Prosodic Cues For Emotion Recognition In Communicator Dialogs

Jeremy C. Ang
jca@icsi.berkeley.edu

December 11, 2002

Contents

1 Introduction

As we strive to make spoken language systems increasingly natural, it becomes clear that systems must recognize not only *what* words a person says, but also *how* the words are spoken—i.e. the user's emotion, as conveyed by speech *prosody*. Yoshimura states it well when he says, "it is essential that machines understand prosodic characteristics which imply a user's various attitude, emotion and intention beyond vocabulary [16]." Emotion recognition has direct consequences for a wide variety of applications, from games and educational software (e.g., to detect if users are enthusiastic or bored), to life-support systems (e.g., to detect panic), to commercial products (e.g., to detect if a user is angry and should be transferred to a human operator). This project focuses on the last type of application, specifically, on the detection of user frustration with a telephone-based dialog system interface. Although the focus is on frustration, note that the method used is general and could be extended to emotion detection involving any type of emotion or domain.

1.1 Previous Work

There has been a considerable amount of previous work in the area of characterizing and detecting emotion in speech [1, 2, 4, 6, 7, 9]. The current study differs from previous work in a number of ways. First, much of the previous work has studied *elicited* emotions. [4, 9] garnered data produced by a small number of actors who are simply instructed to convey the emotion when reading prepared sentences. Elicited data may be ideal for research in areas like descriptive linguistics and speech synthesis, which aim to characterize canonical emotions. For work in recognition of natural emotions across many different speakers, however, it is crucial to use naturally-occurring data. Unfortunately, naturally-occuring data can be difficult to gather in large quantities. Cowie [2] acknowledges these issues, saying "Pure emotion is difficult to study because it is relatively rare and short lived, and eliciting it presents ethical problems. That makes it easy to slip into using simulations as a surrogate, and their ecological validity is highly suspect. There are also warning signs that it may be difficult to generalize findings from that approach. The reason may be that in emotion, as elsewhere, using carefully selected evidence makes it possible to evade issues that rapidly become important in anything but idealized cases. An aspect of that problem is that categorical representations may apply well to archetypal emotions, but much less so elsewhere." This study utilizes a dataset containing a large number of different speakers engaged in a task that itself gives rise to emotion.

Second, previous work has often used methods that are not entirely automatic, assuming correct word transcriptions and features that rely on hand-marked data (such as corrected pitch tracks or specific measurement locations), or relied on very simple prosodic features (e.g., excluding durations) that did not require recognition output. This work is based on the output of a speech recognizer (free recognition, with forced alignment for comparison), and uses

prosodic features that are computed entirely automatically.

Third, unlike studies that examine either emotion or articulation style, or which confound the two, this work aims to determine the association between the two, by including hand-marked articulation style characteristics in the feature database. By including these characteristics (such as hyperarticulation, pausing, or raised-voice) along with the prosodic features, it can be determined which, if any, of the style characteristics are good predictors of emotion, and the relative predictive strength of such features as compared to pure prosodic measurements. That is, the methods for emotion detection are entirely automatic, but we can ask whether there would be added value for emotion detection if automatic detection of articulation style is possible.

A large amount of related work has also been done, however, in the area of *synthesizing* emotion. [10] gives a brief overview of recent emotion synthesis techniques. Two approaches are widely used: formant synthesis and concatenative synthesis. Formant synthesis uses knowledge of acoustic correlates of speech sounds to generate speech. The advantage of the formant synthesis technique is the flexibility of a large number of parameters that can be varied freely. However, speech generated using this technique often sounds mechanical and unnatural, having a "robot-like" quality.

In concatenative synthesis, pre-recorded speech units (e.g. diphones or triphones) are selected and concatenated, then post-processed before output [8]. This method gives far less control over the output of the voice, thus making emotional speech more difficult to generate. Nevertheless, concatenative synthesis outputs sound more human-like and natural.

Both methods use prosodic variation to generate emotional speech, because "global prosodic parameters are often treated as universal or near universal cues for emotion" [10]. Pitch levels, ranges, and slopes, speaking rate and energy, number and duration of pauses, voice quality, and articulation precision are all considered and utilized in various studies on emotion synthesis. These same prosodic features are researched in this study, as well as in other emotion recognition studies [2, 4, 6, 7, 9].

While the features used are similar, the tasks of emotion synthesis and emotion recognition are inherently different. Perhaps there is some overlap in synthesis and recognition in using prosodic cues for certain emotions. For example, anger can be produced or detected by increases in energy, pitch levels, and/or pitch ranges. Still, one cannot simply generalize from synthesis to recognition, mostly because synthesis systems (at least in initial stages of development) generate canonical or explicit emotions. In other words, having a perfectly acceptable emotion synthesizer would still be unlikely to address the large variation on emotional expression found in natural spoken language, which is a core issue for the emotion recognition task.

1.2 Project Overview

This project investigates the use of prosody for the detection of frustration and annoyance in natural human-computer dialog. In addition to prosodic features,

the contribution of language model information and articulation style are examined. The project uses a corpus of human-computer dialog developed under the DARPA Communicator Project, labeled by humans for emotional content and articulation style characteristics. Extracted features include duration, speaking rate, pause, pitch, energy, and spectral tilt features. Features relied on the outputs of a speech recognition system, and their effectiveness in the emotion recognition task is compared to those from forced alignment based on reference transcriptions. Experiments are conducted and evaluated using decision tree classifiers.

1.3 Roadmap to Sections

The following report describes the details of this project. Section 2 discusses the method used to set up the experiments of the project, whose results are in Section 3.

Section 2 discusses where the speech data were obtained, how much of it there was, and how it was processed before use. Additionally, the section describes the labeling of the data, the creation of data sets, and the generation of features.

Section 3 reports the results of experiments and discusses in detail what can be learned or concluded from those results. Section 4 revisits the goal of the project and the differences from previous work in the field. It also summarizes the results and highlights interesting findings. Finally, there is a brief discussion on future directions and extensions, as well as the impact of this work for the field.

2 Method

2.1 Speech Data and Processing

A large, multi-site research and evaluation corpus of human-computer dialog developed under the DARPA Communicator project was chosen for use in this project. The DARPA Communicator project objective is "to support rapid, cost-effective development of multi-modal speech-enabled dialog system with advanced conversational capabilities" [15]. Users called systems built by various sites and made air travel arrangments over the telephone. Although users were not "acting" out any instructed emotions, it is important to note that because users were not making real travel plans, the frequency of frustration was lower than it would have been in real life. For instance, a user is trying to go to Hawaii is unlikely to agree to go to Detroit, yet users often accepted similar changes in plans when the system failed in their original plans. (This observation was noted by the human labelers, who were surprised at the "calmness" which which users accepted system problems, and their willingness to put up with long error-laden interactions.) Thus, these data are less than realistic in terms of the distribution of emotional utterances, but it appears realistic in terms of quality of the speech when emotional utterances did occur. Note that unlike studies using acted emotions, in which emotions are frequent and often exaggerated, this aspect of the data makes the detection task only more difficult than could be expected for a database of real travel planning data.

The corpus data used in this project came from three sources: the University of Colorado (CU) Communicator system, the Carnegie Mellon (CMU) Communicator system, and data from a larger number of sites collected during the June 2000 Communicator evaluation and distributed by NIST. Short calls (calls with fewer than five user utterances) were omitted, because it is unlikely that in such a short exchange the user exhibited emotional responses. The amount of data used in this study and their collection periods are summarized in Table 1. All data were collected over the telephone and sampled at 8 kHz.

Corpus data originating from the NIST collection went through two stages of processing before being used. Much of these data contained long portions of silence before and/or after the user utterance. These long portions of silence slow down the labeling procedure considerably, since labelers have to wait through them while listening to the utterances.

Therefore, a simple energy thresholding scheme was devised to remove initial and final nonspeech regions, carefully chosen to ensure that no actual speech was eliminated. Specifically, the standard deviations of the speech waveforms were calculated every 5ms with 10ms windows, and an empirically determined threshold was used to cut the beginning and end silences in each utterance. Significant periods of beginning or ending silences occasionally remained, due to a click or other noise that surpassed the energy threshold. These remaining silences were found to be acceptable, as the effort required to cut them out without risking removal of actual speech was too great to be worthwhile. This processing reduced the total duration of the NIST data by about 33%.

Table 1: Communicator Data Statistics. Labeled Data account for all the data that were assigned emotion classifications (as described in Section 2.2.1). Used Data describe data used in experimentation (as mentioned in Section 2.4).

Labeled Data					
Source	Dialogs	Utterances	Total Time	Words	Time Period
CU	205	5619	6 h 31 m 26 s	25324	11/29-6/01
CMU	240	8765	2 h 36 m 11 s	12835	1/01-8/01
NIST	392	7515	5 h 28 m 21 s	20234	6/00
Total	837	21899	14 h 35 m 58 s	58393	

Used Data (Originally Agreed)					
				Words	
Source	Dialogs	Utterances	Total Time	Transcript	Recognition
CU	157	4245	3 h 1 m 58 s	11623	11725
CMU	155	3320	1 h 29 m 53 s	8413	9620
NIST	379	5622	4 h 0 m 14 s	16092	16174
Total	691	13187	8 h 32 m 5 s	36128	37519

Used Data (Consensus Version)					
				Words	
Source	Dialogs	Utterances	Total Time	Transcript	Recognition
CU	157	4727	3 h 24 m 12 s	12929	13130
CMU	155	3417	1 h 33 m 35 s	8673	9994
NIST	380	6827	4 h 52 m 42 s	19176	19386
Total	692	14971	9 h 50 m 29 s	40778	42510

After the silence removal stage, each call was volume equalized. Variations from the different sources when they recorded the data resulted in differences in call volumes. Since the original data's volumes varied considerably, human labelers had to adjust the volumes during playback to appropriate listening levels. This readjustment not only slowed productivity, but it also created a danger that the labelers could rate lower energy speech as less frustrated even when the low energy is due to the channel. Volumes were equalized by applying a gain factor to each call so that its standard devation approached an arbitrarily prescribed level. For speech data, equalizing the standard deviation approximately equalizes the average power, assuming the silence/speech ratios are similar.

2.2 Emotion Labeling

Labelers used portions of data from all three sites (CU, CMU, NIST). These three sites were chosen based on the amount of data available to minimize the number of different data and annotation formats that had to be processed (each source used its own conventions).

User utterances were labeled by five students (1 male, 4 female) from UC Berkeley. Because it is undesirable for judgments to rely on linguistics training, labelers came from different disciplines. The goal of the labeling was to come up with a small set of classes for emotion, and an orthogonal set of labels for marking of the articulation style. It was observed in this corpus that occurances of speaker frustration were unrelated to those of hyperarticulation and vice versa. Additionally, hyperarticulation appeared to be speaker specific. As noted in the introduction, this separation of emotion and style is an important aspect of this work, since it allows an assessment of the association between these two logically independent factors. Labeling was done using a modified version of the Rochester Dialog Annotation Tool (DAT) [3]; it displays full utterance transcripts and the labeling choices, allows sound files to be played and comments to be entered, and saves output to SGML.

