Detecting Emotion in Speech: Experiments in Three Domains

Jackson Liscombe
Columbia University
jaxin@cs.columbia.edu

Abstract

The goal of my proposed dissertation work is to help answer two fundamental questions: (1) How is emotion communicated in speech? and (2) Does emotion modeling improve spoken dialogue applications? In this paper I describe feature extraction and emotion classification experiments I have conducted and plan to conduct on three different domains: EPSaT, HMIHY, and ITSpoke. In addition, I plan to implement emotion modeling capabilities into ITSpoke and evaluate the effectiveness of doing so.

1 Introduction

The focus of my work is the expression of emotion in human speech. As normally-functioning people, we are each capable of vocally expressing and aurally recognizing the emotions of others. How often have you been put off by the "tone in someone's voice" or tickled others with the humorous telling of a good story? Though we as everyday people are intimately familiar with emotion, we as scientists do not actually know precisely how it is that emotion is conveyed in human speech. This is of special concern to us as engineers of natural language technology; in particular, spoken dialogue systems. Spoken dialogue systems enable users to interact with computer systems via natural dialogue, as they would with human agents. In my view, a current deficiency of state-of-the-art spoken dialogue systems is that the emotional state of the user is not modeled. This results in non-human-like and even inappropriate behavior on the part of the spoken dialogue system.

There are two central questions I would like to at lest partially answer with my dissertation research: (1) How is emotion communicated in speech? and (2) Does emotion modeling improve spoken dialogue applications? In an attempt to answer the first question, I have adopted the research paradigm of extracting features that characterize emotional speech and applying machine learning algorithms to determine the prediction accuracy of each feature. With regard to the second research question, I plan to implement an emotion modeler – one that detects and responds to uncertainty and frustration – into an Intelligent Tutoring System.

2 Completed Work

This section describes my current research on emotion classification in three domains and forms the foundation of my dissertation. For each domain, I have adopted an experimental design wherein each utterance in a corpus is annotated with one or more emotion labels, features are extracted from these utterances, and machine learning experiments are run to determine emotion prediction accuracy.

2.1 EPSaT

The publicly-available Emotional Prosody Speech and Transcription corpus[1] (EPSaT) comprises recordings of professional actors reading short (four syllables each) dates and numbers (*e.g.*, 'two-thousand-four') with different emotional states. I chose a subset of 44 utterances from 4 speakers (2 male, 2 female) from this corpus and conducted a web-based survey to subjectively label each utterance for each of 10 emotions, divided evenly for valence. These emotions included the positive emotion categories: *confident*, *encouraging*, *friendly*, *happy*, *interested*; and the negative emotion categories: *angry*, *anxious*, *bored*, *frustrated*, *sad*.

Several features were extracted from each utterance in this corpus, each one designed to capture emotional content. Global acoustic-prosodic information – *e.g.*, speaking rate and minimum, maximum, and mean pitch and intensity – has been well known since the 1960s and 1970s

[1] LDC Catalog No.: LDC2002S28.

to convey emotion to some extent (*e.g.*, (Davitz, 1964; Scherer et al., 1972)). In addition to these features, I also included linguistically meaningful prosodic information in the form of ToBI labels (Beckman et al., 2005), as well as the spectral tilt of the vowel in each utterance bearing the nuclear pitch accent.

In order to evaluate the predictive power of each feature extracted from the EPSaT utterances, I ran machine learning experiments using RIPPER, a rule-learning algorithm. The EPSaT corpus was divided into training (90%) and testing (10%) sets. A binary classification scheme was adopted based on the observed ranking distributions from the perception survey: "*not at all*" was considered to be the absence of emotion *x*; all other ranks was recorded as the presence of emotion *x*. Performance accuracy varied with respect to emotion, but on average I observed 75% prediction accuracy for any given emotion, representing an average 22% improvement over chance performance. The most predictive included the global acoustic-prosodic features, but interesting novel findings emerged as well; most notably, significant correlation was observed between negative emotions and pitch contours ending in a plateau boundary tone, whereas positive emotions correlated with the standard declarative phrasal ending (in ToBI, these would be labeled as /H-L%/ and /L-L%/, respectively). Further discussion of such findings can be found in (Liscombe et al., 2003).

