classified.

	happy	sad	afraid	angry
HAPPY	21.7	5.0	15.7	5.0
SAD	5.0	18.3	0.0	5.0
AFRAID	51.7	51.7	65.0	10.0
ANGRY	21.8	25.0	20.0	80.0

The corresponding f1-scores are given in Table 4.24. Angry segments were recognized most

Table 4.24: Precision, recall, and f1-scores for context independent phone models.

	happy	sad	afraid	angry
precision	0.46	0.65	0.36	0.55
recall	0.22	0.18	0.65	0.80
f1	0.30	0.29	0.47	0.65

accurately with an f1-score of 0.65, followed by afraid (0.47), happy (0.30), and sad segments (0.29). Overall, 46.2% of the segments were classified correctly which was about 10% absolute worse than the accuracy we achieved with mean and variance information of the fundamental frequency of the entire utterance. Note that angry and afraid segments were also classified most accurately in that experiment. See Table 4.9 for details.

We think that one of the reasons for the different accuracies of prosodic utterance and phone models lay in context effects on phone models as mentioned above. In addition, note that the parameter estimation of phone models is based on an alignment of the speech signal with the model sequence and is, thus, more susceptible to estimation errors than global utterance models which do not require an alignment. In order to compensate for context effects we used the cluster

algorithm as described in section 3.2.3 to find appropriate context sensitive prosodic models. In this experiment we allowed the following questions about the preceding, the current, and the following phone:

1. the phone's identity
2. place of articulation (bilabial, alveolar, etc.)
3. manner of articulation (fricatives, nasals, liquids, etc.)

In the first step we used these question to cluster models based on their prosodic similarity using the cluster algorithm as described in section 3.2.3. We then trained emotion-specific clustered models in the second step and used them to test for the underlying emotion. The corresponding confusion matrix is given in Table 4.25. The system had a tendency to classify segments as afraid or angry. About 60% of the happy and about 65% of the sad segments were misclassified as either afraid or angry. As a consequence, the f1-scores for happy and sad segments were very

Table 4.25: Confusion matrix based on context dependent phone models relying on mean and variance of the fundamental frequency. Overall, about 49.6% of the utterances were correctly classified.

	happy	sad	afraid	angry
HAPPY	25.0	6.7	18.3	3.3
SAD	16.7	28.3	10.0	5.0
AFRAID	38.3	35.0	56.7	3.4
ANGRY	20.0	30.0	15.0	88.3

low and lay at 0.33. and 0.35, respectively. Angry segments were classified with an accuracy of 0.7, followed by afraid with an score of 0.49, see Table 4.26 for details. The context dependent models outperformed the context independent models in each emotion category and their overall accuracy of 49.6 lay about three percentage points above the context independent modeling approach.

Table 4.26: Precision, recall, and f1-scores for context dependent phone models.

	happy	sad	afraid	angry
precision	0.47	0.47	0.42	0.58
recall	0.25	0.28	0.57	0.88
f1	0.33	0.35	0.49	0.70

Note that we started the clustering with all phone models including noise and consonants models. As a consequence, the first question of the resulting regression tree was to distinguish phones from noise models. A little bit further down the tree, another question distinguished between vowels and consonants. Another interesting observation is that we achieved the best accuracies when we reduced the number of models by increasing the required number of observations for a model. This circumstance indicated that the limiting factor for an additional improvement of phone models was the amount of available training data.

4.2.6 Combining Prosodic Information

In the following experiment, we combined the prosodic features which we explored earlier in this chapter. That is, we had a vector comprising several prosodic observations:

1. four prosodic features characterizing the behavior of the fundamental frequency: mean, variance, and the two jitter features.
2. four prosodic features characterizing the behavior of intensity: mean and variance, and the two tremor features.

In order to use the speaking rate as a discriminate feature, we also modeled the duration of vowels context independently. See the respective experiments above for a detailed descriptions of these features.

Using these nine prosodic features, the system was able to classify 60.4% of the segments in the test set correctly. Note that this percentage of correctly classified segments was about 4% points higher than the best two individual prosodic features: mean and variance of the fundamental frequency. The corresponding confusion matrix is given in Table 4.27. We still found the common confusion of sad and afraid segments. For instance, a third of the sad segments were classified as afraid and 11.7% of the afraid segments as sad.

Table 4.27: Confusion matrix based on prosodic information. Overall, 60.4% of the utterances were classified correctly.

	happy	sad	afraid	angry
happy	55.0	6.7	21.7	6.7
sad	8.3	56.7	11.7	20.0
afraid	18.3	30.0	63.3	6.6
angry	18.3	6.6	3.3	66.7

When we look at the f1-score, given in Table 4.28, we can see that angry segments were recognized best with an f1-score of 0.68. Happy, sad, and afraid segments had an f1-score of 0.58. The combination of prosodic features helped in particular to classify happy, sad, and afraid segments. For these emotions the f1-score was about 8% points higher than the best subset of features studied previously in the corresponding experiments. Note, however, that the overall

Table 4.28: Precision, recall, and f1-scores for the combined prosodic information.

	happy	sad	afraid	angry
precision	0.61	0.59	0.54	0.70
recall	0.55	0.57	0.63	0.67
f1	0.58	0.58	0.58	0.68

accuracy still lay below the performance of humans who were able to classify about 70% of the segments correctly.

Comparing classification accuracies based on prosodic information with accuracies based on spectral information, we see that classification based on spectral information outperformed prosodic information by about 10%. Consult Table 4.7 for the respective experiment involving spectral information. The next section explored the combination of prosodic and spectral information

4.2.7 Combining Spectral and Prosodic Information

In the following experiment, we combined spectral and prosodic information. We used the same nine prosodic features as explored in the previous experiment: mean and variance of the fundamental frequency, two jitter features, mean and variance of the intensity, two tremor features, and speaking rate. Consult the respective experiments above for a more detailed descriptions of these features. Emotion-specific spectral information was captured by adaptation of spectral models. See section 4.2.4 for a more detailed description. The experiment in that section also demonstrated that spectral information and adaptation were a powerful technique to classify emotional speech segments. Remember, using spectral adaptation we classified about 69% of the segments correctly which was about an absolute of 10% better than the classification based on prosodic information. For the current experiment, we used weights determined independently on the development set for the linear combination of spectral and prosodic probabilities. For details see section 4.1.6.

The confusion matrix for the linear combination or spectral and prosodic information is given in Table 4.29. Note that overall, 69.2% of the segments were classified correctly. This accuracy was basically the same accuracy we achieved with spectral information in the first place. Thus, the combination of prosodic and spectral information did not result in an overall improvement.

Table 4.29: Confusion matrix based on the combination of spectral and prosodic information. Overall 69.2% of the utterances were classified correctly.

	happy	sad	afraid	angry
happy	66.7	1.7	15.0	1.7
sad	0.3	56.7	16.7	10.0
afraid	16.7	36.6	66.7	1.6
angry	13.3	5.0	1.6	86.0

Let us compare the current f1-scores, given Table 4.30, with the scores we were able to achieve with spectral information alone, see Table 4.7 in section 4.2.4. We can see that only the f1-score for afraid and happy segments improved from 0.57 to 0.6 and from 0.71 to 0.72, respectively.

Table 4.30: Precision, recall, and f1-scores for the combination of spectral and prosodic information.

	happy	sad	afraid	angry
precision	0.78	0.65	0.55	0.81
recall	0.67	0.57	0.67	0.87
f1	0.72	0.61	0.60	0.84

We list the accuracies for each speaker in Table 4.31 depending on whether classification was based on prosodic or spectral information. An additional row gives the accuracies of the combination of prosodic and spectral information. In general, the classification accuracies largely varied depending upon the speaker. The accuracies for individual speakers ranged from 50% to 68% when relying on prosodic information. For spectral information, the accuracy ranged from

60% to 75%. The main effect of the combination of prosodic and spectral information seemed to

Table 4.31: Speaker-specific classification accuracies based on prosodic and spectral information and their combination.

	A	B	C	D
prosodic	66.6	50.0	68.3	56.7
spectral	60.0	71.7	68.3	75.0
combined	66.1	65.0	78.3	66.7

be that the speaker-specific accuracies became more similar to each other. With the exception of speaker C whose utterances were classified with an accuracy of 78.3%, the remaining accuracy levels lay all at around 66%.

The speaker-specific accuracies of human listeners did not correlate with the accuracies of the classification system. For instance, the utterances of speaker D were classified most accurately by human subjects (75%). The classification by the system achieved only an average accuracy of 66.7%. The utterances of speaker C, in contrast, were classified most accurately by the system (78.3%) but human subjects were only able to achieve an average accuracy of 71.3%. See Table 4.4 for the respective accuracies achieved by human subjects.

In this experiment we linearly combined the probabilities of prosodic and spectral information and chose this overall score to be indicative of the expressed emotion in some utterance. As mentioned above, we determined the weights in this linear combination on an independent development set. In order to assess an upper bound for a combination of prosodic and spectral information we used an oracle which told us when to choose prosodic information and when to choose spectral information to classify a given utterance. The confusion matrix of this oracle experiment is given in Table 4.32. Note that there was still considerable confusion between sad and afraid: 11.7% of the afraid segments were misclassified as sad and 16.7% of the sad segments were classified as afraid. Note that the oracle could perfectly distinguish between afraid and angry segments. Overall, 86.7% of the segments were correctly classified. The corresponding f1-scores

Table 4.32: Confusion matrix based on the combination of spectral and prosodic information using an oracle. Overall 86.7% of the utterances were classified correctly.

	happy	sad	afraid	angry
happy	88.4	5.0	3.3	1.6
sad	3.3	78.3	11.7	3.3
afraid	5.0	16.7	85.0	0.0
angry	3.3	0.0	0.0	95.0

are given in Table 4.33. All emotion classes profited from the oracle. Angry was still classified best with an f1-score of 0.96, followed by sad with an f1-score of 0.89. Afraid and sad were detected with an accuracy of 0.82 and 0.8, respectively. When we compare the oracle with the previous linear combination, we see that the oracle improved the overall accuracy from 69.2% to 87.6%. The linear combination seemed to indicate that prosodic and spectral information were not orthogonal in the Woggles corpus since their combination did not result in an overall improvement. However, this finding was not confirmed by the experiments involving an oracle. Spectral and prosodic information did yield different predictions about the emotion expressed in

Table 4.33: Precision, recall, and f1-scores for the combination of spectral and prosodic information using an oracle.

	happy	sad	afraid	angry
precision	0.90	0.81	0.80	0.97
recall	0.88	0.78	0.85	0.95
f1	0.89	0.80	0.82	0.96

an utterance and knowing when to choose spectral or prosodic information could improve the overall accuracy significantly.

4.2.8 Speaker Independence

In the previous experiments, we were able to classify the emotions in the Woggles corpus with an accuracy comparable to humans subjects. Note, however, that we had a multi-speaker system; that is, we trained and tested on utterances of the same speakers. In the following experiment we investigated next whether the classification accuracies degraded when tested on two speakers who where not in the training set. We then tested spectral and prosodic information separately to see whether the respective models reacted differently on utterances of unseen speakers. In addition, we also examined the linear combination of spectral and prosodic information. For the following experiments we used the same sentences as in the previous test set uttered by two novel speakers. Thus, the current test set comprised 120 utterances.

The confusion matrices are given in Table 4.34. Overall, a system based on spectral information was able to obtain an accuracy of 47%. A system using prosodic models outperformed spectral information and classified 50% of the segments correctly. We found the usual confusion between sad and afraid utterances in both cases. For spectral information, 63.3% of the afraid segments were misclassified as sad and 20% of the sad segments as afraid. On the other hand, this trend was less pronounced for prosodic information: 36.7% of the sad segments were classified as afraid and 10% of the afraid segments were misclassified as sad. Spectral information tended to classify segments as sad whereas prosodic information was inclined to classify segments as happy. Note that neither sad nor afraid segments were classified as angry.

Table 4.34: Confusion matrices for spectral and prosodic information tested on novel two speakers. Overall, spectral information achieved an accuracy of 47% whereas prosodic information obtained an accuray of 50%.

	happy	sad	afraid	angry
HAPPY	60.0	10.0	3.3	10.0
SAD	27.7	63.3	63.3	46.7
AFRAID	6.7	20.0	26.7	6.6
ANGRY	6.6	6.7	6.7	36.7

(b) Spectral Information

	happy	sad	afraid	angry
HAPPY	66.7	16.7	30.0	26.7
SAD	3.3	46.7	10.0	36.6
AFRAID	13.3	36.7	60.0	10.0
ANGRY	16.7	0.0	0.0	26.7

(b) Prosodic Information

The f1-scores for spectral and prosodic information are given in Table 4.35. Happy segments are classified most accurately for both spectral and prosodic information (0.65 and 0.56). The accuracy of happy segments based on prosodic information was very close to the accuracy for afraid segments (0.55). Classification based on spectral information achieved the worst accuracy for afraid (0.33) and classification based on prosodic for angry (0.37). Sad segments could be detected with comparable accuracies both by spectral and prosodic information (0.42 and 0.47).

Table 4.35: Precision, recall, and f1-scores for spectral and prosodic information tested on two novel speakers.

	happy	sad	afraid	angry
precision	0.72	0.32	0.44	0.65
recall	0.60	0.63	0.27	0.37
f1	0.65	0.42	0.33	0.47

(b) Spectral Information

	happy	sad	afraid	angry
precision	0.48	0.48	0.50	0.62
recall	0.67	0.47	0.60	0.27
f1	0.56	0.47	0.55	0.37

(b) Prosodic Information

Regarding prosodic information, this experiment showed that, in particular, sad and afraid segments became more difficult to detect when tested on utterances of novel speakers. See Table 4.28 for the f1-scores of prosodic information evaluated on the original test set. Finally, note that the accuracy dropped from 60% to 50%. For spectral information, the drop was more dramatic. The accuracy fell from 69% to 46% when we switched to the current corpus. The detection accuracies for all four emotions were effected strongly. See Table 4.7 for a comparison.

A classification based on the linear combination of the spectral and prosodic probabilities led to an overall improvement for the current corpus. We used the same weights for the combination as in the previous experiments. The resulting confusion matrix is given in Table 4.36. Overall, 54.2% of the segments were correctly classified. We find the usual confusion between sad and afraid: 36.6% of the sad and 26.6% of the afraid segments are either misclassified as sad or afraid. None of the sad and none of the afraid segments were misclassified as angry.

Table 4.36: Confusion matrix for the combination of prosodic information tested on two novel speakers.Overall, 54.2% of the segments were correctly classified.

	happy	sad	afraid	angry
HAPPY	73.3	6.7	16.7	23.3
SAD	3.3	56.7	26.6	33.3
AFRAID	10.0	36.6	56.7	13.4
ANGRY	13.4	0.0	0.0	30.0

The corresponding f1-scores are given in Table 4.37. The overall accuracy increased from 50% – achieved by prosodic information – to 54% by the combination of prosodic and spectral information. In particular, the emotions happy, sad, and afraid profited from the combination. For instance, the f1-score for afraid rose from 0.37 to 0.52. The f1-score for angry dropped from 0.55 to 0.42.

To summarize, prosodic information was more reliable when testing for speaker independence.

Table 4.37: Precision, recall, and f1-scores for the combination spectral and prosodic information tested on two novel speakers.

	happy	sad	afraid	angry
precision	0.61	0.47	0.49	0.69
recall	0.73	0.57	0.57	0.30
f1	0.67	0.52	0.52	0.42

The drop in accuracy was less pronounced when compared to the drop of the accuracy for spectral information. In addition, classification based on prosodic information performed better than classification based on spectral information. Keep in mind that this was opposite for the first test set in which spectral information outperformed prosodic information. A possible explanation for the brittleness of spectral information when confronted with novel speakers was that the adaptation of spectral models captured very specific idiosyncrasies of the data in the training corpora. Whereas these idiosyncrasies interpolated onto the first test set and led to high accuracies comparable to human performance, these idiosyncrasies did not transfer as well to novel speakers and resulted in a large drop below the accuracies we achieved with prosodic information.

4.2.9 Summary

This section summarizes the results of the experiments carried out with the Woggles corpus. Table 4.38 shows the results of the most important experiments. The best prosodic features

Table 4.38: Overview

No.	Features	Segments	Signal Postprocessing	Test Set	f1-score
1	spectral	utterance		Woggles (1)	0.69
2	F_0, mean/variance	utterance	median smoothing speaker normalization	Woggles (1)	0.56
3	F_0, jitter	utterance		Woggles (1)	0.40
4	Intensity, mean/variance	utterance	voiced segments	Woggles (1)	0.33
5	Intensity, tremor	utterance	voiced segments log	Woggles (1)	0.45
6	Duration	phones		Woggles (1)	0.42
7	2,3,4,5,6			Woggles (1)	0.60
8	1,7			Woggles (1)	0.69
9	1			Woggles (2)	0.47
10	7			Woggles (2)	0.50
11	1,7			Woggles (2)	0.54
12	human			Woggles (1)	0.69
13	1,7 (oracle)			Woggles (1)	0.87

were mean and variance of the fundamental frequency (2) which allowed to classify 56% of the segments correctly. The worst classification results were obtained when we relied on mean and variance of the intensity (4). Only a third of the segments could be classified correctly. Jitter (3), tremor (5), and speaking rate (6) allowed f1-scores of about 0.4. The combination of all nine prosodic features (8) resulted in an f1-score of 0.6 which was 4% absolute better than the f1-score achieved with the mean and variance of the fundamental frequency.

Classification based on a spectral representation of the speech signal achieved an f1-score of 0.69 (1). However, the combination of spectral and prosodic information did not produce an additional improvement of the overall accuracy. Spectral information alone allowed a classification accuracy comparable to human subjects (12). However, knowing when to choose spectral and when to choose prosodic information to detect the emotion expressed in some utterance, an overall accuracy of 0.87 was achieved in the experiment relying on an oracle (13). This finding indicated that spectral and prosodic information yielded to some extend independent information with regard to the expressed emotion in the utterances from the Woggles corpus.

In addition, other experiments were conducted to test spectral and prosodic models on utterances of two novel speakers, that is, speakers who were not in the training set. In this situation, the classification accuracy based on prosodic information fell from 0.6 to 0.5 (10). The accuracy for spectral information fell from 0.7 to 0.46. These results suggested that prosodic models transfered reasonably to novel speakers whereas spectral models focussing too closely on speakers in the training set, did not properly transfer to the utterances of the two novel speakers. This tight focus on the speaker in the training set also explained the high accuracy of 0.69 on the first test

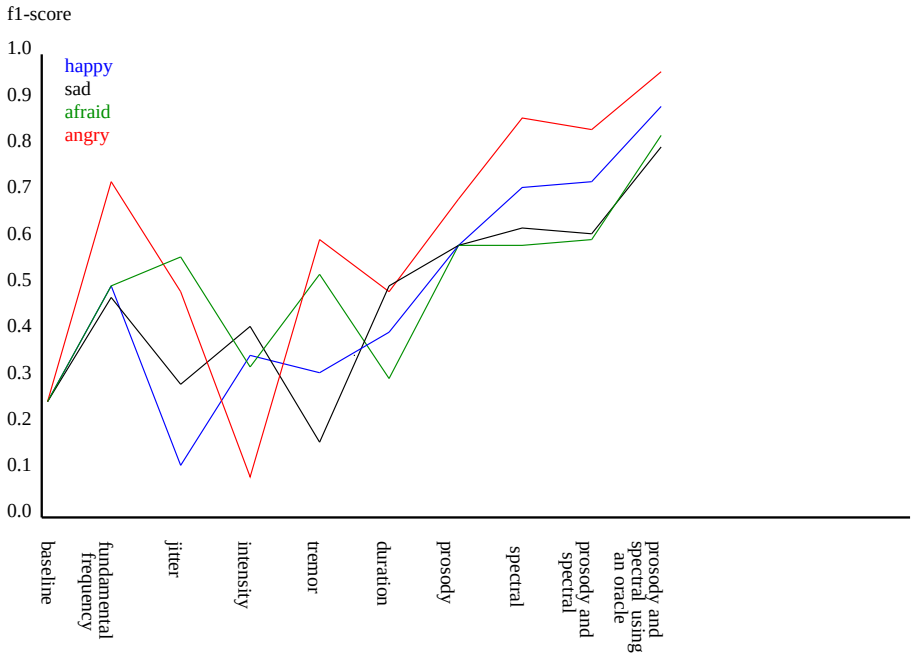


Figure 4.4: Plot of f1-score for prosodic features, the combination of all prosodic features, and the combination of prosodic and spectral information.

set (1).

We also investigated whether we could improve accuracies by modeling other features than the duration feature on the phone model. For this reason we explored context sensitive phone models relying on mean and variance information of the fundamental frequency. While context sensitive models outperformed context dependent models (49.6% and 46.3%), they could still not achieve the accuracy we achieved with this kind of information at the utterance level (55.8%).

