Challenges for Robust Prosody-based Affect Recognition

Heather Pon-Barry, Arun Reddy Nelakurthi

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, Tempe, Arizona, USA
ponbarry@asu.edu, anelakur@asu.edu

Abstract

Prosody-based affect recognition has great potential impact for building adaptive speech interfaces. For example, in intelligent systems for personalized learning, sensing a student's level of certainty, which is often signaled prosodically, is one of the most interesting states to interpret and respond to. However, robust uncertainty recognition faces several challenges, including the lack of gold-standard labels, and differences in expressivity among speakers. In this paper we explore the intersection of these two issues. We have collected a corpus of spontaneous speech in a question-answering task. Three kinds of certainty labels are associated with each utterance. First, speakers rated their own level of certainty. Second, a panel of listeners rated how certain the speaker sounded. Third, an externally crowd-sourced difficulty score is generated for each stimulus (the question). We present a word-level prosodic analysis of individual speaking styles, as they relate to these three different measurements of certainty. Our results suggest that instead of learning one-size-fits-all prosodic models of affect, we might find improvement from learning multiple models corresponding to different speaking styles.

Index Terms: Uncertainty, affect recognition, affect labels, speaking style.

1. Introduction

An exciting goal in human-computer interaction is that of adding human-level emotional behavior to intelligent systems, that is, the ability to perceive a user's emotional state and adaptively respond to it [1]. In speech systems in particular, there has been a lot of work in recent years on detecting a broad spectrum of affective states in speech, including basic emotions [2, 3, 4], frustration [5], charisma [6], uncertainty [7, 8, 9], sleepiness and intoxication [10], and interpersonal stance [11].

There are multiple ways of measuring a speaker's level of certainty. In the existing work on automatic emotion recognition, the most common approach is to measure *perceived* emotion, as annotated by one or more human listeners, producing labels that are by definition subjective [12]. While we treat these labels as a gold standard, we understanding that the subjectivity makes for a challenging classification problem [13]. On the other hand, we can consider *self-reported* certainty, when speakers are asked to rate their own level of certainty. In our prior work, we found that perceived certainty was often *higher* than self-reported certainty [9]. In the same vein, related work on interpersonal stance (friendliness, flirtatiousness, etc.) found that in conversation dyads, self-reported affect was not strongly correlated with perceived affect [11]. In applications such as spoken dialogue systems for tutoring students, we are most interested in knowing a student's *internal* level of certainty.

Prior work has not addressed the question of whether the annotator perceptions or self-reports are an accurate reflection of internal certainty. There is no way to precisely measure internal certainty, but we attempt to address this issue by eliciting speech in a question-answering setting with materials that we hypothesize to be consistently easy or difficult for all individuals. We then generate difficulty scores for each stimulus via crowdsourcing.

In this paper, we present an exploratory analysis of the prosodic characteristics of individual speakers. We find that some speakers produce consistent prosodic expressions of their certainty level, mostly in their pitch, while other speakers show highly inconsistent patterns of speech. Our methodology involves extracting short audio segments of individual words from spoken answers to questions that varying in the speaker's level of certainty. Similar to recent work that has applied principal components analysis (PCA) to large sets of low-level acoustic-prosodic features [14], we identify a set of 10 principal components from a large set of word-level prosodic features. We then use the smaller set of prosodic features to learn several decision trees for each speaker and analyze the manner and consistency of prosodic expression as a way to gauage individual speaking styles.

2. Harvard Uncertainty Speech Corpus

The speech data that we use in this experiment comes from the Harvard Uncertainty Speech Corpus. This section gives an overview of the Harvard Uncertainty Speech Corpus (Section 2.1), the speech elicitation process (Section 2.2), and the methods of annotating and approximating speaker certainty from the hearer's perspective (Section 2.3), the speaker's perspective (Section 2.4), and according the difficulty of the question (Section 2.5).

2.1. Speech Data from Uncertainty Corpus

The Harvard Uncertainty Speech Corpus contains spoken utterances and level of certainty annotations from three question-answering domains [15]. In this paper, we use the *handwritten digit* section of corpus. The utterances were recorded in a lab, in a question-answering setting. The questions and answers are of the form below.

Q: Which train leaves Los Angeles and at what time does it leave?

A: Train seven leaves Los Angeles at 1:27.