2.2.1 Emotion Labels

Based on a number of preliminary studies examining labeling alternatives, the labelers came up with the following scheme for annotation. Every utterance was given one of seven possible emotion labels: NEUTRAL, ANNOYED, FRUSTRATED, TIRED, AMUSED, OTHER, or NOT-APPLICABLE (contained no speech data from the user). NEUTRAL was used for utterances that displayed no particular emotion. Recall that a neutral utterance (as well as any other emotion) could be said in either a natural or a hyperarticulated (robotic or "Star Trek") style. ANNOYED was used for utterances displaying any level of perceptible agitation or impatience, relative to NEUTRAL. FRUSTRATED was used for more extreme forms of annoyance or anger. While ANNOYED and FRUSTRATED could have been grouped together into one class (and indeed this is the grouping used for many of the analyses to follow), at the labeling stage both

Table 2: Frequency of emotion labels. NOT-APPLICABLE cases are waveforms with no user speech; these are excluded in the analyses. Note that low rate of frustration overall is attributable to the fact that users were not making real travel plans, as discussed in the text.

Emotion Class	Instances	Percent
NEUTRAL	41545	83.84%
ANNOYED	3777	7.62%
FRUSTRATED	358	0.72%
TIRED	328	0.66%
AMUSED	326	0.66%
OTHER	115	0.23%
NOT-APPLICABLE	3104	6.26%
Total	49553	100.0%

classes were kept, thus enabling an investigation of detection of FRUSTRATION only, albeit with significantly fewer datapoints (see Table 2). During labeling it was found that some utterances displayed emotional characteristics that did not fit into the "continuum of annoyance" (annoyed or extremely annoyed, i.e. frustrated). These fell into two main classes, which were termed TIRED and AMUSED. In TIRED utterances, the caller sounded apathetic or dejected due to the system interaction, often sighing. This class understandably occurred in extremely long calls, where the user tired of the system but was not angry enough to hang up. AMUSED utterances were often accompanied by laughter, and occurred either when the system unexpectedly correctly recognized the user's input (i.e. the user already had low expectations), or when the system incorrectly responded with a mistake that amused the user. A small set of remaining utterances did not fall into any of the above emotion categories, but were dissimilar to each other; these cases were simply marked as OTHER. Finally, a non-negligible portion of the utterances in the corpus contained no speech from the user, due to problems in the collection. These were labeled as NOT-APPLICABLE and removed them from the analyses.[1]

A total of 49,553 emotion classifications were made on 21,899 utterances from the NIST, CU, and CMU recordings, for an average of 2.26 labelers labeling each utterance. The breakdown of class frequencies is shown in Table 2.

In addition to emotion, each utterance was also labeled for more information, including three important types: articulation style, repeated requests or explicit corrections, and data quality problems. Table 3 lists the additional annotation options, while below the labels found to be important in this study are further described. A full description of all the labels can be found in the Appendix,

[1]The NA waveforms were excluded from the speech recognition experiments, for consistency with results using forced alignment recognition (see section 2.3.) While this is "cheating" since such waveforms could lead to insertions, the problem of empty utterances is considered to be one that should be addressed in system design.

Table 3: Additional Annotation Options

Dialog-Level		
Non-Native Speaker	Global Comment	Accent Type

Utterance-Level		
Repeat-Correction	Data Problems	Self-Talk
Hyperarticulation	Spelling Out	Comment
Pauses Between Words	Final Pitch Rise	
Pauses Between Syllables	Raised Voice	

Section 6.1.

2.2.2 Emotion Label Mapping

Since utterances were labeled by more than one labeler, it was necessary to create a mapping to reduce the multiple labels per utterance into one label used for experiments. This process involved two levels of mapping. The first mapping used three simple rules for reduction. If there was a label that made up the majority, the multiple labels were mapped to that label. If there was a tie between two labels, e.g. 1 to 1 or 2 to 2, the labels were reduced to a 1 to 1 tie and sent to the second level of mapping. The third rule was that ANNOYED and FRUSTRATED were considered on a continuum of annoyance, so an utterance with three labels of NEUTRAL, ANNOYED, and FRUSTRATED would map to ANNOYED. All other combinations of labels were sent to the second level of mapping.

In the second level, certain labels "dominated" other labels. For instance, NOT-APPLICABLE labels dominated all other labels. Therefore, if an utterance had a NOT-APPLICABLE label and a NEUTRAL label, the utterance was mapped to NOT-APPLICABLE. The same is true for the pair of NOT-APPLICABLE and any other label. The NONE label (which was the default label, meaning the utterance was never given a label by the labeler) was dominated by all other labels. As a result, when there was a NONE label and another label for an utterance, the utterance was mapped to the other label. Table 4 shows the order of dominance for all the labels. ANNOYED, FRUSTRATED, AMUSED, and TIRED had equal dominance.

Some utterances were marked with an "x" and ignored in the experiments. This occurred when an utterance had a combination of ANNOYED, FRUSTRATED, AMUSED, and/or TIRED labels (with no majority), or when an utterance only had NONE labels.

Using the "Originally Agreed" and "Consensus Version" data sets as described in 2.4, the distribution of emotion labels after applying both mappings are shown in Table 5.

Table 4: Emotion Label Dominance Order

Most Dominant	1. NOT-APPLICABLE
	2. OTHER
	3. ANNOYED,FRUSTRATED,AMUSED,TIRED
	4. NEUTRAL
Least Dominant	5. NONE

Table 5: Emotion Class Frequency For Experimental Data Sets

"Originally Agreed"						
	Training		Test		Total	
Emotion Class	Freq.	Percent	Freq.	Percent	Freq.	Percent
NEUTRAL	9307	95.36%	3308	96.53%	12615	95.66%
ANNOYED	367	3.76%	94	2.74%	461	3.50%
FRUSTRATED	35	0.36%	7	0.20%	42	0.32%
TIRED	14	0.14%	2	0.06%	16	0.12%
AMUSED	37	0.38%	15	0.44%	52	0.39%
OTHER	0	0.00%	1	0.03%	1	0.01%
TOTAL	9760	100.0%	3427	100.0%	13187	100.0%

"Consensus Version"						
	Training		Test		Total	
Emotion Class	Freq.	Percent	Freq.	Percent	Freq.	Percent
NEUTRAL	9869	88.16%	3442	91.15%	13311	88.91%
ANNOYED	1060	9.47%	276	7.31%	1336	8.92%
FRUSTRATED	125	1.12%	23	0.61%	148	0.99%
TIRED	78	0.70%	9	0.24%	87	0.58%
AMUSED	60	0.54%	24	0.64%	84	0.56%
OTHER	3	0.03%	2	0.05%	5	0.03%
TOTAL	11195	100.0%	3776	100.0%	14971	100.0%

2.2.3 Articulation Style Labels

The same group of five labelers also developed a method for marking articulation style. (While automatic detection of style is an interesting research area, it is not tackled in this project.) After an initial attempt at a single style category, it was obvious that "style" needed to be broken down into the component attributes that labelers were listening for. They arrived at the following binary categories, which unlike emotion labels are not mutually exclusive: hyperarticulation (exaggerated pronunciation of specific phones or syllables), pausing (between words or between syllables in a word), and "raised voice" (an increase perceived loudness or level of vocal effort). These characteristics were annotated at the same time as the emotions were labeled, but labelers were instructed to consider the emotion and style labels as independent.

2.2.4 Repeat-Correction ("REPCO") Labels

Since previous work has described the relationship between articulation style and system errors [5], it was clear that system errors should be included in a study of frustration. It was assumed that some dialog systems will have a good idea of the location of such errors (and indeed there is work on the topic, as noted in the just-cited reference). But since such information was not available in the corpus used, the labelers also labeled these events. The focus was on errors that resulted in a repeat and/or correction by the user. Based loosely on previous work by Kirchhoff [5], utterances were labeled as either not a repeat/correction, a "repeat-or-rephrase-only", a "repeat-or-rephrase-with-explicit-correction", or an "explicit-correction-only".

2.2.5 Data Quality Labels

For *data quality*, properties of the speaker (nonnative, speaker switches, system developer), properties of the speech content (side-talk, joking), and aspects of the recording (noise, system cut-offs) were marked. While joking and system cut-offs were included in the analyses, the other cases were omitted from the present study. In principle, it would be desireable to retain the nonnative speech, which was not infrequent in the CU corpus. But because such speakers (1) were difficult or impossible to judge hyperarticulation for; and (2) were *much* more tolerant of system failures than native speakers (as judged by the nonnatives' much longer calls and low level of frustration), the decision was made to omit them for the sake of data homogeneity. A detailed description of data exclusions is described later in Section 2.4.1.

2.2.6 Labeling Issues

It was found that labeling of emotion as well as articulation style is an inherently difficult task. First, emotion is conveyed on a continuous scale, and for purposes of this work there was a need to come up with discrete labels (alternative approaches such as additional classes or uncertainty labels, did not improve

interlabeler agreement). Second, emotion characteristics vary enormously from person to person, and from context to context. Thus, an issue that arose was whether to label emotion relative to the speaker and previous context, or to use an absolute labeling ignoring both of these factors. The former option was chosen, since that is the most relevant option given the application in mind (detect changes in the current user over the dialog). This also seemed to be warranted because few dialogs began with a frustrated user. Finally, most of the utterances were quite short, often just the word "Yes" or "No", making emotion and style difficult to judge.

2.2.7 Interlabeler Agreement

It is important to measure the reliability of the subjective judgements made by the labelers. If emotion labels made are not reliable (e.g. a computer randomly picks a label), then using the labels and corresponding data for training and testing becomes meaningless. However, if multiple labelers seem to agree consistently in the labels they choose, the confidence in the label increases and the emotion labels seem more reliable. The amount of agreement needed to deem labeled data reliable depends on the difficulty of the labeling task.

The Kappa statistic measures the level of agreement between an arbitrary number of labelers. It considers the proportion of times labelers agree, but also takes into account chance agreement. Essentially, Kappa is the ratio of the proportion of times that the labelers agree (corrected for chance agreement) to the maximum proportion of times that the labelers could agree (corrected for chance agreement), according to the formula:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where k=Kappa, P(A)=proportion of times that the labelers agree, and P(E)= proportion of times that we would expect the labelers to agree by chance.

It is important to know whether the calculated Kappa is greater than that expected by chance. One might expect that Kappa due to chance is 0. However, chance agreement is not constant, so there exists an expected Kappa due to chance. Thus, the significance statistic z is often used, which measures the likelihood a calculated Kappa occurs by chance, where:

$$z = \frac{k}{\sqrt{var(k)}} \tag{2}$$

For large N, the distribution of k is approximated by a Gaussian with zero mean and a variance (the details of which will not be discussed in this report) as described in [11]. See [11] for further reference on Kappa and significance.