In figure 4.4, we plotted the f1-scores of some of the prosodic features investigated in the previous experiments. There are two points we wish to mention. First, note that angry segments were far more detectable than the remaining emotions for all prosodic features with the exception of the mean and variance of intensity. Second, the f1-scores exhibited large oscillations both for prosodic features and emotions. That is, whereas some prosodic features allowed for strong classification of certain emotions, it failed for other emotions. However, if we combined all prosodic features, the f1-scores of the four emotions became very similar. In addition, the graph above also shows that the combination of prosodic and spectral information did not result in an overall improvement of the f1-score. Finally, the graph shows that there is still room for an improvement of the overall accuracies, indicated by the accuracies achieved by the experiment involving an oracle to predict when to choose prosodic and when to choose spectral information.

4.3 Talk Shows and Movies

The second major corpus of this investigation comprised segments from movies and talk shows. The decision to use talk shows and movie screen plays as sources of emotional speech was motivated by three reasons. First, the experiments relying on the Woggles corpus showed that acted speech could be decoded by humans quite reliably. Second, to accurately estimate statistical models, an extensive quantity of training data is needed. By using a films close captions as a first approximation for segmentation, transcription, and tagging, we were able to collect a large supply of emotional speech samples in a relatively short amount of time. Third, even though we did not pursue visual cues in this investigation, this corpus allows the integration of visual information with spectral, prosodic, and verbal information in future work.

4.3.1 The Corpus

The movies and talk shows were down loaded from a Toshiba VCR (M-752) to a Pentium II personal computer equipped with a Crystal Audio System with a sampling rate of 16kHz and 16 bits. The close captions from the video stream were extracted with the Text Grabber VBI Line 21 Video Decoder (GP-500). For the segmentation, transcribing, and tagging we employed five students who were instructed in several training sessions.

Three major steps were involved in tagging these talk shows and movies:

1. **Segmentation**: Transcribers were told to find segmentations which coincided with sentence or utterance boundaries. The expressed emotion within a segment had to be constant. Initial or final noises or silences were to be excluded from the segment.
2. **Transcriptions**: The close captions were the starting point for the transcriptions. Missing noises (human and non-human) or words were added by the transcribers.
3. **Tagging**: Each segment was annotated with three tags. The first tag indicated the gender of the speaker and the second tag the amount of background noise. For the last tag, the emotion expressed by the speaker, we told the transcribers to be as specific as possible and to choose from the emotion tags as given in Table 4.39.

For the tagging process the transcribers relied on either CoolEdit 96 (Syntrillium) or Sound Forge 4.0 (Sonic Foundry) and Sennheiser headsets.

The distribution of emotions within this corpus is given in Table 4.40 in which we considered all segments regardless of the amount of background noise or music. The distribution of emotions changed substantially when we considered only segments suited for acoustic modeling. The exact numbers are given in Table 4.41.

Only utterances with a moderate noise level were used for training and testing spectral and prosodic models. The noise level was ignored for training emotion-specific language models. In order to guarantee a sufficient number of training tokens for the estimation of verbal, spectral, and prosodic models, the experiments were confined to the emotions angry, sad, and neutral. We divided the corpus into three parts: training sets, development set, and two test sets. The

Table 4.39: The list of the ten emotion tags used by the transcribers.

1. neutral (neu): neutral, no noticeable emotion
2. bored (bor) : boredom, disinterest
3. strong joy (stj): happy, exited, laughter, delight
4. weak joy (wkj): slightly happy, liking, love, satisfied, admiration, content
5. sad (sad): sadness,sorrow, depression, remorse, shame, disappointment
6. afraid (afr): upset, worried, fear
7. irony (iro): irony, sarcasm, mockery,
8. angry (ang): anger, reproach, threat, scolding
9. disgust (dis): offense, resentment, disliking, indignation, contempt
10. suspicion (sus): disbelief, doubt, uncertainty, incredulous

Table 4.40: Distribution of speech segments according to their emotion.

tag	neutral	angry	sad	afraid	disgusted	ironic	happy	surprised
# segments	2991	1586	1076	203	26	28	347	19

Table 4.41: Distribution of speech segments according to their emotion suited for training spectral and prosodic models.

tag	neutral	angry	sad	afraid	disgusted	ironic	happy	surprised
#	1344	759	518	4	0	3	19	4

first test set comprised speech segments randomly chosen from several movies. For each emotion category we chose 102 test sentences. The second test set consisted of all segments from the movie "One True Thing". The development set consisted of segments from the movie "Primary Colors" and "Alien".

Since in the first test set the emotions were distributed evenly, the baseline accuracy lay at 33% correctly classified segments which could be achieved by always guessing the same emotion. Extrapolating from the movies within the training set, we could assume that neutral segments occurred more often than angry or sad segments in the second test set. In the training set, about 51% of the segments were neutral. Transferring this ratio to the second test set, the baseline accuracy increased to about 51% by guessing the emotion always to be neutral. Thus, in order to claim that our system achieved a reasonable performance, the system had to obtain an accuracy of at least 51%. The actual distribution of emotion segments is given below in Table 4.42 (a) for the subset suitable for testing acoustic models and (b) for the whole movie.

Table 4.42: Distribution of segments in the second test set from the movie "One True Thing".

	sad	angry	neutral
# segments	67	22	80

(a) suitable for acoustic testing

	sad	angry	neutral
# segments	108	47	174

(b) the whole movie

4.3.2 Human Performance (Intercoder Agreement)

Whereas for the Woggles corpus we could stipulate that we knew the underlying emotion of the utterances, we had no such *a priori* knowledge about the underlying emotions for speech segments within the movie corpus. Thus, assessing the agreement of the transcribers became essential to validate the quality of the corpus and to obtain a performance number for comparing the performance of the system. For the following experiments, we assumed that the original emotion tag given by the transcriber represented the emotion expressed by the actor.

Acoustic and Verbal Information

In the first experiment, we asked four transcribers to classify the utterances of the first test as either angry, sad, or neutral. We used the interface as described in section 4.1.2. The test corpus was randomized, thus no context was given for a particular utterance. The confusion matrix of the four coders is given in Table 4.43. As is apparent from the confusion matrix, transcribers

Table 4.43: Confusion matrix of four coders classifying 306 utterances from the first test set. Overall, about 70% of the segments were classified correctly.

	sad	angry	neutral
SAD	46.6	4.5	5.2
ANGRY	5.7	72.3	3.6
NEUTRAL	47.7	23.2	91.2

experienced problems differentiating neutral and sad segments: 47.7% of the sad sentences were classified as neutral. Subjects had the most problems detecting sad sentences, only 46% of the sad sentences were classified as sad. There was also substantial confusion between angry and neutral since 23.2% of the angry segments were misclassified as neutral.

Table 4.44: Precision, recall, and f1-scores for human subjects classifying the 306 speech segments in the first test set.

	sad	angry	neutral
precision	0.83	0.89	0.56
recall	0.47	0.72	0.91
f1	0.60	0.80	0.70

Table 4.44 presents precision, recall, and f1-scores for all emotions. The f1-score was best for angry segments with 0.8, followed by neutral with an f1-score of 0.70. Sad segments, as already suggested at the discussion of the confusion matrix, were identified with the f1-score of only 0.60.

The accuracy for each of the four coders ranged from 65% to 78%. The overall accuracy was 70%. Note that the baseline for this experiment was 33% achieved by guessing always the same emotion. An overall accuracy of 70% indicated that the emotions expressed in this corpus were detected very well. Remember that the human accuracy on the Woggles corpus was also about 70%. See section 4.2.2 for details.

Acoustic, Verbal and Context Information

With the second experiment we explored the impact of contextual information on the accuracy of emotion detection. Two transcribers listened to all speech segments of the movie "One True Thing" in their natural order. For utterances tagged originally as sad, angry, or neutral, the coders were asked to classify them as either sad, angry, or neutral using the interface as described in section 4.1.2 above. The other two transcribers had no contextual aids since their test set comprised only sad, angry, and neutral utterances in random order of the same movie.

Table 4.45 shows the confusion matrix of the two coders who had to classify the randomized subset of the movie. Similar to the experiment above, neutral seemed to be the default assumption of the coders; 63.6% of the sad and 28.9% of the angry sentences were misclassified as neutral. Sad segments were most difficult to detect, only 32.5% of the sad segments were classified as sad. Overall, 63.4% of the segments were classified correctly which was about 6% lower than for the first test set.

Table 4.45: Confusion matrix of two coders classifying sad, angry, or neutral segments in the movie "One True Thing". Only sad, angry, and neutral segments were played in random order. Overall, about 63.4% of the segments were classified correctly.

	sad	angry	neutral
SAD	32.5	11.6	3.2
ANGRY	3.9	59.4	10.9
NEUTRAL	63.6	28.9	85.9

Table 4.46: Precision, recall, and f1-scores for human subjects who were given no context information.

	sad	angry	neutral
precision	0.78	0.54	0.62
recall	0.32	0.59	0.86
f1	0.46	0.57	0.72

In the second part of this experiment, two different transcribers listened to all speech segments of the movie "One True Thing" in their proper order. For segments previously tagged as sad, angry, or neutral, the subjects were asked to classify them again as either sad, angry, or neutral.

The confusion matrix for the two transcribers is given in Table 4.47. Neutral seemed to be the default assumption of these transcribers as well. 38.1% of the sad and 17.4% of the angry sentence were classified as neutral. Sad sentences were the most difficult to detect, only 57% of the sad sentences were classified as sad. The overall accuracy lay with 72.8% by more than 10% absolute higher than in the previous experiment suggesting that context information did help in this kind of classification task. We expect this trend to be even more obvious for the detection of more subtle emotions such as afraid. Comparing the f1-scores of the second test set with and without context information, we see that context helped to recover, in particular, sad and angry.

Table 4.47: Confusion matrix of two coders classifying sad, angry, or neutral segments in the movie "One true Thing". All speech segments of the movie were played in the proper order. Overall, about 73% of the segments were classified correctly.

	sad	angry	neutral
SAD	57.4	2.9	12.1
ANGRY	4.5	79.7	6.3
NEUTRAL	38.1	17.4	81.6

Table 4.48: Precision, recall, and f1-scores for human subjects who were given context information.

	sad	angry	neutral
precision	0.75	0.70	0.72
recall	0.57	0.80	0.82
f1	0.65	0.74	0.77

Having context information increased the f1-score for sad from 0.46 to 0.65 and for angry from 0.57 to 0.74.

Note, however, that in both experiments above, we compared the classifications of the human subjects to the previously tagged corpus. Tagging the corpus for emotions, the transcribers presumably used context information for a given segment. Thus, the second experiment mimicked this process much closer than the first experiment. The higher accuracy in the second experiment may be partially attributed to this fact. Moreover, we based our results on a total of four coders. As mentioned before, the primary objective of these experiments involving human listeners was not to investigate how humans decode emotional cues but to validate the emotional speech corpora itself and to establish a performance number for assessing the quality of the following classification results.

Verbal Information

While it is not evident how to delete verbal information in speech segments, deleting spectral and prosodic information can be implemented by simply presenting in text form what was said in a given utterance. Thus we can test, how well humans can distinguish among emotions when basing their judgments solely on verbal information.

For the following experiment we asked five subjects to classify all 306 segments from the first test set as either sad, angry or neutral. The confusion matrix of this experiment is given in Table in 4.49. The human subjects seemed to use neutral as a default case since 55.7% of the sad and 31.7% of the angry segments were classified as neutral. Only 30.4% of the sad segments were classified as sad. The subjects' classification accuracies were fairly consistent and ranged from 54-59%. Note that for the first test set the base line was at 33% achieved by voting for the same emotion all the time. Overall, 55.7% of the segments were classified correctly. Compare this performance number to the performance which was achieved if the subjects had the audio information available which was about 70%. Thus, having the audio version instead of only the

Table 4.49: Confusion matrix for humans relying only on a textual representation of the segments in the first training set. Overall, about 55.7% of the segments were classified correctly.

	sad	angry	neutral
SAD	30.4	7.4	9.9
ANGRY	13.9	60.9	14.3
NEUTRAL	55.7	31.7	75.8

textual presentation, humans could improve their accuracy substantially, in this case by about a 15% absolute.

It is also illustrative to examine the precision and recall numbers, in Table 4.50 below, because they show that angry segments were detected actually best with an f1-score of 0.65, followed by neutral segments with an f1-score of 0.54 and sad segments with an f1-score of 0.43.

Table 4.50: Precision, recall, and f1-scores for human subjects classifying speech segments based on a textual representation.

	sad	angry	neutral
precision	0.53	0.61	0.50
recall	0.36	0.69	0.59
f1	0.43	0.65	0.54

With the previous experiments we tried to validate the current corpus. Humans were able to decode emotional cues given in the utterance of this corpus well above chance level. For the first test set, humans achieved an accuracy of about 70% when they could listen to the segments. If only the verbal information was given, that is, the textual representation of what was said in the utterance, the accuracy dropped to about 57% which was still well above the chance level (33%).

We established similar results for the second test set which comprised segments from the movie "One True Thing". Human listeners achieved an accuracy of about 63% when asked to classify a randomized subset of movie segments. Further, when the segments occurred in their natural context and order, the overall accuracy increased to 72.9%. This suggests that additional context information allowed a more accurate classification.

4.3.3 Emotion-Specific Spectral Information

In the first experiment we modeled emotion-specific spectral information using adaptation of spectral models. For a detailed description of this adaptation procedure see sections 3.3 and 4.1.3. Training of emotion-specific spectral models comprised two steps. Starting with the same recognition system as used in the Woggles experiments – see section 4.2.3 – we adapted spectral models in the first step on all training data regardless of the underlying emotion of the respective training samples. With this step we tuned the spectral models to acoustic properties of the speech samples in the current corpus. In the second step, we used the same procedure but adapted spectral models separately on emotion-specific subsets of the speech database using the models trained in the previous step as a starting point. Thus, we obtained three sets of emotion-specific spectral models which we used for the classification in the test phase.

The confusion matrices for both test sets of a classification system based on emotion-specific spectral models are given below in Table 4.51. In 4.51 (a) we give the matrix for the first test

Table 4.51: Confusion matrix using spectral information. Overall 63.9% and 57.1% of the segments in the first and second test set were classified correctly.

	sad	angry	neutral
SAD	52.0	4.0	16.7
ANGRY	8.0	72.0	15.6
NEUTRAL	40.0	24.0	67.7

(a) Test set 1

	sad	angry	neutral
SAD	46.3	4.5	32.9
ANGRY	13.4	77.3	6.3
NEUTRAL	40.3	18.2	60.8

(b) Test set 2

set. The system had – similar to the human subjects – a preference for neutral: 40.0% of the sad and 24.0% of the angry segments were classified as neutral. Overall, 63.9% of the segments were classified correctly.

The confusion matrix for the second test set is given in Table 4.51 (b). Overall, the confusion was similar to the confusion we found with the first test set. The default assumption of the system was also that a segment was neutral: 40.3% of the sad and 18.2% of the angry segments were classified as neutral. Sad and angry segments were distinguished fairly well in both test sets. The confusion between sad and angry segments lay at about 4% in the range of human subjects. Remember also, that human subjects had the tendency to classify segments as neutral, see Tables 4.43, 4.45, and 4.47 for the details. Overall, 57.1% of the segments were classified correctly.

Table 4.52: Precision, recall, and f1-scores for spectral information.

	sad	angry	neutral
precision	0.72	0.76	0.50
recall	0.52	0.72	0.68
f1	0.6	0.74	0.58

(a) Test set 1

	sad	angry	neutral
precision	0.53	0.55	0.61
recall	0.46	0.77	0.61
f1	0.50	0.64	0.61

(b) Test set 2

It is also illustrative to compare the f1-scores of both test sets. First note that the overall f1-score was higher for the first test set (0.63 vs. 0.57). This was not surprising since the segments in the first test set were randomly drawn from movies which were also in the training set. However, this was not the case for the second test set. Angry segments were detected best in both test sets (0.74 and 0.64). Neutral segments achieved an accuracy of about 0.6 for both test sets. Sad segments were classified with an accuracy of 0.6 and 0.5 in the first and second test set, respectively.

Emotion-specific differences in the detection accuracy were also found in our experiments with humans subjects, see Tables 4.44, 4.46, and 4.48 in section 4.3.2. The relationships were identical, that is, angry segments were in general detected best, followed by neutral. Sad segments were detected with the most difficulties, both by the system and by human subjects.

Even though the performance of the system using emotion-specific spectral models lay lower than the respective performance of human subjects, it lay well above chance level. Also note, that humans subjects benefited both from prosodic and verbal cues, cues which we investigated further with the following experiments.

4.3.4 Emotion-Specific Prosodic Information

We investigated several prosodic features and their potential to distinguish among sad, angry, and neutral speech segments in the following experiments.

For each prosodic feature we also showed how its emotion-specific value compared to the values of the remaining emotions. In addition, we compared emotion-specific values with the corresponding values gained in the previous Woggles corpus to check for consistency across corpora.

Fundamental Frequency (Mean and Variance)

Before computing the mean and variance of the fundamental frequency in an utterance, we normalized regarding to the gender of the speaker. Other normalization techniques as discussed in section 4.1.4, turned out to be less accurate. Normalization with respect to the speaker's gender was implemented by using minimum and maximum values to compensate for a non-uniform distribution of emotional speech between men and women. After the removal of spikes, the fundamental frequency was median smoothed. 4.53 (a) gives the confusion matrix of a classification

Table 4.53: Confusion matrix based on the mean and variance of the fundamental frequency. Overall 49.3% and 41.6% of the segments in the first and second test set were classified correctly.

	sad	angry	neutral
SAD	19.8	15.0	11.3
ANGRY	33.7	53.0	12.4
NEUTRAL	46.5	32.0	76.3

(a) Test set 1

	sad	angry	neutral
SAD	12.1	4.5	2.6
ANGRY	50.0	90.9	44.8
NEUTRAL	37.9	4.6	52.6

(b) Test set 2

system, tested on the first test set that used the mean and variance of the fundamental frequency within an utterance as input features. The tendency to classify segments as neutral was also the case in this experiment. 46.5% of the sad and 32.0% of the angry segments were classified as neutral. Only 19.8% of the sad segments were classified as sad. 53% of the angry and 76.3% of the neutral segments were classified correctly. Overall, 49.3% of the segments were classified correctly. 4.59 (b) gives the confusion matrix for the second test set. Segments were classified primarily as angry: half of the sad and 44.8% of the neutral segments were classified incorrectly as angry. Overall, 41.7% of the segments were classified correctly.

The f1-scores are given in Table 4.54. We can see that sad segments were the most difficult to detect in both test sets (0.27 and 0.21). Neutral segments were detected most accurately in both test sets (0.59 and 0.57). Angry segments were detected in the first test set (0.53) better than in the second test set (0.36). Overall, emotional segments were classified more accurately in the first test set.

Table 4.55 gives the relative order of the emotion-specific values of mean and variance of the fundamental frequency. Here, the values for sad segments occupy the extreme positions in these tables. Also, for both mean and variance, sad segments had the largest values while neutral

Table 4.54: Recall, precision, and f1-score for mean and variance of the fundamental frequency.

	sad	angry	neutral
precision	0.43	0.54	0.48
recall	0.20	0.53	0.76
f1	0.27	0.53	0.59

(a) Test set 1

	sad	angry	neutral
precision	0.73	0.23	0.61
recall	0.12	0.91	0.53
f1	0.21	0.36	0.57

(b) Test set 2

segments had the lowest. For mean and variance of the fundamental frequency, the order was

Table 4.55: Relative order of prosodic features.

	sad	angry	neutral
sad		>	>
angry	<		>
neutral	<	<	

(a) utterance F_0 mean

	sad	angry	neutral
sad		>	>
angry	<		>
neutral	<	<	

(b) utterance F_0 variance

identical to the order found within the Woggles corpus, that is sad segments had a higher mean and a larger variance than angry segments, see Table 4.10 for a comparison.

Fundamental Frequency (Jitter)

In the following experiment, we tried to model some of the dynamics of the fundamental frequency. We moved a window over the fundamental frequency to compute the following features:

1. the number of changes from a positive to a negative slope (or vice versa) normalized by the total number of windows and
2. the sum of all χ^2 of all regression slopes normalized by the number of windows.

Note that for the computation of these two feature, the fundamental frequency was not smoothed or normalized. We trained emotion-specific models using these two features on the respective training sets.

The confusion matrices for the first and second test set are given in Table 4.56. For the first test set, the system had the tendency to classify segments as neutral. For instance, 36.6% of the sad and 43.0% of the angry segments were classified as neutral. A different picture emerged, however, from the second test set. Here the system preferred to classify segments as sad. 59.1% of the angry and 39.7% of the neutral segments were classified as sad. It appeared that the roles of sad and angry were reversed in the first and second test set. But overall, 47.7% of the segments in the first and 46.4% of the segments in the second test set were classified correctly.

