PROSODY-BASED AUTOMATIC DETECTION OF ANNOYANCE AND FRUSTRATION IN HUMAN-COMPUTER DIALOG

Jeremy Ang[1,3] *Rajdip Dhillon*[1,3] *Ashley Krupski*[1,3] *Elizabeth Shriberg*[1,2] *Andreas Stolcke*[1,2]

[1]International Computer Science Institute, Berkeley, CA 94704, U.S.A.
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025, U.S.A.
[3]University of California, Berkeley, CA

ABSTRACT

We investigate the use of prosody for the detection of frustration and annoyance in natural human-computer dialog. In addition to prosodic features, we examine the contribution of language model information and speaking "style". Results show that a prosodic model can predict whether an utterance is neutral versus "annoyed or frustrated" with an accuracy on par with that of human interlabeler agreement. Accuracy increases when discriminating only "frustrated" from other utterances, and when using only those utterances on which labelers originally agreed. Furthermore, prosodic model accuracy degrades only slightly when using recognized versus true words. Language model features, even if based on true words, are relatively poor predictors of frustration. Finally, we find that hyperarticulation is not a good predictor of emotion; the two phenomena often occur independently.

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*. 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). In this work we focus on the last type of application, specifically, on the detection of user frustration with a telephone-based dialog system interface. Although we focus on frustration, we note that our method is general and could be extended to emotion detection involving any type of emotion or domain.

There has been considerable past work in the area of characterizing and detecting emotion in speech [1, 2, 3, 4, 5, 6]. The current study differs from from previous work in a number of ways. First, much of the past work has studied *elicited* emotions, 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. This study utilizes a dataset containing a large number of different speakers engaged in a task that itself gives rise to emotion. Second, past 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 locations of specific measurement locations), or relied on very simple prosodic features (e.g., excluding durations) that did not require recognition output. The present 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 speaking style, or which confound the two, in this work we aim to determine the association between the two, by including hand-marked speaking style characteristics in our database. By including the characteristics (such as hyperarticulation, pausing, or raised-voice) along with our prosodic features, we can determine 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, our methods for emotion detection are entirely automatic, but we can ask whether there would be added value for emotion detection if we were able to automatically detect speaking style.

Table 1. Statistics of labeled data

Source	Dialogs	Utterances	Time period
CU	205	5619	11/1999–6/2001
CMU	240	8765	1/2001–8/2001
NIST	392	7515	6/2000
Total	837	21899	-

2. METHODOLOGY

2.1. Speech data

We used a large, multi-site research and evaluation corpus of human-computer dialog developed under the DARPA Communicator project [7]. 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. The 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. The amount of data used in our study and their collection periods are summarized in Table 1. All data were collected over the telephone and sampled at 8 kHz. Roughly 75% of the utterances were used for training; the remaining 25% were used for testing; no dialogs were split between training and test sets.

2.2. Emotion labeling

User utterances were labeled by five students from UC Berkeley. Because we wanted labeling to reflect judgments of the average person, labelers came from different disciplines. Labeling was done using a modified version of the Rochester Dialog Annotation Tool (DAT) [8].

Emotion labels. 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).

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 three further types of information: speaking style, repeated requests or explicit corrections, and data quality problems. For *speaking style* we settled on the following, nonexclusive categories: hyperarticulation (exaggerated pronunciation of specific phones or syllables), pausing (between words or between syllables in a word), and "raised voice" (an increase in pitch, loudness, or both). For *repeated requests or corrections*, we labeled utterances either not a repeat/correction, a "repeat-or-rephrase-only", a "repeat-or-rephrase-with-explicit-correction", or an "explicit-correction-only", based on [9]. For *data quality* we marked 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). While joking and system cut-offs were included in our analyses, we omitted the other cases from the present study. In principle we would have liked 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), we decided to omit them for the sake of data homogeneity.

Labeling Issues. We found that labeling of emotion as well as speaking style is an inherently difficult task. First, emotion is conveyed on a continuous scale, and for purposes of this work we needed 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. We chose the former option, since that is the most relevant option given the application in mind (detect changes in the current user over the dialog). Finally, most of our utterances were quite short, often just the word "Yes" or "No", making emotion and style difficult to judge.