Overall, Kappa is approximately 0.40 with a significance of 26.2 when considering all labeled data and all the emotion classes. When classes are grouped into ANNOYANCE+FRUSTRATION vs. ELSE, Kappa rises to 0.44 with significance 26.5. Kappa for FRUSTRATION vs. ELSE is 0.34 with significance 6.0.

Table 6: Interlabeler Agreement on all labeled data. "A+F" refers to grouping ANNOYED and FRUSTRATED together, whereas "F" refers to the FRUSTRATED utterances only.

	A+F vs. ELSE	F vs. ELSE
Each human with other human, overall	74.0	70.0
Human with human "Consensus" (biased)	84.1	79.9

2.2.8 "Original" and "Consensus" Labels

In a first pass, labelers annotated individually after calibration. The Kappa values above, along with a first-pass interlabeler agreement (even after grouping ANNOYED and FRUSTRATED together) of 74%, were deemed too low for the purposes of this project. However, it must be noted that it appears to be due to the task rather than to the labelers, because agreement among the various pairwise combinations of labelers did not significantly differ, and because agreement did not improve with additional training. Therefore, a second pass of labeling was conducted, which is referred to as "Consensus" labeling, where the two most experienced labelers together relabeled any utterances that original labelers had not agreed on. This was done in context; full calls were displayed and affected utterance were marked. The affected utterances were a small subset given the large number of original agreements on NEUTRAL labels. After consensus labeling, interlabeler agreement jumped to 84%. Table 6 summarizes the interlabeler agreement statisctics on all the labeled data.

It should also be noted that F vs. ELSE classification leads to a 4% reduction in interlabeler accuracy. This may be attributed to the rarity of the FRUSTRATED label (0.72% of all labels), making it less likely that labelers agree on a specific utterance.

2.3 Speech Recognition and Forced Alignment

Both the prosodic and language model features for the modeling used in this project relied on alignment information from a speech recognizer. Rather than use the recognition results from the various Communicator systems (which were not always available), a simplified version of SRI's Hub-5 system for conversational telephone speech [14] was run, using a class-based trigram language model developed for SRI's own Communicator system. This ensured that recognition errors and the specifics of the recognition system (such as the choice of pronunciations) affected data from all sites equally, and removed a potential variable of recognition performance from each site. The word error rates obtained with this system were 29.6% for CMU data, 27.8% for the CU data, and 24.9% for the NIST data (measured on the subset of utterances used in the experiments). To investigate the effects of recognition errors, features based on the reference transcriptions of the users' utterances were computed, via forced alignment to

Table 7: Data Quality labels excluded from analysis

user is a child
testing the system
hardly any speech in call
side talk
unintelligible speech/mumbling
intelligible but muffled/distorted speech
utterance(s) out of order
more than one person talking to system during call
fake accents/mispronunciations/possible speech impediments
transcript is incorrect
noisy/background noise

the waveforms.

2.4 Data Sets

2.4.1 Data Excluded

The labeled data (see Table 1) for the experiments described in Section 3 was organized by eliminating cases deemed as having poor data quality. In Section 2.2, the exclusion of nonnative utterances, speaker switches, system developer utterances, utterances with side-talk, and noisy utterances was mentioned. Table 7 details all the excluded data.

2.4.2 "Originally Agreed" Data Set

Two versions of the remaining labeled data were used to run experiments. "Originally agreed" data included only utterances where the labelers agreed on the emotion class in the first pass of labeling. The details are described in the middle portion of Table 1. Roughly 75% were used for training (9760 utterances); the remaining 25% were used for testing (3427 utterances). No dialogs were split between training and test sets.

2.4.3 "Consensus Version" Data Set

The "Consensus version" uses more data than the "Originally agreed" because it includes the utterances that went through a second pass of labeling by consensus labelers. The bottom portion of Table 1 describes the exact amount of data used for this version. Similar to in the "Originally agreed" data set, the data were split into training and test with a 3 to 1 ratio (11195 to 3776). The dialogs were split in the same way in both data sets.

2.5 Classifiers

Decision trees were used as classifiers, employing a brute-force iterative feature selection algorithm to find a minimal set of useful features and avoid the problem of greedy search. Because of the large skew in class sizes, the data were downsampled to equal class priors to allow the tree maximum sensitivity to features. This approach, when used in multiple experiments (varying the downsampling random seed each time), proved superior to not downsampling and also to upsampling.

Because of the fairly limited size of the emotional-utterance corpus, results were obtained as linear averages from 20 separate experiments per condition, each with a different random downsampling of the training data. This is needed due to the small data sizes for the emotion classes.

A detailed example of how 20 experiments were run and averaged to obtain the results is shown in the Appendix, Section 6.4.

2.6 Features

Three types of features were investigated in this work:

1. Prosodic features.
2. Language model features.
3. Other features.

These are described below, and detailed in Appendix Section 6.2.

2.6.1 Prosodic Features

The following types of prosodic features were extracted: duration and speaking rate features, pause features, pitch features, energy features, and spectral tilt features. These feature types are similar to those that have been investigated in previous work on emotion recognition. For example, [2] states that pitch, duration, intensity, and spectral makeup are relevant. [7] considers pitch contours and short-time average power envelopes. Speech power and pitch are examined by [9], among other features. [4] measures smoothed pitch features, derivatives of pitch, and rhythm features like speaking rate.

Duration Features *Duration features* included various statistics involving the average vowel duration in the utterance and the maximum and average durations of the normalized (for true or recognized phone identity) phones in the utterance. These features should help detect emotion by exploiting a correlation between a user's frustration with the computer system and their response by speaking slower so the "dumb" computer can understand. The maximum duration statistic could find frustration in certain emphasized words of an utterance, e.g. "Noooooooo, I want to go to Newark." The average statistics could detect a general slowdown in the utterance. The normalizations are calculated

through dividing the phone by the average duration for that phone throughout the Communicator data.

Speaking Rate Features *Speaking rate features* consisted of a syllable rate feature approximated by counting the number of vowels (for true or recognized words) and dividing by the duration of the utterance. This is an approximation assuming that syllables on average contain one vowel. Like the duration statistics, this feature hopes to capture the overall pace of an utterance, where a slowdown may indicate an emotional response.

Pause Features *Pause features* included the utterance "speech percentage," the duration of the longest pause, and the number of long pauses inside an utterance. Speech percentage was calculated by dividing the duration of all the speech in an utterance by the duration of speech plus duration of interior pauses. Long pauses were considered as pauses greater than 70ms in length. Lengthier pauses between words could also indicate an emotional response.

Pitch Features *Pitch features* were based on post-processed F0 output using a stylization and regularization algorithm based on an updated version of software from work by Sönmez et al. [13, 12].

Specifically, initial F0 data were gathered using the get_f0 utility from ESPS. These raw F0 data contain irregularities of the pitch tracker such as offshoots or pitch halving/doubling. Pitch halving/doubling is treated with a median filter after applying a lognormal tied-mixture to fit the frame-level pitch values [13]. Then, a stylized contour is obtained by fitting a piecewise linear model to the estimated pitch values over voiced regions. Refer to [13, 12] for more details.

Two versions of pitch features were used, one based on data from all utterances in a call, and one using only the first five utterances. The latter, which turned out to be nearly as good as the full-call version, allows for online emotion detection (especially since users are rarely frustrated during the first five utterances). Pitch features included raw and speaker-normalized minimum and maximum utterance pitch, as well as the maximum pitch taken within the region of the longest normalized vowel, and slope information at various locations. Higher pitch values could hint at an emotional response, just as in human to human arguments voices tend to be raised. Larger variation in pitch, tracked by higher pitch slopes, could also indicate emotion.

Energy Features *Energy features* recorded the average RMS energy during voiced frames, as well as the maximum and average RMS energy during the longest normalized vowel. Utterances of annoyance or frustration could possibly be recognized from higher RMS energies, as people tend to raise the volume of their voices when displaying these emotions.

Spectral Tilt Features *Spectral tilt features* attempted to find the spectral weighting, or "tilt," of speech. Spectral tilt features included the average of

the first cepstral coefficient, the average slope of the linear fit to the magnitude spectrum, and the average difference in the sum of log energies in low and high frequency regions—all taken over the longest normalized vowel.

2.6.2 Language Model Features

A class-based trigram model was trained from the words in each of the classes (using the same word classes as used in the recognizer), and computed log likelihoods according to the models for each of the test utterances. For convenience and to best assess the joint contribution of language model and prosodic features, the language model features were added to the prosodic decision trees. Two types of language model features were investigated, the *difference of log likelihoods of the two classes*, and the coarser feature of the *sign of the likelihood difference*.

The difference of log likelihoods of the two classes was heavily used by the decision trees, but led to poor results on the test data, clearly showing overfitting. This feature was eliminated in favor of a more coarse feature, the *sign* of the likelihood difference, which did not show overfitting problems.

2.6.3 Other Features

In addition, three other feature types were included: dialog position features, a feature recording repeated attempts and explicit corrections, and articulation style features.

Dialog Position Features *Dialog position features* were recorded in three ways: number of system prompts, number of utterance responses, and number of words before an utterance. These can be assumed to be automatically obtained by a system. The feature recording the number of words was based on true words, which is not automatic, but this feature was not used in the decision trees. Thus, having a feature based on recognized words would likely have similar results.

Repeat/Correction Feature The *Repeat/Correction feature* marked if the utterance was a repeat or rephrasing of a previous utterance, or whether there was an explicit correction of the system response to previous utterances. (Refer to Section 2.2.4.) This is of course nontrivial to obtain, but the detection of repeats and corrections is considered as a separate problem and one in which many systems already have some ability to detect.

Articulation Style Features *Articulation style features* were based on labels such as hyperarticulation, pausing between words or between syllables in a word, and "raised voice." These were described in Section 2.2.3.

3 Results and Discussion

The results of this work are reported and discussed in five sections. The first three involve emotion recognition performance by humans and by machine. These results are based on forced alignments (from true words) for feature processing. The next section looks at how performance is affected by features based on true words as opposed to those calculated from recognition outputs. Lastly, the details of feature usage are shown and the implications of the results are discussed.

Table 8 summarizes the experimental results. The columns of the table differentiate the type of emotion recognition task. The first column of results (including accuracies and efficiencies) are from the ANNOYED+FRUSTRATED versus ELSE task as described in Section 3.2, using features that relied on true words (forced alignment information based on reference transcripts). The second column records experimental results of the same task, but using features that relied on ASR output information. The third and fourth columns similarly report results, except they are from the FRUSTRATED versus ELSE task (described in Section 3.3). Results are given in both accuracy (percentage of correct decisions) and efficiency (reduction in class entropy provided by the model).

