A COMPARISON OF CLASSIFIERS FOR DETECTING EMOTION FROM SPEECH

Izhak Shafran

zakshafran@jhu.edu
Center for Language and Speech Processing
Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21209

*Mehryar Mohri**

mohri@cs.nyu.edu
Courant Institute of Mathematical Sciences
New York University
719 Broadway, 12th Floor
New York, NY 10003, USA

ABSTRACT

Accurate detection of emotion from speech has clear benefits for the design of more natural human-machine speech interfaces or for the extraction of useful information from large quantities of speech data. The task consists of assigning, out of a fixed set, an emotion category, e.g., *anger*, *fear*, or *satisfaction*, to a speech utterance. In recent work, several classifiers have been proposed for automatic detection of a speaker's emotion using spoken words as the input. These classifiers were designed independently and tested on separate corpora, making it difficult to compare their performance.

This paper presents three classifiers, two popular classifiers from the literature modeling the word content via n-gram sequences, one based on an interpolated language model, another on a mutual information-based feature-selection approach, and compares them with a discriminant kernel-based technique that we recently adopted. We have implemented these three classification algorithms and evaluated their performance by applying them to a corpus collected from a spoken-dialog system that was widely deployed across the US. The results show that our kernel-based classifier achieves an accuracy of 80.6%, and outperforms both the interpolated language model classifier, which achieved a classification accuracy of 70.1%, and the classifier using mutual information-based feature selection (78.8%).

1. INTRODUCTION

Detecting emotion from speech can be viewed as a classification task. It consists of assigning, out of a fixed set, an emotion category e.g., *joviality*, *anger*, *fear*, or *satisfaction*, to a speech utterance. Accurate detection of emotion from speech has clear benefits for the design of more natural human-machine speech interfaces or for the extraction of useful information from large quantities of speech data. It can help design more natural spoken-dialog systems than those currently deployed in call centers or used in tutoring systems. The speaker's emotion can be exploited by the system's dialog manager to provide more suitable responses, thereby achieving better task completion rates. Emotion detection can also be used to rapidly identify relevant speech or multimedia documents from a large data set.

Several techniques for detecting emotion from speech have been recently described [4, 12, 1, 11, 9, 6, 5, 13, 8, 10]. But the relative performance of these techniques has not been measured since the experiments reported by the authors were carried out on distinct corpora. The results reported are not always indicative of the performance of these techniques in real-world applications. Some are based on the unrealistic assumption that the word transcription of the spoken utterance is given in advance. Others are derived from experiments with speech data produced by professional actors expressing distinct emotion categories.

This paper compares several techniques for detecting emotion by evaluating their performance on a common corpus of speech data collected from a deployed customer-care application (HMIHY 0300). Emotion detection classifiers can use diverse information sources, e.g., acoustic or lexical information. To use a common set of input features, we compared classifiers using spoken words as the input. We present a comparison of three classification algorithms that we have implemented: two popular classifiers from the literature modeling the word content via n-gram sequences, one based on an interpolated language model [6], another on a mutual information-based (MI-based) feature-selection approach [9, 8], and compare them with a discriminant kernel-based technique that we recently adopted [2, 13].

We first give a brief description of the three classifiers evaluated (Section 2) and then present the results of our experiments (Section 3).

2. CLASSIFIERS

In spoken-dialog applications, a speaker's emotion may vary during the course of the interaction, but the dialog-manager processes the speaker's input only after each turn of the dialog. Thus, the problem of detecting emotion can be formulated as that of assigning an emotion category e to each utterance. Two main types of information sources can be used to identify the speaker's emotion: the word content of the utterance and acoustic features such as pitch range. Previous work, including our own experiments with the data used in our experiments suggests that the word content of

*Much of this work was done when this author was affiliated with AT&T Labs – Research.

an utterance is a better indicator of emotion than speaking style [10, 13].[1] Other information sources such as the history beyond the current dialog turn can be captured through the dialog state and used for emotion detection. The classification algorithms presented in this section only use the spoken words at each turn of the dialog, which may be available as a unique word sequence or as a word lattice generated by an automatic speech recognizer, but in most cases they can be enriched to use other information sources. In the following, we describe the three classifiers compared in our experiments.

2.1. Interpolated Language Model Classifier

A classifier for detecting emotion using an interpolated language model was described by [6]. This section gives a brief description of that classification technique.