2.2 HMIHY

"How May I Help You[SM]" (HMIHY) is a natural language human-computer spoken dialogue system developed at AT&T Research Labs. The system enables AT&T customers to interact verbally with an automated agent over the phone. Callers can ask for their account balance, help with AT&T rates and calling plans, explanations of certain bill charges, or identification of numbers. Speech data collected from the deployed system has been assembled into a corpus of human-computer dialogues. The HMIHY corpus contains 5,690 complete human-computer dialogues that collectively contain 20,013 caller turns. Each caller turn in the corpus was annotated with one of seven emotional labels: *positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative*[2]*, very other negative*. However, the distribution of the labels was so skewed (73.1% were labeled as *positive/neutral*) that the emotions were collapsed to *negative* and *non-negative*.

In addition to the set of automatic acoustic-prosodic features found to be useful for emotional classification of the EPSaT corpus, the features I examined in the HMIHY corpus were designed to exploit the discourse information available in the domain of spontaneous human-machine conversation. Transcriptive features – lexical items, filled pauses, and non-speech human noises – we recorded as features, as too were the dialogue acts of each caller turn. In addition, I included contextual features that were designed to track the history of the previously mentioned features over the course of the dialogue. Specifically, contextual information included the rate of change of the acoustic-prosodic features of the previous two turns plus the transcriptive and pragmatic features of the previous two turns as well.

The corpus was divided into training (75%) and testing (25%) sets. The machine learning algorithm employed was BOOSTEXTER, an algorithm that forms a hypothesis by combining the results of several iterations of weak-learner decisions. Classification accuracy using the automatic acoustic-prosodic features was recorded to be approximately 75%. The majority class baseline (always guessing *non-negative*) was 73%. By adding the other feature-sets one by one, prediction accuracy was iteratively improved, as described more fully in (Liscombe et al., 2005b). Using all the features combined – acoustic-prosodic, lexical, pragmatic, and contextual – the resulting classification accuracy was 79%, a healthy 8% improvement over baseline performance and a 5% improvement over the automatic acoustic-prosodic features alone.

2.3 ITSpoke

This section describes more recent research I have been conducting with the University of Pittsburgh's Intelligent Tutoring Spoken Dialogue System (ITSpoke) (Litman and Silliman, 2004). The goal of this research is to wed spoken language technology with instructional technology in order to promote learning gains by enhancing communication richness. ITSpoke is built upon the Why2-Atlas tutoring back-end (VanLehn et al., 2002), a text-based Intelligent Tutoring System designed to tutor students in the domain of qualitative physics using natural language interaction. Several corpora have been recorded for development of ITSpoke, though most of the work presented here involves tutorial data between a student and human tutor. To date, we have labeled the human-human corpus for anger, frustration, and uncertainty.

As this work is an extension of previous work, I chose to extract most of the same features I had extracted from the EPSaT and HMIHY corpora. Specifically, I extracted the same set of automatic acoustic-prosodic features, as well as contextual features measuring the rate of change of acoustic-prosodic features of past student turns. A new feature set was introduced as well, which I refer to as the breath-group feature set, and which is an automatic method for segmenting utterances into intonationally meaningful units by identifying pauses using background noise estimation. The breath group feature set

[2] 'Other negative' refers to any emotion that is perceived negatively but is not anger nor frustration.

comprises the number of breath-groups in each turn, the pause time, and global acoustic-prosodic features calculated for the first, last, and longest breath-group in each student turn.

