ICSLP 2002 (In Press)

COMBINING ACOUSTIC AND LANGUAGE INFORMATION FOR EMOTION RECOGNITION

Chul Min Lee, Shrikanth S. Narayanan

University of Southern California
Speech Analysis and Interpretation Laboratory
Electrical Engineering and IMSC
Los Angeles, CA

Roberto Pieraccini

SpeechWorks International
New York, NY

ABSTRACT

This paper reports on emotion recognition using both acoustic and language information in spoken utterances. So far, most previous efforts have focused on emotion recognition using acoustic correlates although it is well known that language information also conveys emotions. For capturing emotional information at the language level, we introduce the information-theoretic notion of 'emotional salience'. For acoustic information, linear discriminant classifiers and k-nearest neighborhood classifiers were used in the emotion classification. The combination of acoustic and linguistic information is posed as a data fusion problem to obtain the combined decision. Results using spoken dialog data obtained from a telephone-based human-machine interaction application show that combining acoustic and language information improves negative emotion classification by 45.7% (linear discriminant classifier used for acoustic information) and 32.9%, respectively, over using only acoustic and language information.

1. INTRODUCTION

The importance of emotion recognition in human speech, e.g. automatic dialog systems in call centers, has increased in recent days to improve both the naturalness and efficiency of human-machine interactions [1]. Automatic dialog systems with the ability of recognizing emotions can comfort callers by changing the response accordingly or passing the calls over to human operators. Automatic emotion recognizers are systems that assign category labels to emotion states. While cognitive theory in psychology argues against such categorical labeling [2], it provides a pragmatic choice, especially from an 'engineering standpoint'.

In this paper, we favor the notion of application dependent emotions, and thus focus on a reduced space of emotions, in the context of developing algorithms for conversational interfaces. In particular, we focus on recognizing 'negative' and 'non-negative' emotions from speech data. The detection of negative emotions can be used as a strategy to improve the quality of the service in call center applications. In previous work, we presented results for emotion recognition based on acoustic information [3]. Here, we propose combining both acoustic and language information in a principled manner to detect two emotion states in spoken dialog.

Acoustic correlates related to prosody of speech, such as pitch, energy, and speech rate of the utterances, have been used for recognizing emotions [4, 5]. But, additional linguistic information would be useful; for example, the use of swear words, and the repetition of the same sub-dialog [6]. A scheme to combine 'content-based' information with acoustic features was proposed in [7], in which the authors used details about topic repetition as their 'language' information.

In this paper, we combine the emotion information conveyed by words (and sequence of words) with that from acoustic features. People tend to use specific words to express their emotions in spoken dialogs because they have learned how some words are related to the corresponding emotions. In this regards, for example, psychologists have tried to identify the language of emotions by asking people to list the English words that describe specific emotions [8]. Such results would be useful for identifying emotional keywords; our interest is in associating emotions to words in spoken language and it is highly domain dependent. We focus on categorizing negative emotions using data obtained from callers communicating with automatic dialog systems. We obtained the emotional 'keywords' in this data by calculating the emotional salience of the words in the data corpus. The salience of a word in emotion recognition can be defined as mutual information between a specific word and emotion category. Similar ideas have been used in natural language acquisition [9]. In other words, salience of a word is a measure of how much information the word provides about the emotion category.

We, next, consider the problem of combining acoustic and linguistic information for emotion recognition. This can be cast as a data fusion problem. Here, the acoustic and linguistic information streams are assumed independent and that each independent decision rule is known. Because we have two emotion classes, the problem is posed as a binary hypothesis test.

The rest of the paper is organized as follows: Section 2 describes the data corpus used. In Section 3 we explain how to identify the emotionally salient words in the data corpus and make a decision; the decision combination scheme based on acoustic and linguistic information is described in Section 4. Section 5 presents the experimental results, and discussion of the results is in Section 6.

2. DATA CORPUS

The speech data used in the experiments were obtained from real users engaged in a spoken dialog with a machine agent over the telephone for a call center application deployed by SpeechWorks [10]. To provide reference data for automatic classification ex-

periments, the data were independently tagged by two human listeners. Only those data that had complete agreement between the taggers (about 65% of the data) were chosen for the experiments reported in this paper. After the database preparation, we obtained 665 utterances for female speakers with 532 non-negative and 133 negative utterances and 514 for male (392 non-negative and 122 negative emotion-tagged utterances).