In this experiment, we examine the first two words of such utterances, for example, "train seven" or "train two".

The Harvard Uncertainty Speech Corpus contains the audio corresponding the to answers (not the questions). A notable feature of the utterances in the corpus is that when a speaker is uncertain, the uncertainty can be attributed to a particular *word or phrase* in the utterance. The entire corpus contains 1700 utterances, roughly 150 minutes of speech. The handwritten digit section of the corpus contains 1100 utterances, about 90 minutes of speech. Detailed descriptions of the corpus are available in previously published works [9, 16].

2.2. Background on Method of Speech Elicitation

The speech elicitation materials are designed in a way that controls the difficulty of the stimulus. This is achieved by asking participants to engage in a task that necessitates speaking a spontaneous utterance that incorporates reading handwritten digits that vary in how legible they are. The digit images are drawn from the MNIST database of handwritten digits [17]. The materials for eliciting speech are designed so that participants would speak the selected MNIST digit aloud in the context of answering a question. The handwritten digit images are embedded in illustrations of train routes connecting two U.S. cities, where the handwritten digits indicate the train number. An example illustration is shown in Figure 1.

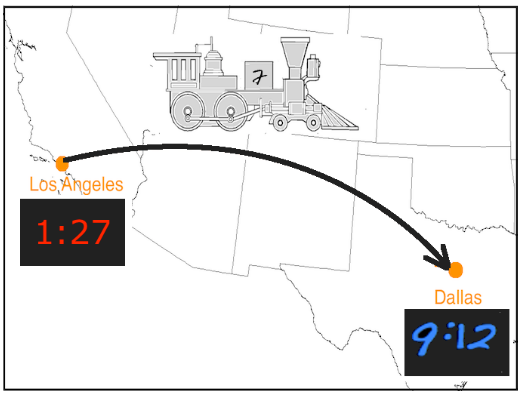

Figure 1: Example speech elicitation illustration featuring an ambiguous handwritten digit image, the train number.

We collected speech from twenty-two native English speakers. At the start of the data collection experiment, participants read a task scenario explaining why they are deciphering handwritten train conductor notes and answering questions about them. For each train route illustration, participants are asked a single question. The participants respond aloud, speaking spontaneously. Their word choice is influenced by a warm-up task where they are given answers to read aloud. This lets us have influence over the length and lexical content of the utterances without the participant explicitly reading aloud.

The method for eliciting uncertain speech is a modification the method used in a previous collection of affective speech [9]. In that work, we did not attempt to control the speaker's level of certainty. As a result, there was no way to verify whether a speaker's self-reported level of certainty was an accurate reflection of his or her actual certainty.

2.3. Certainty Labels from Hearer's Perspective

Each utterance in the corpus is annotated with the level of certainty from a hearer's perspective. We collected annotations from a panel of six human judges. Every annotator listened to and rated the entire set of 1100 utterances. They rated level of certainty on a 1 to 5 scale (1 = very uncertain, 5 = very certain). They did not see any contextual information such as the handwritten images. For each utterance, we consider the mode (average) of the six annotator labels to be the certainty label from the hearer's perspective. The distribution of certainty labels from the hearer's perspective in the corpus is shown in Figure 2.

The agreement among the six annotators highlights the subjective nature of the hearer-centric affect labeling paradigm. Across all pairs of annotators, we find an average pairwise agreement of 54.3%, average Cohen's kappa of 0.235, and average Spearman correlation coefficient of 0.494. If we look only at the pair of annotators with the highest agreement, we see much higher values: pairwise agreement of 74.1%, Cohen's kappa of 0.407, and Spearman correlation of 0.62.

2.4. Certainty Labels from Speaker's Perspective

Each utterance in the corpus is annotated with the level of certainty from the speaker's perspective. The speakers are asked, "How certain were you about the answer you just gave?" during the speech elicitation process. They rate their level of certainty on a 1 to 5 scale (1=very uncertain, 5=very certain). The distribution of certainty labels from the speaker's perspective in the corpus is shown in Figure 2.