Table 4.56: Confusion matrix based on intensity features. Overall 47.7% and 46.4% of the segments in the first and second test set are classified correctly.

	sad	angry	neutral
SAD	44.6	28.0	12.4
ANGRY	18.8	29.0	17.5
NEUTRAL	36.6	43.0	70.1

(a) Test set 1

	sad	angry	neutral
SAD	60.6	59.1	39.7
ANGRY	28.8	31.8	21.8
NEUTRAL	10.6	9.1	38.5

(b) Test set 2

However, if we look at the corresponding f1-scores, given in Table 4.57, we see a more consistent picture. In both test sets, the f1-scores for sad and neutral were very similar. Neutral segments had in both sets the highest f1-scores (0.56 and 0.51), followed closely by the f1-scores for sad segments (0.48 and 0.53). Angry segments were detected with more difficulty. The corresponding f1-scores were 0.35 and 0.22.

Table 4.57: Recall, precision, and f1-scores for jitter features.

	sad	angry	neutral
precision	0.53	0.45	0.46
recall	0.45	0.29	0.70
f1	0.48	0.35	0.56

(a) Test set 1

	sad	angry	neutral
precision	0.48	0.16	0.77
recall	0.61	0.32	0.38
f1	0.53	0.22	0.51

(b) Test set 2

The explanation for the high f1-scores for sad and neutral segments and the low score for angry segments could be found when we look at the relative positions of the emotion-specific values of these two features which are given in Table 4.58. For both features, the sad values were larger than the corresponding values for angry and neutral. The neutral values always remained smaller. Finally, note that the relative positions of the sad and angry values to each other were

Table 4.58: Relative order of prosodic features.

	sad	angry	neutral
sad		>	>
angry	<		>
neutral	<	<	

(a) normalized number of changes

	sad	angry	neutral
sad		>	>
angry	<		>
neutral	<	<	

(b) normalized χ^2

identical to the positions of these features in the Woggles corpus, given in Table 4.13 in section 4.2.5.

Intensity (Mean and Variance)

Intensity had not been a strong cue in the experiments using the Woggles corpus. See section 4.2.5 for the corresponding experiments. We speculated that this was the case because of the recording conditions for the Woggles corpus. The recording conditions prohibited very dynamical utterances, thus forcing the drama students to use a relative similar intensity for all emotional variations of a given sentence. We had no such constraint in the current corpus and expected intensity to become a much more reliable cue for detecting emotions.

Because we had less control over the recording conditions for the current corpus we explored some normalization techniques. The most successful technique used in the results reported below implied a movie specific normalization of the intensity. Only energy in voiced segments was considered, and the minimum and maximum values were used for normalization. Normalization using minimum/maximum values turned out to be superior to normalization by average since emotions were not necessarily uniformly distributed within a movie.

Using a system relying only on information about intensity mean and variance within an utterance, we were able to correctly classify 58.4% of the segments in the first test set. The corresponding confusion matrix is given in Table 4.59 (a). Most of the confusion took place between sad and neutral: 45.4% of the neutral segments were classified as sad and 13.9% of the sad segments were classified as neutral. The confusion matrix for the second test set is given in

Table 4.59: Confusion matrix based on intensity features. Overall 58.4% and 61.4% of the segments in the first and second test set were classified correctly.

	sad	angry	neutral
SAD	75.2	19.0	45.4
ANGRY	10.9	60.0	15.4
NEUTRAL	13.9	21.0	39.2

(a) Test set 1

	sad	angry	neutral
SAD	37.9	0.0	17.9
ANGRY	1.5	68.2	2.6
NEUTRAL	60.6	31.8	79.5

(b) Test set 2

Table 4.59 (b). Similar to the first test set, most of the confusion took place between neutral and sad: 17.9% of the neutral segments were classified as sad, and 60.6% of the sad segments were classified as neutral. Note that the confusion was marginal between sad and neutral. No angry segment was classified as sad and only 1.5% of the sad segments were classified as angry. Overall, 61.4% of the segments in the second test set were classified correctly.

Table 4.60: Recall, precision, and f1-scores for mean and variance of the intensity.

	sad	angry	neutral
precision	0.55	0.70	0.52
recall	0.75	0.60	0.39
f1	0.63	0.65	0.39

(a) Test set 1

	sad	angry	neutral
precision	0.64	0.83	0.57
recall	0.38	0.68	0.79
f1	0.48	0.75	0.66

(b) Test set 2

Intensity was an insightful cue for detecting angry segments. The f1-score for angry segments was 0.65 for the first and 0.75 for the second test set, higher than for neutral (0.39 and 0.66) and

for sad segments (0.63 and 0.48).

The relative order of intensity mean and variance among the emotions sad, angry and neutral – given in Table 4.61 below – can be succinctly summarized. For sad segments both intensity mean and variance were smaller than for angry or neutral segments. The opposite was the case for angry segments. Their mean and variance values were larger than for both neutral and sad segments. This order was consistent with the order we found for angry and sad segments within the Woggles corpus. See Table 4.16 in section 4.2.5.

Table 4.61: Relative order of prosodic features.

	sad	angry	neutral
sad		<	<
angry	>		>
neutral	>	<	

(a) utterance intensity mean

	sad	angry	neutral
sad		<	<
angry	>		>
neutral	>	<	

(b) utterance intensity variance

This experiment substantiated the previous speculation that the recording conditions for the Woggles corpus were the cause for the poor performance of intensity in that corpus. Differences in intensity turned out to be a quite reliable indicator for the emotion expressed in an utterance.

Intensity (Tremor)

The previous experiment showed that intensity mean and variance of an utterance allowed for a clear discrimination among the emotions well above chance level. In the following experiment, we explored two additional features derived from intensity. We moved a window over the voiced segments in an utterance and computed the following two features:

1. the numbers of changes from a positive to a negative slope (or vice versa) normalized by the total number of windows and

2. the sum of all χ^2 of all regression slopes normalized by the number of windows.

We obtained two very different pictures for the first and second test set. For the first test set, 52.7% of the segments overall were classified correctly. But for the second test set, the accuracy plunged to 26.5%. The system had the tendency to classify segments of the first test set as neutral and for the second set as angry. For instance, 46.5% of the sad segments and 28.0% of the angry segments were misclassified as neutral in the first test set. In the second test set, 78.8% of the sad and 78.2% of the neutral segments were classified as angry. The corresponding confusion matrices are given in Table 4.62.

Table 4.62: Confusion matrix based on tremor information. Overall 52.7% of the segments in the first and 26.5% of the segments in the second test set were classified correctly.

	sad	angry	neutral
SAD	16.8	11.0	8.2
ANGRY	36.6	61.0	10.3
NEUTRAL	46.5	28.0	81.5

(a) Test set 1

	sad	angry	neutral
SAD	10.6	0.0	2.6
ANGRY	78.8	100.0	78.2
NEUTRAL	10.6	0.0	19.2

(b) Test set 2

The tendencies to classify segments as either neutral or angry were reflected in the corresponding f1-scores given in Table 4.63. For the first test set, the f1-score for neutral was 0.63, followed by 0.59 for angry. Sad achieved only an f1-score of 0.23. We had the same relative order for the second test set where neutral was detected best, followed by angry and sad. However, the overall accuracy was much lower. For neutral, we had an f1-score of 0.3, and for sad and angry f1-scores of 0.19 and 0.28, respectively.

Table 4.63: Recall, precision, and f1-scores for tremor features.

	sad	angry	neutral
precision	0.47	0.56	0.51
recall	0.17	0.61	0.81
f1	0.25	0.59	0.63

(a) Test set 1

	sad	angry	neutral
precision	0.78	0.16	0.68
recall	0.11	1.0	0.19
f1	0.19	0.28	0.30

(b) Test set 2

The relative positions of the emotion-specific value of the two tremor features are given in Table 4.64. For the first feature, the neutral value was larger than the values of the remaining emotions. The sad value was the smallest. However, for the second feature, the neutral value again occupied an extreme position. This time, the sad value was smaller than the other values. The value for angry was larger than the values of the remaining emotions. Note that the relative positions of the sad and angry values of the first feature for this corpus are the same as in the Woggles corpus, given in Table 4.19. That is, the angry value was larger than the sad value. For the second feature, we did not get a consistent picture. For the current corpus, the angry value was larger than the sad value while in the Woggles corpus the opposite was the case. It is not

Table 4.64: Relative order of prosodic features.

	sad	angry	neutral
sad		<	<
angry	>		<
neutral	>	>	

(a) normalized number of changes

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(b) normalized χ^2

quite clear whether the two tremor features yield reliable information for the discrimination of emotions. For the Woggles corpus and for the first test set in the current experiment, tremor information allowed the discrimination of emotions well above chance level. However, the tremor values failed on the second test set. Moreover, the emotion-specific values were not consistent across corpora.

Speaking Rate (Phone Duration)

Speaking rate was a reliable feature for the detection of emotions in the Woggles corpus, see section 4.2.5 for details. We operationalized speaking rate as state occupancy in suprasegmental hidden Markov states which were specified to model phones. We computed this state occupancy only for vowels.

Table 4.65 gives the confusion matrix of a system relying only on emotions-specific phone durations. For both test sets, the system preferred to classify segments as angry. For instance, in the first test set 56.4% of the sad and 71.1% of the neutral segments were misclassified as angry.

Table 4.65: Confusion matrix based on speaking rate information. Overall 39.6% and 23.2% of the segments in the first and second test set were classified correctly.

	sad	angry	neutral
SAD	19.8	10.0	10.3
ANGRY	56.4	80.0	71.1
NEUTRAL	23.8	10.0	18.6

(a) Test set 1

	sad	angry	neutral
SAD	11.9	13.6	11.4
ANGRY	68.7	59.1	65.8
NEUTRAL	19.4	27.3	22.8

(b) Test set 2

Due to the system's strong preference to classify segments as angry, the corresponding f1-scores – given in Table 4.66 – lay in most of the cases below chance level. For instance, in the first test set, the f1-scores for sad and neutral were 0.28 and 0.24. Only angry, with an f1-score of 0.52 lay above chance level. In the second test set, not even angry lay above chance level.

Table 4.66: Recall, precision and f1-scores for speaking rate.

	sad	angry	neutral
precision	0.50	0.39	0.35
recall	0.20	0.80	0.19
f1	0.28	0.52	0.24

(a) Test set 1

	sad	angry	neutral
precision	0.40	0.12	0.49
recall	0.18	0.59	0.23
f1	0.18	0.20	0.31

(b) Test set 2

The reason for this poor performance of the duration feature can be found when we look at the relative positions of this feature for each emotion. Since sad segments were in general produced with a lower speaking rate than angry or neutral segments, the corresponding value was higher than for angry or neutral. Angry segments were produced fastest since the value for angry was smaller than for the remaining emotions. The relative order between sad and angry was the same as in the the Woggles corpus, see Table 4.22 for details. There, sad segments were also uttered slower than angry segments. However, if we have a closer look, we see that the difference in the speaking rate was less pronounced for the current corpus. While the average vowel durations in sad segments were longer than in angry segments in the Woggles corpus, this was true in only for 58.8% of the vowels in the current corpus. The differences between the other

Table 4.67: Relative order of speaking rate.

	sad	angry	neutral
sad		>	>
angry	< (58.8%)		<
neutral	< (64.7%)	> (58.8%)	

pairs were also not very pronounced. This situation explains the overall poor performance of a speaking rate feature to discriminate among the emotions in the current corpus. We thought that there are two reasons for this poor performance:

- The current corpus comprised segments from various movies with very different acoustic properties. Even though we used adaptation of spectral models to tune the recognition system to these peculiarities, the large acoustical variance may have prevented an optimal adaptation. The Woggles corpus, in contrast, was recorded in a way to ensure an acoustic similarity among the utterances. As a consequence, the alignment of the transcription with the speech signal may not have been as accurate in the current corpus as in the previous experiments using the Woggles corpus. Note that the computation of state occupancy which we employed as an operationalization of speaking rate was based on this alignment.

  An indication for the increased acoustic diversity of the current corpus compared to the Woggles corpus was the discrepancy in the classification accuracies based on spectral information. Remember that for the Woggles corpus, we were able to achieve an accuracy of about 70% while for the current corpus the corresponding accuracy was only at about 60%. This difference became more evident when we looked at the corresponding baselines achieved by always guessing the same emotion. For the Woggles corpus the baseline lay at 25% and for the first test set of the current corpus by 33%.

- In addition to the acoustic variance among the movies and talk shows, we also had to cope with variances due to of different speaking styles and dialects. For the Woggles corpus, in contrast, the recording conditions produced a homogeneous speaking style among the participating drama students.

Even though the speaking rate failed to indicate the expressed emotion in this experiment, we know from the previous Woggles experiment that speaking rate can be a reliable feature if we can compensate for variance arising from different speaking styles or dialects and can guarantee an accurate extraction of the durations of segments.

Combining Prosodic Information

Whereas in the previous experiments we investigated individual prosodic features, the following experiment combined the eight following prosodic features: mean and variance of the fundamental frequency and intensity, two jitter features, and two tremor features. Consult the previous experiments for a more detailed descriptions of these features.

The confusion matrix for the first and second test set is given in Table 4.68. The overall accuracy was 61.4% and 62.7% for the first and second test set, respectively. Remember that we achieved an accuracy of 58% and 61% in a previous experiment with just intensity mean and variance. Thus, we gained about 2% absolute by the combination of prosodic features. For both test sets, most of the confusion took place between sad and neutral. For instance, in the first test set 39.6% of the sad segments were classified as neutral and 15.5% of the neutral segments were classified as sad. In the second test set, 57.7% of the neutral segments were classified as sad and 13.6% of the sad segments as neutral.

Table 4.68: Confusion matrix based on eight prosodic features. Overall 61.4% and 62.7% of the segments in the first and second test set were classified correctly.

	sad	angry	neutral
SAD	47.5	15.0	15.5
ANGRY	12.9	67.0	14.5
NEUTRAL	39.6	18.0	70.1

(a) Test set 1

	sad	angry	neutral
SAD	84.9	22.7	57.7
ANGRY	1.5	77.3	2.6
NEUTRAL	13.6	0.0	39.7

(b) Test set 2

When we look at the f1-scores, given in Table 4.69, we see that the f1-scores in both test sets for angry were higher than for the remaining emotions (0.69 and 0.81). In the first test set, sad segments were detected with an accuracy of 0.54 and in the second set with 0.65. Neutral was detected better in the first test set than in the second set (0.61 and 0.53).

Table 4.69: Recall, precision, and f1-scores for combined prosodic information.

	sad	angry	neutral
precision	0.62	0.71	0.54
recall	0.48	0.67	0.70
f1	0.54	0.69	0.61

(a) Test set 1

	sad	angry	neutral
precision	0.53	0.85	0.478
recall	0.85	0.77	0.40
f1	0.65	0.81	0.53

(b) Test set 2

With the combination of prosodic features we obtained an accuracy which was comparable to the accuracy we achieved with spectral information in a previous experiment, see section 4.3.3. For the second test set, the accuracy of the combination of prosodic features was higher than the accuracy obtained by spectral information (0.63 vs. 0.47). In the next experiment we combined the prosodic features explored in this section with spectral information to see whether the combination would result in an overall improved performance.

Combining Prosodic and Spectral Information

In the following experiment we combined prosodic and spectral information. We chose the same eight prosodic features as in the previous experiment: mean and variance of the fundamental frequency and the intensity, two jitter, and two tremor features. See the experiments above for a more detailed descriptions of these features, also see section 4.3.3 for a description of the way how we model spectral information in the current corpus. We combined linearly the prosodic and spectral probabilities using weights determined independently on a development set.

The combination of prosodic and spectral information resulted in an overall f1-score of 0.68 for the first and 0.63 for the second test set. If we look at the confusion matrices, given in Table 4.70, we see that most of the confusion took place between sad and neutral segments. For instance, in the first test set 37.7% of the sad segments were misclassified as neutral and 16.5% of the neutral segments as sad. In the second test set, 57.7% of the neutral segments were classified as sad and 13.6% of the sad segments as neutral. Angry segments were detected quite accurately. Note that of the angry segments in the second test set none were misclassified as sad.

Table 4.70: Confusion matrix based on prosodic and spectral information. The overall f1-score for the first test set is 0.68 and for the second test set 0.63.

	sad	angry	neutral
SAD	57.4	8.0	16.5
ANGRY	4.9	75.0	11.3
NEUTRAL	37.7	17.0	72.2

(a) Test set 1

	sad	angry	neutral
SAD	84.8	22.7	57.7
ANGRY	1.5	77.3	2.6
NEUTRAL	13.6	0.0	39.7

(b) Test set 2

The relative high confusion between sad and neutral segments was reflected in the corresponding f1-scores given in Table 4.71 for both test sets. The f1-scores for sad and neutral were lower than for angry. For instance, for the first test set, the f1-scores for sad and neutral were 0.63 while for angry the f1-score was 0.79.

Table 4.71: Recall, precision, and f1-scores for the combination of prosodic and spectral information.

	sad	angry	neutral
precision	0.71	0.75	0.72
recall	0.57	0.75	0.72
f1	0.63	0.79	0.63

(a) Test set 1

	sad	angry	neutral
precision	0.53	0.85	0.78
recall	0.85	0.77	0.40
f1	0.65	0.81	0.53

(b) Test set 2

If we compare the f1-scores we obtained considering solely prosodic and spectral information, we see that the combination resulted in an overall moderate improvement. See Tables 4.51 and 4.69 for the corresponding f1-scores. We achieved f1-scores of 0.63 and 0.57 for spectral information and f1-scores of 0.61 and 0.63 for prosodic information for the first and second test set.

We also assessed an upper bound of pooling spectral and prosodic information by using an oracle which decided when to rely on spectral information and when to when to rely on prosodic information. The corresponding confusion matrices are given in Table 4.72. In the first test set, most of the confusion took place between sad and neutral (22.7% and 11.3%) and angry and neutral (16.0% and 8.2%). The same was true for the second test set. For instance, 9.1% of the angry segments were confused with sad segments and 30.8% of the neutral segments were classified as neutral. However, note that there was far less confusion than in the experiment above in which we combined linearly the spectral and prosodic probabilities. As a consequence, the classification accuracies were significantly higher as well. For the first test, we were able to classify about 78% of the segments correctly and about 79% for the second test set.

Table 4.72: Confusion matrix based on prosodic and spectral information using an oracle. The overall f1-score for the first test set is 0.78 and for the second test set 0.79.

	sad	angry	neutral
SAD	74.3	4.0	11.3
ANGRY	3.0	80.0	8.2
NEUTRAL	22.7	16.0	80.4

(a) Test set 1

	sad	angry	neutral
SAD	89.4	9.1	30.8
ANGRY	1.5	90.9	2.5
NEUTRAL	9.1	0.0	66.7

(b) Test set 2

The corresponding f1-score are given in Table 4.73. Using an oracle improved the classification accuracy for all emotions in both test sets. Angry was classified most accurately in both test sets (0.84 and 0.89), followed by sad (0.79 and 0.78) and neutral (0.73 and 0.76).

Table 4.73: Recall, precision, and f1-scores for the combination of prosodic and spectral information using an oracle.

	sad	angry	neutral
precision	0.83	0.88	0.67
recall	0.74	0.80	0.80
f1	0.79	0.84	0.73

(a) Test set 1

	sad	angry	neutral
precision	0.69	0.87	0.90
recall	0.89	0.91	0.67
f1	0.78	0.89	0.76

(b) Test set 2

To summarize, we were able to show that the linear combination of prosodic and spectral probabilities improved the overall classification accuracy for the first test set. No such improvement could be demonstrated for the second test set. We assessed an upper bound of the combination of prosodic and spectral information by using an oracle in order to predict when to rely on prosodic and when rely on spectral information to classify an utterance with regard to the expressed emotion. This upper bound lay at least 10% points higher than the accuracy we achieved with a linear combination.

4.3.5 Emotion-Specific Verbal Information

One interesting property of the current movie corpus was that we could investigate emotion specific verbal cues. As described in section 3.1, we used language models to capture emotion-specific verbal information. Even though the corpus comprised segments from about twenty movies and talk shows, its size was still small compared to corpora typically used to train language models. The details of the distribution of segments, words, and word types in the various sets are given in in Table 4.74. The word counts refer to sentences after function words were deleted. The training set, given in Table 4.74 (a), comprised about 5,400 segments in total, about 30,000 words and about 3,000 word types. Because emotional segments were not distributed uniformly throughout a movie, more neutral than sad or angry segments were available. The first test set, given in Table 4.74 (b), was the test set we used throughout the previous experiments, and comprised segments randomly drawn from several movies. Within this corpus, emotional segments were distributed uniformly. The second test set, given in Table 4.74 (c), comprised sad, angry, and neutral segments without any substantial background noise from the movie "One True Thing" and was used in the previous experiments as well. Finally, Table 4.74 (d) informs about all sad, angry, and neutral segments in the movie "One True Thing" regardless of the amount of background noise. This last test set comprised a total of 329 segments, a third were sad and about a sixth were angry.