The problem of emotion detection can be formulated as a classical maximum a posteriori decoding. Let $w = w_1 \cdots w_k$ denote the sequence of words spoken, E a finite set of emotion categories, $e \in E$ an emotion category, and $\mathbf{P}(e \mid w)$ the probability of e given that the sequence w was spoken. The problem consists of finding $\hat{e}$ as defined by:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}} \mathbf{P}(e \mid w) \tag{1}$$

Using Bayes' rule, $\mathbf{P}(e \mid w)$ can be rewritten as: $\frac{\mathbf{P}(w|e)\mathbf{P}(e)}{\mathbf{P}(w)}$. Since $\mathbf{P}(w)$ does not depend on e, the problem can be reformulated as:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}} \mathbf{P}(w \mid e)\, \mathbf{P}(e) \tag{2}$$

where $\mathbf{P}(e)$ is the a priori probability of observing e and $\mathbf{P}(w \mid e)$ the probability of the sequence w given that an emotion category e has been expressed. $\mathbf{P}(e)$ can be estimated by the frequency of e in the training data. For each emotion category $e \in E$, $\mathbf{P}(w \mid e)$ can be modeled by an n-gram statistical grammar $\hat{\mathbf{P}}(w \mid e)$ using a standard smoothing technique such as that the Katz back-off technique [7]. The data available for each emotion category e may be too small to train a robust statistical model for $\mathbf{P}(w \mid e)$. To cope with this problem, one can interpolate the model $\hat{\mathbf{P}}(w \mid e)$ with a general n-gram model $\tilde{\mathbf{P}}(w)$ trained on the data for all emotion categories $e \in E$ using a standard linear interpolation with parameter λ, for $i = 1 \ldots k$:

$$\overline{\mathbf{P}}(w_i \mid E) = \lambda \hat{\mathbf{P}}(w_i \mid e) + (1 - \lambda)\tilde{\mathbf{P}}(w_i) \tag{3}$$

where $\overline{\mathbf{P}}(w_i \mid E)$ is the model resulting from the interpolation. λ determines the trade-off between the two models and can be selected to maximize the likelihood.

The system presented by [6] was designed for classification into five emotion categories:

$$E = \{Anger, Fear, Satisfaction, Excuse, Neutral\}$$

and was based on a unigram model ($n = 1$). For the creation of that system, various special-purpose and language-specific pre-processing procedures (*stemming*, *stopping*, and *compounding*) were also applied to the lexical input, some of them manually. A part-of-speech tagger was used to help with stemming, a stop-list of about hundred words was used to filter out high-frequency words (stopping), and a list of twenty compound words was used to compensate for the limited span of the unigram models used (compounding).

2.2. MI-based Feature-Selection Classifier

Another classifier for emotion detection is presented by [8]. This section briefly describes the component using lexical information for identifying categories. Its main ingredient is the use of mutual information for feature selection.

The main idea behind this feature selection is that not all words, or sequences of words, are relevant attributes for predicting the emotion category of an utterance. To select the most relevant words, the mutual information criterion can be used. The *saliency* of a word w_0 for predicting emotion categories is thus defined as the mutual information between $P(e)$, the probability of an emotion category, and the conditional probability of e given the presence of the word w_0 in the spoken utterance:[2]

$$\mathrm{sal}(w_0) = \sum_{e \in E} \mathbf{P}(e \mid w_0) \log \frac{\mathbf{P}(e \mid w_0)}{\mathbf{P}(e)} \tag{4}$$

This can be used for feature selection when the features used for prediction of the emotion category are words or sequences of words. A subset S of words or word sequences with the highest saliency can be selected as features. In the case of words, a modified version of the maximum a posteriori procedure can be used to determine the emotion category associated with the sequence of words spoken $w = w_1 \cdots w_k$:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}} \Pi_{i=1}^{k} \mathbf{P}(w_i \mid e)\, \delta_S(w_i)\, \mathbf{P}(e) \tag{5}$$

where δ_S is the characteristic function of the set S.

One problem with this approach is that the key measure, $\mathrm{sal}(w_0)$, is not reliable for infrequent words. For example, words that occur only once in the training set and happen to be tagged with a specific emotion category have a high empirical saliency, but this may not generalize to the new occurrences in the test set. Some heuristics can be used to remove such words from the set and improve the robustness of the classifier.

2.3. Kernel-Based Discriminant Classifier

A general framework, *rational kernels*, was recently introduced to extend kernel-based statistical learning techniques

[1] We describe in a more extensive and forthcoming study how acoustic and lexical information can be naturally combined within a common framework to create a classifier more powerful than one based on any one of these information sources alone.