3.1 Prediction of Human Labels by Humans

The first two rows of Table 8 show interlabeler agreement for the versions of data as described in Section 2.4. The accuracies reported are based on a random selection of the pair of emotion labels per utterance, done in the following way. For the first row of accuracies, if an utterance was labeled by more than two labelers (a small minority of utterances), two emotion labels are randomly picked for that utterance and used for the interlabeler agreement calculation. The table reports the average of many instances of interlabeler agreement (enough for the average accuracy to settle down) calculated in this way. Similarly, for the "Consensus" human with human agreement, the consensus label was paired with a random selection of the original emotion labels. This pair was used for the interlabeler agreement calculation, and the results in the table reflect the average of many instances of this calculation.

We notice that with consensus, agreement increases approximately 10% from the baseline (83.9% from 72.6% for A+F vs. ELSE, 77.3% from 68.8% for F vs. ELSE). Additionally, we see that agreement in the F vs. ELSE case is lower by about 5% than in the A+F vs. ELSE case. This is consistent with the results using all the labeled data discussed in Section 2.2. The difference between the two results is that the ones reported in this section are based on the "Originally Agreed" and "Consensus Version" data sets, whereas those in Table 6 are based on all the labeled data.

Table 8: Summary of experimental results. "A" = Annoyed; "F" = Frustrated; "STYLE" = articulation style features; "REP" = repeat/correction features; "LM" = language model features; "Consensus version" = emotion labels arrived at after labelers resolved any disagreements; "Originally agreed" = subset of utterances on which individual labelers had agreed on first labeling pass; "Acc" = accuracy (linear average of 20 separate experiments); "Eff" = efficiency (linear average of 20 experiments). Note: LM features were computed for the first task only, although in principle could be computed for both. Accuracies reflect equal class distributions in the test set through downsampling.

	A+F vs. ELSE				F vs. ELSE			
	True words		ASR words		True words		ASR words	
	Acc	Eff	Acc	Eff	Acc	Eff	Acc	Eff
Human with other human, overall	72.6				68.8			
Human with human "Consensus" (biased)	83.9				77.3			
Consensus version, [All Features]	80.2	32.7			93.2	67.2		
Originally agreed, [All Features]	85.4	47.2			91.8	63.3		
Consensus version, [no STYLE] (baseline)	75.2	21.2	75.1	21.9	86.4	46.5	87.0	49.5
Originally agreed, [no STYLE]	80.0	32.0	78.5	28.2	86.4	44.6	85.7	46.9
Consensus version, [no STYLE, no REP]	71.1	14.6	70.7	14.8	84.2	39.7	86.7	47.9
Originally agreed, [no STYLE, no REP]	77.1	23.0	74.5	18.6	80.4	31.8	83.6	39.6
Consensus version, [REP *only*]	69.8	12.8			76.6	21.1		
Originally agreed, [REP *only*]	74.7	18.5			85.4	14.3		
Consensus version, [LM *only*]	65.6	3.8						
Originally agreed, [LM *only*]	64.5	-0.9						

3.2 Task 1: Prediction of Annoyance and Frustration versus Else by Machine

Experiments were run with two basic classification tasks. The first task involved classifying ANNOYANCE+FRUSTRATION versus ELSE. The ELSE class contained all remaining emotion types (NEUTRAL, plus the small amounts of other emotions such as TIRED, AMUSED, and OTHER, to account for all datapoints). Table 8 shows the results of this task under the "A+F vs. ELSE"

column, using both true words and recognized words. The different rows in the table show results for different experiment conditions, in which both the source of the predicted emotion labels and the features available to the decision tree were varied. In the "Consensus Version" experiments, the model predicted the labels resulting from the consensus labeling pass; in the "Originally Agreed" experiments, only the subset of utterances for which individual labelers had been in agreement on the first pass of labeling were included.

Looking at the ANNOYANCE+FRUSTRATION vs. ELSE experiments, we can draw several conclusions. First, we see that the baseline experiment (Consensus version, no STYLE features) with an accuracy of 75.2% shows better prediction of human consensus labels than individual human labelers do with each other (72.6%).

When the dialog state (repeat/correction) feature is excluded, the results are slightly worse (71.1% for tree versus 72.6% for human to human). Since the repeat/correction feature is the only feature used here that is not automatically obtained, this result shows that automatic emotion recognition for this task can perform comparably with humans. Section 3.5 discusses the implications of this further.

We also see that when considering only the utterances on which labelers originally agreed, performance consistently improves by 5–6% (except for the language model only experiment). This improvement is expected, since presumably labelers agreed on cases that were more clear-cut prosodically. In other words, these utterances showed emotions more explicitly.

The repeat/correction feature always increases performance, sometimes by up to 4%. Again this is expected, since users are typically more frustrated after system errors. Articulation style features increase performance relative to the baseline prosodic tree by about 5% from baseline. Potential candidates for this improvement include hyperarticulation, pauses, and raised-voice features.

When using the repeat/correction feature only while excluding all other features, performance drops about 6% from the baseline. When only using the language model features, accuracies drop by 15-20% from the baseline. Note the negative efficiency reported in the last row of Table 7. Since results reported are optimized for accuracy of the decision trees, small negative efficiencies are possible.

Figure 1 shows the ROC curves for the five experiments on the ANNOYED+FRUSTRATED vs. ELSE task using different feature sets available to the classifier. ROC curves plot the tradeoff between false alarms and false rejections by plotting false alarm rate in the x-axis against the rate of correct detection in the y-axis. A point on the ROC curve is found by setting a decision threshold for the two-choice classification task and determining the amount of false alarms and correct detections. The curve is then obtained by sweeping across different thresholds.

An ideal ROC curve would hit the upper left corner of the graph, where there is perfect detection with no false alarms. Therefore, the higher and farther to the left a curve is, the better. Examining Figure 1, we see that the best curve is obtained from the experiment where all features are available (the purple

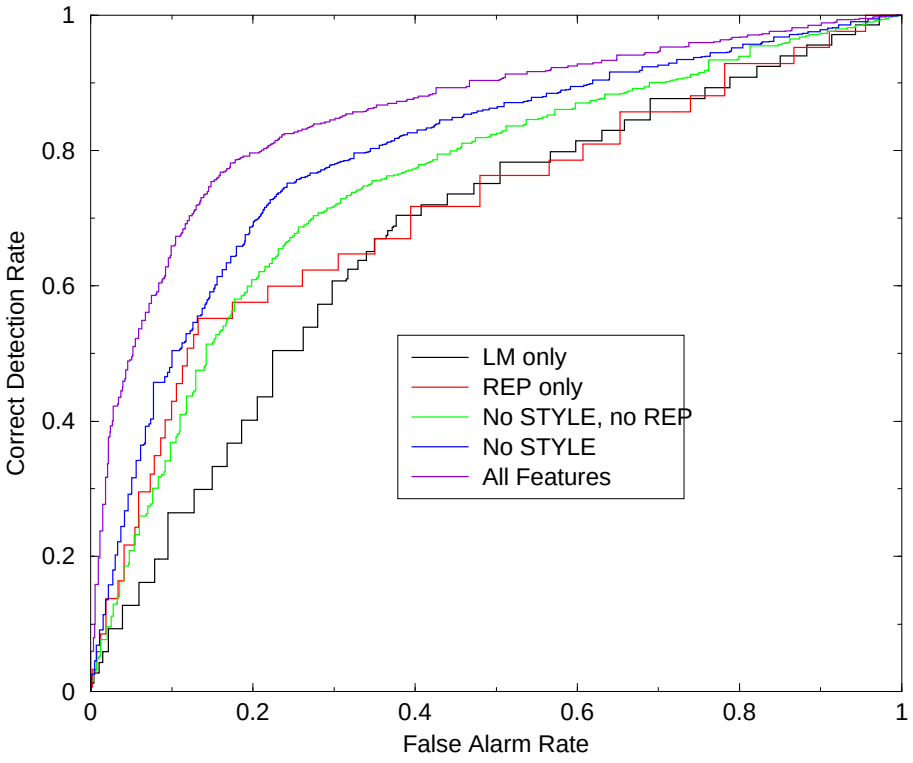

Figure 1: ROC curves of ANNOYED+FRUSTRATED vs. ELSE task using different feature sets.

line). In general, the curves show the trend seen in Table 8, where the order of performance (from better to worse) is using all features, using all but articulation style features, using all but articulation style and the repeat/correction features, using just the repeat/correction feature, and finally, using just the language model features.

3.3 Task 2: Prediction of Frustration versus Else by Machine

The second task involved classifying FRUSTRATION versus ELSE. As in the experiment in Section 3.2, the ELSE class contained all the remaining emotion types. The right-most columns of Table 8 show the results of this task.

The first task yielded significantly more data in the emotion class, since ANNOYED was much more frequent than FRUSTRATED, as can be inferred from Table 2. This second task aimed to detect only extreme cases, which would be predicted to be an easier task. However, this experiment involved very little data, and thus only cautious conclusions can be drawn. One of these is that the performance on this task is consistently and significantly better than on the ANNOYANCE+FRUSTRATION vs. ELSE classification (by an average of about 9%). This follows the prediction that extreme cases of FRUSTRATION are easier to detect. Additionally, the relative increases or decreases in performance based on feature sets used follow the general trend of the Task 1 experiments. For example, the addition of articulation style features increases performance by 5-7%, which is similar to the 5% mentioned for the first task. More specific observations will be discussed when investigating feature usage in Section 3.5.

Figure 2 summarizes the experimental results. Looking closely, we can see that relative performance is very consistent based on features available, regardless of task or data set. The only inconsistency is with the repeat/correction feature for Task 2 on the "Originally Agreed" data set (the right-most hollow triangle). Notice from Table 8, however, that the efficiency (14.3) is quite low. This is likely due to the fact that the repeat/correction feature is a categorical feature, which is not evenly distributed over the data (repeats and corrections often occur back to back). This can lead to a mismatch of data in the test and training sets when the sets are so small (minority class has 35 instances in training, as seen in Table 5), which in turn leads the results that do not follow trend. With the consensus version data (hollow circles), however, the efficiency is higher and the accuracy follows trend because it has more data (125 instances of the minority class in training set).

3.4 Effect of True Words versus Recognition Output

All the above experiments are based on forced alignments for feature processing. In parallel experiments using automatic recognition outputs, accuracies were only 0.1-2.6% worse in the ANNOYANCE+FRUSTRATION vs. ELSE task, and slightly better in the FRUSTRATION vs. ELSE tasks, as shown in Table 8. The ROC curves in Figure 3 also show the similarity in the parallel experiments.

