CAN PROSODY INFORM SENTIMENT ANALYSIS? EXPERIMENTS ON SHORT SPOKEN REVIEWS

François Mairesse, Joseph Polifroni†*

Nokia Research Center Cambridge
4 Cambridge Center, Cambridge, MA 02142, USA

Giuseppe Di Fabbrizio

AT&T Labs - Research, Inc.
Florham Park, NJ 07932, USA

ABSTRACT

While most online content is created using textual interfaces, recent improvements in speech recognition accuracy allows the creation of content *through speech*. This technology allows users to share reviews about entities of interest without any delay, using mobile devices. This paper builds on the previous work on textual sentiment analysis to investigate whether information in the speech signal can be used to predict sentiment from short spoken reviews. For this purpose we collected a short spoken reviews from 84 speakers. Results show that models trained on features characterizing the review's pitch significantly outperform a majority class baseline, *without textual information*. When taking text-based sentiment predictions into account, our results suggest that prosody can alleviate the effect of speech recognition errors on sentiment detection, however a larger dataset is needed to test whether this can be done without harming performance on low word error rates.

Index Terms— sentiment analysis, opinion mining, prosody

1. INTRODUCTION

Online content is typically generated using text-based interfaces, with users entering reviews, comments, and status updates via keyboard. The growing number of mobile devices enabled with large vocabulary automatic speech recognition (ASR) technology makes a new input modality available for consumer-generated media: voice. Not only can reviews be entered more easily, content created on-the-fly is likely to reflect the user's experience more accurately, as the information is still fresh in the user's mind.

In this paper we report on recent experiments in processing spoken reviews for automatic sentiment detection. We look at the sentiment classification performance using ASR hypotheses compared with that for human transcriptions. In order to fully exploit the information contained in the speech signal, we also examine the use of prosody, something unavailable in text reviews, as a feature in sentiment classification.

The paper is organized as follows. Section 2 describes related work in sentiment detection from text and emotion recognition in speech. Section 3 describes the collection of a corpus of spoken reviews, as well as the collection of a larger set of short textual reviews to train a baseline model (see Section 4). Section 5 details our ASR engine for decoding spoken reviews. Section 6 describes our experimental framework, and Section 7 presents our sentiment classification results. Finally, Section 8 discusses implications of our results and concludes this paper.

*Now working at Vlingo Corporation, Cambridge, MA.
†Now working at Quanta Research Center, Cambridge, MA.

2. RELATED WORK

Traditionally, work on sentiment detection has focused on text, specifically consumer-generated reviews written for websites providing information about a variety of consumer products. The initial work in this field focused on predicting polarity from user reviews, i.e., whether a particular review was primarily positive or negative in its assessment [1]. As the technology matured, it became possible to determine a more fine-grained rating, indicating a scale of sentiment [2], as well as to predict sentiment toward individual attributes of the entity of interest [3]. While this line of research has focused only on written reviews, we have shown in recent work that sentiment can be reliably extracted from ASR hypotheses [4].

In parallel, the problem *emotion recognition* has recently received much of attention in the speech community. While it differs from *sentiment analysis*, we believe that similar techniques can be used to model both problems. Research into the automatic detection of emotion within speech varies by the type and number of emotions targeted for classification, the type of database (e.g., acted vs. spontaneous), and, the classification scheme and features used (see [5] for a survey). For the most part, the speech signal itself has been used to derive the features for classification ([6, 7], *inter alia*), however adding linguistic information has been shown to improve performance [8]. Furthermore, Schuller *et al.* have found that there is little degradation in performance between the use of human and ASR transcriptions of speech [9].

In this work, we examine the use of prosodic features for *sentiment analysis*, in combination with linguistic features derived from automatically recognized speech. The following section describes how we collected sentiment-labeled speech data.

3. DATA COLLECTION FOR SPOKEN REVIEW CLASSIFICATION

Since we do not know of any existing sentiment-labeled speech dataset, we designed an experiment to elicit short spoken reviews from a large set of speakers. We collected spoken review summaries from 84 participants who were given mobile phones instrumented with data collection software. They were asked to use the phones to answer questions about a restaurant where they had eaten, with the only restriction on recruitment being that the experience had occurred within the preceding two weeks. To ensure that at least some of our data were reflective of the contexts in which mobile phones are used, we asked 25% of our users to record their utterances on a sidewalk next to a busy street.

