Identification of Confusion and Surprise in Spoken Dialog using Prosodic Features

Rohit Kumar [†], *Carolyn P. Rosé* [†], *Diane J. Litman* [‡]

[†] Language Technologies Institute, Carnegie Mellon University
[‡] Department of Computer Science, University of Pittsburgh
Pittsburgh, USA
rohitk, cprose @ cs.cmu.edu | litman @ cs.pitt.edu

ABSTRACT

Sensitivity to a user's emotional state offers promise in improving the state of the art in spoken dialog systems. In this work, we attempt to detect the speaker's states of confusion and surprise using prosodic features from his/her utterances. We have collected a corpus of utterances in realistic settings using an experimental methodology aimed at eliciting *confusion* and *surprise* from users. Classification experiments have yielded up to a 27.2% improvement over baseline performance using F0 and power features. We achieved the greatest success at classification of emotions that were most successfully elicited.

Index Terms: emotion detection, realistic settings, data collection methodology, self-report

1. INTRODUCTION

As a part of an effort to make spoken dialog systems increasingly natural, it is beneficial to enable these systems not only to recognize the content encoded in a user's response, but also to extract information about the emotional state of the user by analyzing how those responses have been spoken. This additional information can then be used by the dialog management framework to avoid misunderstandings and for recovering from errors [1]. Also, information about a user's emotional state has specific uses in certain task domains like tutoring [2], tele-marketing and health counseling [3].

In this paper, we present our work towards identifying the user's emotions of *confusion* and *surprise*. We describe a method for eliciting these emotions from the user in a range of realistic situations and report classification accuracies on the task of identifying these reported emotions in the speech signal itself. Our work is motivated by the possibility of using the resultant classifier in guiding a dialog system to decide whether to repeat its previous turn, ask for explicit / implicit confirmation or proceed to the next item in the task agenda. Batliner et. al. [1] describes the architecture of one such dialog system.

A considerable amount of work has been done in integrating emotion recognition from speech both in the frameworks of automatic speech recognition [4] and spoken dialog systems [5] [6]. In [7], authors report high accuracies in detecting annoyance and frustration using prosodic and language-based models trained on data that was manually annotated with categorical labels of emotions. Taking an alternative stand on emotion recognition, Batliner et. al. [1] propose to look for indicators of trouble in communication as indirect evidence of emotional responses from users.

One contribution of this work is an innovative experimental paradigm in which we use self-report questionnaires as a gold standard for training our classifiers rather than data that has been annotated by experimenters. This approach is advantageous in that it eliminates the need for the time consuming process of developing a reliable coding scheme and then doing the annotation by hand. Furthermore, we have reason to believe these labels provided by self-report questionnaires are a more accurate reflection of emotions experienced by the user during their interaction with the system. Adequate reliability on annotating emotions is difficult to achieve [7] particularly when annotators are asked to judge infrequent emotions like *confusion* and *surprise* rather than more typical ones like *anger* and *boredom*.

The remainder of this paper is organized as follows. We first describe our data collection infrastructure and process, as well as report on the success of our emotion elicitation method. Next we describe the features we extract from the speech signal. We then report our success at training classifiers based on these features. Finally, we conclude this paper with a summary of this investigation and scope for future work.

2. DATA COLLECTION

Most of the work in emotion detection from speech has been done using corpora collected from deployed voice-based systems [5] [7] or systems developed for specific applications [2]. In contrast to this, we create a corpus of utterances by employing prompts intentionally designed to elicit our target emotions. This is similar in spirit to prior work where emotions were elicited through interactions with computer games [8].

2.1 Emotions of Confusion and Surprise

Although *Confusion* and *Surprise* are common terms for some of the emotions experienced during interaction with a spoken dialog system (SDS), we find the notions of *Clarity* and *Appropriateness* intuitively related to *Surprise*, and yet more specific. Also, we relate Certainty to the *absence of Confusion*. Henceforth, we work with *Uncertainty*, *lack of Clarity* and *Inappropriateness* as the emotions of Confusion and Surprise. We offer further evidence in support of the use of these three emotions in section 2.4.1. We refer to the three collectively as our target emotions hereafter.