3. EMOTIONAL SALIENCE

The strategy here is to "spot keywords" for improving the recognition of emotions. To identify the keywords in the utterances, we adopted the information-theoretic concept of salience; a salient word with respect to a category is one which appears more often in that category than at in other parts of the corpus and is considered as a distance measure from the null words of which the relative frequency in each class is the same. We used a salience measure to find the keywords that are related to emotions in the speech data. While listening to the data for tagging the emotion classes, the listeners reported that they tended to feel negative emotions if they heard certain words in the utterances e.g., "No" or swear words. People tend to use certain words more frequently in expressing their emotions because they have learned the connection between the certain words and the related emotions. This is a topic well-studied in psychology [8].

For calculating emotional salience, first we denote the words in the utterances by $W = \{w_1, w_2, \ldots, w_n\}$ and the set of emotion classes by $E = \{e_1, e_2, \ldots, e_k\}$ (here $k = 2$, negative and non-negative), and then the self mutual information is given by [11]

$$i(w_n, e_k) = \log_2 \frac{P(e_k|w_n)}{P(e_k)} \tag{1}$$

where $P(e_k|w_n)$ is the posterior probability that an utterance contain word w_n implies emotion class e_k, and $P(e_k)$ denotes the prior probability of that emotion. We can see that if the word w_n in an utterance highly correlates to an emotion class, then $P(e_k|w_n) > P(e_k)$, and $i(w_n, e_k)$ is positive. Whereas, if the word w_n makes a class e_k less likely, $i(w_n, e_k)$ is negative. If there is no effect by the word, $i(w_n, e_k)$ will be zero because $P(e_k|w_n) = P(e_k)$. The emotional salience of a word for emotion category is defined as mutual information between a specific word and emotion class,

$$sal(w_n) = I(E; W = w_n) = \sum_{j=1}^{k} P(e_k|w_n) i(w_n, e_k) \tag{2}$$

That is, emotional salience is a measure of the amount of information that a specific word contains about the emotion category. Illustrative examples of salient words in the data corpus are given in Table 1. Emotion here represents the one maximally associated with the given word. After identifying the salient words, we removed all the proper nouns such as names of person and places since they may not convey any emotions on their own. Salience of a word can, however, be extended to include a word pair or a word triplet. For example, the word "Damn" would be followed by "It" rather than "Damn" itself, and thus we may build salient word pairs. However, we focus on single words in this paper. Such extensions will be explored in future work.

Word	Salience	Emotion
You	0.73	negative
What	0.66	negative
No	0.56	negative
Damn	0.47	negative
Computer	0.47	negative
Delayed	0.26	non-negative
Baggage	0.25	non-negative
Right	0.01	non-negative

Table 1. A partial list of salient words in the data. "Emotion" represents maximally correlated emotion class given words, i.e., the emotion class that maximizes the posterior probability of emotion given a word

4. DECISION METHODS ON ACOUSTIC AND LANGUAGE INFORMATION

For the decision/classification using acoustic features, we used two methods, namely linear discriminant classifiers (LDC) and k-nearest neighborhood (k-NN) classifiers, and the results were reported in our previous study [3]. Briefly, LDC classifies test data after estimating the mean of each class using training data, and k-NN classifiers is a memory-based classifier and its classification is based on majority vote in k number of nearest neighborhood of test data.

When any of the salient words obtained in Section 3 is in the test data, it can be evident that the utterance with those words will belong to the indicated emotion class. We can measure how evident the utterances belong to emotion classes by the posterior probability of emotion given the salient word, $P(E|W)$. If there are several salient words, we multiplied the posterior probability for each word. And the decision is made according to,

$$\arg\max_{e_k} \prod_{\text{salient words}} P(e_k|w) \tag{3}$$

4.1. Combination of Acoustic and Language Information

Let E_0 and E_1 denote non-negative and negative emotions, respectively. We consider the problem of combining acoustic and language information at the decision level [12], and assume they are statistically independent to each other. The decision rule is given by

$$\begin{array}{ll} \frac{P(E_1|A,W)}{P(E_0|A,W)} > 1, & \text{decide } E_1 \\ \text{otherwise}, & \text{decide } E_0 \end{array} \tag{4}$$

where E represents emotion class, A stands for acoustic information, and W denotes language information. Using Bayes' rule,

$$P(E|A,W) = P(E|W)\frac{P(A|E,W)}{P(A,W)} \tag{5}$$

$$\propto P(E|W)P(A|E) \tag{6}$$

In Eq. 6, we drop the normalization factor and use the prior knowledge that W does not affect A. Because of the separation of the posterior probability in Eq. 5 into acoustic and language only, we can make a decision in each information stream as:

$$d_i = \begin{cases} -1, & \text{if } E_0 \text{ is declared} \\ +1, & \text{else} \end{cases} \tag{7}$$