Table 4.74: Distribution of segments, words, and word types for training and test sets. The second test set comprised all angry, sad, and neutral segments in the movie "One True Thing". The subset of these segments which had no background noise is given (d). All counts are based on sentences in which function words were deleted beforehand.

	sad	angry	neutral
# segments	968	1635	2817
# words	5014	10299	16725
# types	630	1582	3201

(a) Training set

	sad	angry	neutral
# segments	102	102	102
# words	495	555	620
# types	241	305	282

(b) Test set 1

	sad	angry	neutral
# segments	67	22	80
# words	263	140	416
# types	132	94	233

(c) Test set 2 (no background noise)

	sad	angry	neutral
# segments	108	47	174
# words	443	314	961
# types	168	103	471

(d) Test set 2

The deletion of function words turned out to improve the accuracy of the subsequent emotion discrimination. Moreover, discrimination based on bigram back-off models tended to be more accurate than discrimination on unigrams or trigrams. Since training of emotion-specific language models had to be based on such a small data set, we could only expect that the most obvious verbal cues were modeled.

Modeling Verbal Information (Test Set 1)

We started exploring emotion-specific verbal information with the first test set. We trained emotion-specific back-off bigrams on emotion-specific training sets. Note that we first deleted all function words from these sets. Consult sections 3.1 and 4.1.5 for a detailed description for the training and testing procedure. We then used these bigrams to score the segments in the test set after all function words were deleted as well. We chose the score of the highest scoring emotion-specific language model to be indicative of the expressed emotion. The corresponding confusion matrix is given in Table 4.75. As observed several times before in previous experiments

Table 4.75: Confusion matrix using back-off bigram language models. Overall, 46.7% of the segments in the first test set were classified correctly.

	sad	angry	neutral
SAD	41.2	22.5	30.4
ANGRY	5.9	33.3	3.9
NEUTRAL	52.9	44.2	65.7

on acoustic and prosodic information, the system had a preference to classify segments as neutral. For instance, 52.9% of the sad and 44.2% of the angry segments were classified as neutral. This preference to classify segments as neutral reflected most likely the situation that more training data was available to train neutral language models than sad or angry models. The system was able to detect a third of the angry segments. It classified only 5.9% of the sad and 3.9% of the neutral segments as angry.

The precision, recall and the f1-scores are given in Table 4.76. Note that for all three emotion

Table 4.76: Precision, recall and f1-score for the first test set using verbal information.

	sad	angry	neutral
precision	0.44	0.77	0.40
recall	0.41	0.33	0.66
f1	0.42	0.47	0.50

classes, the f1-score lay above the baseline of 0.33. Remember that we also asked human subjects to classify these segments based only on a textual presentation. The corresponding results were given in Table 4.50. In that experiment, the f1-score for sad segments was 0.43 and was very similar to the current f1-score for sad (0.42). The same was the case for neutral segments. The system achieved an f1-score of 0.5 and human listeners obtained and f1-score of 0.54. The largest difference between human subjects and the system lay in the detection of angry segments. Human subjects achieved their highest f1-score of 0.65 for angry segments. The system, in contrast, could only achieve a score of 0.47. Overall, the system was able to classify 46.7% of the segments correctly whereas humans achieved a correct classification for 55.7% of the segments.

Modeling Verbal Information (Test Set 2)

With the next experiment, we explored verbal cues in the second test set comprising segments from the movie "One True Thing". For training the emotion-specific bigram models, we used the training set as described above and the first test set. The training and test procedure were the same as described above.

We tested on two test sets. The first set comprised only the sad, angry, and neutral segments of the movie which were free of any substantial background noise. The second set consisted of all sad, angry, and neutral segments in the movie regardless of the amount of background noise. The reason for this separation into two test sets was to investigate the impact of the combination of spectral, prosodic, and verbal information in subsequent experiments. The combination of these three sources, however, could only be tested on the segments free of background noise.

The confusion matrix for the second test set comprising segments free of any background noise is given in Table 4.77 (a). The overall performance of 46.7% was very similar to the first test set which was given in Table 4.75 above. The overall default was to classify a segment as neutral: 52.9% of the sad and 40.9% of the angry segments were classified as neutral. For the second test

Table 4.77: Confusion matrix using back-off bigram language models for the movie "One True Thing". Overall 46.7% and 47.4% of the segments in the first and second test set were classified correctly.

	sad	angry	neutral
SAD	46.3	40.9	38.7
ANGRY	2.9	18.2	6.3
NEUTRAL	50.8	40.9	55.0

(a) Test set 2 (no background noise)

	sad	angry	neutral
SAD	41.7	34.0	32.8
ANGRY	1.8	12.8	6.3
NEUTRAL	56.5	53.2	60.9

(b) Test set 2

set, comprising all sad, angry, and neutral segments, we had a similar confusion pattern. Most of the confusion took place between sad and neutral: 56.5% of the sad segments were classified as neutral and 32.8% of the neutral segments were classified as sad. Overall, about 47% of the segments were classified correctly.

Precision, recall, and f1-scores for both test sets are given in Table 4.78. The f1-scores for sad

Table 4.78: Precision, recall, and f1-scores for the second test set using verbal information

	sad	angry	neutral
precision	0.44	0.36	0.51
recall	0.46	0.18	0.55
f1	0.45	0.24	0.53

(a) Test set 2 (no background noise)

	sad	angry	neutral
precision	0.38	0.32	0.55
recall	0.42	0.13	0.61
f1	0.40	0.18	0.58

(b) Test set 2

segments were similar for both test sets (0.45 and 0.42). Neutral segments were detected with the highest accuracy, their f1-scores was 0.53 and 0.58. Angry segments were detected with an

f1-score of only 0.21 and 0.18. For all test sets, the overall detection rate was with about 47% fairly consistent and lay well above chance level.

Examples of Verbal Information

In order to explore the verbal cues in the current corpus, we ranked the emotion-specific language models. Note, however, that the most likely bigram within an emotion-specific language model was not necessarily the bigram which helped to discriminate among the emotions. We were interested in those bigrams which were very likely in one emotion-specific language model and very unlikely in the remaining models and thus discriminated among the respective emotions.

In order to find these bigrams, we introduced the following simple measure. We were interested in those bigrams $w_{i-1}w_i$ modeling emotion e which satisfied the constraints as given with equations 4.23 and 4.24:

$$P_e(w_i \mid w_{i-1}) > \alpha \tag{4.23}$$

$$\frac{\sum_{e' \in E} P_{e'}(w_i \mid w_{i-1})}{n} > \beta P_e(w_i \mid w_{i-1}) \tag{4.24}$$

where α and β were values to be determined empirically, E denotes the set of emotions we are trying to discriminate, and n is the cardinality of E. With equation 4.23 we guaranteed that the bigram $w_{i-1}w_i$ is likely to occur in a corpus consisting of segments expressing emotion e. With equation 4.24 we guaranteed that the likelihood of the bigram $P_e(w_i \mid w_{i-1})$ is larger than the average likelihood of the bigram $w_{i-1}w_i$.

In order to illustrate this measure and to exemplify the most obvious verbal cues in the current corpus, we listed the first twenty emotion-specific bigrams in Table in 4.79. Note that we deleted function words before applying the language models to compute the score. We could use the lists of bigrams in Table 4.79 to speculate about the lexical cues modeled by the emotion-specific bigrams. In particular, the bigrams modeling verbal cues encoding anger comprise certain lexemes, metaphorizations, and intensification. The bigram "*objection honour*" on rank 18 was a consequence of several movies taking place in a court room. The bigram "*factory played*" originated from the movie "Roger and Me". For bigrams modeling cues signalling sadness, the high number of the words "*no*" and "*love*" was surprising. The best way to summarize the bigrams signalling neutral, was that they really do not invoke any particular emotion. They sounded more affirmative than the bigrams signalling sadness.

As a side remark, we tried to use the measure introduced in equations 4.23 and 4.24 in an additional experiment to classify segments according to the expressed emotion. We were not able to achieve any significant improvements on the results of the experiments reported above.

To summarize, even though the emotion-specific language models had to be trained on relatively small corpora consisting of about 5,000 to 17,000 words, they modeled some emotion-specific lexical cues and were able to discriminate among sad, angry, and neutral with an accuracy of about 47%, well above chance level. Note, also, that human subjects performing a similar task, that is, classifying sentences based on verbal cues only, achieved an accuracy of 55.7%.

With the following experiment we explored whether the combination of verbal and non-verbal

Table 4.79: The top twenty emotion-specific bigrams using the distance measure as defined in euquations 4.23 and 4.24. Remember that function words were deleted before the computation of the emotion-specific language model scores.

	sad	angry	neutral
1	no mommy	fuck fuck	now ever
2	love want	shut fuck	so ever
3	mean love	shut fucking	got got
4	love just	fuck fucking	so see
5	no only	factory played	now so
6	want want	fucking bullshit	about well
7	sorry going	fuck harold	so got
8	no know	hate hate	know about
9	no want	fucking leave	so so
10	no sorry	shut god	about know
11	no mollly	fucking shit	right got
12	going so	get fuck	now right
13	sorry no	stop fucking	right now
14	no doctor	stop fucking	just got
15	how going	tell fuck	no time
16	no nothing	knows hell	so think
17	no really	tell fuck	right so
18	no good	objection honour	yes know
19	how how	fuck talking	well let
20	no going	hey fuck	right just

information resulted in a synergy effect and led to an overall improvement in the detection of the emotional state of the speaker.

4.3.6 Combining Prosodic and Verbal Information

In the following experiment, we combined prosodic and verbal information. We modeled prosodic information based on eight features: mean and variance of the fundamental frequency and the intensity, two jitter, and two tremor features. Consult the corresponding experiments above for a detailed description of these features. We used emotion-specific back-off bigrams to model verbal information. See section 4.3.5 for a detailed description. We combined prosodic and spectral probabilities linearly with weights determined empirically on an independent development set.

The corresponding confusion matrices for the first and second test set are given in Table 4.80. Overall, about 62% and 63% of the segments in the first and second test set were classified correctly. For the first test set, the system tended to classify segments as neutral: 40.6% of the sad and 15.0% of the angry segments were classified as neutral. For the second test set, the system tended to classify segments as sad since 22.7% of the angry and 57.7% of the neutral segments were confused with sad.

Table 4.80: Confusion matrix based on eight prosodic features and verbal information. Overall about 62% of the segments in the first and 63% of the segments in the second test set were classified correctly.

	sad	angry	neutral
SAD	46.5	15.0	15.5
ANGRY	12.9	70.0	14.4
NEUTRAL	40.6	15.0	70.1

(a) Test set 1

	sad	angry	neutral
SAD	86.4	22.7	57.7
ANGRY	1.5	77.3	2.6
NEUTRAL	12.1	0.0	39.7

(b) Test set 2

The corresponding f1-scores are given in Table 4.81. Angry segments achieved the best f1-scores: 0.71 and 0.81 for the first and second test set, respectively. For the first test set, angry was followed by neutral with an f1-score of 0.62. For the second test set, sad segments (0.66) were classified better than neutral (0.53).

If we compared the current f1-scores to the f1-scores we obtained by relying only on prosodic features – consult Table 4.69 – we saw only a marginal improvement. Remember that the overall f1-scores for prosodic information were about 0.61 for the first and 0.63 for the second test set.

Table 4.81: Recall, precision, and f1-scores using both prosodic and verbal information.

	sad	angry	neutral
precision	0.61	0.72	0.55
recall	0.47	0.70	0.70
f1	0.53	0.71	0.62

(a) Test set 1

	sad	angry	neutral
precision	0.53	0.85	0.40
recall	0.86	0.77	0.40
f1	0.66	0.81	0.53

(b) Test set 2

This experiment demonstrated that verbal and prosodic cues in the decoding of emotions are

not necessarily orthogonal. The combination of prosodic and verbal information did not improve the overall f1-score when compared to the accuracy we obtained in a previous experiment by relying just on prosodic information.

4.3.7 Combining Spectral and Verbal Information

When we combined verbal and prosodic information in the previous experiment, the gain in classification accuracy was only moderate. In this experiment we pooled verbal and spectral information to see whether this combination led to an overall improvement. We linearly combined the spectral and verbal probabilities using weights which we determined on an independent development set.

The confusion matrices for the current experiment are given in Table 4.82. Overall, about 65% of the segments in the first and 59% in the second test set were classified correctly. For the first test set, the system tended to classify segments as neutral since 42.6% of the sad and 19.0% of the angry segments were classified as neutral. The same trend was true for the second test set: 39.4% of the sad and 18.2% of the angry segments were classified as neutral. In addition, more than half of the neutral segments were confused with angry. The corresponding f1-scores of

Table 4.82: Confusion matrix based on spectral and verbal information. The overall f1-score for the first test-set was 0.65 and for the second set 0.59.

	sad	angry	neutral
SAD	47.5	7.0	15.5
ANGRY	9.9	74.0	10.3
NEUTRAL	42.6	19.0	74.2

(a) Test set 1

	sad	angry	neutral
SAD	46.9	4.5	30.8
ANGRY	13.5	77.3	51.2
NEUTRAL	39.4	18.2	64.0

(b) Test set 2

the linear combination of verbal and spectral probabilities are given in Table 4.83 below. Angry segments achieved the highest f1-scores with 0.76 and 0.65 for the first and second test set, followed by neutral segments (0.62 and 0.63) and sad segments (0.56 and 0.51).

If we compared the current f1-scores with the scores of spectral information, given in Table 4.52, we saw that the combination seemed to improve the correct classification of angry and neutral. For instance, for the first test set, the f1-score for angry improved form 0.74 to 0.76 and for neutral from 0.58 to 0.62 whereas the f1-score for sad segments dropped from 0.6 to 0.56. Overall, the combination amounted to an improvement of about 2% absolute over a system based only on spectral information.

Table 4.83: Recall, precision, and f1-scores using spectral and verbal information.

	sad	angry	neutral
precision	0.69	0.79	0.54
recall	0.48	0.74	0.74
f1	0.56	0.76	0.62

(a) Test set 1

	sad	angry	neutral
precision	0.55	0.57	0.62
recall	0.47	0.77	0.64
f1	0.51	0.65	0.63

(b) Test set 2

To summarize, the combination of verbal information with either prosodic or spectral information resulted only in moderate improvements for both test sets. The next experiment explored

whether there was an overall improvement when combining spectral, verbal, and prosodic information.

4.3.8 Combining Spectral, Prosodic, and Verbal Information

In the final experiment using the current movie corpus, we combined spectral, prosodic, and verbal information. Similar to the previous experiments, we combined the probabilities of each of these three model sets linearly with weights determined on an independent development set. For the second test set, we were not able to achieve any improvements over the best results achieved earlier. For the first test set, however, we were able to improve slightly by about 1% to 69% correctly classified segments. The confusion matrix for the first test is given in Table 4.84. The system had the tendency to classify segments as neutral: 35.6% of the sad and 17.0% of the angry segments were classified as neutral. We found this kind of confusion in earlier experiments as well. Remember that the best previous accuracy was 68% using prosodic and

Table 4.84: Confusion matrix based on prosodic, spectral, and verbal information. Overall, about 69% of the segments in the first set were classified correctly.

	sad	angry	neutral
SAD	58.4	8.0	15.5
ANGRY	5.9	75.0	10.3
NEUTRAL	35.6	17.0	74.2

spectral information. See Tables 4.70 and 4.71 for the respective numbers of that experiment. The accuracy for angry segments stayed constant across these two experiments.

Table 4.85: Recall, precision, and f1-scores using verbal, spectral, and prosodic information for the first test set.

	sad	angry	neutral
precision	0.72	0.82	0.58
recall	0.58	0.75	0.74
f1	0.64	0.79	0.65

The accuracies for sad and neutral segments increased slightly. The f1-scores for the combination of spectral, prosodic and, and verbal cues are given in Table 4.85. Angry segments were classified most accurately with an f1-score of 0.79. The f1-scores for sad and and neutral were 0.64 and 0.65, respectively.

We also computed the classification accuracies pretending we knew when to choose spectral, prosodic, or verbal information. The corresponding confusion matrices of these oracle experiments are given in Table 4.86. In the first test set, the only noteworthy confusion takes place between sad and neutral (13.9%) and between angry and neutral (9.8%). We find even less confusion in the second test set.

As a consequence of the marginal confusion between the emotions the corresponding f1-scores, given in Table 4.87, were very low. For the first test set, angry segments were classified with an accuracy of 0.9, followed by sad (0.86) and neutral (0.84). For the second test set, the f1-score for all three emotion lay at around 0.9. Overall, 86.6% of the segments in the first and 90.4% of the segments in the second test set were classified correctly. Recall that the upper bounds of the combination of spectral and prosodic information were 78% and 79% for the first and second test

Table 4.86: Confusion matrix based on prosodic and spectral information using an oracle. The overall f1-score for the first test set is 0.68 and for the second test set 0.63.

	sad	angry	neutral
SAD	84.3	3.9	7.8
ANGRY	1.9	86.3	2.9
NEUTRAL	13.8	9.8	89.2

(a) Test set 1

	sad	angry	neutral
SAD	97.0	9.1	12.8
ANGRY	0.0	90.9	2.6
NEUTRAL	3.0	0.0	84.6

(b) Test set 2

Table 4.87: Recall, precision, and f1-scores for the combination of prosodic and spectral information using an oracle.

	sad	angry	neutral
precision	0.88	0.95	0.79
recall	0.84	0.86	0.89
f1	0.86	0.90	0.84

(a) Test set 1

	sad	angry	neutral
precision	0.84	0.91	0.97
recall	0.97	0.91	0.85
f1	0.90	0.91	0.90

(b) Test set 2

set. We established these upper bounds also by using an oracle predicting when to choose the spectral and when to choose the prosodic score. Thus, the additional verbal information allowed an improvement of about 10% absolute for both test sets.

The gain in the overall accuracy by the linear combination of spectral, prosodic, and verbal information was marginal. However, the current overall accuracy of 69% lay just 1% absolute lower than the accuracy humans achieved. We did not expect the classification system to outperform humans on the classification of emotional speech segments.

The experiments relying on an oracle indicated that spectral, prosodic, and verbal information was orthogonal to some extend. The fact that our approach of linear combining spectral, prosodic, and verbal probabilities fell about 10% short of the bound established by the oracle experiment suggested that alternative approaches might bridge the gap.

4.3.9 Summary

In this section we investigated spectral, prosodic, and verbal cues in sad, angry, or neutral speech segments drawn from movies and talk shows. We conducted several experiments involving human subjects to control the quality of the cues signalling an emotion. When the human subjects listened to the speech segments, they classified correctly about 70% of the segments. However, in the case when the subjects based their judgement only on a textual representation, their accuracy dropped to about 56% correctly classified segments. An additional pilot experiment showed that context information helped the classification. In this experiment, the first group of two subjects had to judge the emotion expressed in a randomly drawn utterance from the movie "One True Thing". The second group judged the same segments which, in contrast, were represented in their original order and context. As a result, the second group outperformed the first one significantly. As mentioned earlier, we conducted the experiments involving human listeners to control the quality of the respective corpus and to establish a classification accuracy which we could use to evaluate the accuracy of spectral, prosodic, or verbal cues. Note that these experiments were not intended to explore how humans decode emotions expressed.

In Table 4.88 we list the results of some of the experiments we carried out in this section. All cues were evaluated on two test sets. The first test set comprised speech segments randomly drawn from several movies. The second test set comprised all relevant speech segments from the movie "One True Thing".

The first experiment investigated spectral information which we modeled by cepstral coefficients, power, delta power, and delta-delta power (1). We trained emotion-specific spectral models using adaptation techniques. Using these codebooks for classification, we were able to classify 63% of the segments in the first and 57% of the segments in the second test set correctly.

The next experiments explored several prosodic cues. Classification based on the mean and variance of the fundamental frequency classified 49% of the first and 41% of the second test set correctly (2). Jitter information classified 47% of the first and 46% of the second test set correctly (3). The most reliable prosodic cue was the mean and variance of the intensity (4). Using these two features, we were able to classify 58% of the segments in the first and 61% of the segments in the second test set correctly. Remember that these two features failed to indicate the emotion for the segments in the Woggles corpus. We speculated then that this failure was due to the recording conditions which forced the actors' intensity to lie within a certain range. In order to become reliable indicators, intensity had to be normalized with regard to the movie. In addition, we only considered intensity in voiced regions. We studied two additional features based on intensity. Tremor information allowed to classify more than half of the segments in the first test set correctly (5) whereas it failed completely to indicate the expressed emotions in the second test set. We also explored the potential of the speaking rate to indicate the expressed emotions for the segments in the current corpus (6). We discovered, in contrast to the Woggles corpus, speaking rate was not a reliable indicator. We attributed this shortcoming to the acoustical diversity of the corpus which did not a allow a very accurate alignment of the utterance with the signal, a prerequisite for the estimation of the speaking rate. In addition, different speaking styles and dialects complicated the estimation of the speaking rate parameters. In a final experiment we combined the above mentioned prosodic parameters with the exception of speaking rate. Here, the combination of these prosodic features resulted in an improvement of 2-3% absolute for the two test sets.