There were a total of nine questions in all, the first seven being factual questions about the restaurant itself (i.e., name, location, and cuisine type), as well as a series of questions in which the users were asked to assign a numeric rating to the overall experience as well as to individual attributes for each restaurant (i.e., food quality,

service, and ambiance). In one of the final two questions the user was asked to review the restaurant in his/her own words, with no restriction placed on length or content. The final question also asked for a shorter review: "If you posted status updates or tweets, what would you say about the experience?". The answer to this question is referred to as a *short review* throughout the rest of this paper. In these utterances, users summed up the experience in short, pithy statements, similar to a social network status update or a tweet.

The rest of this paper focuses on predicting sentiment from such short reviews, because (a) we believed that they would contain a larger proportion of sentiment-related linguistic and prosodic cues; and (b) they match what we would expect from our target application in which users share their reviews on-the-fly using their mobile device. While our dataset only includes the short reviews, it is important to note that the answers to all questions are used for speaker normalization (see Section 6).

To reduce data sparsity, mixed reviews are considered as negative. Reviews of restaurants that received an overall ranking of four or five out of five were assigned to the *positive* class; those below four were assigned to the *negative* class, resulting in 52 positive and 32 negative reviews.

While this data collection effort produced real reviews in realistic conditions, time constraints associated with recruiting subjects and providing (and retrieving) handsets prevented us from collecting large datasets needed to train robust models for this paper. Small datasets are especially problematic in discrete feature spaces such as required for text-based classification. In order to get sufficient data to train text-based models, we set up a text-based data collection to collect realistic data, i.e., close to our target application but without having to record speech. For this, subjects were asked to (a) read a review from the we8there.com website; (b) write down a short (fewer than 160 characters) summary of that review that they could imagine speaking upon leaving the restaurant; and (c) indicate whether the overall sentiment of the summary was negative, positive, or both. To ensure a balanced dataset, only reviews with original ratings either below 2 or above 4 out of 5 were presented, in equal proportions. The data collection was crowdsourced through Amazon Mechanical Turk, which resulted in a corpus of 3,268 textual review summaries produced by 384 individual annotators, out of which 1055 were rated as negative, 1600 as positive, and 613 as mixed.

4. TEXT-BASED CLASSIFICATION BASELINE

Since a large part of an utterance's sentiment is conveyed through text, our first focus was to build a baseline text-based sentiment classifier, with the goal of classifying ASR outputs. Following on previous work on text-based sentiment analysis [1], we trained a support vector machine (SVM) classifier on the textual reviews described in Section 3, in order to predict whether a review is positive, negative, or both from binary n-gram feature indicators (with $n = 1, 2, 3$). We found that a linear kernel produced the best results on our dataset, yielding a 76% accuracy over a 10-fold cross-validation, whereas the majority class baseline produces 49%. In the following sections, we use the text-based classifier trained on the full corpus of 3,268 textual review summaries.

5. SPEECH RECOGNITION FOR SPOKEN REVIEWS

The AT&T Watson [10] speech recognizer was used to automatically convert each spoken review summary to text. Due to the scarcity of speech data available for the restaurant review domain, we had to rely on text sources to create a language model appropriate for this task. As baseline, we mined approximately 87k reviews describing more than 6k restaurant businesses from the citysearch.com website (*CitySearch*). Each document was split into sentences, tokenized and normalized to better mimic the length and the format of the transcribed spoken data. A similar data preparation procedure was applied to the other datasets described in Table 1. The *AMT text* dataset is composed of short text reviews summarized by the Amazon turkers described in Section 3. The *GoodRec* data is a set of short restaurant and bar recommendations mined from the goodrec.com website. Bigram Katz's backoff language models were created from each dataset. The final language model was composed by linear interpolation of the three language models, each weighted by the coefficient λ in Table 1, which was estimated by minimizing the perplexity of a development set. To match the mobile acoustic characteristics, all experiments used a triphonic HMM acoustic model originally developed for a voice search mobile service. When the resulting language model is tested with the spoken review summaries described in section 3, the overall word accuracy is 56.8%. In fact, an error analysis shows that most of the mistakes can be attributed to out of vocabulary proper names, such as the restaurant mentioned in the reviews (e.g., *at Oleana / not only on a*), and various other misspellings and transcription inconsistencies. However, this should be considered as bootstrapping system created without any speech data from the target domain and easy to improve when more specific data is collected.