Classification Method		Error,%
Acoustic Only	LDC	30.0
	kNN(k=3)	33.25
Linguistic Only		24.35
Combination	LDC	19.25
	kNN(k=3)	21.0

(a)

Classification Method		Error,%
Acoustic Only	LDC	29.5
	kNN(k=3)	28.0
Linguistic Only		33.77
Combination	LDC	24.75
	kNN(k=3)	24.0

(b)

Table 2. Classification error results for acoustic, linguistic features and the combination of acoustic and linguistic features. We randomly select the 100 training and 20 test data samples for both acoustic and linguistic information for each emotion class; the salient words are obtained from the training data only. And then the results were obtained by averaging 10 independently sampled test data. (a) represents the results in female data and (b) represents the results in male data.

Classification Method		Error,%
Acoustic Only	LDC	30.0
	kNN(k=3)	33.25
Linguistic Only		17.42
Combination	LDC	12.75
	kNN(k=3)	15.25

(a)

Classification Method		Error,%
Acoustic Only	LDC	29.5
	kNN(k=3)	28.0
Linguistic Only		31.92
Combination	LDC	19.5
	kNN(k=3)	19.5

(b)

Table 3. Classification error results for acoustic, linguistic features and the combination of acoustic and linguistic features. We use all the data including female and male data to obtain the salient words in language information represented by 'linguistic only' in the table. And then the results were obtained by averaging 10 independently sampled test data. (a) represents the results in female data and (b) represents the results in male data.

where $i = 0, 1$.

The decision of combined features can be implemented as a logical function [12] and we adopted an "OR" logical combiner, i.e., if either acoustic or language features declared its emotional class to be E_1, then the combined decision is also declared E_1. The combined decision rule, therefore, is given by

$$d = \begin{cases} +1, & \text{if } d_0 + d_1 \geq 0 \\ -1, & \text{else} \end{cases} \quad (8)$$

5. EXPERIMENTAL RESULTS

For acoustic information, we used two pattern classification methods to classify the emotion states conveyed by the utterances: one is LDC and the other is a k-NN classifier. Acoustic features comprise utterance-level statistics obtained from pitch (F0) and energy of the speech data. These include mean, median, standard deviation, maximum, and minimum for F0, and mean, median, standard deviation, maximum, and range (maximum-minimum) for energy information. The parameter of the k-NN classifier, k, was set to be three for both female and male data.

Two training scenarios were considered. In the first one, the training data set and test data set were selected 10 times from the data pool in a random manner. Each training set had 200 utterances (100 utterances from each emotion class), and test set had 40 utterances; 20 from each class. In the second scenario, all the data including both female and male data were used for estimating emotional salience of words. The training data for the acoustic information, and the test data were the same as for scenario 1. The goal here was to explore the role of "out of vocabulary" problem in training data.

The probability $P(E|W)$ for each salient word was estimated by smoothed relative frequencies. Then the decision was made by comparing $P(E|W)$ in the test utterances using Eq. 3. The same test data was used in the decision making for both acoustic and language information. Finally, the combined decision for test data was made using Eq. 8. Experiment results are shown in Tables 2 and 3. In Table 2, the emotional salience of words was estimated by 200 training data randomly selected from all the data pool in each gender, and Table 3 shows the results when the emotional salience of words was decided by all the data (1179 utterances). The results for female and male data are separated into (a) and (b) in each table. The error represents the misclassification error rate averaged over 10 independently chosen test data.

Overall, the results show that we can improve the performance of emotion recognizer significantly by combining acoustic and language information. When we partition the data into training and test for language information, the results for language information only case are worse than those obtained by training using all the available data for estimating the salient words. First, this points out that the training data in the language level is rather sparse and has significant consequences for detection. At the same time, using the training data for testing has the danger of overfitting. This is, in fact, illustrated by the 'linguistic only' results in Table 3. When we look over the list of emotionally salient words, many words come from the female data and; therefore, the results from language information in this case indicates overfitting.