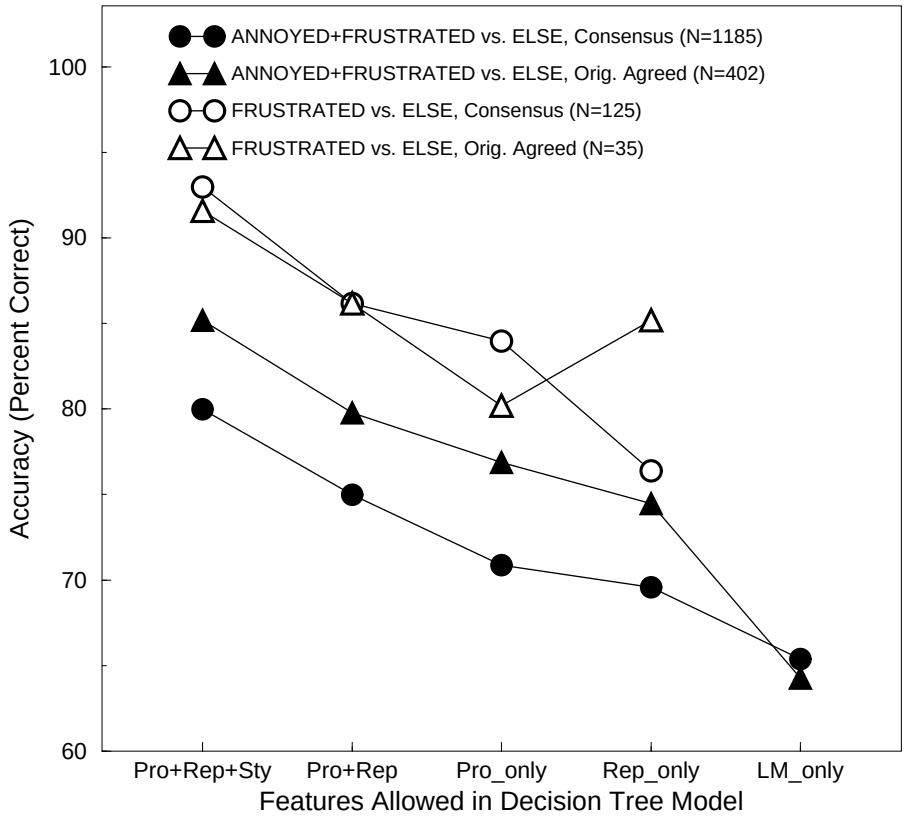

Figure 2: Comparison of annoyance and frustration detection with different input features. Values plotted are accuracies of machine prediction of the noted type of human label (either Consensus or Originally Agreed). Pro = prosody, Sty = Style, Rep = repetition/correction feature, N = size of emotion class in training set.

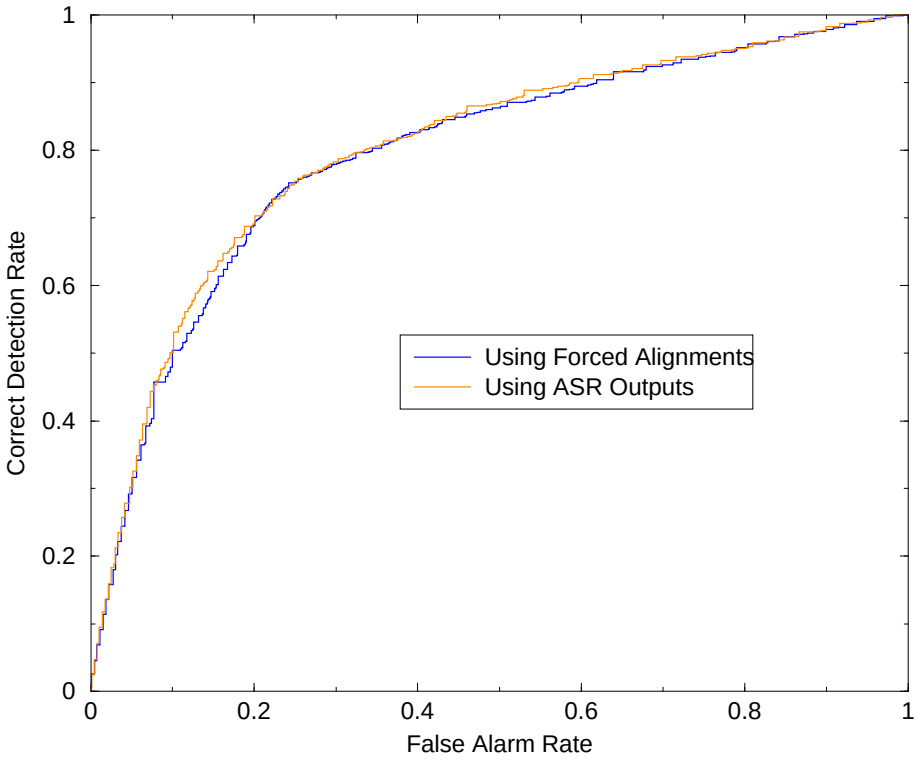

Figure 3: ROC curves of ANNOYED+FRUSTRATED vs. ELSE task based on forced alignments and ASR outputs for feature processing.

These results imply that for this (and possibly other) emotion recognition tasks based on whole utterances, highly accurate word recognition is not necessarily a requirement. This is also implied by the observation that the prosodic features used were largely over whole utterances as opposed to over words. That is, emotional expression can be seen as a longer term phenomena, which lessens the need for correct word recognition. Overall, language model contributions had little effect as well.

3.5 Feature Usage Analysis

Overall feature usage for baseline experiment of the ANNOYED + FRUSTRATED versus ELSE task consisted of five main types of features. Table 9 shows these five feature types and the specific feature usages of each. Feature usage is reported as the percentage of decisions for which the feature type is queried; thus features higher in the tree have higher usage than those lower in the tree. The most-queried feature type, temporal features, represented roughly

Table 9: Feature Usage of the baseline experiment for the ANNOYED + FRUSTRATED versus ELSE Task (i.e. Consensus Version, [no STYLE]). Descriptions of features can be found in Appendix Section 6.2.

Duration, Speaking Rate and Pause Features	**28.18%**
MAXPHDUR_N	15.61%
SYLRATE	6.26%
VOWELDUR_DNORM_E_5	2.68%
SYLRATE_DNORM_E_5	1.83%
SPCHPCT_DNORM_E	0.95%
SPCHPCT_DNORM_E_5	0.64%
PAUSE7_COUNT	0.16%
SPCHPCT	0.05%
Pitch Features	**26.57%**
MAXF0_IN_MAXV_N	7.36%
MAXF0	7.04%
MINF0TIME	4.52%
MAXF0TIME	2.34%
MAXF0RISE_DNORM_E	1.75%
MINF0_BASELN	1.17%
LASTF0_BASELN	0.76%
MAXF0_POS	0.64%
LASTSLOPE	0.42%
FIRSTF0_BASELN	0.30%
RISERATIO_DNORM_E_5	0.22%
MAXF0_TOPLN	0.05%
Repeat/Correction Feature	**26.30%**
REPCO	26.30%
Energy Features	**11.37%**
RMS_DNORM_E	11.29%
AVGRMS_IN_MAXV_N_DNORM_E	0.08%
Dialog Position Features	**7.59%**
UTTPOS	4.65%
SYSPOS	2.94%

28% of total usage. The features in this category were mainly normalized duration and speaking rate features, including features normalized by only the first five utterances in the call. Longer durations and slower speaking rates were associated with frustration. Pitch features represented about 27% of total usage, and included the maximum F0 in the longest vowel, the maximum overall F0, the times that the maximum and minimum F0s occurred, the maximum speaker-normalized F0 rise, and the distance of various F0 statistics from the speaker baseline. All were associated with frustration when their values were high. The repeat/correction feature represented roughly 26% of total usage as well, with (as expected) more frustration after system errors. The speaker-normalized RMS energy accounted for 11% of the usage, and the remaining 8% of usage was from features tracking the number of dialog exchanges between the user and system occurring before the utterance in question.

Looking closely at the feature usage of the features selected by the tree, only the repeat/correction (REPCO) feature is not obtained automatically. As mentioned earlier in 2.2, however, researchers like Kirchhoff [5] are working on automatic detection of this type of feature. Therefore, the results obtained could be approximated automatically to some degree as research continues in this area.

Table 10 shows the feature usage of the baseline experiment for the FRUSTRATED versus ELSE task. When comparing the usages of this task with the ANNOYED+FRUSTRATED versus ELSE task, a significant difference can immediately be seen. There is a large increase the usage of pitch features when classifying frustration. Accordingly, the other feature types decrease in usage, yet remain significant. There are two possible reasons for this difference. First, when users get frustrated, the pitch levels and/or contours of their voices change significantly, to a much greater extent than when they are simply annoyed. As a result, the decision tree can pick up on these pitch feature differences so the usage of these features increases dramatically. Second, it is possible that frustrated users become so emotional they cease trying to speak slowly so the system can understand. Instead, they resort to their natural expression of frustration, which causes the temporal features of duration, speaking rate, and pause to carry less information on the emotional state of the user. Consequently, the decision tree shifts its usages to other feature types that are more useful for the task. This reason alone cannot account for the differences in feature usage between the two tasks, however, because while it addresses why pitch features increase in usage and temporal features decrease, it does not address the fact that the other feature types decrease in usage (they would increase if this was the sole reason for changes in usage). Therefore, it is likely that a combination of the two is at work to account for the differences in feature usage. Refer to Appendix Section 6.3 for feature usage details for other experiments report in Table 8.

The experiments also showed that among the articulation style features, raised voice and hyperarticulation are helpful predictors for emotion. Pauses between syllables and words were not useful. Though hyperarticulation is helpful in predicting frustration, it is not equivalent to frustration, or else it would

Table 10: Feature Usage of the baseline experiment for the FRUSTRATED versus ELSE Task (Consensus Version, [no STYLE]). Descriptions of features can be found in Appendix Section 6.2.

Pitch Features	**60.58%**
MAXF0	26.44%
MAXF0_IN_MAXV_N	23.65%
LASTF0_BASELN	3.12%
MAXF0_TOPLN	2.43%
LASTSLOPE	1.91%
MINF0TIME	1.25%
MINF0_BASELN	0.74%
RISERATIO_DNORM_E_5	0.69%
LASTF0	0.20%
FIRSTF0_BASELN	0.15%
Duration, Speaking Rate and Pause Features	**16.39%**
MAXPHDUR_N	14.03%
SPCHPCT	1.72%
PAUSE7_COUNT	0.38%
SPCHPCT_DNORM_E	0.26%
Repeat/Correction Feature	**11.63%**
REPCO	11.63%
Energy Features	**9.26%**
MAXRMS_IN_MAXV_N	8.69%
RMS_DNORM_E	0.57%
Dialog Position Features	**2.15%**
SYSPOS	2.15%

be a much more useful predictor. This indicates that people can hyperarticulate when calm and they can be frustrated and not hyperarticulate.