Dataset	Sentences	Vocabulary	λ
CitySearch	651,850	124,730	0.672
AMT text	5,283	7,499	0.189
GoodRec	16,6074	3,6347	0.139

Table 1. Datasets used to train the ASR language model and optimized interpolation weights.

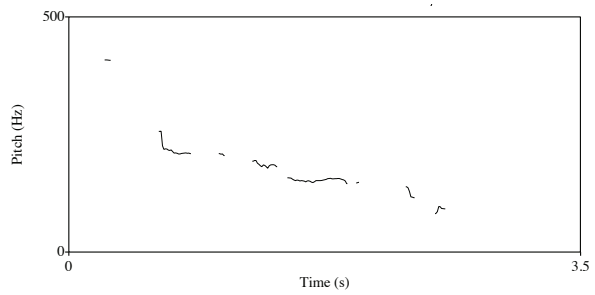
6. SENTIMENT ANALYSIS FROM ACOUSTIC FEATURES

As we are not aware of any existing study on the prosody of opinion, we adopt an empirical approach to feature and model selection. We extract acoustic features from the spoken review summaries using the openEAR/openSMILE toolkit [11]. This toolkit was designed for emotion recognition; however we believe that its features are relevant for sentiment analysis. We extracted 988 features from openEAR's `emobase` feature set, which computes values for the following characteristics of the signal over a 25 ms window, every 10 ms: Intensity; Loudness; 12 MFCC; Pitch (F0); Probability of voicing; F0 envelope; 8 Line spectral frequencies; and Zero-crossing rate.

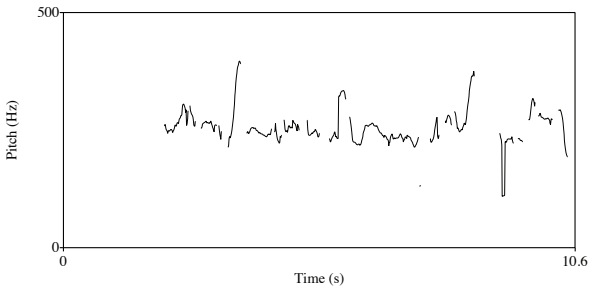
For each of these characteristics, delta regression coefficients are computed. The feature set consists of the following functions of the values of each characteristics as well as the delta coefficients: Max/min values; Relative position of max/min; Range, arithmetic mean; 2 linear regression coefficients and linear/quadratic error; Standard deviation, skewness, kurtosis; Quartile 1, 2, 3, and 3 inter-quartile ranges.

As most reviews consist of a sentence or two, we first extract acoustic features over the whole review. However we noticed that the end of the review summary is often likely to express a strong sentiment, hence we also extract the same features over the last 5 seconds of the spoken review. Figures 1(a) and 1(b) shows an example of the F0 contour for two sample reviews. In the context of a speech recognition system, word alignment and sentence boundary detection could also be used to inform the feature extraction process; however we leave this as future work.

Since the expression of opinion varies highly between individuals, we train two types of models: a *speaker-independent* model



(a) *'It's a nice restaurant but a little disappointing.'*



(b) *'Had a great time at Tapeo trying their authentic Spanish tapas, I really enjoyed the goat cheese entrees and had a great time with friends!'*

Fig. 1. Example F0 contours for two spoken reviews.

trained on the raw acoustic features detailed above, as well as a *speaker-dependent* model trained on the difference between the raw feature values and an estimate of the feature values of generic sentiment-free speech for that speaker. The sentiment-free speech is approximated by taking the average feature values over all the speech samples collected for that speaker, which include a longer review as well as utterances describing the restaurant's name and the rating of individual attributes (9 sound samples in total, see Section 3). Manual inspection of a portion of those samples suggested that they do not convey sentiment as strongly as the short reviews.

In our experiments we compared different learning algorithms trained using the Weka toolbox [12], including logistic regression, AdaBoost, a C4.5 decision tree and an SVM classifier with a radial-basis function (RBF) kernel. Unless mentioned otherwise all parameters are set to their default values. In order to reduce data sparsity issues, we also investigate different feature selection schemes. Our automated feature selection algorithm selects features with an information gain ratio above 0.1 over the training data.