2.5. Certainty Labels from Image Difficulty Score

We attempt to control the speaker's actual level of certainty by designing stimuli that are uniformly difficult or easy and we then use crowdsourcing to obtain a difficulty score for each stimulus. Each utterance in the corpus has a legibility score associated with the handwritten digit (the train number) that was used to prompt the question. We used Amazon's Mechanical Turk [18, 19] to collect human judgements from which we generate image legibility scores. Mechanical Turk is an online labor market that facilitates the assignment of human workers to quick and discrete *human intelligence tasks* (HITs). We showed Turkers a digit image and instructed them to identify the digit using a drop-down menu. Each digit was labeled by 100 human workers. Details of the HIT design are available in previously published work [16].

The legibility score for each image is defined as 1 minus the Shannon entropy of the human label distribution:

$$\text{Legibility} = 1 - \Big[-\sum_{i=1}^{N} P(x_i) log P(x_i) \Big]$$

Thus, scores fall in the range [0,1]. A score of 1 has an entropy of 0 and indicates high legibility (all 100 people choose the same label). The handwritten digit in Figure 1 has a legibility score of 0.75. The distribution of difficulty scores (legibility scores) for the stimuli used in eliciting the speech data is shown in Figure 2.

3. Experiment and Results

In this experiment we analyze the speaking styles of individual speakers. We explore whether these individuals are prosodically expressive regarding level of certainty, and if they display consistency in their prosodic expression. Because of design of the corpus, each speaker utters phrases such as "train one" or "train two" multiple times, with differing levels of certainty. Figure 3 shows the spectrograms of three utterances from the same speaker saying "train two" while feeling uncertain, neutral, and certain. Because the corpus contains such sets of lexically-identical phrases, in this experiment we compare

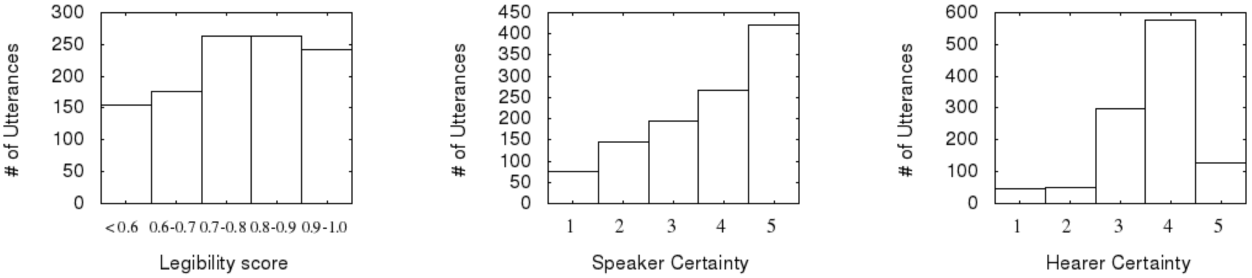

Figure 2: Three histograms: (left) distribution of difficulty scores for the stimuli that prompted the utterances in the corpus, (middle) distribution of certainty labels from the speaker's perspective, (right) distribution of certainty labels from the hearer's perspective.

word-level prosodic features. However, of the ten possible digits, only some are repeated with enough frequency to be analyzed. From the larger corpus, we identify a set of speakers and digits such that,

- each speaker utters each digit 3 or more times, *and*
- for each speaker-digit combination, the certainty labels are distributed among certain, neutral, and uncertain.

This yields a set of 408 utterances representing eight speakers and six digits. Table 1 shows the utterance counts for these eight speakers. For example, in our corpus, speaker a says "train one" five times and says "train two" eight times.

Table 1: Number of utterances that contain the phrases, "train one", "train two", "train three", "train five", "train seven", and "train nine", for a subset of speakers in the corpus.

	Num instances of "*train...*" per speaker						
Speaker	*one*	*two*	*three*	*five*	*seven*	*nine*	**Total**
a	5	8	4	7	8	8	**40**
b	4	8	6	6	9	10	**43**
c	4	11	4	6	6	12	**43**
d	3	8	5	6	8	12	**42**
e	3	10	5	6	7	8	**39**
f	4	7	7	5	7	11	**41**
g	3	6	5	6	9	8	**37**
h	6	7	3	7	9	7	**39**

3.1. Unit of Analysis

Because the corpus contains repeated instances of specific words, spoken with different levels of certainty, we perform prosodic analysis at the *word level*. The segments of interest are the train numbers, which correspond to the MNIST handwritten digits. The word-level audio segments are generated semi-automatically. We use the CMU Sphinx speech recognition toolkit to automatically transcribe each utterance and generate word alignments. The audio segments are manually verified and errors are manually corrected.