[2] Note that the mutual information criterion could be used for emotion detection by selecting the emotion category that maximizes the mutual information between itself and word sequences. However, this approach is not practical when a relatively limited amount of training data is available.

to the analysis of variable-length sequences or, more generally, weighted automata. Rational kernels are efficient to compute and can be combined with support vector machines (SVMs) to form powerful discriminant classifiers for a variety of text and speech processing tasks [2]. This section gives a brief overview of a specific family of rational kernels, n-gram kernels, that were used in our experiments.

A rational kernel can be viewed as a similarity measure between two sequences or weighted automata. One may, for example, consider two utterances to be similar when they share many common n-gram subsequences. This can be extended to the case of weighted automata or lattices over the alphabet Σ in the following way. A word lattice A can be viewed as a probability distribution P_A over all strings $s \in \Sigma^*$. Modulo a normalization constant, the weight assigned by A to a string x is $[\![A]\!](x) = -\log P_A(x)$. Denote by $|s|_x$ the number of occurrences of a sequence x in the string s. The expected count or number of occurrences of an n-gram sequence x in s for the probability distribution P_A is:

$$c(A, x) = \sum_s P_A(s)|s|_x \tag{6}$$

Two lattices generated by a speech recognizer can be viewed as similar when the sum of the product of the expected counts they assign to their common n-gram sequences is sufficiently high. Thus, we define an n-gram kernel k_n for two lattices A_1 and A_2 by:

$$k_n(A_1, A_2) = \sum_{|x|=n} c(A_1, x)\, c(A_2, x) \tag{7}$$

The kernel k_n is a positive definite symmetric rational kernel or equivalently verifies the Mercer condition [2], a condition that guarantees the convergence of training for discriminant classification algorithms such as SVMs. Furthermore, it can be computed efficiently using weighted transducer algorithms [2]. The sum of two kernels k_n and k_m is also a positive definite symmetric rational kernel [2]. Thus, we can define an *n-gram rational kernel* K_n as the positive definite symmetric rational kernel obtained by taking the sum of all k_m, with $1 \leq m \leq n$:

$$K_n = \sum_{m=1}^{n} k_m \tag{8}$$

The feature space associated with K_n is the set of all m-gram sequences with $m \leq n$. These kernels can be combined with other families of positive definite symmetric kernels, e.g., *polynomial kernels* of degree p defined by $(K + a)^p$, to define more complex kernels, which we used in our experiments.

3. EXPERIMENTS AND COMPARISON

To compare the emotion detection classifiers described in the previous section in a real-world task, we evaluated their performance on data extracted from a deployed customer-care system, the AT&T "How May I Help You" system (HMIHY 0300).

Classifier	Accuracy
a) Interpolated language model classifier	70.1
b) MI-based feature-selection classifier	78.8
c) Kernel-based classifier with one-best	79.9
d) Kernel-based classifier with lattices	80.6

Table 1. Comparison of several classifiers based on spoken word sequences for detecting emotion. The word sequences or lattices used were the output of an automatic speech recognition system.

The corpus used consisted of 5147 utterances from 1854 speakers. The emotion category of the speaker for each utterance was originally tagged into one of seven emotion categories [13]. For this study, they were grouped into only two categories – negative and non-negative. This is similar to the categories used in the experiments carried out by [8]. The utterances were presented to human annotators in the order of occurrence, thus they had the advantage of knowing the context beyond the utterance being labeled. Note, this is unlike some of the previous studies [8, 6] where the utterances were presented in a random order. On the average, the utterances were about 15 words long. A subset of 448 utterance was used for testing on which two human labelers were in full agreement. For further details on the corpus used and the consistency of annotations, refer to [13].

The input to the classification task consisted of the word sequences or lattices generated by an automatic speech recognition system whose word error rate on this data set was 37.8%. This contrasts with some previous work, e.g., [6, 8], where manual transcriptions were used instead.

The classifiers were tested using a range of parameters. The interpolated language model classifier was evaluated with the interpolation parameter varying from 0.5 to 1 in steps of 0.1. Both the interpolated language model classifier and the rational kernel-based discriminant classifier were evaluated with n-gram orders of one to five. The mutual-information-based feature-selection classifier was evaluated only with unigrams due to the unreliable estimates of infrequent contexts in limited training data. This classifier has two other parameters that were varied to determine their best performance, namely, the cardinality of the set of salient words S and the count threshold for ignoring infrequent words.