I used the WEKA machine learning software package to classify whether a student answer was perceived to be *uncertain*, *certain*, or *neutral*[3] in the ITSpoke human-human corpus. As a predictor, C4.5, a decision-tree learner, was boosted with AdaBoost, a learning strategy similar to the one presented in Section 2.2. The data were randomly split into a training set (90%) and a testing set (10%). The automatic acoustic-prosodic features performed at 75% accuracy, a relative improvement of 13% over the baseline performance of always guessing *neutral*. By adding additional feature-sets – contextual and breath-group information – I observed an improved prediction accuracy of 77%. Thus indicating that breath-group features are useful. I refer the reader to (Liscombe et al., 2005a) for in-depth implications and further analysis of these results. In the immediate future, I will extract features previously mentioned in Section 2.2 as well as the exploratory features I will discuss in the following section.

3 Work-in-progress

In this section I describe research I have begun to conduct and plan to complete in the coming year, as agreed-upon in February, 2006 by my dissertation committee. I will explore features that are not well studied in emotion classification research, primarily pitch contour and voice quality approximation. Furthermore, I will outline how I plan to implement and evaluate an emotion detection and response module into ITSpoke.

3.1 Pitch Contour Clustering

The global acoustic-prosodic features used in most emotion prediction studies capture meaningful prosodic variation, but are not capable of describing the linguistically meaningful intonational behavior of an utterance. Though phonological labeling methods exist, such as ToBI, annotation of this sort is time-consuming and must be done manually. Instead, I propose an automatic algorithm that directly compares pitch contours and then groups them into classes based on abstract form. Specifically, I intend to use partition clustering to define a disjoint set of similar prosodic contour types over our data. I hypothesize that the resultant clusters will be theoretically meaningful and useful for emotion modeling. The similarity metric used to compare two contours will be edit distance, calculated using dynamic time warping techniques. Essentially, the algorithm finds the best fit between two contours by stretching and shrinking each contour as necessary. The score of a comparison is calculated as the sum of the normalized real-valued distances between mapped points in the contours.

[3] With respect to certainness.

3.2 Voice Quality

Voice quality is a term used to describe a perceptual coloring of the acoustic speech signal and is generally believed to play an important role in the vocal communication of emotion. However, it has rarely been used in automatic classification experiments because the exact parameters defining each quality of voice (*e.g.*, creaky and breathy) are still largely unknown. Yet, some researchers believe much of what constitutes voice quality can be described using information about glottis excitation produced by the vocal folds, most commonly referred to as the glottal pulse waveform. While there are ways of directly measuring the glottal pulse waveform, such as with an electroglottograph, these techniques are too invasive for practical purposes. Therefore, the glottal pulse waveform is usually approximated by inverse filtering of the speech signal. I will derive glottal pulse waveforms from the data using an algorithm that automatically identifies voiced regions of speech, obtains an estimate of the glottal flow derivative, and then represents this using the Liljencrants-Fant parametric model. The final result is a glottal pulse waveform, from which features can be extracted that describe the shape of this waveform, such as the Open and Skewing Quotients.

3.3 Implementation

The motivating force behind much of the research I have presented herein is the common assumption in the research community that emotion modeling will improve spoken dialogue systems. However, there is little to no empirical proof testing this claim (See (Pon-Barry et al., In publication) for a notable exception.). For this reason, I will implement functionality for detecting and responding to student emotion in ITSpoke (the Intelligent Tutoring System described in Section 2.3) and analyze the effect it has on student behavior, hopefully showing (quantitatively) that doing so improves the system's effectiveness.

Research has shown that frustrated students learn less than non-frustrated students (Lewis and Williams, 1989) and that human tutors respond differently in the face of student uncertainty than they do when presented with certainty (Forbes-Riley and Litman, 2005). These findings indicate that emotion plays an important role in Intelligent Tutoring Systems. Though I do not have the ability to alter the discourse-flow of ITSpoke, I will insert active listening prompts on the part of ITSpoke when the system has detected either frustration or uncertainty. Active listening is a technique that has been shown to diffuse negative emotion in general (Klein et al., 2002). I hy-

pothesize that diffusing user frustration and uncertainty will improve ITSpoke.