2.2 Methodology and Infrastructure

In our experimental methodology, we invite participants to interact with a SDS. They are told that this would help in improving the system. The SDS is scripted to elicit a subset of our target emotions at each turn in the interaction. After the participants interact with the system, they are asked to complete a self-report questionnaire, which asks them about the arousal of our target emotions at each turn of their interaction with the system on a single dimensional scale associated with each target emotion. As mentioned earlier, we use these reported values of arousal as labels to train classifiers.

2.2.1 Voice recording over Telephone

During the first stage of data collection, participants are asked to call a voice-based system using a fixed line telephone. The system consists of a VoiceXML based interaction script deployed on a voice platform. We have chosen to base the interaction script on a scenario of a customer survey about grocery stores. The script has 16 turns of system initiated interactions between the system and the participant. The system does not do any sophisticated dialog management, but it is programmed to reprompt in case of no response from the participant. The participant's response at each turn is recorded and stored.

2.2.2 Self-Report Questionnaire

During the second stage, the participants complete a self-report questionnaire using a web-based interface. The participants fill in their age, gender, native language and ethnicity at the start of the questionnaire. Then the interface takes the participants through their interaction with the system one turn at a time. At each turn they are allowed to listen to the system's utterance and their response at that turn as many times as they wish. The participants are asked to answer the following questions corresponding to each of our target emotions on a scale of 0 to 5 for each turn.

- How clear was the intention of system's question to you?
- How appropriate did you think the system's question was as a part of this survey?
- How certain did you feel about your response to the question?

The value of zero for each scale corresponds to total absence of clarity, appropriateness and certainty respectively, and the value of 5 corresponds to perfect clarity, appropriateness and certainty.

2.3 Eliciting confusion and surprise

In the interaction script, 11 of the 16 system's turns are crafted to elicit one or more of our target emotions from the participant. These turns represent common mistakes made by

- Authors of voice applications while designing the system. e.g., too long prompts (turn 3); use of unusual words and phrases (turn 4).
- Behavior Indicative of faulty technology. e.g., unexpected prompt due to ASR failure (turn 5).

Further, in order to keep the elicitation scenario realistic, the participant's expected responses at all of the turns are designed to be similar to responses in most common spoken dialog applications. For example: Yes/no, Digits and Short utterances like proper nouns, time, numbers, etc. The system's turns and emotions expected to be elicited by each turn are shown in Table 1. We evaluate the success of this interaction script in eliciting the expected emotions in section 2.4.2.

Table 1: System's turns in the interaction script and corresponding expected target emotions

Turn	Prompt [Target Emotion(s) expected to be Elicited]
1	Please tell the name of the grocery store or super market you most frequently shop at. [NONE]
2	Are you satisfied with the customer service provided at your most frequently used Grocery Store? [NONE]
3	Please answer the following question about the store on a scale of 1 to 10, where 1 is the worst performance and 10 is the best. If you do not have any particular opinion you can say No Opinion. Do you understand the scoring Scheme? [UNCERTAIN, UNCLEAR]
4	What is your score for the confectionary produce in the store? [UNCERTAIN, UNCLEAR]
5	Too bad you do not have an opinion. Do you feel bad about it? [INAPPROPRIATE]
6	How much money have you spent on soda pop in the past year? [UNCERTAIN]
7	Tell us something in particular that you like about the store you usually shop at? [UNCERTAIN]
8	All light plagues pushed his sensitive computers. [UNCLEAR, INAPPROPRIATE, UNCERTAIN]
9	Do you think the aisles in the store are conveniently ordered? [NONE]
10	How would you score the ordering of the aisles in the store? [NONE]
11	Have you ever noticed the yellow elephants in the ice cream section checking out the different flavors? [INAPPROPRIATE, UNCERTAIN]
12	How many mystery novels do you read every year? [INAPPROPRIATE]
13	Do you think the tellers at the check out counter are attractive enough? [INAPPROPRIATE]
14	What time do you find most convenient for shopping at the store? [NONE]
15	If Jane has two dogs and each one needs to be walked for an hour, how many hours does Jane have to walk? [INAPPROPRIATE, UNCERTAIN]
16	Would you like it if we provided free coupons to the participants of this survey? [INAPPROPRIATE]