4 Conclusions

This project investigated the feasibility of recognizing emotion using automatic means, unlike previous work that often used methods that are not entirely automatic. The outputs of a speech recognizer were used instead of hand-marked reference transcripts, as well as prosodic features that were *automatically* extracted and normalized based on the recognizer alignments. The design included results for both forced alignment and free recognition to enable comparison. In addition, different ways of annotating emotion (individual and consensus) were used, and many additional annotations were included for analysis of the relationship between prosodic and other features (as well as to make the data maximally useful for other types of research). The project also uses naturally occuring emotional data as opposed to the elicited or acted emotional data of previous work. Taken from users making air travel arrangements over the telephone, the data more closely resemble what would be found in real-life applications. This project also determined to find the relationship between articulation style and emotion, using data where the two were labeled independently.

From the experiments conducted, several important conclusions can be drawn:

1. When classifying ANNOYED and FRUSTRATED versus ELSE (Task 1), human prediction results in a 72.6% accuracy. Meanwhile, the baseline experiment of machine prediction performs at 75.2% accuracy. The same experiment, without the repeat/correction feature, performs at 71.1%. These results show that machine prediction can in fact compete with human prediction, where the machine outperforms humans (75.2% to 72.6%) with the repeat/correction feature, while an entirely automatic system performs slightly worse (72.6% to 71.1%).

2. The FRUSTRATED versus ELSE task (Task 2) consistently gives better results than Task 1 by an average of about 9%, giving evidence that machines can better discern more extreme forms of emotion.

3. Experiments using recognized words as opposed to true words leads to a 0.1-2.6% degradation in accuracies in the first task. With Task 2, experiments using recognized words give slightly better accuracies. These results imply that emotional expression can be seen as a longer term phenomena (utterance-level as opposed to word-level), which lessens the need for correct word recognition.

4. Feature usage analysis of the baseline experiment of Task 1 shows the importance of duration, speaking rate, and pause features (28%), pitch features (27%) and the repeat/correction feature (26%). In Task 2, pitch features become much more important in classification, jumping to 60% usage.

5. The experiments showed that raised voice and hyperarticulation are helpful predictors for emotion, while pauses between syllables and words were

not useful. Hyperarticulation is not equivalent to frustration, however, even though it helps predict frustration. Thus, people can hyperarticulate when calm and not hyperarticulate when frustrated.

While this project explored many different areas of the emotion recognition task, much more remains to be done. The Communicator corpus, although worthwhile because of its naturally occuring emotional data, lacks the quantity of emotional data for much experimental analysis. Finding or creating a corpus with more emotional data is a key future development in this work. In addition, a corpus with a greater amount of realistic frustration would be useful. The users recorded in the Communicator corpus were not making real travel plans, so they seemed to display more patience and willingness to cope with system errors than users genuinely making travel plans on the system. Therefore, the frequency of frustration was lower than would be expected.

The techniques used here can be generalized from frustration recognition to any form of emotion recognition, so experiments involving other emotional classes are another area for future research. Future research could explore multi-class performance as well, since this research solely focused on two-class experiments.

Another area of future work involves more research into prosodic features for the task. This project simply touched the surface of many features. For example, while spectral tilt was included in the experiments, the features were very simple (average over longest normalized vowel). It was no surprise that such simple features were not used by the decision trees in the experiments. Many other features using the spectral tilt data, such as maxes, mins, ranges, and slopes over different regions of the utterance (e.g. overall, vowel portions, end) could prove useful if generated and used. Future research could include a similar investigation into the other feature types as well.

This project used decision tree classifiers, while many other options are available. The use of trees was motivated by ease of analysis, as the project was in the initial stages of finding useful features, which would be less apparent using other methods of classification. However, one could research into those other methods, such as neural nets or Gaussian mixtures.

In general, this project suggests that machines can perform emotion recognition on par with human ability, at least for the types of emotion explored here. This observation opens the possibility for a wide variety of applications, from educational software to games to customer support applications. Since perfect word recognition does not seem essential to the emotion recognition task, applications where ASR is errorful, not available, or even unfeasible can still potentially use emotion recognition. Furthermore, emotion recognition could possibly even aid in speech recognition. For instance, having emotional information in an utterance could prove useful in feature extraction, where a recognition system is aware of pitch or duration shifts due to emotional state, thereby producing "cleaner" features for classification.

5 Acknowledgements

Thanks must first go to Jesus Christ, my Lord and Savior, who has continued to provide all the grace and strength I need daily. Through my graduate career, He has proven Himself faithful, as always. Thanks to Elizabeth Shriberg and Andreas Stolcke, who have provided much direction and work for the project. Thanks to Nelson Morgan for bringing me into his research group and being my advisor. I also thank Ashley Krupski, Kai Filion, Kattya Baltodano, Mercedes Carter, and Raj Dhillon for data labeling, Harry Bratt and Kemal Sönmez for developing the pitch stylizer used in feature computation, the Communicator teams at CU, CMU, Lucent, and SRI for providing the data to this project, and the many others who helped here and there along the way (Barry, David, Don, Adam, et al.). This work was funded by the DARPA ROAR program under contract N66001-99-D-8504, by NASA award NCC 2-1256, by NSF STIMULATE grant IRI-9619921, and by the DARPA Communicator project at ICSI and U. Washington. The views herein are those of the author and do not reflect the policies of the funding agencies.

6 Appendix

6.1 Descriptions of Additional Annotations

6.1.1 Dialog-Level Annotations

Accent Type Type of accent (e.g. British, Chinese, New York) that the labeler noted.

Global Comment Comments that pertained to the entire call.

Non-Native Speaker Describes whether or not speaker was a native English speaker.

6.1.2 Utterance-Level Annotations

Comment Comments on the utterance that labelers felt were pertinent. Complete comment list described in Table 11.

Data Problems Problems with the data, such as cut-off speech, long silences, or background noise.

Final Pitch Rise Marks a noticeable rise in pitch at the end of the utterance.

Hyperarticulation Denotes whether the speaker "overenunciates" words or parts of words in the utterance.

Pauses Between Words When the utterance had unnaturally long pauses between words.

Pauses Between Syllables When the utterance had pauses between syllables of a word.

Raised Voice Marked when the overall perceived loudness or level of vocal effort of the utterance is noticeably raised.

Repeat-Correction Marked if the utterance was a repeat or rephrasing of a previous utterance, or whether there was an explicit correction of the system response to previous utterances.

Self-Talk The user is talking to him/herself or someone else. In general, this is marked when the user is not addressing the system.

Spelling Out Utterance has words or numbers spelled out. (This is useful for detecting corrections and repeated attempts.)

6.2 Description of Features

The following section describes features used in this project. Some features, such as the Dialog ID or Utterance ID, are part of the original data. Others are labeled by humans in our project (e.g. the emotion labels, hyperarticulation, and repeat correction). The majority of features, however, are automatically extracted for our project.

Table 11: Comment List

user is a child
testing the system
hardly any speech in call
shaky voice
under the influence
upset
side talk
unintelligible speech/mumbling
intelligible but muffled/distorted speech
utterance(s) out of order
more than one person talking to system during call
fake accents/mispronunciations/possible speech impediments
happy
user comments on system
utterance is irrelevant/random
joking
transcript is incorrect
confused
raised voice
noisy/background noise
long silences
speech-cut-off
miscellaneous
laughing
sarcastic
satisfied
lengthened speech
polite
amused/surprised

Table 12: Emotion Labels Key. The '?' is used to denote uncertainty, where the labeler had a harder time deciding on the emotion for an utterance.

EMOTION	Tag
neutral	n
neutral?	nq
annoyed	a
annoyed?	aq
amused/surprised	as
amused/surprised?	asq
angry/frustrated	af
angry/frustrated?	afq
disappointed/tired	dt
disappointed/tired?	dtq
other	o
other?	oq
NA	NA
NA?	NAq
NONE	NO
NONE?	NOq

6.2.1 Utterance Information Features

Dialog ID (DIALOGID) Unique identifier for the dialog. DIALOGID is in the form of ####_##_##_##_######## for NIST database dialogs. For the CMU data, it is ########-###. For the COLORADO data, it is sls-########-### or sls-########-###b.

Emotion Labels with context (EMOTION_WC) String of emotion labels, with labels separated by an '_'. The labels follow the mapping described in Table 12.

Labelers (LABELER_WC) String of labelers, with names separated by an '_'. The order of the labelers corresponds with the order of their labels in EMOTION_WC.

Utterance ID (UTTID) Unique identifier for the utterance. UTTID is in the form of ####_##_##_##_########_usr### for NIST utterances. For the CMU data, it is ########-###-### or ########-###-###_###. For the COLORADO data, it is sls-########-###-### or sls-########-###b-###.

Words of Utterance (WORDS) The words of the utterance based on the reference files in the Communicator databases. Each word is separated by an '_'. If the utterance contained no words, the field contains a single '_'.

6.2.2 Dialog Position Features

System Position (SYSPOS) Number of system prompts in the call before the utterance. The first utterance in the call has a SYSPOS of 1, with the SYSPOS incremented by one for each of the following utterances.

Utterance Position (UTTPOS) Like SYSPOS, except utterances of the call with no words uttered (according to the reference files in the Communicator database) do not increment UTTPOS.

Word Position (WORDPOS) The position of the first word in the utterance with respect to the call (i.e. WORDPOS is the number of words uttered before the current utterance plus one), according to the reference files in the Communicator databases. Noises, laughs, coughs and the like are not counted, but rejects (such as word fragments) are included.

6.2.3 Duration Features

Maximum Normalized Phone Duration (MAXPHDUR_N) The maximum phone duration in the utterance obtained in a similar fashion to the average.

Average Normalized Phone Duration (PHDUR_N) Average phone duration normalized (through division) with the average statistics for the phones over the Communicator database.

Relative Longest Normalized Vowel Position (V_N_POS) Ignoring beginning and end silence (according to the forced alignment), the relative position of the onset of the longest normalized vowel in an utterance. For example, if V_N_POS is 0.0, the longest normalized vowel begins at the beginning of the utterance. If 0.5, it begins in the middle.

Average Vowel Duration (VOWELDUR) Using the phone level forced alignments, the average duration of vowels (aa, ae, ah, ao, aw, ay, eh, er, ey, ih, iy, ow, oy, puh, uh, uw, ax).

6.2.4 Speaking Rate Features

Syllable Rate (SYLRATE) Syllable rate is approximated by dividing the number of vowels by the duration of the utterance (ignoring reject phones because they are indistiguishable between consonants and vowels).

6.2.5 Pause Features

Longest Pause Duration (MAXPAUSE) Maximum length of pause in the interior of an utterance.

Number of Long Pauses (PAUSE7_COUNT) Number of interior pauses longer than 70ms.

Speech Percentage (SPCHPCT) Percentage of an utterance that is speech. Calculated by dividing the duration of all the speech by the duration of speech plus duration of interior pauses.

6.2.6 Pitch Features

First Slope in LNV (1STSLP_IN_MAXV_N) The first F0 slope in the longest normalized vowel (LNV) region when excluding frames where the probability of pitch is less than the probability of halving or doubling.