3.2. Prosodic features

Initially, we extract 230 prosodic features from each audio segment. We use the openSMILE feature extraction toolkit [20] with the `emobase` config file. The features include low-level descriptors (F0, F0-envelope, intensity, loudness, voice quality, and zero-crossing rate), functionals, and delta regression coefficients for smoothed feature contours.

We use principal component analysis to identify 10 principal prosodic components of the digit word segments in our data (using the entire corpus—word segments from all utterances of all speakers). PCA is performed using the WEKA toolkit [21]. The ranked results are aggregated and a set of 10 principal components are identified for further analysis. The resulting 10 features are listed below (delta features indicated by ^{d}).

1. F0 average^d
2. F0 range^d
3. F0 slope^d
4. F0 skewness^d
5. F0 envelope max^d
6. Intensity average^d
7. Intensity skewness^d
8. Intensity minimum
9. Probability of voicing^d
10. Zero-crossing rate^d

3.3. Speaker analysis

In order to understand how these features are related in predicting the level of uncertainty in utterances, we have made use of decision tree learning. Considering certainty labels from the speaker's perspective (3 classes), we learn separate decision tree classifiers for each speaker-word combination. That is, we learn a decision tree for [$speaker = a$, $word =$ "one"], [$speaker = a$, $word =$ "two"], and so on. In total, we learn six decision trees for each speaker. The maximum depth of decision trees is 3. We used the WEKA toolkit [21] implementation of C4.5 algorithm (J48).

For each speaker, we evaluate whether the learned decision criteria are consistent across all six words. In other words, we ask: for speaker a, are the informative prosodic features consistent for the word "one", the word "two", the word "three" and so on. We then do the same analysis for certainty labels from the hearer's perspective, and for the difficulty score approximation of certainty. Table 2 shows the speaker-specific consistency results. The 10 prosodic features are collapsed into three groups: pitch (#1-#5), intensity (#6-#8), and voice (#9-#10). Separate results are shown for the three approximations of certainty: labels from the speaker's perspective, labels from the hearer's perspective, and difficulty of the question (legibility

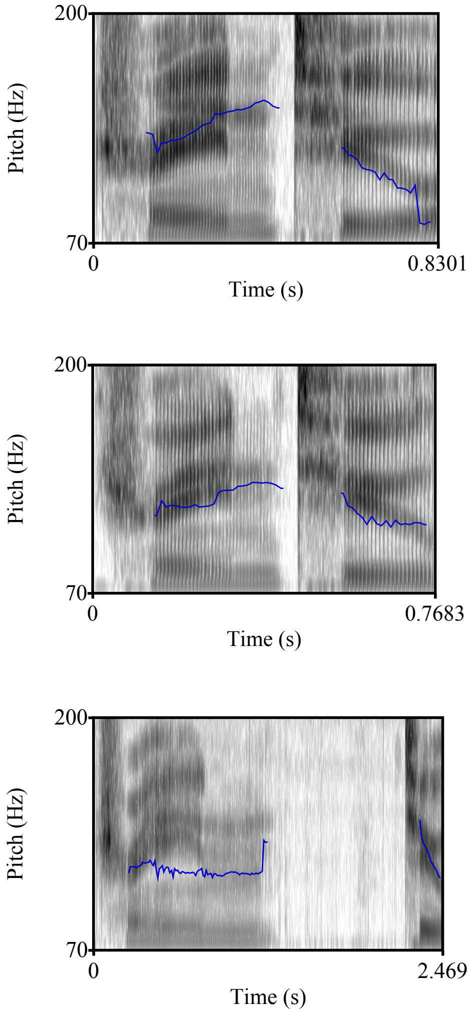

Figure 3: Three instances of a single speaker saying "train two" with varying affect: certain (top), neutral (middle) and uncertain (bottom). The pitch estimate (blue line) is overlaid atop the spectrogram.

score). For speakers that used consistent modes of prosodic expression, checkmarks indicate the class of prosodic feature that distinguished the certain, neutral, and uncertain words.