6. DISCUSSION

In this paper, we explored automatic recognition of negative emotions in speech signals using data obtained from a real-world application. Both acoustic and language information were used for the emotion recognition. The results show that significant improvement can be made combining acoustic and language information compared with the results with acoustic information only. Table 2, which gives the results where the emotionally salient words were estimated from a small portion of the data, the relative improvements obtained by combining acoustic and language information were 35.8% for LDC and 36.8% for k-NN in female data, and

16.1% for LDC and 11.1% for k-NN in male data compared with the results obtained using acoustic information only. When combining the female and male results, the improvements are 26.0% and 23.8% for LDC, and 25.6% and 21.3% for k-NN classifiers over using just acoustic and language information, respectively. When salience of words were estimated from all the data, the improvements were 57.5% for LDC and 54.1% for k-NN for the female data, and 33.9% for LDC and 30.4% for k-NN for the male data, again compared with the results using acoustic information only. When combining the female and male results, the improvements are 45.7% and 32.9% for LDC, and 42.3% and 25.7% for k-NN classifiers over just acoustic and language information, respectively. There are several issues that need to be further explored in the future.

First of all, data sparsity is even more a stringent problem for linguistic modeling than at the acoustic level since acoustic and linguistic data are at 2 different scales. In the test phase using language information, many utterances were left undecided due to the fact that the words in certain utterances were not in the list of salient words seen in the training data, even one or more words were apparently related to emotion classes. To explore this problem, we need to experiment on the dependence of language information on the number of salient words and increasing the amount of data in the data corpus. We also need to study effective smoothing techniques to deal with sparsity.

Secondly, in this paper we estimated the emotional salience calculation at a single word level; however, the emotional salience should be extended to word pairs or word sequences. That may lead to a more reasonable estimation of the emotional salience in the sense that human beings can incorporate word sequences to judge emotion states. This should be possible,again, with a larger corpus.

The third issue is that there is previous research on collecting words related to emotion states, the so called 'language of emotion' [8, 13]. If we can combine those word lists as the emotional language lexicon, we may build a more general 'emotional language model'. This is also related to the first issue of the data sparsity since if we can generate a general model of emotional lexicon of a language, we can easily combine it with the domain data in estimating the salience of words. The problem is, however, that most of the words in the lists are generic rather than specific; therefore, we need to find out how to match/adapt the words in the wordlists with the word in the real-world data (especially for a specific application domain).

The fourth issue that should be further explored is how to best combine acoustic and language information. In this paper, we proposed it as a data fusion problem and combined informaion at the decision level using a logical "OR" function. However, there are several other possible combination schemes, e.g., feature level combination or giving different weights to acoustic and language information in Eq. 6. The weights would be determined by confidence score of the acoustic and language decision or relative effects on the decisions, and the formula can be described as:

$$d = \begin{cases} +1, & \text{if} \quad \lambda_1 \log \frac{P(E_1|W)}{P(E_0|W)} + \lambda_2 \log \frac{P(A_1|W)}{P(A_0|W)} \geq Th \\ -1, & \text{else} \end{cases} \tag{9}$$

where λ_1 and λ_2 represent the relative importance in the decision made by language and acoustic information only, and Th is a threshold.

The last issue is about classification methods. Since emotion states do not have clear-cut boundaries, we need to explore and develop the classification methods to deal with this vague boundary problem. This line of study may also give light on integrating other dialog information to improve emotion recognition.

7. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Sig. Proc. Mag.*, vol. 18(1), pp. 32–80, Jan 2001.

[2] A. Ortony, G.L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge Univ. Press, UK, 1988.

[3] C.M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. Automatic Speech Recognition and Understanding*, Dec 2001.

[4] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Artificial Neu. Net. In Engr.(ANNIE '99)*, 1999.

[5] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *ICSLP '96*, Philadelphia, PA, 1996.

[6] S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S.S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[7] A. Batliner, K. Fischer, R. Huber, J. Spiker, and E. Noth, "Desperately seeking emotions: Actors, wizards, and human beings," in *Proc. ISCA Workshop on Speech and Emotion*, 2000.

[8] R. Plutchik, *The Psychology and Biology of Emotion*, HarperCollins College, New York, NY, 1994.

[9] A. Gorin, "On automated language acquisition," *J. Acoust. Soc. Am.*, vol. 97(6), pp. 3441–3461, 1995.

[10] SpeechWorks, "http://www.speechworks.com/index flash.cfm," .

[11] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, 1991.

[12] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-22, pp. 98–101, 1986.

[13] The Balanced Affective Word List Project, "http://www.sci.sdsu.edu/cal/wordlist," .