Second to Last Slope of Fitted F0 (2nd2LASTSLOPE) The second to last slope of the stylized F0 (as described in 2.6.1) when excluding frames where the probability of pitch is less than the probability of halving or doubling.

First Fitted F0 (FIRSTF0) Excluding frames where the probability of pitch is less than the probability of halving or doubling, the first (nonzero) F0 value.

Average Fitted F0 (FITTEDF0) The average fitted F0 value over voiced frames where the pitch probability is greater than the probability of halving or doubling. This is based on the stylization software used. (Refer to 2.6.1 for details.)

Last Fitted F0 (LASTF0) Excluding frames where the probability of pitch is less than the probability of halving or doubling, the last (nonzero) F0 value.

Last Slope of Fitted F0 (LASTSLOPE) The last slope of the stylized F0 when excluding frames where the probability of pitch is less than the probability of halving or doubling.

Last Slope in LNV (LASTSLP_IN_MAXV_N) The last F0 slope in the longest normalized vowel (LNV) region when excluding frames where the probability of pitch is less than the probability of halving or doubling.

Maximum Fitted F0 (MAXF0) Excluding frames where the probability of pitch is less than the probability of halving or doubling, the maximum fitted F0 value.

Maximum Negative Slope of Fitted F0 (MAXF0FALL) When excluding frames where the probability of pitch is less than the probability of halving or doubling, the maximum negative slope of the stylized F0.

Maximum Slope of F0 (MAXF0RISE) Excluding frames where the probability of pitch is less than the probability of halving or doubling, the maximum slope of the stylized F0.

Time when Maximum F0 occurred (MAXF0TIME) Relative to the beginning of the utterance, the time when the maximum fitted F0 occurred.

Maximum F0 during the LNV (MAXF0_IN_MAXV_N) In the longest normalized vowel (LNV) region, the maximum fitted F0 when excluding frames where the probability of pitch is less than the probability of halving or doubling.

Relative Maximum F0 Position (MAXF0_POS) Ignoring beginning and end silence (according to the forced alignment), the relative position of the maximum fitted F0 in the utterance. For example, if MAXF0_POS is 0.0, the maximum fitted F0 begins at the beginning of the utterance. If 0.5, it begins in the middle.

Minimum Fitted F0 (MINF0) Excluding frames where the probability of pitch is less than the probability of halving or doubling, the minimum (nonzero) fitted F0 value.

Time when Minimum F0 occurred (MINF0TIME) Relative to the beginning of the utterance, the time when the minimum fitted F0 occurred.

Slope with Most Frames in LNV (MOSTSLP_IN_MAXV_N) The F0 slope having the most frames in the longest normalized vowel (LNV) region when excluding frames where the probability of pitch is less than the probability of halving or doubling.

Percentage of Pitch Rise Frames (RISERATIO) The number of frames where the pitch rises divded by the number of frames where the pitch rises or falls, excluding frames where the probability of pitch is less than the probability of halving or doubling.

6.2.7 Energy Features

Average RMS in the LNV (AVGRMS_IN_MAXV_N) When excluding frames where the probability of pitch is less than the probability of halving or doubling, the average RMS energy in the longest normalized vowel (LNV) region.

Maximum RMS in the LNV (MAXRMS_IN_MAXV_N) In the longest normalized vowel (LNV) region, the maximum RMS energy when excluding frames where the probability of pitch is less than the probability of halving or doubling.

Average RMS Energy (RMS) Average RMS energy over voiced frames where the pitch probability is greater than the probabilty of halving or doubling. This is based the stylization software used.

6.2.8 Spectral Tilt Features

First Cepstral Coefficient (VCEP1) The average of the first cepstral coefficient over the longest normalized vowel.

Slope of Linear Fit to Magnitude Spectrum (VLINTILT) The average slope (times -1) of linear fit to the magnitude spectrum of each frame over the longest normalized vowel.

Difference in Log Energies in Frequency Bands (VTILT) After taking the magnitude spectrum for each vowel frame, the log energies are summed in two regions, below 1000Hz and between 1000Hz and 4000Hz. The sum from the upper frequencies is subtracted from the sum from the lower frequencies. This difference is averaged over the longest normalized vowel.

6.3 Feature Usage Tables

Tables 13 through 18 are feature usage tables for experiments whose results were reported in Table 8 in Section 3.

6.4 Example of Average of Twenty Experiments

Table 19 shows the twenty different experiments conducted for the baseline experiment of the ANNOYED + FRUSTRATED vs. ELSE task. The results in Table 8 reflect the linear average of these individual experiments, and the feature usage statistics are reported in the left column of Table 15.

Table 13: Feature Usage of "Consensus version, [All Features]"

A+F vs. ELSE		F vs. ELSE	
Articulation Style	**38.14%**	**Articulation Style**	**57.96%**
RAIVO	26.47%	RAIVO	52.59%
HYPER	11.20%	HYPER	5.37%
PAWRD	0.47%		
		Duration, Speaking Rate	
Repeat/Correction	**18.75%**	**and Pause**	**16.09%**
REPCO	18.75%	MAXPHDUR_N	15.73%
		SYLRATE_DNORM_E_5	0.36%
Pitch	**17.61%**		
MAXF0_IN_MAXV_N	7.53%	**Pitch**	**14.30%**
MAXF0RISE_DNORM_E	2.60%	MAXF0	6.40%
MAXF0	1.92%	MAXF0_IN_MAXV_N	2.93%
MINF0TIME	1.32%	MAXF0_TOPLN	1.72%
MAXF0_POS	1.27%	LASTF0_BASELN	1.45%
MAXF0TIME	0.59%	LASTSLOPE	1.00%
MINF0_BASELN	0.57%	MINF0TIME	0.80%
LASTF0_BASELN	0.50%		
1STSLP_IN_MAXV_N	0.42%	**Repeat/Correction**	**6.89%**
LASTSLOPE	0.37%	REPCO	6.89%
FIRSTF0_BASELN	0.33%		
RISERATIO_DNORM_E_5	0.20%	**Dialog Position**	**2.75%**
		SYSPOS	2.07%
Duration, Speaking Rate		UTTPOS	0.67%
and Pause	**15.56%**		
MAXPHDUR_N	6.03%	**Energy**	**2.02%**
SYLRATE	3.68%	MAXRMS_IN_MAXV_N	1.46%
VOWELDUR_DNORM_E_5	3.59%	AVGRMS_IN_MAXV_N	
SPCHPCT_DNORM_E	1.21%	_DNORM_E	0.56%
PAUSE7_COUNT	0.49%		
SYLRATE_DNORM_E_5	0.48%		
SPCHPCT_DNORM_E_5	0.08%		
Dialog Position	**8.42%**		
UTTPOS	5.61%		
SYSPOS	2.82%		
Energy	**1.51%**		
RMS_DNORM_E	1.20%		
MAXRMS_IN_MAXV_N	0.31%		

Table 14: Feature Usage of "Originally agreed, [All Features]"

A+F vs. ELSE		F vs. ELSE	
Articulation Style	**37.64%**	**Articulation Style**	**83.33%**
RAIVO	29.49%	RAIVO	83.33%
HYPER	6.93%		
PAWRD	1.22%	**Pitch**	**14.68%**
		MAXF0	7.19%
Pitch	**22.04%**	MAXF0_IN_MAXV_N	3.02%
MAXF0_IN_MAXV_N	12.21%	MAXF0_POS	1.67%
MINF0_BASELN	3.69%	FIRSTF0_BASELN	1.31%
MAXF0	2.67%	PDIFF_MAXF0_BASELN	0.88%
MAXF0TIME	0.91%	MINF0	0.63%
MAXF0RISE_DNORM_E	0.78%		
MAXF0_POS	0.56%	**Energy**	**1.98%**
RISERATIO_DNORM_E_5	0.44%	MAXRMS_IN_MAXV_N	1.98%
LASTSLOPE	0.42%		
MINF0TIME	0.33%		
PLOGRATIO_MAXF0_IN _MAXV_N	0.03%		
Repeat/Correction	**14.88%**		
REPCO	14.88%		
Language Model	**9.75%**		
POST_AN_SIGN	5.76%		
POST_AN_LOGRATIO	3.99%		
Duration, Speaking Rate and Pause	**7.94%**		
MAXPHDUR_N	4.09%		
VOWELDUR_DNORM_E_5	0.99%		
SYLRATE	0.95%		
SPCHPCT	0.83%		
SYLRATE_DNORM_E_5	0.80%		
SPCHPCT_DNORM_E	0.28%		
Dialog Position	**4.64%**		
SYSPOS	2.41%		
UTTPOS	2.24%		
Energy	**3.11%**		
RMS_DNORM_E	2.13%		
AVGRMS_IN_MAXV_N _DNORM_E	0.72%		
MAXRMS_IN_MAXV_N	0.27%		

Table 15: Feature Usage of "Consensus version, [no STYLE]"

A+F vs. ELSE		F vs. ELSE	
Duration, Speaking Rate and Pause	**28.18%**	**Pitch**	**60.58%**
MAXPHDUR_N	15.61%	MAXF0	26.44%
SYLRATE	6.26%	MAXF0_IN_MAXV_N	23.65%
VOWELDUR_DNORM_E_5	2.68%	LASTF0_BASELN	3.12%
SYLRATE_DNORM_E_5	1.83%	MAXF0_TOPLN	2.43%
SPCHPCT_DNORM_E	0.95%	LASTSLOPE	1.91%
SPCHPCT_DNORM_E_5	0.64%	MINF0TIME	1.25%
PAUSE7_COUNT	0.16%	MINF0_BASELN	0.74%
SPCHPCT	0.05%	RISERATIO_DNORM_E_5	0.69%
		LASTF0	0.20%
Pitch	**26.57%**	FIRSTF0_BASELN	0.15%
MAXF0_IN_MAXV_N	7.36%		
MAXF0	7.04%	**Duration, Speaking Rate and Pause**	**16.39%**
MINF0TIME	4.52%	MAXPHDUR_N	14.03%
MAXF0TIME	2.34%	SPCHPCT	1.72%
MAXF0RISE_DNORM_E	1.75%	PAUSE7_COUNT	0.38%
MINF0_BASELN	1.17%	SPCHPCT_DNORM_E	0.26%
LASTF0_BASELN	0.76%		
MAXF0_POS	0.64%	**Repeat/Correction**	**11.63%**
LASTSLOPE	0.42%	REPCO	11.63%
FIRSTF0_BASELN	0.30%		
RISERATIO_DNORM_E_5	0.22%	**Energy**	**9.26%**
MAXF0_TOPLN	0.05%	MAXRMS_IN_MAXV_N	8.69%
		RMS_DNORM_E	0.57%
Repeat/Correction	**26.30%**		
REPCO	26.30%	**Dialog Position**	**2.15%**
		SYSPOS	2.15%
Energy	**11.37%**		
RMS_DNORM_E	11.29%		
AVGRMS_IN_MAXV_N _DNORM_E	0.08%		
Dialog Position	**7.59%**		
UTTPOS	4.65%		
SYSPOS	2.94%		