2.4 The Corpus

Following the methodology discussed above, we have collected data from 17 participants (10 Females, 7 Males) who are mostly 18 - 26 year old native English speaking university students. All participants completed the telephonic interaction followed by the self-report. The resulting 272 utterances of participant responses were manually checked and the ones without any audible speech or with missing self report values were eliminated. Finally a corpus of 257 utterances was used for the classification experiments described below.

2.4.1 Correlations between our Target Emotions

In order to validate our formulation of *Confusion* and *Surprise* as three specific target emotions, we check for correlations between the reported values for the scales related to the three target emotions to a notion of *Surprise*. Due to the intuitive notion that the participants may be surprised either due to perceived inappropriateness or lack of clarity, we estimated lack of surprise by selecting the lesser of the two reported values for appropriateness and clarity.

The notion that *clarity* and *appropriateness* are related to *lack of surprise* is supported by the relatively high correlation between *appropriateness* and *lack of surprise* (R.$_{\text{Squared}}$=90.6%) and *clarity* and *lack of surprise* (R.$_{\text{Squared}}$=65.2%). Furthermore, because there is only a moderate correlation between *clarity* and *appropriateness* (R.$_{\text{Squared}}$=52.5%), which is lower than the correlations between either of them and *lack of Surprise*, it seems justified and more perspicuous to treat them as separate emotional responses related to lack of surprise rather than conflating them into a single *surprise* category.

We observe a relatively weak correlation between *lack of surprise* and *certainty* (R.$_{\text{Squared}}$=31%) which justifies treating them independently. All correlations were calculated using regression analyses on the reported values of the three target emotions and the calculated value for *lack of surprise*. All correlations are significant ($p < 0.001$).

Table 2: Most frequent values for each scale at each turn (Number in bracket indicates frequency of the value)

Turn	Appropriate	Certain	Clear	Results
1	**5** (13)	**5** (11)	**5** (15)	-
2	**5** (16)	**5** (13)	**5** (13)	-
3	**5** (10)	**5** (14)	**5** (13)	Failed
4	**3** (8)	**5** (5)	**2,5** (4)	Partial
5	**0** (10)	**0,2,4,5** (3)	**0** (5)	Success
6	**4** (5)	**3** (5)	**5** (7)	Partial
7	**4,5** (8)	**4,5** (6)	**5** (8)	Failed
8	**0** (16)	**0** (15)	**0** (17)	Success
9	**5** (10)	**5** (9)	**5** (13)	-
10	**5** (10)	**4** (7)	**5** (13)	-
11	**0** (13)	**0** (5)	**0** (8)	Success
12	**0** (9)	**5** (11)	**5** (8)	Success
13	**0** (7)	**5** (7)	**5** (8)	Success
14	**5** (11)	**5** (7)	**5** (12)	-
15	**0** (14)	**5** (6)	**0,5** (4)	Partial
16	**5** (8)	**5** (7)	**5** (9)	Failed

2.4.2 Evaluating our Elicitation approach

To evaluate our approach for eliciting target emotions, we compared the expected emotion at each turn to the reported values of each emotion for 257 utterances. Table 2 shows the most frequently reported value for all 3 scales at each turn. We consider a turn to be successful if the most frequently reported value for the expected emotion (as per Table 1) is either 0 or 1. We consider it a partial success if the most frequently reported value for the expected emotion is 2 or 3 and values of 4 and 5 are taken as failure in eliciting corresponding emotion. For example turn 12 is considered a success because it is expected to elicit inappropriateness and the most frequently reported value for appropriateness at turn 12 is zero.