7. EVALUATION RESULTS

7.1. Quantitative results

A first result is that none of the speaker-independent models outperform the majority class baseline. This is likely to be due to the large inter-speaker variation and the lack of contextual information such as gender and age. Speaker normalization improves performance overall, although the decision tree and SVM models trivially learn to return the majority class. Adaboost is the best performing learning algorithm on our dataset, with a 67.8% accuracy. This corresponds to a 5.9% absolute increase over the majority class baseline (61.9%), without modelling any textual content. We therefore use Adaboost with speaker-dependent features throughout the rest of this paper.

Since openEAR computes a large number of acoustic features, feature selection has an important role to play in order to reduce data sparsity issues. Results in Table 2 show that the information gain criteria applied greedily to the full features set only increases performance by 1%. We manually selected different subsets of prosodic features, and found that using only features characterizing F0 improves performance, resulting in a 72.9% classification accuracy from prosody only. A paired t-test shows that the 11% performance increase over the baseline is significant ($p < .05$, two-tailed). This shows that pitch contains useful information for discriminating sentiment, as illustrated by the two F0 contours in Figure 1. Interestingly, performing automated feature selection on F0 features did not improve performance.

Features	**Accuracy**
Majority class baseline	61.9
All acoustic features	66.7
Automated feature selection	67.8
F0 features only	**72.9**
Feature selection on F0 features	71.1

Table 2. Classification accuracy over a 10-fold cross-validation using Adaboost and different speaker-dependent feature sets. Significant improvements over the baseline are in bold ($p < .05$).

Feature combination	**ASR**	**Trans.**
1. Majority class baseline	61.9	61.9
2. Text prediction only/no acoustic features	75.0	**84.4**
3. Automatically selected acoustic features	68.9	77.8
4. F0 features only	**72.6**	**81.0**
5. Automatically selected F0 features only	**82.5**	**81.0**

Table 3. Classification accuracy over a 10-fold cross-validation when including a text-based prediction feature based on the ASR output (*ASR*) or human transcript (*Trans*). Accuracies significantly higher than the baseline are in bold ($p < .05$, two-tailed).

Since sentiment is a semantic concept, we expect text-based models to outperform models relying solely on prosody. However our hypothesis is that combining prosodic and textual information improves overall performance. We transcribed each review summary and computed sentiment predictions for each transcription using the text-based model described in Section 4. We also decoded each spoken summary using the ASR engine described in Section 5, and derived sentiment predictions from each first-best hypothesis. In order to test whether prosodic information improves performance over text-only information, we added a feature indicating the text-based sentiment prediction to the prosody-informed Adaboost model, and compared its performance with using text-based models only.[1]

The second row in Table 3 shows that the text-based model produces an accuracy of 75.0% on ASR outputs and 84.4% on human transcripts. The latter significantly outperforms the majority class baseline ($p < .05$, two-tailed). This result suggests that sentiment detection is robust to ASR errors, since the noise introduced by the ASR only yields a 9.4% decrease in sentiment classification accuracy. Results in Table 3 show that combining ASR information with the automatically selected set of features does not improve over the text-based baseline (row 3, first column). If we assume error-free ASR outputs, we find that acoustic features generally confuse the classifier (cf. rows 3-5 vs. row 2 in the second column). Interestingly, rows 4 and 5 in Table 3 show that adding F0-related features to ASR-based predictions produces the only models which signif-

[1] We did not train text-based models on the 84 transcripts because of the high sparsity of n-gram feature counts.

	Rule condition			Class	α
1	if Δ(F0 at quartile 2 - F0 at quartile 1 in last 5 secs)	$\leq$	-74.0	*neg*	1.2
2	if Δ(F0 at quartile 3 - F0 at quartile 2)	$\leq$	20.3	*neg*	1 .1
3	if Δ(standard deviation of F0)	$\leq$	-23.2	*neg*	.82
4	if Δ(minimum of F0 delta)	$\leq$	-13.3	*pos*	.67

Table 4. Subset of rules learned by the Adaboost model trained on speaker-dependent F0 features (α=rule weight, *neg*=negative class, *pos*=positive class). The Δ symbol represents the difference between the raw feature value and the average value computed on the speaker's sentiment-free speech.

icantly outperform the majority class baseline. While F0 features taken all together are informative, row 5 shows that further automatic feature selection yields the best performance on ASR inputs, resulting in an 7.5% increase over predictions made from ASR information only. Although this increase was not significant over the 10 cross-validation folds, the resulting accuracy of 82.5% is only 3.9% lower than the accuracy reported in the absence of ASR errors (row 2, second column), suggesting that prosody can alleviate the effect of ASR errors for sentiment detection. However, a larger dataset is needed to test whether this can be done without harming performance on low word error rates.