Table 16: Feature Usage of "Originally agreed, [no STYLE]"

A+F vs. ELSE		**F vs. ELSE**	
Duration, Speaking Rate		**Pitch**	**72.41%**
and Pause	**28.22%**	MAXF0_IN_MAXV_N	31.20%
MAXPHDUR_N	19.83%	MAXF0	17.17%
SYLRATE	4.34%	FIRSTF0_BASELN	12.64%
SPCHPCT_DNORM_E	2.52%	MINF0TIME	5.59%
SYLRATE_DNORM_E_5	0.65%	MAXF0_TOPLN	5.34%
SPCHPCT	0.44%	MINF0	0.46%
PAUSE7_COUNT	0.43%		
		Duration, Speaking Rate	
Pitch	**26.79%**	**and Pause**	**20.41%**
MAXF0_IN_MAXV_N	13.54%	MAXPHDUR_N	12.37%
MAXF0TIME	4.03%	SPCHPCT	5.00%
MAXF0	1.81%	PAUSE7_COUNT	3.04%
RISERATIO_DNORM_E_5	1.58%		
MAXF0_POS	1.43%	**Energy**	**4.20%**
LASTF0_BASELN	1.35%	RMS_DNORM_E	2.37%
MINF0TIME	1.06%	MAXRMS_IN_MAXV_N	1.83%
MAXF0RISE_DNORM_E	0.85%		
1STSLP_IN_MAXV_N	0.52%	**Repeat/Correction**	**2.00%**
MAXF0_TOPLN	0.41%	REPCO	2.00%
FIRSTF0_BASELN	0.19%		
MINF0_BASELN	0.04%	**Dialog Position**	**0.99%**
		SYSPOS	0.99%
Repeat/Correction	**25.22%**		
REPCO	25.22%		
Energy	**13.72%**		
RMS_DNORM_E	12.54%		
MAXRMS_IN_MAXV_N	0.79%		
AVGRMS_IN_MAXV_N			
_DNORM_E	0.40%		
Dialog Position	**4.10%**		
UTTPOS	2.56%		
SYSPOS	1.54%		
Language Model	**1.95%**		
POST_AN_SIGN	1.95%		

Table 17: Feature Usage of "Consensus version, [no STYLE, no REP]"

A+F vs. ELSE		F vs. ELSE	
Pitch	**45.26%**	**Pitch**	**72.30%**
MAXF0_IN_MAXV_N	16.88%	MAXF0_IN_MAXV_N	31.65%
MAXF0	5.26%	MAXF0	23.20%
LASTF0_BASELN	3.94%	LASTF0_BASELN	8.13%
MINF0TIME	3.94%	RISERATIO_DNORM_E_5	3.24%
MINF0_BASELN	3.73%	LASTSLOPE	1.37%
MAXF0RISE_DNORM_E	3.29%	FIRSTF0_BASELN	1.12%
MAXF0_POS	2.47%	MINF0_BASELN	0.95%
MAXF0TIME	1.99%	MAXF0TIME	0.76%
LASTSLOPE	1.26%	MINF0TIME	0.75%
RISERATIO_DNORM_E_5	0.77%	1STSLP_IN_MAXV_N	0.41%
MAXF0_TOPLN	0.64%	MINF0	0.35%
LASTF0	0.34%	PLOGRATIO_MAXF0_IN _MAXV_N	0.19%
FIRSTF0_BASELN	0.31%	LASTF0	0.19%
PLOGRATIO_RISERATIO	0.20%		
PDIFF_LASTF0_BASELN	0.14%		
1STSLP_IN_MAXV_N	0.11%		
Duration, Speaking Rate and Pause	**27.49%**	**Duration, Speaking Rate and Pause**	**23.85%**
MAXPHDUR_N	23.20%	MAXPHDUR_N	17.80%
SYLRATE	2.24%	SPCHPCT_DNORM_E	4.36%
SPCHPCT_DNORM_E	1.31%	SPCHPCT	1.69%
VOWELDUR_DNORM_E_5	0.53%		
SYLRATE_DNORM_E_5	0.16%	**Energy**	**2.12%**
PAUSE7_COUNT	0.05%	MAXRMS_IN_MAXV_N	2.02%
		AVGRMS_IN_MAXV_N _DNORM_E	0.1%
Dialog Position	**10.75%**	**Dialog Position**	**1.01%**
SYSPOS	6.12%	SYSPOS	1.01%
UTTPOS	4.63%		
		Spectral Tilt	**0.73%**
Energy	**9.81%**	VTILT_DNORM_E	0.73%
RMS_DNORM_E	9.22%		
AVGRMS_IN_MAXV_N _DNORM_E	0.39%		
MAXRMS_IN_MAXV_N	0.20%		
Language Model	**6.69%**		
POST_AN_LOGRATIO	6.69%		

Table 18: Feature Usage of "Originally agreed, [no STYLE, no REP]"

A+F vs. ELSE		F vs. ELSE	
Pitch	**47.32%**	**Pitch**	**63.48%**
MAXF0_IN_MAXV_N	26.50%	MAXF0_IN_MAXV_N	46.69%
MAXF0	5.48%	MAXF0	7.50%
MAXF0TIME	5.40%	MAXF0_TOPLN	5.00%
MINF0_BASELN	2.40%	MAXF0_POS	1.89%
LASTF0_BASELN	2.18%	LASTF0	1.64%
MAXF0RISE_DNORM_E	2.14%	MINF0	0.76%
LASTSLOPE	1.63%		
MINF0TIME	0.91%	**Duration, Speaking Rate and Pause**	**28.12%**
MAXF0_POS	0.58%	MAXPHDUR_N	26.90%
PDIFF_LASTF0_BASELN	0.11%	SPCHPCT	0.71%
		SPCHPCT_DNORM_E	0.51%
Duration, Speaking Rate and Pause	**33.97%**	**Energy**	**6.31%**
MAXPHDUR_N	28.38%	MAXRMS_IN_MAXV_N	2.58%
SYLRATE	4.88%	AVGRMS_IN_MAXV_N _DNORM_E	2.20%
SPCHPCT	0.46%	RMS_DNORM_E	1.53%
SPCHPCT_DNORM_E	0.16%		
SPCHPCT_DNORM_E_5	0.10%	**Dialog Position**	**2.08%**
		SYSPOS	2.08%
Dialog Position	**10.45%**		
SYSPOS	6.55%		
UTTPOS	3.90%		
Energy	**6.98%**		
RMS_DNORM_E	5.16%		
AVGRMS_IN_MAXV_N _DNORM_E	1.56%		
MAXRMS_IN_MAXV_N	0.26%		
Language Model	**1.28%**		
POST_AN_SIGN	1.28%		

Table 19: Detailed Experimental Results of ANNOYED + FRUSTRATED vs. ELSE, "Consensus version, [no STYLE]"

Expt.	Seed #	Accuracy	Efficiency	Best Tree
1	9952	77.93%	23.87%	9952-453
2	94777	76.59%	23.07%	94777-361
3	84954	75.75%	23.08%	84954-212
4	80450	74.58%	21.83%	80450-514
5	74685	75.25%	20.52%	74685-374
6	70233	75.59%	23.58%	70233-379
7	61078	76.09%	23.38%	61078-260
8	59567	75.08%	21.76%	59567-379
9	33978	68.56%	11.47%	33978-284
10	33911	76.76%	24.33%	33911-215
11	33508	75.42%	19.79%	33508-216
12	29587	77.42%	24.78%	29587-572
13	26587	75.59%	21.06%	26587-395
14	260	74.75%	21.16%	260-282
15	20652	75.75%	22.81%	20652-256
16	20374	76.09%	21.26%	20374-350
17	17582	73.58%	17.21%	17582-638
18	1700	77.42%	24.27%	1700-220
19	1398	73.08%	18.20%	1398-207
20	10389	73.08%	16.09%	10389-222

Average Accuracy	75.22%
Accuracy Std. Dev.	2.01
Average Efficiency	21.18%
Efficiency Std. Dev.	3.23

References

[1] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desparately seeking emotions, or: Actors, wizards, and human beings. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 195–200, Belfast, September 2000.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, January 2001.

[3] The multiparty discourse group. http://www.cs.rochester.edu/research/cisd/resources/damsl/, April 2002.

[4] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Proceedings from ICSLP 96*, volume 3, pages 1970–1973, Philadelphia, PA, USA, October 1996.

[5] K. Kirchhoff. A comparison of classification techniques for the automatic detection of error corrections in human-computer dialogues. In *Proceedings of the NAACL Workshop on Adaptation in Dialogue Systems*, Pittsburgh, PA, June 2001.

[6] C. M. Lee, S. Narayanan, and R. Pieraccini. Recognition of negative emotions from the speech signal. In *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, December 2001.

[7] Tsuyoshi Moriyama and Shinji Ozawa. Emotion recognition and synthesis system on speech. In *Proceedings from IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 840–844, June 1999.

[8] Iain R. Murray, Mike D. Edgington, Diane Campion, and Justin Lynn. Rule-based emotion synthesis using concatenated speech. In *Proceedings of the ISCA Workshop on Emotion and Speech*, pages 173–177, New Castle, Northern Ireland, September 2000.

[9] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. *Neural Computing and Applications*, 9(4):290–296, 2000.

[10] Marc Schröder. Emotional speech synthesis: A review. In *Proceedings from Eurospeech*, volume 1, pages 561–564, Aalborg, Denmark, 2001.

[11] Sidney Siegel and Jr. N. John Castellan. *Nonparametric Statistics For The Behavioral Sciences*. McGraw-Hill, Inc., second edition, 1988.

[12] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In R. H. Mannell and J. Robert-Ribes, editors, *Proceedings from ICSLP*, volume 7, pages 3189–3192, Sydney, December 1998. Australian Speech Science and Technology Association.

[13] M. Kemal Sönmez, Larry Heck, Mitchel Weintraub, and Elizabeth Shriberg. A lognormal tied mixture model of pitch for prosody-based speaker recognition. In *Proceedings from Eurospeech*, volume 3, pages 1391–1394, Rhodes, Greece, 1997.

[14] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng. The sri march 2000 hub-5 conversational speech transcription system. In *Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[15] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garafolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Proceedings from Eurospeech*, pages 1371–1374, Aalborg, Denmark, 2001.

[16] Takashi Yoshimura, Satoru Hayamizu, Hiroshi Ohmura, and Kazuyo Tanaka. Pitch pattern clustering of user utterances in human-machine dialogue. In *Proceedings from ICSLP 96*, volume 2, pages 837–840, Philadelphia, PA, USA, October 1996.