We see that 5 turns are successful at eliciting their expected emotion. Only 3 turns failed at eliciting the expected emotion. We do not consider the 5 turns which are not expected to elicit any emotion in Table 2. They are all successful.

Further, Table 3 shows the evaluation of our elicitation scenario for each of our target emotions. Inappropriateness was elicited successfully 6 out of 7 times. Uncertainty was poorly elicited being successful only 1 out of 7 times. We suspect that people conceal their uncertainty in order to promote their positive face. Finally they were only 3 turns where we tried to elicit lack of clarity and we were successful at only 1 of those turns.

Table 3: Emotion-wise elicitation success

	Expected	**Success**	**Partial**	**Failure**
Inappropriate	7 turns	6 turns	0 turns	1 turns
Uncertain	7 turns	1 turns	3 turns	3 turns
Unclear	3 turns	1 turns	1 turns	1 turns

3. FEATURES FOR IDENTIFYING OUR TARGET EMOTIONS

Each of the 257 utterances in our corpus is represented by a vector of 52 features. These features are comprised of gender, expected participant response and prosodic features related to F0, power and duration.

The expected participant response feature was assigned four different nominal values based on the turn at which the utterance was recorded. The four values are Yes-No, Digit, Short-Utterance-numbers, Short-Utterance-other. 7 out of the 16 turns expect a Yes-No responses; 4 turns expect a non-numeric short utterance; 3 turns expect a numeric short utterance and 2 turns expect a digit.

We extracted F0 and Power contours with fixed frame size for each utterance using Wavesurfer [9]. The following features were calculated using the contours. The numbers in brackets indicate number of features. All features are computed automatically.

Pitch Features: We calculated Average, Maximum, Minimum and Range of F0 (4) and delta F0 (4). Also we computed normalized Maximum, Minimum and Range of F0 (3) for each utterance. The normalization was done with the average F0 of all the utterances of the participant. All F0 parameters are computed only over voiced frames in the utterances. The voiced frames are identified while computing F0 in Wavesurfer.

Power features were computed similar to pitch features. They include Average, Maximum, Minimum and Range of Power (4) and delta Power (4). Normalized Maximum, Minimum and Range of power (3) are included. These features are computed considering both the voiced and unvoiced frames. We added total power, total power in voiced frames and ratio of voiced power to total power to this set of features (3). Also, a ten point power contour (10) and its normalized form (10) were included. The ten point power contour is contour is computed by dividing the utterance in 10 uniformly sized segments and average power of each segment was then included in the contour.

Duration features included total duration of the utterance, duration of voicing and duration of initial silence (3). Duration of initial silence was computed automatically using heuristics on power and F0 contour. We included ratios of duration of voicing and duration of initial silence to total duration (2) as features.

4. CLASSIFICATION EXPERIMENTS

Using the features mentioned in section 3, we trained a classifier with support vector machines (SVM) for predicting the reported values of our target emotions using the Weka toolkit [10]. The classifiers were trained on two sets of labels.

- For one of the sets, we used the emotion values reported by the participant as labels. These labels have 6 distinct values (0,1,2,3,4,5) for each emotion. We refer to this as the *Non-Aggregated* set of labels.
- Another set of classifiers were trained on aggregated labels obtained by clustering the 6 labels into 2 distinct binary values (0,1). Utterances with reported values between 0 and 2 were labeled 0. Others with reported values between 3 and 5 were labeled 1. This set is referred to as the *Aggregated* set of labels.

We consider the frequency of the majority class as the baseline for these experiments. Majority class is chosen to be the most frequently occurring label. In order to compare the performance of the classifiers with humans, we asked two humans to listen to 105 utterances from 7 of the participants and report the values for our target emotions using the same interface as used by the participants. They were not allowed to listen to the system's turns. In Table 4, we report classification accuracies for each of our target emotions. The rows list accuracies from the Majority classifier (Baseline), Best of the two humans (Human) and the SVM Classifier both for the non-aggregated and aggregated set of labels. All classifiers were tested using a 10-fold cross validation on the whole corpus of 257 utterances.