We also compared the relative order of the emotion-specific values of all the prosodic features

to each other. As it turned out, for sad and angry the relative order of the values of the prosodic features was the same as in the Woggles corpus. For instance, sad segments had in both corpora a larger mean and variance in their fundamental frequency. For intensity, the mean and variance for sad segments was lower than for angry segments.

Table 4.88: Overview

No.	Features	Segments	Signal Postprocessing	Test Set	f1-score
1	spectral	utterance		1/2	0.64/0.57
2	F_0, mean/variance	utterance	median smoothing gender normalization	1/2	0.49/0.41
3	F_0, jitter	utterance		1/2	0.47/0.46
4	Intensity, mean/variance	utterance	voiced segments normalized	1/2	0.58/0.61
5	Intensity, tremor	utterance	voiced segments log	1/2	0.52/0.25
6	Duration	phones	vowels only	1/2	0.39/0.23
7	Verbal	bigram	function words deleted	1/2	0.46/0.47
8	2, 3, 4, 5, 6			1/2	0.61/0.63
9	7, 8			1/2	0.62/0.63
10	1, 7			1/2	0.65/0.59
11	1, 8			1/2	0.68/0.63
12	8, 10			1/2	0.69/0.63
13	humans			1	0.7
14	1, 8 (oracle)			1/2	0.78/0.79
15	1, 7, 8 (oracle)			1/2	0.87/0.90

The current corpus allowed us to explore verbal cues as well (7). We trained emotion-specific language models (bigrams back-off models) and applied them to classify segments according to the expressed emotions. Using these emotion-specific language models we were able to classify about 46% of the segments in the first and second test set correctly. Human subjects were able to classify 56% of the segments correctly if their judgment was based on a transcription of the utterance.

In several experiments we pooled spectral, prosodic, or verbal information by combining their probabilities linearly with weights determined on an independent development set. The gains in the overall accuracy were moderate. For instance, combining spectral and verbal information amounted to an improvement of 2% absolute (10). Combining prosodic and verbal information did result in an improvement of about 1% for the first test set but no improvement was achieved for the second corpus (9). The combination of spectral and prosodic information improved the accuracy to 68% for the first test set (11). No improvement was achieved for the second test set. Remember that the combination of prosodic and spectral information in the Woggles corpus also did not yield significant improvements. See section 4.2.7 for details. Pooling all three kinds of information, we achieved an accuracy of 69% for the first test set (12) whereas for the second test set, we did not achieve any improvements. However, the performance of spectral information or the linearly combined prosodic information lay already close to the performance of human subjects. The failure of a linear combination of spectral, prosodic, and verbal probabilities

to improve the overall classification accuracies does not necessarily entail that the respective information is not orthogonal. We assessed the upper bound for pooling spectral, prosodic, and verbal information by using an oracle predicting when to choose the spectral, prosodic, or verbal information (14 and 15). The respective upper bounds lay about 10% absolute over the accuracies achieved with a linear combination (11 and 12).

In Figure 4.5, we plotted the emotion-specific f1-scores for several of the experiments based on the first test set. Remember that the baseline was 33% achieved by guessing always the same emotion.

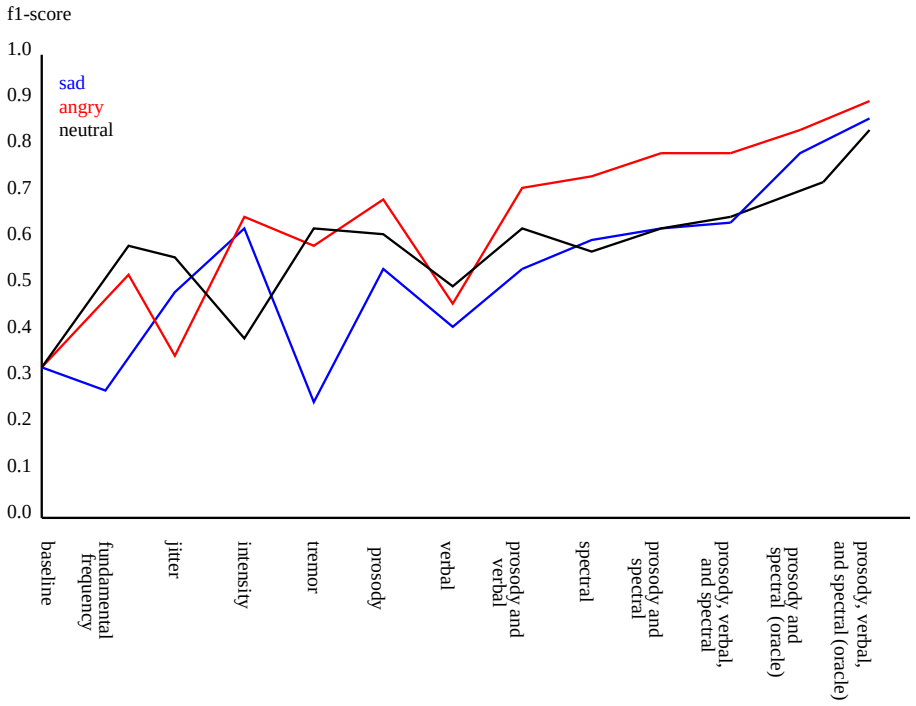


Figure 4.5: Plot of f1-score for prosodic features, the combination of all prosodic features, the combination of prosodic and spectral information, the combination of prosodic, spectral, and verbal information, and two oracle experiments. Note that the f1-scores are based on the first test set.

In general, angry segments were detected best. Looking at the experiments investigating prosodic information, we can see that certain prosodic features yielded very good accuracies for certain emotions while breaking down completely on others. This was the case for all prosodic features until their combination. The combination of prosodic features did not only improve the overall accuracy but also resulted in comparable accuracies for each emotion. We observed a similar behavior of prosodic information in the Woggles corpus. See Figure 4.4 in section 4.2.9 for details. The graph also shows the moderate improvements when we combined spectral, prosodic, and verbal information. The last two data points indicate upper bounds on the combination of spectral, prosodic, and verbal information. We established these upper bounds by an oracle predicting when to choose the spectral, prosodic, or verbal information.

4.4 Prosodic Cues In Chinese and German

With the following experiments we investigated whether emotion-specific prosodic properties of English transfered onto other languages such as Spanish or German. We used prosodic models which were trained on English movies and talk shows. See section 4.3 for details. Note that we only investigated global prosodic features such as mean and variance of the intensity and the fundamental frequency of utterances. Other prosodic features such as speaking rate could not be explored since no alignment path of the signal and the utterance was available.

4.4.1 Spanish

The first corpus consisted of segments from the movie "Johnny Cien Pesos". The test set comprised 90 segments. Within the test set the emotion tags were distributed uniformly. Thus the baseline performance was 33%. We computed eight prosodic features: mean and variance of the fundamental frequency, normalized by gender, mean and variance of the intensity, two jitter and two tremor features. Note that intensity was normalized with respect to the movie. For a detailed description of the prosodic features consult sections 4.1.4 and 4.2.5.

Using prosodic models previously trained on English segments of movies and talk shows, we tested the 90 segments in this Spanish corpus. The confusion matrix of this test is given in Table 4.89. The system had a tendency to classify segments as sad. For instance, 64.3% of the angry and 25.9% of the neutral segments were misclassified as sad. There was very little confusion between angry and neutral. Only 7.1% of the angry segments were classified as neutral and no neutral segment was classified as angry. Overall, 50.6% of the segments were classified correctly. Looking at the f1-scores, given in Table 4.90, we see that neutral segments were detected most

Table 4.89: Confusion matrix based on fundamental frequency and intensity. Overall, close to 50% of the segments were classified correctly.

	sad	angry	neutral
Sad	50.0	64.3	25.9
Angry	7.1	28.6	0.0
Neutral	42.9	7.1	74.1

accurately with an f1-score of 0.66. The f1-scores of both sad and angry lay with 0.42 above chance level. The relative order of the eight prosodic features are given in Table 4.91 and 4.92.

Table 4.90: Precision, recall and f1-scores for prosodic information.

	sad	angry	neutral
precision	0.36	0.80	0.59
recall	0.50	0.29	0.74
f1	0.42	0.42	0.66

Let us look at the first table in which we display the order of the four features based on the

fundamental frequency. Let us also recall the relative positions of these four features in the corpus comprising segments from movies and talk shows, see Tables 4.55 and 4.58. Note that for all four features in both corpora the values for neutral segments were always smaller than for sad and angry segments. However, the positions of sad and angry to each other changed across these two corpora. While sad values were larger than the corresponding angry values in the movie and talk show corpus they were smaller in the current Spanish corpus. The picture regarding

Table 4.91: Relative order of F_0 features.

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(a) F_0 mean

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(b) F_0 variance

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(c) normalized number of changes

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(d) normalized χ^2

to the four intensity features is less coherent. The positions of these four features are given in Table 4.92 for the current Spanish corpus and in Tables 4.61 and 4.62 for the movie corpus. In general, it was the case that in both corpora the values for angry segments were larger than for sad segments, the only exception being the third feature (c) in the current corpus. Also, the positions of angry and neutral to each other were consistent across both However, this was not the case for sad and neutral. While the eight prosodic feature did not behave identically on both

Table 4.92: Relative order of intensity features.

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(a) Intensity mean

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(b) Intensity variance

	sad	angry	neutral
sad		>	>
angry	<		<
neutral	<	>	

(c) normalized number of changes

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(d) normalized χ^2

corpora, they still allowed to classify Spanish speech segments well above chance level. About 50% of the segments were classified correctly. Remember that the system was able to classify correctly about 60% of the segments in the test set of the English movie corpus using the very same prosodic features.

4.4.2 German

We also tested prosodic models trained on the English movie corpus on segments from the movie "Das Boot". This German corpus comprised a total of 84 segments: 27 sad, 35 angry, and 22 neutral segments. We used the same 8 prosodic features as in the preceding experiment: mean and variance of the fundamental frequency and intensity, two jitter, and two tremor features.

The resulting confusion matrix is given in Table 4.93. Overall, only 39.3% of the segments were classified correctly which was about a drop of 10% absolute compared to the experiment testing Spanish data and about a drop of 20% absolute compared to the experiment with English data, see Tables 4.3.4 in sections 4.3.4 and 4.4.1, respectively. Thus, the English prosodic models did not extrapolate to the German data. However, it is interesting to see that none of the sad and none of of the neutral segments were misclassified as angry.

Table 4.93: Confusion matrix based on eight prosodic features. Overall, only about 39% of the segments were classified correctly.

	sad	angry	neutral
Sad	14.8	51.4	4.5
Angry	0.0	22.9	0.0
Neutral	85.2	25.7	95.5

The high confusion among the emotions resulted in very low f1-scores. Sad had an f1-score of only 0.16, angry of 0.37. Neutral segments had the highest f1-score with 0.56. The reason

Table 4.94: Precision, recall, and f1-scores for prosodic information.

	sad	angry	neutral
precision	0.17	1.0	0.4
recall	0.15	0.23	0.96
f1	0.16	0.37	0.56

for the failure of the English prosodic models to classify the German segments can be found if we look at the relative positions of the emotion-specific values for the respective eight features, given in Tables 4.95 and 4.96. None of the positions of the emotion-specific values based on fundamental frequency agreed completely with the positions of these features in English, see Tables 4.55 and 4.58 in section 4.3.4. However, some of the positions of the emotion-specific intensity based values agreed with the positions of these features in English, see Tables 4.61 and 4.64 in section 4.3.4. For instance, both for German and English it was the case that neutral segments were louder than sad segments and that angry segments were still louder than neutral segments. German and English also agreed on the tremor features.

Table 4.95: Relative order of F_0 features.

	sad	angry	neutral
sad		<	<
angry	>		>
neutral	>	<	

(a) F_0 mean

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(b) F_0 variance

	sad	angry	neutral
sad		<	<
angry	>		>
neutral	>	<	

(c) normalized number of changes

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(d) normalized χ^2

Table 4.96: Relative order of intensity features.

	sad	angry	neutral
sad		<	<
angry	>		>
neutral	>	<	

(a) Intensity mean

	sad	angry	neutral
sad		<	<
angry	>		>
neutral	>	<	

(b) Intensity variance

	sad	angry	neutral
sad		>	<
angry	<		<
neutral	>	>	

(c) normalized number of changes

	sad	angry	neutral
sad		<	>
angry	>		>
neutral	<	<	

(d) normalized χ^2

4.4.3 Summary

The size of the underlying corpora of this two pilot experiments above prevents any general claims about the universality of prosodic cues. More languages with substantial larger data sets must be explored to warrant any such claims. We also have to be careful about the origin of the speech segments studied. All the corpora consisted of speech segments produced by actors performing in movies which were produced with the idea in mind to be exported to foreign countries. In addition, actors had certainly prior exposure to foreign languages.

However, besides all the caveats mentioned above, some of the intensity features, in particular, exhibited a consistent behavior across languages. For instance, the values of the intensity mean and the tremor features were consistent for sad and angry across English, Spanish, and German.

4.5 Spanish Call Home

With the Spanish Call Home corpus we explored whether we could detect emotions in spontaneous Spanish telephone conversations. Note that this corpus, in contrast to the previous corpora, did not rely on actors. Instead, people were granted free long distance telephone calls if they allowed their conversations to be recorded. In the following experiments we modeled only emotion-specific spectral information. Moreover, we confined the experiments to binary classification tasks. That is, we tried to distinguish between emotion pairs: sad vs. weak joy and sad vs. strong joy. We expected spectral differences between these emotion pairs to be the largest. In addition, only these emotions occurred in the corpus frequently enough to allow a reliable estimation of the model parameters.

4.5.1 The Corpus

The Spanish Call Home corpus comprised a total 120 conversations. We annotated a subset of 39 conversations with emotion tags using the list of tags described previously in section 4.3.1. The distribution of speech act segments in these 39 conversations is given in Table 4.97. Note that

Table 4.97: Distribution of speech acts according to their emotion tag where *neu* indicates neutral sentences, *wkj* weak joy, *sad* sad, *afr* afraid, *stj* strong joy, *iro* irony, *bor* bored, *sus* suspicion, and *ang* angry.

tag	neu	wkj	sad	afr	stj	dis	iro	bor	sus	ang
#	9868	2470	341	170	160	97	93	46	44	15

only segments tagged as neutral, sad, afraid, and weak and strong joy occurred more than 150 times in the corpus. In general, it can be said that corpus did not comprise emotional segments in a density we had encountered in the previous corpora.

The speech style in the Spanish Call Home corpus was extremely informal and, moreover, comprised several South American dialects combined with foreign words, both English and local. The foreign line quality was often poor and there was a lot of background noise such as babies crying or kitchen noises.

For the following experiments we used the Janus Recognition Toolkit which was trained on a subset of 80 conversations and additional training data from Call Friend Spanish and the Ricardo database. The word error rate of the trained system was 61.1% which was comparable to results reported from other sites, such as 57.4% from BBN, 57.5% from SRI, and 61.3% from NSA.

4.5.2 Intra- and Intercoder Tagging Agreement

The conversations were segmented into segments using a schema as described in (Finke et al., 1998). For annotating the resulting segments with emotion tags we employed two transcribers, both natives from Chile. In order to validate the quality of the expression of emotions and

the consistency of the tagging process, we conducted two experiments in which we tested the agreement between the two transcribers and between a transcriber and herself.

Intracoder Agreement

To test for intracoder agreement we asked transcriber A to tag again three conversations she had tagged about 2 months ago. She was told not to consult her previous tagging. The confusion matrix for the two tagging sessions of transcriber A is given in Table 4.98. The baseline was

Table 4.98: Intracoder Confusion Matrix for Transcriber A. Overall, transcriber agreed on 83.9 of her tags.

$A_1 \setminus A_2$	neu	wkj	sad	afr	stj	dis	iro	bor	sus	ang	Tot
neu	576	65	8	8		16	2		2		677
wkj	24	93			3				1		121
sad			9			1					10
afr				13							13
stj		3			2						5
dis						19					19
iro	2	4					8				14
bor											0
sus									3		3
ang									3		3
Tot	602	165	17	21	5	36	10	0	6	0	862

78.5% or 69.8% depending which of the two tagging sessions we assume to be correct. The baseline could be achieved by guessing always the most frequent tag, that is neutral. As the confusion matrix shows, most of the confusion took place between weak joy and neutral: 65 of the segments tagged as weak joy were previously tagged as neutral. And 24 segments, previously tagged as weak joy were tagged as neutral in the second session.

Intercoder Agreement

In order to test for agreement between transcribers, we asked transcriber B to tag 5 conversations previously tagged by transcriber A. The confusion matrix is given in Table 4.99. Similar to the intracoder agreement, most of the confusion between coders A and B took place among the tags neutral (neu) and weak joy (wkj). 54 of the segments tagged as weak joy by tagger A were classified as neutral by coder B. And 21 of the segments classified as neutral by coder B were tagged as weak joy by the other coder.

The baseline of this agreement test was 71.1% or 68.7% – depending whether we chose the tags of transcriber A or B to be correct. The baseline could be achieved by always guessing the most frequent emotion tag which was neutral for both tagging sessions. Note that a different set of conversations was used in the intracoder and intercoder experiment. Thus, the agreement numbers are difficult to compare. However, in both experiments the actual agreement lay higher

Table 4.99: Intercoder Confusion Matrix. Overall,the two transcribers agreed on 92.6% of the tags.

$A \setminus B$	neu	wkj	sad	afr	stj	dis	iro	bor	sus	ang	Tot
neu	1143	54	8	5		4		1	2		1217
wkj	21	316			12	1	2				352
sad	9		36	1							46
afr	1			22					1		24
stj		2			35						37
dis	2		1			13					16
iro							15				15
bor								2			2
sus									2		2
ang										1	1
Tot	1176	372	45	28	47	18	17	3	5	1	1712

than the baseline, suggesting some consistency in the tagging process. In addition, both experiments above also validated the expression of emotion in this corpus. Both transcribers were able to decode some of the emotions expressed in the conversations well above chance level.

4.5.3 Spectral Cues

We trained emotion-specific models on 23 conversations to capture spectral differences. We used these emotion-specific spectral models to classify the speech segments in the remaining 16 conversations. See section 4.1.3 for a more detailed description of the training and testing procedure. In order to have a sufficient number of training tokens for the parameter estimation, we confined the following experiments to the emotions weak and strong joy and sad. We also expected the spectral differences between these emotion pairs to be the largest.

In the first experiment we tried to distinguish between sad and weak joy. The corresponding confusion matrix is given in Table 4.100. Overall, 94.5% of the segments were classified correctly which was above the baseline of 92.6% achieved by guessing always weak joy. The precision,

Table 4.100: Confusion Matrix for weak joy and sad. Overall, 94.5% of the segments were classified correctly.

	wkj	sad
Wkj	96.5	36.7
Sad	3.5	63.3

recall, and f1-scores are given in Table 4.101. Since only 3.5% of the weak joy segments were misclassified as sad, the f1-score for weak joy lay at 0.97. However, since 36.7% of the sad segments were classified as neutral, the corresponding f1-score lay at only 0.64. In the next experiment we tested whether spectral information could distinguish between sad and strong joy. The confusion matrix for this emotion pair is given in Table 4.102. Note that none of the sad

Table 4.101: Recall, Precision, and f1-scores for spectral information.

	wkj	sad
precision	0.97	0.61
recall	0.96	0.63
f1	0.97	0.64

segments was misclassified. However, 69.5% of the strong joy segments were classified as sad. Overall, 64% of the segments were classified correctly. The baseline is 51.8% achieved again by always guessing the most frequent tag, i.e. strong joy (stj). The corresponding f1-score are given

Table 4.102: Confusion Matrix for Strong Joy and Sad. Overall, 64% of the segments were classified correctly.

	stj	sad
STJ	30.5	0
SAD	69.5	100

in Table 4.103. Since none of the sad segments was misclassified, we had a perfect recall-value for sad. Sad segments were classified with an accuracy of 0.73, strong joy achieved an accuracy of 0.46.