7.2. Qualitative analysis

A benefit of rule-based models is that they provide information about how the features are used for making predictions. The Adaboost algorithm learns a sequence of rules returning a class and an associated weight. Classifying an unseen instance requires summing the weights of the classes returned by the triggered rules, and predicting the class with the highest weight.

Table 4 illustrates a subset of the rules learned by the speaker-dependent model trained on F0 features without textual information, which yields a 72.9% accuracy (see Table 2). Rules 1 and 2 carry the most weight in the model. They indicate that a small inter-quartile range—either below or above the most frequent pitch value, respectively—is indicative of negative sentiment. In other words, this rule suggests that negative content tends to be uttered monotonously, especially during the last 5 seconds (Rule 1). Similarly, Rule 3 indicates that a pitch sample whose standard deviation is at least 23 Hz lower than the average standard deviation in sentiment-free speech is more likely to contain negative sentiment. Rule 4, on the other hand, suggests that the presence in the utterance of a low F0 derivative—i.e., a steep pitch drop—is indicative of positive sentiment.

Overall, it is interesting to note that 40% of the rules learned by the model include a feature computed over the last 5 seconds of the review, confirming our hypothesis that many sentiment-related prosodic cues are located towards the end of the utterance. However, whether this finding holds for longer summaries remains to be tested.

8. DISCUSSION AND CONCLUSION

This paper presents a first attempt at modelling prosody for sentiment detection. We find that features characterizing F0 carry enough information to significantly outperform a majority class baseline without using any textual information. If the utterance's text is known, we find that adding prosodic features confuses the classifier, however not significantly. On the other hand, if only the *ASR hypothesis* is known, we observe that prosody improves performance over a model relying solely on the text-based prediction. While this increase is not significant over 10 cross-validation folds, the addition of prosodic features results in a significant performance increase over the majority class baseline. Since there is a high level of redundancy between text-based and prosodic sentiment cues, in future work we are planning to extend our dataset and study the effect of prosodic features on instances for which the text-based classifier is not confident.

Our results show that sentiment detection can be robust to ASR errors, since we only observe a 9.4% decrease despite the relatively high word error. To investigate the type of words that were associated with ASR errors, we collected a list of 199 common stopwords and found that 41.6% of the deleted words belonged to that list, and that 59.1% of the insertions and 33.9% of the substitutions produced a word in that list. This suggests that robustness to ASR noise comes from the fact that words indicating sentiment tend to be recognized correctly, possibly because of the emphasis typically associated with meaningful content words.

This preliminary work has treated prosodic features and text-based features *independently*, however we believe that overall performance can be improved by training models from word-dependent prosodic features derived from a forced alignment of the ASR engine. Additionally, future work should investigate models with more fine-grained temporal dependencies, such as Hidden Markov Models or Conditional Random Fields.

9. REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of EMNLP*, Philadelphia, PA, 2002, pp. 79–86.

[2] B. Pang and L. Lee, "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the Annual Meeting of the ACL*, 2005, pp. 115–124.

[3] B. Snyder and R. Barzilay, "Multiple aspect ranking using the good grief algorithm," in *Proceedings of HLT-NAACL*, 2007, pp. 300–307.

[4] J. Polifroni, S. Seneff, S. R. K. Branavan, C. Wang, and R. Barzilay, "Good grief I can speak it! Preliminary experiments in audio restaurant reviews," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, Berkeley, CA, 2010.

[5] 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, no. 9-10, pp. 1062–1087, 2011.

[6] S. Yacoub, S. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems," in *Proceedings of Eurospeech*, 2003, pp. 1–4.

[7] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," in *Proceedings of ICME*, 2002.

[8] C. M. Lee and S. S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 293–303, 2005.

[9] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: Putting ASR in the loop," in *Proceedings of ICASSP*, 2009, pp. 4585–4588.

[10] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tür, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar, "The AT&T WATSON Speech Recognizer," in *Proceedings of ICASSP*, 2005.

[11] F. Eyben, M. Wallmer, and B. Schuller, "openEAR - introducing the Munich open-source emotion and affect recognition toolkit," in *Proceedings of HUMAINE ACII*, Amsterdam, 2009.

[12] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, CA, 2005.