"Original" and "Consensus" Labels. In a first pass, labelers annotated individually after calibration. Interlabeler agreement (even after grouping ANNOYED and FRUSTRATED together) was only about 71%, with a Kappa of 0.47. We deemed this too low for our purposes, but note that it appears to be due to the task rather than to our labelers, because agreement among the various pairwise combinations of labelers did not significantly differ, and because agreement did not improve with additional training. We therefore conducted a second pass of labeling, which we refer to as "Consensus" labeling, in which the two most experienced labelers together relabeled any utterances that original labelers had not agreed on.

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%

2.3. Speech recognition and forced alignment

Both the prosodic and language model features for our modeling relied on alignment information from a speech recognizer. Rather than use the recognition results from the various Communicator systems (which were not always available), we ran a simplified version of SRI's Hub-5 system for conversational telephone speech [10], 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. 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 our experiments). To investigate the effects of recognition errors, we also computed features based on the reference transcriptions of the users' utterances, via forced alignment to the waveforms.

2.4. Prosodic model

Prosodic features. We extracted the following types of features: duration and speaking rate features, pause features, pitch features, energy features, and spectral tilt features. *Duration features* included the maximum and average durations of the normalized (for true or recognized phone identity) vowels or phones in the utterance. *Speaking rate features* included the number of vowels divided by the duration of the utterance. *Pause features* included the ratio of speech to pause time, the duration of the longest pause, and the number of long pauses inside an utterance. *Pitch features*, which proved to be quite useful, were based on post-processed F0 output using a stylization and regularization algorithm based on an updated version of [11]. Pitch was further post-processed using a lognormal tied mixture model of F0 that provides estimates of an individual speaker's pitch range [11]. We used two versions, 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 slopes at various locations. *Energy features* included the maximum or average RMS energy during voiced frames and during the longest normalized vowel, normalized by the mean and variance

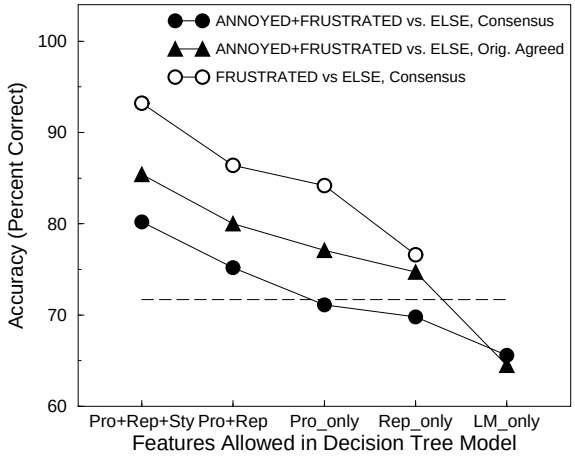

Fig. 1. Comparison of annoyance and frustration detection with different input features. Pro = prosody, Sty = Style, Rep = repetition/correction feature. The dashed line indicates accuracy for human interlabeler agreement on the first task.

of energy over the whole call (or over the first 5 utterances only). *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. In addition to the prosodic features, two *nonprosodic features* were included: the position of the utterance in the dialog, and the labels described earlier for repeated attempts and explicit corrections. Position can be assumed to be automatically obtained by a system; repeats and corrections are of course not as easy to obtain, but we consider their detection a separate problem and one in which many systems already have some ability to detect.

We used decision trees as our 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 our class sizes, we downsampled our data 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. In testing, we used all the data, but weighted the class accuracies to simulate equal classes.

2.5. Language model features

We trained a class-based trigram model 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, we added the language model features to the prosodic decision trees. We tried two types of language model features. One feature, 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. We eliminated this feature in favor of a more coarse feature, the *sign* of the likelihood difference, which did not show overfitting problems.