After collecting data from an emotion-enabled ITSpoke I will compare evaluation metrics with those of a control study conducted with the original ITSpoke system. One such metric will be learning gain, the difference between student pre- and post-test scores and the standard metric for quantifying the effectiveness of educational devices. Since learning gain is a crude measure of academic achievement and may overlook behavioral and cognitive improvements, I will explore other metrics as well, such as: the amount of time taken for the student to produce a correct answer, the amount of negative emotional states expressed, the quality and correctness of answers, the willingness to continue, and subjective post-tutoring assessments.

4 Contributions

I see the contributions of my dissertation to be the extent to which I have helped to answer the questions I posed at the outset of this paper.

4.1 How is emotion communicated in speech?

The experimental design of extracting features from spoken utterances and conducting machine learning experiments to predict emotion classes identifies features important for the vocal communication of emotion. Most of the features I have described here are well established in the research community; statistic measurements of fundamental frequency and energy, for example. However, I have also described more experimental features as a way of improving upon the state-of-the-art in emotion modeling. These exploratory features include breath-group segmentation, contextual information, pitch contour clustering, and voice quality estimation. In addition, exploring three domains will allow me to comparatively analyze the results, with the ultimate goal of identifying universal qualities of spoken emotions as well as those that may particular to specific domains. The findings of such a comparative analysis will be of practical benefit to future system builders and to those attempting to define a universal model of human emotion alike.

4.2 Does emotion modeling help?

By collecting data of students interacting with an emotion-enabled ITSpoke, I will be able to report quantitatively the results of emotion modeling in a spoken dialogue system. Though this is the central motivation for most researchers in this field, there is currently no definitive evidence either supporting or refuting this claim.

References

M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, 2005. *Prosodic Typology – The Phonology of Intonation and Phrasing*, chapter 2 The original ToBI system and the evolution of the ToBI framework. Oxford, OUP.

J. R. Davitz, 1964. *The Communication of Emotional Meaning*, chapter 8 Auditory Correlates of Vocal Expression of Emotional Feeling, pages 101–112. New York: McGraw-Hill.

Kate Forbes-Riley and Diane J. Litman. 2005. Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus. In *Proceedings of 6th SIGdial Workshop on Discourse and Dialogue,*, Lisbon, Portugal.

J. Klein, Y. Moon, and R. W. Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2):119–140, February.

V. E. Lewis and R. N. Williams. 1989. Mood-congruent vs. mood-state-dependent learning: Implications for a view of emotion. *D. Kuiken (Ed.), Mood and Memory: Theory, Research, and Applications, Special Issue of the Journal of Social Behavior and Personality*, 4(2):157–171.

Jackson Liscombe, Jennifer Venditti, and Julia Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Eurospeech*, Geneva, Switzerland.

Jackson Liscombe, Julia Hirschberg, and Jennifer Venditti. 2005a. Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech*, Lisbon, Portugal.

Jackson Liscombe, Guiseppe Riccardi, and Dilek Hakkani-Tür. 2005b. Using context to improve emotion detection in spoken dialogue systems. In *Proceedings of Interspeech*, Lisbon, Portugal.

Diane Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Proceedings of the 4th Meeting of HLT/NAACL* (Companion Proceedings), Boston, MA, May.

Heather Pon-Barry, Karl Schultz, Elizabeth Owen Bratt, Brady Clark, and Stanley Peters. In publication. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education (IJAIED)*.

K. R. Scherer, J. Koivumaki, and R. Rosenthal. 1972. Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. *Journal of Psycholinguistic Research*, 1:269–285.

K. VanLehn, P. Jordan, and C. P. Rose. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proceedings of the Intelligent Tutoring Systems Conference*, Biarritz, France.