Table 4.103: Recall, Precision, and f1-scores for spectral information.

	stj	sad
precision	1.0	0.57
recall	0.31	1.0
f1	0.46	0.73

4.5.4 Summary

The experiments in this section were definitely handicapped by the very limited number of segments expressing emotions in the telephone conversations. Note that, both parties of the telephone conversations were aware that their conversation was recorded and used for research. Due to this condition, the number of emotional segments was most likely low and the emotions expressed were, in addition, only moderate. We, therefore, confined our experiments to binary classification tasks. That is, we conducted experiments to distinguish between sad and weak joy and between sad and strong joy. We adapted models to capture emotion-specific spectral differences. Using these emotion-specific models we were able to distinguish in both cases among the emotions with an accuracy which lay above the baseline.

This experiment indicated that it is possible to use spectral information to detect emotions in spontaneous speech in telephone conversations. We also tried to use verbal and prosodic information to classify speech segments. However, we were not able to achieve a performance

accuracy above chance. We think that there were several reasons for this poor performance of prosodic and verbal information. First, it was very difficult to estimate reliably the fundamental frequency due to the often poor line quality and the fact that telephone speech is sampled with 8kHz. Training emotion-specific language models requires substantial data which was not available in the current corpus. The limited amount of training data was one of the reasons why verbal information failed to indicate the emotion. In general, it can be said that the telephone conversations took place between family members or people who knew each other very well and for a long time. We hypothesize that because of this familiarity both members could rely on idiosyncratic hints and very subtle cues which are obviously more difficult to detect and subject to future research.

Chapter 5

Summary

In this investigation, we explored verbal and non-verbal cues in the communication of emotions. We focused our attention on three domains: verbal cues, spectral cues, and prosodic cues. We studied their potential to discriminate among several emotions (happy, sad, angry, afraid, and neutral) from speech segments of four different corpora.

The first corpus comprised 50 English sentences portrayed by nine drama students in happy, sad, afraid, and angry variations. The second corpus comprised English speech segments from talk shows and movies. The third corpus consisted of segments from Spanish and German movies. With this third corpus we studied whether prosodic models trained on English data were also able to discriminate emotions in Spanish or German speech segments. Since only sad, angry, and neutral segments occurred frequently enough in the second and third corpus to warrant a reliable parameter estimation of emotion-specific models, we confined the experiments which use these two corpora to those three emotions. With the last corpus, we explored whether spectral information could be used to detect emotions in speech segments from spontaneous Spanish telephone conversations. Note that this fourth corpus did not rely on actors. To our knowledge these four corpora constitute the largest collection of emotional speech data available and the collection, transcription, and tagging was a substantial part of this investigation.

We conducted several experiments to control the quality of these corpora. In these experiments we asked human subjects to listen to segments from these corpora and to classify them according to the expressed emotions. In all of these studies, the subjects could consistently recover the emotions expressed by the actors or by the participants in the telephone conversation. For the first corpus, for instance, subjects were able to classify about 70% of the speech segments correctly where most of the remaining confusion took place between sad and afraid segments. We conducted a similar experiment using the second corpus. Human subjects were also able to classify about 70% of the segments correctly. However, in this case, they had only to discriminate between sad, angry, and neutral and, in addition, could rely on verbal cues which were absent in the first corpus. We also tested the ability of subjects to classify segments based on verbal information only, that is, the textual representation of what was said in the utterance. The accuracy dropped by about 14%. Only 56% of the segments were classified correctly.

In our investigation we captured emotion-specific verbal information with bigram back-off

models. Using these language models to distinguish between sad, angry, and neutral segments, we were able to classify about 46% of the segments in the test set correctly which was about a drop of 10% absolute compared to the accuracy human subjects were able to achieve on this task. Note that these language models were trained on relatively small corpora and were only able to model the most obvious verbal cues. In order to approach human accuracy more training data and refined modeling techniques are required. A possible extension might be the inclusion of decision trees which are able to ask questions about syntactic constructions and the discourse.

We modeled spectral information by means of cepstral coefficients and we used codebook adaptation to train emotion-specific models. In general, it was the case that this technique captured emotion-specific spectral differences quite well, and the classification of speech with respect to the expressed emotion was fairly accurate, sometimes approaching human performance. For instance, in the first corpus, we were able to classify 69% of the segments correctly. With the second corpus comprising segments from movies and talk shows, classification based on spectral information achieved an accuracy of 60%.

We also applied emotion-specific spectral information to the classification of segments from Spanish telephone conversations. We tried to distinguish between sadness and weak joy and between sadness and strong joy. In both cases, the classification accuracy was better than chance.

We paid special attention to prosodic information and investigated several individual prosodic features and their potential to discriminate among the emotions expressed in the respective corpora. We found nine prosodic features to be most reliable:

- The mean and variance of the fundamental frequency within an utterance which we normalized depending on the sex of the speaker or depending on the speaker herself.
- Two jitter features which we computed by moving a window over the fundamental frequency to compute its smoothness.
- The mean and variance of the intensity within the utterance which we normalized with regard to the movie or talk show. In addition, we only considered intensity in voiced regions within the speech segment.
- Two tremor features which we computed by moving a window over the intensity to compute its smoothness.
- The speaking rate which we modeled with vowel durations.

Using these nine features we trained emotion-specific prosodic models. Based on these prosodic models we were able to classify 60% of the segments in the test set of the first corpus correctly. We could achieve about the same accuracy on the test sets of the second corpus. It was interesting to observe that individual prosodic features yielded very different accuracies for certain emotions. For instance, prosodic models based only on mean and variance of the intensity within the utterance were able to classify sad and angry segments fairly accurately. But these two features failed completely to detect neutral segments. However, combining all nine prosodic features not only resulted in an overall better classification accuracy but also all emotions were detected with comparable accuracies. This was the case for both the first and the second corpus.

We also studied the relative order of the emotion-specific values of these nine prosodic features to each other. As it turned out, in most of the cases the relative order of the emotion-specific values of these features was preserved across corpora. For instance, the intensity of sad segments was lower than the intensity of angry segments in both corpora. The consistency of the relative order of these values across corpora was another indication that the respective prosodic features yielded reliable cues for the communication of emotions. Some prosodic features, in particular features based on intensity, produced even consistent values across languages. That is, the relative order of the values of these features was the same for English, Spanish, and German. Note that the nine prosodic features as described above were similar to prosodic features in other studies (Amir and Ron, 1998; Banse and Scherer, 1996).

We also investigated prosodic information other than duration at the phone level. The motivation behind these studies was to compensate for phone intrinsic prosodic properties. Low vowels, for instance /a/, have an intrinsic lower fundamental frequency than high vowels, such as /i/. However, modeling these intrinsic prosodic properties explicitly with prosodic phone models did not result in an overall improvement of the classification accuracy. We speculated that phone intrinsic prosodic properties were overwritten by the overall variance in the data due to different speakers, speaking styles, and different ways to encode an emotion.

While we did explore a wide range of prosodic features, some prosodic features remain subject of future research. For instance, we did not consider pauses which is considered to be a reliable indicator for sadness. The reason for not considering pauses in our experiments was that the transcription protocol used pauses as an indicator for a speech segment boundary.

We found prosodic features pertaining to the whole utterance to be the most reliable indicators for the expressed emotion. These global features, however, failed when an emotion was signalled more subtly. Listening to the segments which were misclassified in our experiments, we found that emotions were sometimes expressed by the prosodic modifications of a single word which render the respective word more salient. Prosodic modifications pertaining to a single word within an utterance comprising several words could not have been picked up by the above global prosodic features. Thus, we need a more fine grained modeling of prosodic parameters at the phone, syllable, or word level. Note that the suprasegmental hidden Markov model which we introduced in this investigation as a tool to model prosodic events, allows the modeling of prosodic events at these levels. However, when modeling more fine grained prosodic events at the phone, syllable, or word level, we have to account for the multi-functionality of prosody. Prosody is not only used to signal emotions but implements a variety of linguistic functions. There might be several reasons for a word to be salient. A word's saliency might indicate a certain emotion, it might indicate that the word constitutes new and important information in the current discourse, or it might simply indicate some segmentation information. Thus, modeling prosodic events at these levels requires an integrated approach which reflects that prosody simultaneously implements several linguistic functions. We explored the impact of the expression of certain emotions on the phone level and we were able to show that context dependent prosodic phone models classified about half of the segments in the Woggles test set correctly. Modeling other linguistic functions of prosody at this level as well, should improve the overall accuracy and should finally facilitate the modeling of more subtle cues in the communication of emotions. This, however, is subject to future research.

The linear combination of spectral, prosodic, and verbal scores did not necessarily result in an overall improvement of the classification accuracy. In fact, the gains were moderate and confined to one or two percent absolute. However, in most of our experiments, the accuracy of

a classification system based on either spectral or prosodic information approached closely the accuracy of humans. We think that acoustic and verbal information becomes more orthogonal if we consider a larger variety of emotions. For instance, the high confusability of afraid and sad segments in the Woggles corpus both by the system and human listeners was probably due to the total lack of verbal information. We also established upper bounds on the combination of spectral, prosodic, and verbal information using an oracle which predicted when to choose spectral, prosodic, or verbal scores. These upper bounds lay about 10% absolute higher than the accuracies achieved with linear combinations of the respective scores. This indicates that different approaches to combine spectral, prosodic, and verbal information might be promising and result in an additional overall improvement.

Prosodic information was more robust than spectral information when confronted with novel movies or novel speakers. When confronted with utterances of novel speakers, the accuracy of a classification system based on spectral information dropped from 70% to only 46% correctly classified segments from the Woggles test set while the classification system based on prosodic information fell from 60% to 50%. Similar results were found for the movie corpus. When confronted with the novel movie "One True Thing" the accuracies of spectral models fell from 63% to 57%. At the same time, classification based on prosodic information stayed nearly constant.

There are several apparent ways to improve and to extend the existing approach. The most obvious course would seek to integrate visual information into the classification process. This visual information could comprise information about facial movements but also about certain body postures and gestures. Other additional information sources could include information about heartbeat, body temperature, and skin resistance. These last three sources will become even more interesting when their assessment can be performed in a non-intrusive way. Physiological information, in general, will also help to detect stress or cognitive overload more reliably (Fay and Middleton, 1941; Mendoza and Carballo, 1998a; Protopapas and Lieberman, 1997; Mendoza and Carballo, 1998b; Laukkanen et al., 1996; Griffin and Williams, 1987; Streeter et al., 1983; Simonov and M.V.Frolov, 1973). Other directions might include the automatic detecion of lies (Hollien, Geison, and Hicks, 1987; Disner, 1982).

In our investigation we considered only a very limited number of emotions: happy, sad, afraid, and angry. The reason for this limitation was that sufficient training data was only available for these emotions. Future research on the expression of emotions has to extend this list with less frequent emotions. This, of course, means that more transcribed and tagged data is needed. Moreover, it would also be useful to have information about the intensity of an expressed emotion, for instance, is the individual happy or very happy, annoyed or furious? A dimensional view on emotions might prove beneficial in this context. It might also be the case that by considering a more comprehensive list of emotions and their intensity, verbal, spectral, and prosodic information becomes more orthogonal and the combination proves to be more fruitful than in the experiments in this investigation. However, this is subject to future research.

In conclusion, this investigation revealed that verbal and non-verbal cues can be used to detect the emotion expressed in an utterance. Our classification system based on verbal, spectral, and prosodic information was able to achieve a classification accuracy comparable to the performance of human listeners.

By integrating this classification system into human-computer interfaces we would be able to endow computers with the ability to detect the emotion expressed by a human user and to act accordingly. Humans would not be faced with computers which emote heavily but ignore quite impolitely the user's feelings. Since humans tend to extrapolate their interpersonal communication skills to the interaction with computers, a computer interface capable of both the expression and detection of emotions would accommodate this tendency and result in a more user friendly interface. However, if computers can become our friends then there is probably also potential for them to become our enemies.

Appendix A

The Meeting Browser

One possible application of the automatic detection of emotions is the Meeting Browser (Waibel et al., 1998). The Meeting Browser allows to peruse through a meeting which is represented along a horizontal time line. Along this time line additional information be displayed synchronously:

- the current speaker
- the line of sights of people participating in the meeting.

The Meeting Browser allows queries about this kind of information. In addition it is possible to display certain discourse information, such speech acts and so on (Finke et al., 1998). Other interesting information might consist of the emotional states of the discourse participants. It would allow queries for certain conditions, such as, when is speaker A angry at speaker B and talks about topic C.

In order to demonstrate the capabilities of both the Meeting Browser and the automatic detection of emotions, we integrated the system for the emotion detection into the Meeting Browser allowing the Browser to display all of the emotional states of the speakers. A snapshot of this Meeting Browser application is given in Fig. A.1 for the movie "A Few Good Men". The display of the Meeting Browser comprises three major windows. The two windows at the bottom display the close captions of what is said annotated by the emotion the speaker is in. The right window shows the corresponding video segment. The top window displays the whole meeting/movie at once using emotion labels which extend over whole speech segments. We use black to indicate neutral segments. Angry segments are represented by red and sad segments by blue. This window also allows to browse through the meeting and select segments of interests. When a segment of interest is selected, the Meeting Browser plays both audio and video in the bottom right window while highlighting the words uttered in the bottom left window. At the same moment, the time cursor – represented by the vertical green bar in the top window – slides through the selected segment.

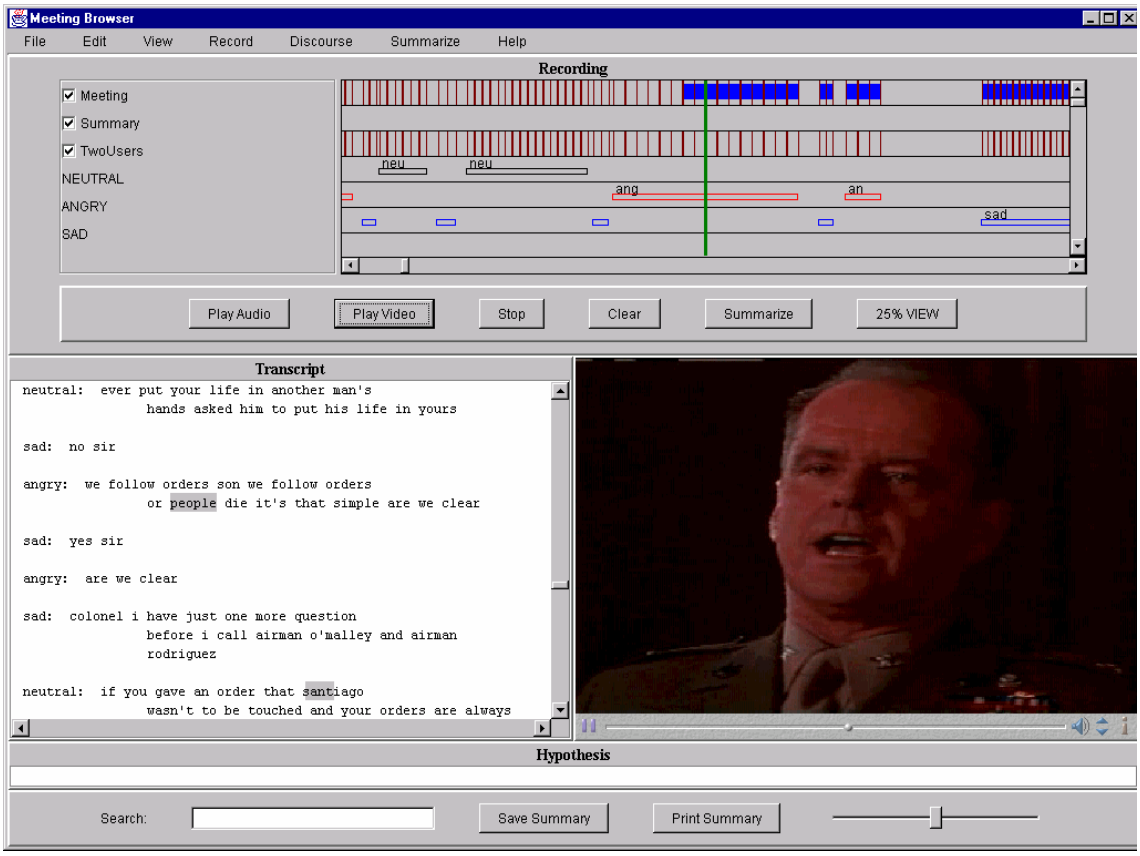


Figure A.1: Snapshot of the Meeting Browser displaying the movie "A Few Good Men". The left major window shows the close captions annotated by an emotion tag as detected by the system. The right window shows the video. Neutral speech segments are marked as black, angry segments as red, and sad segments as blue in slideable top window.

Appendix B

The JANUS Prosodic Toolkit

The Janus Prosodic Toolkit (JPTk) allows for the integration of prosodic information into the Janus recognition system. JPTk comprises several new modules which facilitate the extraction of prosodic features from the speech signal and the training of prosodic models. Prosodic information can be used to train models to detect the prosodic properties of emotional speech, syntactic or discourse boundaries, or stressed and unstressed words and so on.

Following the JANUS object oriented software design idea, prosodic modules are implemented as high level objects which can be accessed interactively at the Tcl/Tk level (Ousterhout (1994)). Prosodic objects start their life under-specified. Their concrete instantiation is achieved by configuring these objects at the Tcl/Tk level. This open design approach allows to experiment with different configurations for training and testing without the need to consult the underlying C-code. JPTk also makes use of the AWT of Tcl/Tk. Graphical interfaces allow an easy first glimpse at the data, for example, the contour of the fundamental frequency or the intensity. JPTk is available for the following platforms: Sun Solaris, Linux, Alpha, and Hewlet Packard.

Within a speech recognition system, acoustic events are normally observed at a rate of 10ms. This rate is not appropriate for observing prosodic events which require substantial context information. Ideally, we would like to observe prosodic events in segments spanning phones, syllables, words, or even whole utterances. Within JPTk, it is possible to define these kind of segments in the dictionary by using lexical tags. For instance, if we want to investigate prosodic phone models we would specify every phone in the dictionary as segment final:

```
{AND} {{AE phoneFinal} {N phoneFinal} {DD phoneFinal}}
```

This encoding might look redundant at first sight but it is consistent with the encoding you have to use when you investigate, for example, syllable based prosodic segments:

```
{AND} {{AE}  {N} {DD syllableFinal}}
```

or word based segments:

```
{AND} {{AE} {N} {DD wordFinal}}
```

It is, of course, possible to model different segment sizes at the same time, for instance, phone and word segments:

```
{AND} {{AE phoneFinal} {N phoneFinal} {DD phoneFinal wordFinal}}
```

Note that these final tags have to be defined in the corresponding dictionary tag set. There is a special predefined tag called `uttFinal` which denotes the end of the whole utterance segment.

These segment final tags play a crucial role in the specification and training of prosodic models. Each prosodic model is associated with a set of prosodic observations. To be consistent with the JANUS terminology we will use the term *features* to refer to these observations. Within a feature description the above mentioned segment final tags are used to associate prosodic segments with prosodic features. Thus, it is possible to have different features for differently bounded prosodic segments.

Note that segment final tags only specify the segment boundary. They do not specify which actual prosodic model instantiates the respective segment. The mapping from a segment boundary to a prosodic model is achieved by a regression tree in the following way. We determine properties within the context surrounding the segment boundary by using what we call property functions. The result is a property vector which serves as the input to a regression tree. The questions within this regression tree refer to positions within this property vector, and depending on the answer, we traverse the tree. When we arrive at a leaf node, an integer pointer refers to the actual model index. This setup allows prosodic models to be clustered and, moreover, to be context dependent.

Note that this appendix is not a JPTk manual. Please consult the online documentation for a more detailed description of the prosodic modules and their usage. The objective of this appendix is to illustrate the possibilities of JPTk and how it was used to model emotion specific prosodic models.

Throughout this chapter we will develop an example program which illustrates how to specify, train, and test context independent prosodic phone models. During this illustration we sometimes have to rely on traditional JANUS modules, for example, the dictionary or the feature description set. We assume that the reader is familiar with these objects since we will not explain in detail their specification.

B.1 Prosodic Feature Set

Tags within the dictionary are used to define segment boundaries of prosodic models. The same tags are used to define the corresponding prosodic features for these models. The specification of prosodic features is similar to the specification of acoustic features in the acoustic feature set, `FeatureSet`. There are, however, several important differences:

- A typical specification in the acoustic feature description comprises the input and output

features and a function responsible for transforming the input to the output feature. In contrast, a specification within the prosodic feature description comprises an additional segment final tag, indicating that this feature belongs to segments bounded by this tag.

- The feature description set is evaluated only once, namely at the very beginning of the decoding process. In contrast, the prosodic feature set is evaluated every time we leave a segment. To be precise, each time we leave a segment as marked by the segment final tag, we evaluate that part of the description which specifies prosodic features for this very segment.
- You can use features from the acoustic feature set within the prosodic feature set but not vice versa.