Table 1 summarizes the results of our experiments. With the interpolated language model classifier, unigram models performed as well as the higher-order n-gram models. The best results were obtained with the interpolation parameter $\lambda = 0.8$. As an alternative to model interpolation, we also experimented with the standard count merging of the n-gram counts of the specific and general models, but this did not lead to any improvement over the results obtained with model interpolation. The MI-based feature-selection classifier yielded significantly better results than the interpolated language model classifier: an improvement of the classification accuracy by 7.7% absolute. This suggests that feature selection plays a crucial role for emotion detection

with a classifier based on an n-gram model since that is the key difference between the interpolated language model classifier and the MI-based feature-selection classifier. This result was obtained when infrequent words below 8 occurrences were ignored in computing mutual information, and when the set of salient words S was reduced to the top 350 most salient words. While S included words such as *disconnect*, *good*, *yes*, *correct* and *cancel* that could be viewed by humans as indicative of an emotion category for the corpus used, it also contained a number of seemingly uninformative words such as *hi*, *couple*, *see* and *name*.

The classifier based on rational kernels combined with SVMs outperformed the previous two classifiers with an accuracy gain of 1.1% absolute over the best one of them. This could be further improved by using the full word lattices generated by the speech recognition system (80.6% accuracy). The best result was obtained with an n-gram kernel of order four ($n = 4$). The design of the kernel-based classifier does not rely on the definition of a specific subset of words since that can introduce a bias. Moreover, the generalization bounds for SVMs do not depend on the dimension of the feature space. The results show the benefits of the use of a kernel-based large-margin classification system that can be used with the word lattices generated by a speech recognition system.

4. CONCLUSIONS

We presented a comparison of three automatic classification algorithms for detecting emotion from a speaker's word content. The results reflect the performance of these classifiers in a real-word task since the data used in our experiments was extracted from a deployed customer-care system, (HMIHY 0300). They demonstrate that the discriminant classifier based on rational kernels outperforms the two other popular classification techniques.

There are many other rational kernels, e.g., complex gappy n-gram kernels or *moment kernels*, kernels exploiting higher-order moments of the distribution of the counts of sequences [3], that could be explored and that could perhaps further improve the classification accuracy in these experiments. The kernel framework also provides a flexible way of using other information sources for emotion detection. We are showing in a forthcoming and longer article how acoustic information can be combined with the lexical information within this framework and lead to significant improvements. Features related to the dialog or semantic features could also be used to improve the accuracy of emotion detection.

5. ACKNOWLEDGMENT

We thank Michael Riley for his multiple contributions to the comparison presented in this paper, which included the formulation of the design principles for annotation of the corpus used, the classifier based on acoustic features [13], general insights into the problem of emotion detection, and help with improving a previous draft of this paper.

6. REFERENCES

[1] A. Batliner, K. Fisher, R. Huber, J. Spilker, and E. Noth. Desperately seeking emotions: actors, wizards, and human beings. In *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 195–200, 2000.

[2] C. Cortes, P. Haffner, and M. Mohri. Rational Kernels: Theory and Algorithms. *Journal of Machine Learning Research (JMLR)*, 5:1035–1062, 2004.

[3] C. Cortes and M. Mohri. Distribution Kernels Based on Moments of Counts. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*, Banff, Alberta, Canada, July 2004.

[4] F. Dellaert, T. Polzin, and A. Waibel. Recognizing Emotions in Speech. In *Proceedings of the International Conference on Spoken Language Processing*, volume 3, pages 1970–1973, Philadelphia, PA, 1996.

[5] L. Devillers and I. Vasilescu. Prosodic cues for emotion characterization in real-life spoken dialogs. In *Proceedings of European Conference on Speech Communication and Technology*, 2003.

[6] L. Devillers, I. Vasilescu, and L. Lamel. Emotion detection in task-oriented dialog corpus. In *Proceedings of the IEEE Int'l Conference on Multimedia*, 2003.

[7] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Speech and Audio Processing*, volume 35 (3), pages 400–01, 2003.

[8] C. M. Lee and S. S. Narayanan. Towards detecting emotions in spoken dialogs. In *IEEE Transactions on Speech and Audio Processing*, 2004 (to appear).

[9] C. M. Lee, S. S. Narayanan, and R. Pieraccini. Combining acoustic and language information for emotion recognition. In *Proceedings of the International Conference on Spoken Language Processing*, 2002.

[10] D. J. Litman and K. Forbes-Riley. Predicting emotions in spoken dialog from multiple knowledge sources. In *Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 2004.

[11] V. A. Petrushin. Emotion recognition in speech signal: Experimental study, development, and application. In *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[12] T. Polzin and A. Waibel. Detecting Emotions In Speech. In *Proceedings of Cooperative Multimodal Communication*, 1998.

[13] I. Shafran, M. Riley, and M. Mohri. Voice Signatures. In *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003.