Table 4: Classification Accuracies

Target Emotion	**Inappropriateness**	**Uncertainty**	**Lack of Clarity**
Classifier	**Non - Aggregated Labels**		
Baseline	36.5759	42.0233	49.4163
Human	36.1905	50.4762	53.8462
SVM	45.9144	43.5798	51.3619
	Aggregated Labels		
Baseline	58.7549	75.0973	68.0934
Human	62.8571	79.0476	70.1923
SVM	74.7082	75.0973	67.7043

From Table 4, we observe that *Inappropriateness* is identified more accurately than *lack of Clarity* and *Uncertainty* in both the sets of labels. However accuracy is low despite about 27.2% improvement over the baseline. Also we observe that our models perform better than human classifiers for both set of labels for *Inappropriateness*.

Classification accuracies for *Uncertainty* and *lack of Clarity* are worse than human classification in all cases and slightly better to worse compared to the baseline. We attribute this to a failure to elicit a sufficient number of clear examples of *Uncertainty* and *lack of Clarity* as show in Table 3. We believe this problem would be rectified by collecting more data.

5. CONCLUSIONS & FUTURE WORK

In this paper, we proposed to identify confusion and surprise in spoken utterances from users of a dialog system. We have collected data with the intention of eliciting three different emotions relating to confusion and surprise. We were successful in eliciting the emotion of *Inappropriateness* in user utterances and have been able to classify it with 27.2% better accuracy than human classifier and baseline. Although we were able to elicit *Uncertainty* and *lack of Clarity* with one of our designed prompts, we were not able to elicit enough examples in this short data collection to learn a reliable model.

We need to revise the interaction script to ensure successful elicitation for all of our target emotions and collect more data to reliably train the classifiers. Furthermore, we plan to experiment with lexical features extracted from an ASR. In connection with our motivation, we intend to explore issues related to using the resulting classifiers with existing spoken dialog systems.

5. ACKNOWLEDGEMENTS

This project is supported by ONR Cognitive and Neural Sciences Division, Grant number N000140410107. We want to thank Mr. S. P. Kishore for valuable input during this work.

6. REFERENCES

[1]. Batliner, A., Fischer, K., Huber, R., Spilker, J., and Noth, E., "How to find trouble in communication", *Speech Communications, Vol. 40, 2003, p117-143*.

[2]. Litman, D., Forbes-Riley, K., "Recognizing emotions from student speech in tutoring dialogs", *Proc. of IEEE ASRU Workshop, St. Thomas, 2003*.

[3]. Bickmore, T., "Relational Agents: Effecting changes through Human Computer Relationships", *Ph.D. thesis, MIT Media Arts and Science, 2003*.

[4]. Bosch, L., "Emotions, speech and the ASR Framework", *Speech Communications, Vol. 40, 2003, p213-225*.

[5]. Lee, C. M., and Narayanan, S. S., "Towards Detecting Emotions in Spoken Dialogs", *IEEE Trans. on Speech and Audio Processing, Vol. 13-2, 2005, p293-303*.

[6]. Polzin, T. S., and Waibel, A., "Emotion-sensitive Human-Computer Interfaces", *Proc. of ISCA Workshop on Speech and Emotions, Newcastle, 2000, p201-206*.

[7]. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A., "Prosody-based automatic detection on annoyance and frustration in human computer dialog", *Proc. of ICSLP, Denver, 2002, p2037-2040*.

[8]. Johnstone, T., "Emotional speech elicited using computer games", *Proc. of ICSLP, Philadelphia, 1996, p1989-1992*

[9]. Beskow, J., Sjlander, K., "Wavesurfer–an open source speech tool", *Proc. of ICSLP, Beijing, 2000, p464-467*.

[10]. Witten, I. H., Frank, E., "Data Mining: Practical machine learning tools and techniques", *2^{nd} Edition, Morgan Kaufmann, San Francisco, 2005*.