Since, it is possible to use features defined in the acoustic feature set within the prosodic set, you have to specify the former as an extra argument when creating the latter:

```
% Creating a prosodic feature set
% ---------------------------------------
% 'acousticFeatSet' is a FeatureSet object and
% 'prosodicFeatDescFile.txt' is the file containing the
%  actual prosodic feature description
ProsodicFeatureSet prosodicFeatDesc acousticFeatureSet \
                   -desc prosodicFeatDescFile.txt
```

The following functions are defined for the computation of prosodic features:

- `moments`:
- `range` :
- `cut` :
- `smooth` :
- `length` :
- `lin` :
- `concat` :
- `pitch` :

It is possible to define prosodic features for several differently bounded segments within one prosodic feature set. Note that the last feature defined for some segment specification implements the prosodic features for the respective models.

The actual feature description should be specified within a file which has to be specified with the `-desc` argument when we create a prosodic feature set (see above). For our example, we specify the following prosodic features for prosodic phone models:

```
% Prosodic Feature Set Description File
% -------------------------------
% note PITCH and POWER are define in the acoustic feature
% description file

% prosodic features for phoneSegments
$fes moments phoneFinal phonePitchMoms PITCH
$fes cut     phoneFinal phonePitchMean phonePitchMoms 0 1
$fes moments phoneFinal phonePowerMoms POWER
$fes cut     phoneFinal phonePowerMean phonePowerMoms 0 1
$fes concat  phoneFinal phoneFeatures  phonePitchMean phonePitchMean
```

Thus prosodic phone models will be based on two features, the mean of the fundamental frequency (`PITCH`) and the mean of the energy (`POWER`).

B.1.1 Property Set

Prosodic models are not atomic but are characterized by a vector of properties. This property vector is computed by functions which are specified in a property set. A property set comprises a list of functions which are evaluated at every segment boundary. The values of these functions make up the property vector which is the input to the regression tree for the final mapping to a prosodic model index.

JPTk offers the following property functions:

1. `lexTag`: this function requires one string argument indicating the lexical tag you want ask about. For example, if you want to model syllables, you might want to add lexical tags such as `nucleus`, `coda`, or `onset`:

   ```
   {AND} {{AE nucleus}  {N coda} {DD coda syllableFinal}}
   ```

   Using the function `lexTag` you can ask whether a certain phone within a syllable is annotated with a certain tag.

2. `phoneIdx`: this function returns the index of the current phone within the phone set as specified by the additional argument. This phone set has to refer to a set within the `PhonesSet` of the dictionary.

By using the optional argument `-pos` within a property function it is possible to compute properties of the context surrounding the current segment. For example, `-pos -1` refers to the previous phone. In case no context is available, the property functions return `-1`.

Creating a property set is fairly straightforward because the definition does not require any additional objects as arguments:

```
 % Creating a property set
PropertySet phonePropertySet
```

To continue our example, a specification of the property set is given below:

```
% Property set for prosodic phone models
prop1 phoneIdx Phones
```

By using the property function `phoneIdx` in this way, we assign a unique integer to each phone. We load this specifiction with the `read` function:

```
% Reading specification of properties from file 'phoneProps.txt'
phonePropertySet read phoneProps.txt
```

B.1.2 Question Set

A question within a question set refers to a position in the property vector as computed by the property functions within a property set. Questions return boolean values only.

Creating a question set is fairly straightforward. No additional objects are required as arguments.

```
% Creating a question st
QuestionSet phoneQuestionSet
```

The syntax for specifying a question is as follows. The beginning integer refers to the position within the property vector the question refers to. It follows the compare operator which can be either =, <, or >. The last value specifies the value we are asking for in that position within the property vector.

In our ongoing example, we specified only one property for phones. That is, the specification of the actual questions is as follows:

```
% Question set for prosodic phone models
0  = 0.000000
0  = 1.000000
0  = 2.000000
0  = 3.000000
0  = 4.000000
0  = 5.000000
0  = 6.000000
0  = 7.000000
0  = 8.000000
0  = 9.000000
0  = 10.000000
0  = 11.000000
0  = 12.000000
```

```
0  = 13.000000
0  = 14.000000
0  = 15.000000
0  = 16.000000
0  = 17.000000
0  = 18.000000
0  = 19.000000
0  = 20.000000
0  = 21.000000
0  = 22.000000
0  = 23.000000
0  = 24.000000
0  = 25.000000
0  = 26.000000
0  = 27.000000
0  = 28.000000
0  = 29.000000
0  = 30.000000
0  = 31.000000
0  = 32.000000
0  = 33.000000
0  = 34.000000
0  = 35.000000
0  = 36.000000
0  = 37.000000
0  = 38.000000
0  = 39.000000
0  = 40.000000
0  = 41.000000
0  = 42.000000
0  = 43.000000
```

Assuming we stored the above specifications in a file named `phoneQuestions.txt`, we can read them in by the following command:

```
% Reading specification of questions from file 'phoneQuestions.txt'
phoneQuestionSet read phoneQuestions.txt
```

B.1.3 Regression Tree

Using a question set and a regression tree we can finally map from properties (property vector) to an actual prosodic model. At each segment boundary, we first compute the properties of the current segment. Starting at the root node, we then traverse the tree, asking node specific questions about these properties. Depending on the boolean answer we pursue the left node (false) or the right node (true). Arriving at a leaf node, an additional argument, `-model`, specifies the index of the corresponding prosodic model.

Creating a regression tree is fairly straightforward because it requires only a question set as

an additional argument:

```
% Creating a regression tree
RegTree phoneRegTree phoneQuestionSet
```

Regression trees are defined top-down, beginning with the root node `ROOT`. A typical specification starts with a mother node, followed by an integer index pointing to a question within the corresponding question description, followed by two daughter nodes. The first daughter, points to a node which is followed up when the corresponding question returns false. The second daughter points to a node which is followed up when the question returns true. The absence of a daughter is indicated by a hyphen. Thus, leaf nodes have two hyphens. Moreover, leaf nodes have an additional argument `-model` whose integer value specifies the index of the corresponding prosodic model. Finally, because leaf nodes have no daughters, there is no question to be asked. The absence of question is indicated by a `-1`.

Continuing our example of prosodic phone models, the specification of the regression tree is as follows:

```
% Regression tree for prosodic phone models
ROOT            {0} node-AE leaf-AA
node-AE         {1} node-AH leaf-AE
leaf-AA         {-1} - - -model 0
node-AH         {2} node-AO leaf-AH
leaf-AE         {-1} - - -model 1
node-AO         {3} node-AW leaf-AO
leaf-AH         {-1} - - -model 2
node-AW         {4} node-AX leaf-AW
leaf-AO         {-1} - - -model 3
node-AX         {5} node-AXR leaf-AX
leaf-AW         {-1} - - -model 4
node-AXR        {6} node-AY leaf-AXR
leaf-AX         {-1} - - -model 5
node-AY         {7} node-B leaf-AY
leaf-AXR        {-1} - - -model 6
node-B          {8} node-CH leaf-B
leaf-AY         {-1} - - -model 7
node-CH         {9} node-D leaf-CH
leaf-B          {-1} - - -model 8
node-D          {10} node-DD leaf-D
leaf-CH         {-1} - - -model 9
node-DD         {11} node-DH leaf-DD
leaf-D          {-1} - - -model 10
node-DH         {12} node-DX leaf-DH
leaf-DD         {-1} - - -model 11
node-DX         {13} node-EH leaf-DX
leaf-DH         {-1} - - -model 12
node-EH         {14} node-ER leaf-EH
leaf-DX         {-1} - - -model 13
node-ER         {15} node-EY leaf-ER
```

```
leaf-EH         {-1} - - -model 14
node-EY         {16} node-F leaf-EY
leaf-ER         {-1} - - -model 15
node-F          {17} node-G leaf-F
leaf-EY         {-1} - - -model 16
node-G          {18} node-HH leaf-G
leaf-F          {-1} - - -model 17
node-HH         {19} node-IH leaf-HH
leaf-G          {-1} - - -model 18
node-IH         {20} node-IX leaf-IH
leaf-HH         {-1} - - -model 19
node-IX         {21} node-IY leaf-IX
leaf-IH         {-1} - - -model 20
node-IY         {22} node-JH leaf-IY
leaf-IX         {-1} - - -model 21
node-JH         {23} node-K leaf-JH
leaf-IY         {-1} - - -model 22
node-K          {24} node-L leaf-K
leaf-JH         {-1} - - -model 23
node-L          {25} node-M leaf-L
leaf-K          {-1} - - -model 24
node-M          {26} node-N leaf-M
leaf-L          {-1} - - -model 25
node-N          {27} node-NG leaf-N
leaf-M          {-1} - - -model 26
node-NG         {28} node-OW leaf-NG
leaf-N          {-1} - - -model 27
node-OW         {29} node-P leaf-OW
leaf-NG         {-1} - - -model 28
node-P          {30} node-R leaf-P
leaf-OW         {-1} - - -model 29
node-R          {31} node-S leaf-R
leaf-P          {-1} - - -model 30
node-S          {32} node-SH leaf-S
leaf-R          {-1} - - -model 31
node-SH         {33} node-T leaf-SH
leaf-S          {-1} - - -model 32
node-T          {34} node-TD leaf-T
leaf-SH         {-1} - - -model 33
node-TD         {35} node-TH leaf-TD
leaf-T          {-1} - - -model 34
node-TH         {36} node-UH leaf-TH
leaf-TD         {-1} - - -model 35
node-UH         {37} node-UW leaf-UH
leaf-TH         {-1} - - -model 36
node-UW         {38} node-V leaf-UW
leaf-UH         {-1} - - -model 37
node-V          {39} node-W leaf-V
leaf-UW         {-1} - - -model 38
node-W          {40} node-Y leaf-W
leaf-V          {-1} - - -model 39
```

```
node-Y          {41} node-Z leaf-Y
leaf-W          {-1} - - -model 40
node-Z          {42} -      leaf-Z
leaf-Y          {-1} - - -model 41
leaf-Z          {-1} - - -model 42
```

Assuming that the above specification is stored in a file named `phoneRegTree.txt`, we can read this file in with the following command:

```
% Reading specification of properties from file 'phoneRegTree.txt'
phoneRegTree read phoneRegTree.txt
```

The regression tree above, combined with the property set, as defined in Section B.1.2, and the question set, as defined in Section B.1.1, merely implements a bijective mapping from a phone index to a prosodic model index. However, the regression tree, question and property sets allow to use maximum likelihood clustering techniques to develop more accurate prosodic models which can, in addition, be context sensitive. For the theoretical background see Section 3.2.3 and for implementational issues consult the online documentation.

B.2 Prosodic Hierarchy Putting Everything Together

It is convenient to be able to refer to all prosodic models corresponding to the same boundary tag, for example, all prosodic models instantitiated be the `phoneFinal` tag. In addition, note that we are left with several loose ends. For example, prosodic models are not yet linked with a regression tree, a dictionary, a property set or a prosodic feature set. To wrap up these loose ends we introduce the concept of a prosodic hierarchy into JPTk.

A prosodic hierarchy needs access to several JANUS objects which therefore have to be provided as additional arguments when we create a hierarchy:

```
% Creating a prosodic hierarchy
Hierarchy prosodicHierarchy dictionary prosodicFeatDesc
```

To add prosodic models – corresponding to the same segment boundary tag – to this hierarchy we specify:

```
% adding a prosodic level to the prosodic hierarchy
prosodicHierarchy add phoneModels phoneFinal phoneRegTree \
                      phonePropDesc
```

The argument following the function name `add`, i.e. `phoneModels`, refers to the name which will allow the reference to all prosodic phoneModels. The remaining arguments refer to the segment

final tag, the corresponding regression tree, and the property set, respectively. Thus, we can find the segment final tag in the tag set of the dictionary used, presumably in that very dictionary, the prosodic feature description, and, finally, in the `add` message above.

We are still left with the specification of prosodic models. The specification of a prosodic model comprises the model name, the number of prosodic features and the number of mixtures to model these prosodic features, followed by a list of the model parameters (weight, variance, and mean). For example, let us assume we want to train prosodic phone models. Because we have yet no model parameters we initialize everything to zero:

```
AA  2 1 {0.0 0.0 0.0 0.0 0.0}
AE  2 1 {0.0 0.0 0.0 0.0 0.0}
AH  2 1 {0.0 0.0 0.0 0.0 0.0}
AO  2 1 {0.0 0.0 0.0 0.0 0.0}
AW  2 1 {0.0 0.0 0.0 0.0 0.0}
AX  2 1 {0.0 0.0 0.0 0.0 0.0}
AXR 2 1 {0.0 0.0 0.0 0.0 0.0}
AY  2 1 {0.0 0.0 0.0 0.0 0.0}
B   2 1 {0.0 0.0 0.0 0.0 0.0}
CH  2 1 {0.0 0.0 0.0 0.0 0.0}
D   2 1 {0.0 0.0 0.0 0.0 0.0}
DD  2 1 {0.0 0.0 0.0 0.0 0.0}
DH  2 1 {0.0 0.0 0.0 0.0 0.0}
DX  2 1 {0.0 0.0 0.0 0.0 0.0}
EH  2 1 {0.0 0.0 0.0 0.0 0.0}
ER  2 1 {0.0 0.0 0.0 0.0 0.0}
EY  2 1 {0.0 0.0 0.0 0.0 0.0}
F   2 1 {0.0 0.0 0.0 0.0 0.0}
G   2 1 {0.0 0.0 0.0 0.0 0.0}
HH  2 1 {0.0 0.0 0.0 0.0 0.0}
IH  2 1 {0.0 0.0 0.0 0.0 0.0}
IX  2 1 {0.0 0.0 0.0 0.0 0.0}
IY  2 1 {0.0 0.0 0.0 0.0 0.0}
JH  2 1 {0.0 0.0 0.0 0.0 0.0}
K   2 1 {0.0 0.0 0.0 0.0 0.0}
L   2 1 {0.0 0.0 0.0 0.0 0.0}
M   2 1 {0.0 0.0 0.0 0.0 0.0}
N   2 1 {0.0 0.0 0.0 0.0 0.0}
NG  2 1 {0.0 0.0 0.0 0.0 0.0}
OW  2 1 {0.0 0.0 0.0 0.0 0.0}
P   2 1 {0.0 0.0 0.0 0.0 0.0}
R   2 1 {0.0 0.0 0.0 0.0 0.0}
S   2 1 {0.0 0.0 0.0 0.0 0.0}
SH  2 1 {0.0 0.0 0.0 0.0 0.0}
T   2 1 {0.0 0.0 0.0 0.0 0.0}
TD  2 1 {0.0 0.0 0.0 0.0 0.0}
TH  2 1 {0.0 0.0 0.0 0.0 0.0}
UH  2 1 {0.0 0.0 0.0 0.0 0.0}
UW  2 1 {0.0 0.0 0.0 0.0 0.0}
V   2 1 {0.0 0.0 0.0 0.0 0.0}
W   2 1 {0.0 0.0 0.0 0.0 0.0}
```

```
Y   2 1 {0.0 0.0 0.0 0.0 0.0}
Z   2 1 {0.0 0.0 0.0 0.0 0.0}
```

To complete the example, we still have to read in the prosodic model specification. Assuming that we saved the above prosodic model specification in a file name `phoneModels.txt`, we can write:

```
% read in prosodic phone models
prosodicHierarchy:phoneModels read phoneModels.txt
```

The linear order of prosodic models within the file matters. A leaf node of a regression tree specifies with its `-model` argument the index of the respective prosodic model and this index refers to a model's position within the list of all prosodic models.

The estimation of these model parameters takes place when we train prosodic models. In our example, the first parameter will be based on the distribution of the fundamental frequency (`PITCH`), the second parameter on the distribution of energy (`POWER`) in the speech samples within the training set. Remember, both features were defined in the prosodic feature description for these prosodic segments.

Note that in this example, the prosodic hierarchy comprises only the phone level. Additional levels such as a syllable or a word level can be added accordingly.

This concludes the initialization of prosodic models within the prosodic toolkit of Janus. The following sections, first, describe briefly how prosodic models can be trained, and, second, point to possible applications.

B.3 Training Prosodic Models

Training of prosodic models is based on a forced alignment of the speech signal with a text transcription of what was said in the respective utterance. Based on this alignment we know the start and end times of prosodic segments and we can compute and accumulate the corresponing prosodic features. For maximum-likelihood training we then compute mean and variance of these prosodic features to estimate the prosodic model's parameter.

Note that a hidden Markov model, `HMM`, requires information about prosodic segment boundaries. Thus, we have to provide the prosodic hierarchy as an additional argument when we create an `HMM`:

```
% Creating a hidden Markov model with additional information
% about prosodic segment boundaries
HMM hmm dictionary -hierarchy prosodicHierarchy
```

We start the training of prosodic models by creating accumulators. By referring to the respective prosodic level within the hierarchy we can achieve this by:

```
prosodicHierarchy:phoneModels createAccus
```

After each forced alignment we accumulate based on the resulting alignment path. We have to provide this alignment path, `path`, as an additional argument to the accumulator function `accu`:

```
prosodicHierarchy:phoneModels accu path hmm
```

We compute the alignment path and accumulate prosodic features for each utterance in the training set. To estimate the model parameter, we compute mean and variance for each prosodic feature for each prosodic model with the `update` function:

```
prosodicHierarchy:phoneLevel update
```

We might want to save the accumulators for further processing. For sure, we want to save the prosodic models to a file.

```
% Saving accus to file 'phoneLevelAccus.txt'
prosodicHierarchy:phoneModels save  phoneLevelAccus.txt
% Saving prosodic phone models to file 'phoneLevelModels.txt'
prosodicHierarchy:phoneModels write phoneLevelModels.txt
```

Note that the accuracy of the alignment has an immediate impact on the accuracy on the parameter estimation of the prosodic models. Thus, if the alignment is based on an inaccurate speech recognition system or an inaccurate transcription the resulting prosodic models will be flawed.

As shown above, we can supply the hidden Markov model with prosodic information by using the `-hierarchy` argument. Once you have estimated prosodic models you can use this prosodic models to participate in the computation of the alignment path, meaning the path will be based on acoustic and prosodic model information (or just prosodic information!). The corresponding command is:

```
path viterbiProsody hmm
```

Given trained prosodic models we can compute the log-likelihood of the signal with the function `score`:

```
prosodicHierarchy:phoneLevel score path
```

If the prosodic hierarchy comprises more than one level we use

```
prosodicHierarchy score path hmm
```

to compute the combined likelihood.

B.4 Using Prosodic Models

Due to the open design of the prosodic objects, prosodic models can be used in several ways:

- **Yes/no-question:** Using for example, the `linReg` function within the prosodic feature set to determine the final rise of the fundamental frequency, one can train prosodic models which discriminate between yes/no-questions and statements.
- **Segmentation:** Syntactic boundaries are often marked prosodically, for example by, silence, lowering of the fundamental frequency and intensity, or a preboundary lengthening. All this information can easily be extracted by the functions within prosodic feature set.
- **Stress detection:** A stressed word within a utterance is marked by an increased fundamental frequency, intensity, or lengthening when compared to other words within the utterance. The prosodic feature description set allows access to all these features.
- **Speech recognition:** Prosodic models can be used in two ways within a speech recognition system. First, prosodic models can be used within the Viterbi to find the most likely path for a given utterance with the speech signal. Thus the training of acoustic models is moderated by prosodic information. Second, prosodic models can be used within the lattice rescoring process within the decoding step.
- **Emotion detection:** As demonstrated in this investigation, prosodic models can be used to model emotion specific prosodic cues.

Using Tcl/Tk all this functionality can be implemented without ever consulting or changing the underlying C-code.

Appendix A

The Woggles Corpus

Are you angry at me?
Are you happy?
Are you my friend?
Are you talking to me?
Be my friend, Shrimp.
Bear, don't defend Shrimp.
Can I help you?
Don't be angry.
Don't be scared of Wolf.
Follow me.
Go through the chute.
Go to sleep now.
Hey Wolf, you wouldn't want to play, would you?
How come you are moping?
I am angry at Wolf now.
I am attacking Wolf because he is attacking me.
I am not happy.
I am not playing with you because I'm scared of you.
I am playing with you because I like you.
I am scared of Bear since he attacked me.
I can't talk now, I am going to dance on the pedestals.
I don't think I want to play follow the leader with you.
I don't understand what's going on.
I like to go through the chute.
I never get tired.
I want to be friends.
I want to see you go through the chute again.
I want to sleep.
I would love to play follow the leader.
I'm angry because you won't play.
I'm not angry at you.
I'm not moping.
I'm sorry. Don't be sad.

```
Is there a problem?
Play follow the leader with me.
Please go away from me.
Please help me stop Wolf from attacking me.
Shrimp is a good friend of mine.
Talk to me.
Try to be happy.
Want to dance on the pedestals with me?
What do you want?
Who are you talking to?
Why are you attacking Shrimp?
Why don't you lead?
Wolf, don't attack Shrimp.
Would you like to play follow the leader?
Yes, I'm talking to you.
You look sad; is there anything I can do to help?
Shrimp, don't attack Wolf.
```

References

Amir, N. and S. Ron. 1998. Towards an automatic classification of emotions in speech. In *Proc. of ICLSP*, Sydney.