3. EXPERIMENTS AND RESULTS

Experiments were run with two basic classification tasks: ANNOYANCE+FRUSTRATION vs. ELSE, and FRUSTRATION vs. ELSE. In both cases, the ELSE class contained all remaining emotion types (NONE, plus the small amounts of other emotions such as TIRED, AMUSED, and OTHER, since we wanted to account for all datapoints). The first task allowed us to use significantly more emotional data, as can be inferred from Table 2. The latter task aimed to detect only extreme cases of anger. Results for both tasks, using both true and recognized words, are shown in Table 3. The different rows in the table show results for different experiment conditions, in which we varied both the source of the predicted emotion labels and the features available to the decision tree. 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 pass 1 were included. Note that the latter case is expected to show better results, since presumably labelers agreed on cases that were more clear-cut prosodically. Results are given in both accuracy (percentage of correct decisions) and "efficiency" (reduction in class entropy provided by the model). Because of the fairly limited size of our emotional-utterance corpus, we report results averaged from 20 separate experiments for each condition, each with a different random downsampling of the training data.

Looking first at the ANNOYANCE+FRUSTRATION vs. ELSE experiments, as summarized by the middle column of Table 3, we can draw several conclusions. First, we see that the baseline experiment (Consensus version, no STYLE features) at 75.2%, shows better prediction of human consensus labels than individual human labelers do with each other (72.6%). When we exclude the dialog state (repeat/correction) feature, the results are slightly worse (71.1% for tree versus 72.6% for human to human.) 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.). 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. Speaking style features also increase performance relative to the baseline prosodic tree. Potential candidates for the improvement include hyperarticulation, pauses, and raised-voice features; the actual contributing feature is discussed in the section on feature usage below. The FRUSTRATION vs. ELSE 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%).

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 3. 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.

Overall feature usage for the ANNOYED+FRUSTRATED versus ELSE task used five main types of features. We report feature usage 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 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 utter-

Table 3. Summary of experimental results. "STYLE" = speaking 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 simulated equal class distributions in the test set through sample weighting.

	ANNOY.+FRUST. vs. ELSE				FRUST. vs. ELSE			
	True words		ASR words		True words		ASR words	
	Acc	Eff	Acc	Eff	Acc	Eff	Acc	Eff
Each 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						

ances in the call. Longer durations and slower speaking rates were associated with frustration. Pitch features represented about 26% of total usage, and included the maximum F0 in the longest vowel, the maximum overall F0, the times that the maximum and minimum F0s occured, 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 thus far.

The experiments showed that among the speaking style features, only raised voice is a helpful predictor for emotion. Hyperarticulation and pauses between syllables and words were not useful. This indicates that it is not crucial to detect hyperarticulation for emotion detection, and confirms our initial decision to treat the two phenomena as separate. However, our prosodic features could be useful in detecting hyperarticulation itself, although this remains an interesting open question for further study.

4. ACKNOWLEDGMENTS

Kai Filion, Mercedes Carter, and Kattya Baltodano participated in the first data labeling pass. Harry Bratt and Kemal Sönmez developed the pitch stylizer used in feature computation. We thank the Communicator teams at CU, CMU, Lucent, and SRI for providing the data to our project, and Eric Fosler-Lussier, Katrin Kirchhoff, and Mari Ostendorf for valuable discussions. 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 authors and do not reflect the policies of the funding agencies.

5. REFERENCES

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

[2] T. Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech", *in Proceedings from IEEE International Conference on Multimedia Computing and Systems*, vol. 1, pp. 840–844, June 1999.

[3] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks", *Neural Computing and Applications*, vol. 9, pp. 290–296, 2000.

[4] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech", in H. T. Bunnell and W. Idsardi, editors, *Proc. ICSLP*, vol. 3, pp. 1970–1973, Philadelphia, Oct. 1996.

[5] 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*, pp. 195–200, Belfast, Sep. 2000.

[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, Dec. 2001.

[7] 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 P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, pp. 1371–1374, Aalborg, Denmark, Sep. 2001.

[8] "The Multiparty Discourse Group", http://www.cs.rochester.edu/-research/cisd/resources/damsl/, April 2002.

[9] 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 20001.

[10] 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.

[11] 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, *Proc. ICSLP*, vol. 7, pp. 3189–3192, Sydney, Dec. 1998. Australian Speech Science and Technology Association.