4. Discussion

We see two primary observations from this exploratory analysis. First, among the prosodic features that we analyzed, features related to pitch are the strongest differentiators between certain and uncertain affect, voice features are second strongest. Second, this analysis, though preliminary, suggests that some speakers consistently display their certainty through their prosody while others are inconsistent. We hypothesized that inconsistent speakers would be inconsistent across all three certainty metrics (speaker, hearer, and legibility). The results show only a small amount of support for this: speakers f and h are inconsistent under both the speaker and hearer metrics. The fact that there is no overlap between the inconsistent speakers in the bottom section of Table 2 and the other two sections indicates that the difficulty score metric may be too coarsely defined (with two binary classes), or that it may not be as aligned with speaker and hearer certainty labels as we had posited.

Table 2: Speaker-specific prosodic modes for conveying uncertainty. Certainty is approximated in three ways: the speaker's perspective, the hearer's perspective, and the difficulty of the question.

	Speaker							
	a	b	c	d	e	f	g	h
Certainty labels: *speaker*								
Pitch	✓		✓	✓	✓			
Intensity					✓			
Voice	✓		✓		✓			
Inconsistent		✗				✗	✗	✗
Certainty labels: *hearer*								
Pitch		✓	✓		✓		✓	
Intensity								
Voice			✓					
Inconsistent	✗			✗		✗		✗
Difficulty of question								
Pitch	✓	✓		✓		✓	✓	✓
Intensity								
Voice								
Inconsistent			✗		✗			

We see that some speakers, e.g., speakers a and c, have similar manners of conveying certainty. Our next steps involve a clustering analysis to explore whether natural clusters can explain the variation seen among those speakers who convey their certainty in their prosody.

5. Conclusion

This paper presents an exploratory analysis of the prosodic characteristics of individual speaking styles, as they relate to three different measurements of certainty. We find that some speakers have consistent ways of conveying their level of certainty prosodically, while other speakers are inconsistent. Among the prosodic signals, pitch-related features are the strongest. Across the three different measures of certainty: speaker's perspective, hearer's perspective, and item difficulty, we find more varying speaker behaviors, suggesting the need for further analysis.

This work is of broad relevance to researchers studying affect recognition. Robustly recognizing affect, and especially subtle affective-cognitive states such as uncertainty, faces many challenges. It is not surprising that speakers have different ways of prosodically expressing affect. In this paper, we show that some speakers produce consistent prosodic signals of certainty, mostly in their pitch, while other speakers show highly inconsistent patterns of speech. Instead of using the same techniques for detecting affect in all speakers, there is great potential utility in adaptive affect detection. For example, if a person is very inconsistent in their speech signals, then an intelligent, multimodal system should direct its inference efforts toward signals from other modalities such as lexical content or facial expressions. On the other hand, if a speaker is prosodically expressive, adaptive systems in the future may dynamically determine which prosodic signals to weight more strongly.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, January 2001.

[2] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[3] R. Fernandez and R. Picard, "Classical and novel discriminant features for affect recognition from speech," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 473–476.

[4] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[5] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002, pp. 2037–2040.

[6] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *Proceedings of Interspeech*, 2005, pp. 513–516.

[7] J. Liscombe, J. Hirschberg, and J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 1837–1840.

[8] H. Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 74–77.

[9] H. Pon-Barry and S. M. Shieber, "Recognizing uncertainty in speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 251753, 2011.

[10] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in *Proceedings of Interspeech*, 2011, pp. 3201–3204.

[11] R. Ranganath, D. Jurafsky, and D. A. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Computer Speech and Language*, vol. 27, no. 1, pp. 89–115, 2012.

[12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[13] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[14] N. G. Ward and A. Vega, "A bottom-up exploration of the dimensions of dialog state in spoken interaction," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Seoul, South Korea, 2012, pp. 198–206.

[15] H. Pon-Barry, S. M. Shieber, and N. Longenbaugh, "Eliciting and annotating uncertainty in spoken language," in *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC)*, 2014.

[16] H. Pon-Barry, "Inferring speaker affect in spoken natural language communication," Ph.D. dissertation, Harvard University, 2013.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[18] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[19] W. Mason and S. Suri, "Conducting behavioral research on Amazon's Mechanical Turk," *Behavior Research Methods*, vol. 44, pp. 1–23, 2011.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia, MM '10*, 2010.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.