Bakis, R. 1976. Continuous speech word recognition via centisecond acoustic states. In *Proc. ASA Meeting*, Washington, DC, April.

Ball, G. and J. Breese. 1998. Emotion and personality in a conversational character. In *Proc. of wecc-98, Workshop on Embodied Conversational Characters*, Tahoe City, California. AAAI, ACM/SIGCHI.

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

Barry, W.J. 1981. Prosodic functions revisited again! *Phonetica*, 28:320–340.

Beckman, M. 1986. *Stress and Non-stress Accent.* Dordrecht: Foris Publication.

Bezooijen, R. Van. 1984. *The chararcteristic and recognizability of vocal expressions emotions.* Dordrecht: .

Bonebright, T.L., J.L. Thompson, and D.W. Leger. 1996. Gender stereotypes in the expression and perception of vocal affect. *Sex Roles*, 34:429–445.

Breiman, L., J.H. Friedman, R.A. Ohlsen, and C.J. Stone. 1984. *Classification and Regression Trees.* Chapman & Hall.

Cahn, J. E. 1990. Generating expression in synthesized speech. Technical report, Speech Reserch Group, Media Laboratory, MIT.

Cairns, D.A. and J.H.L. Hansen. 1994. Nonlinear analysis and classification of speech under stressed conditions. *Journal of the Acoustical Society of America*, 96(6):3392–3400.

Campbell, N. 1995. Prosodic influence of segmental quality. In *Proc. of European Conf. on Speech Communication and Technology*, pages 1011–1014, Madrid.

Cassell, J., Bickmore M, M. Billinghurst, L. Cambbell, K. Chang, Vihjálmsson H, and Yan H. 1998. An architecture for embodied conversational characters. In *Proc. of wecc-98, Workshop on Embodied Conversational Characters*, Tahoe City, California. AAAI, ACM/SIGCHI.

Cassell, J., C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. 1994. ANIMATED CONVERSATION: Rule-based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. In *Proceedings of SIGGRAPH '94.*

Chen, L.S., T.S. Huang, T. Miyasato, and R. Nakatsu. 1998. Multimondal human emotion-expression recognition. In *Proc. of International Conference on the Automatic Face and Gesture Recognition*, pages 366–371, Nara. Japan. IEEEE Conmputer Society.

Chung, G. and S. Seneff. 1997. Hierarchical duration modelling for speech recognition using the ANGIE framework. In *Proc. of Eurospeech*, pages 1475–1478, Rhodes, Greece. ESCA.

Clark, E. 1990. On the pragmatics of contrast. *Journal of Child Language*, 17:417–431.

Cowie, R. and E. Douglas-Cowie. 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Proc. of ICLSP*, pages 1989–1992.

Crystal, T.H. and A.S. House. 1988. Segemental durations in connected-speech signals. *Journal of the Acoustical Society of America*, 835:1553–1573.

Darwin, C. 1998. *The Expression of the Emotions in Man and Animals.* Third edition. Oxford University Press.

Davis, P.J., S.P. Zhang, A. Winkworth, and R. Bandler. 1996. Neural control of vocalization: Respiratory and emotional influences. *Journal of Voice*, 10(1):23–38.

Davis, S.B. and P. Mermelstein. 1990. Comparison of parametric representations for monosyllabic word recognition in continously spoken sentences. In A. Waibel and K.F. Lee, editors, *Readings in Speech Recognition.* Morgan Kaufmann, pages 65–74.

Davitz, J.R. 1964. Auditory correlates of vocal expressions of emotional meanings. In J.R. Davitz, editor, *The Communication of Emotional Meaning.* McGraw-Hill, New York, pages 101–112.

Davitz, J.R. 1969. *The Language of Emotion.* New York: Acacdemic Press.

de Saussure, F. 1816. *Cours de linguistique general.* Payot.

Dellaert, F., T.S. Polzin, and A. Waibel. 1996. Recognizing emotions in speech. In *Proc. ICSLP*, Philadelphia PA, USA.

Disner, S.F. 1982. Stress evaluation and voice lie detection: A review. *UCLA Working Papers in Phonetics*, 54:78–92.

Fairbanks, G. and L.W. Hoaglin. 1941. An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monographs*, 8:85–90.

Fay, P. and W. Middleton. 1941. The ability to judge truth-telling, or lying, from the voice as transmitted over a public address system. *Journal of General Psychology*, 24:211–215.

Fiehler, R. 1990a. Emotionen und Konzeptualisierungen des Kommunikationsprozesses. *Grazer Linguistische Studien 33/34 Spache: Emotion*, pages 63–74.

Fiehler, R. 1990b. *Kommuikation und Emotion. Theoretische und empirische Untersuchungen zur Rolle der Emotionen in der verbalen Interaktion.* Berlin: de Gruyter.

Finke, M. 1999. Personal communication.

Finke, M., M. Lapara, A. Lavie, L. Levin, L. Mayfield Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. 1998. Clarity: Inferring discourse structure from speech. In *Proc. of the AAAI 98 Spring Symposium.*

Fonagy, I. 1978. A new method of investigating the perception of prosodic features. *Language and Speech*, 21:34–49.

Frick, R. 1985. Communicating emotion. The role of prosodic features. *Psychological Bulletin*, 97(3):412–429.

Fudge, E. 1970. Phonological structures and "expressiveness". *Journal of Linguistics*, 6:161–188.

Gales, M.J.F. 1996. The generation and use of regression class trees for mllr adaption. Technical Report CUED/F-INFENG/TR 263, Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England, August.

Givon, T. 1991. Markedness in grammar: Distributional, communicative and cognitive correlates of syntatic structure. *Studies in Language*, 15(2):335–370.

Greasley, P., J. Setter, M. Waterman, C. Sherrad, P. Roach, S. Arnfield, and D. Horton. 1995. Representation of prosodic and emotional features in a spoken language database. In *Proceedings of the XIII International Congress of Phonetic Sciences*, Stockholm.

Griffin, G. and C. Williams. 1987. The effects of different levels of task complexity on three vocal measures. *Aviation Space and Environmental Medicine*, 58(1165-1170).

Gupta, V., M. Lenning, P. Mermelstein, P. Kenny, P.F. Seitz, and D.O. Shaughnessy. 1992. Use of mimimum duration and ernergy contour for phonemes to improve large vocabulary isolated-word recognition. *Computer Speech and Language*, 6:345–359.

Hansen, J.H.L. 1992. Morphological constrained feature enhancement with adaptive cepstral compensation (mce-acc) for speech recognition in noise and Lombard effect. Technical Report DSPL-92-5, Department of Electrical Engineering, Duke University, Durham, North Carolina 277708-0291.

Hansen, J.H.L. and M.A. Clements. 1993. Source generator equailization and enhancement of spectral properties for robust speech recognition in noise and stress. Technical Report DSPL-93-2, Department of Electrical Engineering, Duke University, Durham, North Carolina 277708-0291.

Hansen, J.H.L. and B.D. Womack. 1996. Feature analysis and neural network-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processining*, 4(4):307–313.

Healey, J., J. Seger, and R. W. Picard. 1999. Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural situations. Technical Report 483, M.I.T. Media Laboratory Perceptual Computing Section.

Hess, W. 1983. *Pitch Determination of Speech Signals : Algorithms and Devices*. Springer series in information sciences. Berlin; New York: Springer-Verlag.

Heuft, B., T. Portele, and M. Rauth. 1996. Emotions in time domain synthesis. In *Proc. of ICSLP*, pages 1974–1977.

Hirschberg, J. and D. Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3).

Hollien, H., L. Geison, and J.W.Jr. Hicks. 1987. Voice stress evaluators and lie detection. *Journal of Forensic Sciences*, 32:405–418.

Hübler, A. 1998. *The Expressivity of Grammar*. Berlin, New York: Mouton de Gruyter.

Irvine, J.T. 1982. Language and affect: Some cross-cultural issues. In *Contemporary Perceptions of Language:Interdiscplinary Dimensions*. Georgetown Univeristy Press, Washington D.C., pages 31–47.

Isbister, K. and C. Nass. 1998. Personality in conversational characters: Building better digital interaction partners using knowledge about human personality preferences and perceptions. In *Proc. wecc-98, Workshop on Embodied Conversational Characters*, Tahoe City, California. AAAI, ACM/SIGCHI.

Jelinek, F. 1976. Continuous speech recogntition by statistical methods. In *Proc. IEEE*, April.

Jelinek, F. 1998. *Statistical Methods for Speech Recognition*. The MIT Press.

Johnstone, T. 1996. Emotional speech elicited using computer games. In *ICSLP*, pages 1985–1988.

Juang, B.H. and L.R. Rabiner. 1985. Mixture autoregressive hidden Markov models for speech signals. In *IEEE Trans. Acoustic, Speech, Signal. ASSP-33*, pages 1404–1413.

Kannan, A., M. Ostendorf, and J.R. Rohlicek. 1994. Maximum likelihhod clustering of Gaussians for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(3):453–455.

Katz, G.S. 1997. *A Quantitative Study of Vocal Acoustics in Emotional Expression.* Ph.D. thesis, University of Pittsbugh.

Katz, G.S., J.F. Cohn, and C.A. Moore. 1996. A combination of vocal f_0 dynamic and summary features discriminates between three pragmatic categories of infant directed speech. *Child Development*, 67.

Katz, S. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(4):400–401.

Kießling, A. 1997. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Aachen: Shaker Verlag.

Knapp, M.L. and J.A. Hall. 1997. *Nonverbal communication in human interaction.* Fort Worth: Harcourt Brace College Publishers.

Knower, F.H. 1941. Analysis of some experimental variations of simulated vocal expressions of the emotions. *Journal of Social Psychology*, 14:369–372.

Kompe, R. 1996. *Prosody in Speech Understanding Systems.* Ph.D. thesis, Technische Fakultät der Universität Erlangen-Nürnberg, Germany.

Ladd, D.R. 1983. Peak features and overall slope. In A. Cutler and D.R. Ladd, editors, *Prosody: Models and Measurements.* Springer, Berlin, pages 39–52.

Laukkanen, A.M., E. Vilkman, P. Alku, and H. Oksanen. 1996. Physical variations related to stress and emotional state: A preliminary study. *Journal of Phonetics*, 24:313–335.

Lavie, A., A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan. 1997. Janus III: Speech-to-speech translation in multiple languages. In *Proc. of ICASSP*, Munich.

Lazarus, Richard S. 1994. *Emotion and Adaptation.* Oxford University Press.

Lea, W. A. 1990. Prosodic aids to speech recognition. In W. A. Lea, editor, *Trends in Speech Recognition.* Prentice-Hall Inc., Englewood Cliffs, NJ, pages 166–205.

Legetter, C.J. and P.C. Woodland. 1994. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG/TR 181, Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England, June.

Lehiste, I. 1970. *Suprasegmentals.* Cambridge, MA: MIT-Press.

Levinson, S.E. 1986. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1(1):28–45, March.

Lieberman, P. and S.B. Michaels. 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America*, 34:922–927.

Mathesius, C. 1964. Verstärkung und Emphase. In J. Vachek, editor, *A Prague School Reader in Linguistics*. Indiana University Press, Bloomington, pages 426–432.

McCauley, L., B. Gholson, X. Hu, and A. Graesser. 1998. Delivering smooth tutorial dialogue using a talking head. In *Proc. of wecc-98, Workshop on Embodied Conversational Characters*, Tahoe City, California. AAAI, ACM/SIGCHI.

Medan, Y., E. Yair, and D. Chazan. 1991. Super resolution pitch discrimination of speech signals. *IEEE Transactions on Signal Processing*, 39(1):40–48, January.

Mendoza, E. and G. Carballo. 1998a. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice*, 12(263-273).

Mendoza, E. and G. Carballo. 1998b. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice*, 12:263–273.

Moriyama, T., H. Saito, and S. Ozwa. 1997. Evaluation of the relationship between emotional concepts and emotional parameters on speech. In *Proc. of ICASSP*, pages 1431–1434.

Mozziconacci, S.J.L. 1998. *Speech variability and emotion: Production and perception.* Ph.D. thesis, Eindhoven, The Netherlands.

Mozziconacci, S.J.L. and D.J. Hermes. 1999. Role of intonation patterns in conveying emotion in speech. In *Proc. of ICPhS99*, San Francisco, USA.

Murray, I.R. and J.L Arnott. 1993. Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America*, 2:1097–1108.

Murray, I.R. and J.L. Arnott. 1996. Synthesizing emotions in speech: Is it time to get excited? In *Proc. ICLSP*, pages 1816–1819, Philadelphia PA, USA.

Nakatsu, R. 1997. Image/speech processing that adopts an artistic approach - towards integration of art and technology. In *Proc. of ICASSP*, pages 207–210.

Olveres, J., M. Billinghurst, J. Savage, and A. Holden. 1998. Intelligent expressive avatars. In *Proc. of wecc-98, Workshop on Embodied Conversational Characters*, Tahoe City, California. AAAI, ACM/SIGCHI.

Ortony, A. and T. Turner. 1990. What is basic about basic emotions? *Psychological Review*, 97:315–331.

Ostendorf, M., V. Digalakis, and O.A. Kimball. 1997. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transacations on Speech and Audio Processing.*

Ousterhout, John K. 1994. *Tcl and the Tk Toolkit.* Addison-Wesley.

Paul Ekman, Richard J. Davidson (Editor). 1994. *The Nature of Emotion : Fundamental Questions.* Series in Affective Science. Oxford University Press.

Picard, R. W. and J. Healey. 1997. Affective wearables. Technical Report 432, M.I.T. Media Laboratory Perceptual Computing Section.

Protopapas, A. and P. Lieberman. 1997. Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America*, 101(4):2267–2277.

Rabiner, L. and B.-H. Juang. 1993. *Fundamentals of Speech Recognition.* Signal Processing Series. Upper Saddle River, New Jersy 07458: Prentice Hall Inc.

Rabiner, L. and R. Schafer. 1978. *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall Inc.

Reeves, B. and C. Nass. 1996. *The Media Equation.* Cambridge University Press.

Ries, K. 1997. A class based approach to domain adaptation and constraint integration for empirical m-gram models. In *Proc. of Eurospeech.*

Riseberg, J., J. Klein, R. Fernandez, and R. W. Picard. 1997. Frustrating the user on purpose: Using biosignals in a pilot study to detect the user's emotional state. Technical Report 458, M.I.T. Media Laboratory Perceptual Computing Section.

Roessler, R. and J.W. Lester. 1976. Voice predicts affect during psychotherapy. *Journal of Nervous and Mental Disease*, 163:166–176.

Russel, M.J. and R.K. Moore. 1985. Explicit modeling of state occupancy in hidden markov models for automatic speech recognition. In *Proc. ICASSP.*

Sato, J. and S. Morshiama. 1996. Emotion modeling in speech production using emotion space. In *Proc. IEEE Int. Workshop on Robot and Human Communication*, pages 472–477, Tsukuba, Japan, Nov.

Schafer, R.W. and L.R. Rabiner. 1990. Digital representation of speech signals. In A. Waibel and K.F. Lee, editors, *Readings in Speech Recognition.* Morgan Kaufmann, pages 49–64.

Scheirer, J., R. Fernandez, and R.W. Picard. 1999. Expression glasses: A wearable device for facial expression recognition. Technical Report 484, M.I.T. Media Laboratory Perceptual Computing Section.

Scherer, K.R. 1971. Randomized splicing: A note on a simple technique for masking speech content. *Journal of Experimental Research in Personality*, 5.

Scherer, K.R. 1974. Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. In S. Weitz, editor, *Non-verbal Communication.* Oxford University Press, New York, pages 105–111.

Scherer, K.R. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99:143–165.

Scherer, K.R., R. Banse, H.G Wallbott, and T. Goldbeck. 1991. Vocal cues in emotion encoding and decoding. *Motivation & Emotion*, 2(15):123–148.

Scherer, K.R. and G. Bergmann. 1984. Vocal communication. *The German Journal of Psychology*, 8(1):57–90.

Scherer, K.R. and P. Ekman. 1984. *Approaches to emotion.* Hillsdale, N.J.: Erlbaum.

Scherer, K.R., D.R. Ladd, and K.E.A. Silverman. 1984. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustic Society of America*, 76:1346–1356.

Scherer, K.R. and J.S. Oshinsky. 1977. Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1:331–346.

Scherer, K.R., H.G. Wallbott, F.J. Tolkmitt, and G. Bergmann. 1985. *Die Stressreaktion: Physiologie und Verhalten.*

Secreset, B.G. and G.R. Doddington. 1983. An integrated pitch tracking algorithm for speech systems. In *Proc. of ICASSP.*

Seymore, K. and R. Rosenfeld. 1997a. Large-scale topic detection and language model adaptation. Technical Report Report CMU-CS-97-152, Carnegie Mellon University, June.

Seymore, K. and R. Rosenfeld. 1997b. Using story topics for language model adaptation. In *Proc. of Europseech.*

Siegman, A.W. and S. Boyle. 1993. Voices of fear and anxiety and sadness and depression - the effects of speech rate and loudness on fear and anxiety and sadness and depression. *Journal of Abnormal Psychology*, 102:430–437.

Simonov, P.V. and M.V.Frolov. 1973. Utilization of human voice for estimation of man's emotional stress and state of attention. *Aerospace Medicine*, 44:256–258.

Streeter, L.A., N.H. MacDonald, W. Apple, R.M. Krauss, and K.M. Galotti. 1983. Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, 73:1354–1360.

Sulc, J. 1977. To the problem of emotional changes in human voice. *Activas Nervosa Superior*, 19:215–216.

Tartter, V.C. and D. Braun. 1994. Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, 96:2101–2107.

Thymé-Gobbel, A.E. 1998. Emotion and style in Call Friend data. Technical report, Natural Speech Technologies, Inc., March.

Tischer, B. 1993. *Die vokale Kommunikation von Gefühlen.* Fortschritte der psychologischen Forschung, volume 18. Weinheim: Psychologie Verlags Union.

Tosa, N. and R. Nakatsu. 1996. Life-like communication agent - emotion sensing character "mic" and feeling character "muse". In *Proc. of Multimedia '96*, Hiroshima, Japan.

Uldall, E. 1960. Attitudinal meanings conveyed by intonation contours. *Language and Speech*, 3:223–234.

Utsuki, N. and N. Okamura. 1976. Relationship between emotional state and fundamental frequency of speech. Technical Report 16, Reports of the Aeromedical Labaratory, Japan Air Self-Defense Force.

Volek, B. 1987. *Emotive Signs in language and semantic functioning of derived nouns in Russian.* Amsterdam etc.: Bejamins.

Vroomen, J., R. Collier, and S. Mozzicanacci. 1993. Duration and intonation in emotional speech. In *Proc. of Eurospeech*, pages 577–580, Berlin, Germany.

Vyzas, E. and R.W. Picard. 1998. Affective pattern classification. Technical Report 473, M.I.T. Media Laboratory Perceptual Computing Section.

Vyzas, E. and R.W. Picard. 1999. Offline and online recognition of emotional expression from physiological data. Technical Report 488, M.I.T. Media Laboratory Perceptual Computing Section.

Waibel, A., M. Bett, M. Finke, and R.Stiefelhagen. 1998. Meeting Browser: Tracking and summarising meetings. In *Proc. of the DARPA Broadcast News Workshop.*

Walbott, H.G. and K.R. Scherer. 1986. Cues and channels in emotion recognition. *Journal of Personality and Social Psychologie*, 51:690–699.

Watson, D. and A. Tellegen. 1985. Towards a consensual structure of mood. *Psychological Bulletin*, 96(2):219–235.

Wightman, C.W., S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price. 1991. Segmental duration in the vicinity of prosodic phrase boundaries.

Womack, B.D. and J.H.L Hansen. 1995. Stress independent robust HMM speech recognition using neural network stress classification. In *Proc. of Eurospeech.*

Womack, B.D. and J.H.L Hansen. 1996. Improved speech recognition via speaker stress directed classification. In *Proc. of ICASSP.*

Zeppenfeld, T., M. Finke, K. Ries, M. Westphal, and A. Waibel. 1997. Recognition of conversational telephone speech using the Janus speech engine. In *Proc. of ICASSP*, Munich , Germany.

Zhan, P., M. Westphal, M. Finke, and A.Waibel. 1997. Speaker normalization and speaker adaptation - a combination for conversational speech recognition. In *Proc. of Eurospeech.*

Zwicker, E. and H. Fastl. 1990. *Psychoacoustics. Facts and Models.* Series in Information Science, volume 22. Berlin: Springer.

Zwicker, E. and R. Feldkeller. 1967. *Das Ohr als Nachrichtenempfänger.* Stuttgart: Hirzel.