Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
4(1), 2016, 19-25

ISSN 2347 – 7393

Original Research Article

RECOGNIZING EMOTION IN SPEECH COMBINING LDC AND K-NN WITH RESPECT TO BODO LANGUAGE

Bimal Kumar Kalita, Research Scholar and Pran Hari Talukdar, Professor

Dept. of Instrumentation & USIC, Gauhati University

Abstract: This paper works on the techniques for automated categorization of spoken sounds based on the emotional condition of the speaker. The information employed for the study comes from a corpus of selected natural dialogs recorded in different emotions deployed by Speech Works. In this study, Gaussian class-conditional likelihood distribution with respect to linear discriminant classification (LDC) and K-nearest neighborhood (K-NN) schemes are used to classify utterances into basic emotion states, positive and negative. The utterance level statistics of the fundamental frequency and the energy of the speech signal are used by the classifiers. The promising first selection and forward feature selection are used for feature selection to improve classification performance. The dimensionality of the features reduced to maximize classification accuracy. Improvements achieved up to 87%.

Keywords- Bodo, LDC, K-NN, Gaussian**.**

1. Introduction:

The significance of emotion detection in human speech, e.g. computerized dialog techniques in call centers, has increased in current days to get better both the naturalness and effectiveness of human-machine interactions [1]. Computerized dialog techniques with the knack of recognizing emotions could comfort callers by changing the response accordingly or passing the calls over to human operators. Computerized emotion recognizers were techniques that assign category tags 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 function 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'

For Correspondence:
bimal.kumarkalita@gmail.com
Received on: January 2016
Accepted after revision: February 2016

emotions from speech data. The detection of negative emotions could be used as a strategy to get better the quality of the service in call center applications. Here, we propose combining acoustic and language / linguistic information in a principled way 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 [3, 4]. But, additional linguistic information would be useful; for instance, the work of swear words, and the repetition of the same sub-dialog [5]. A scheme to combine 'content-based' information with acoustic features was proposed in [6], in which the authors used details about topic repetition as their 'language / linguistic' information. In this paper, we combine the emotion information conveyed by words (and sequence of words) with that from acoustic features. People tend to work specific words to express their emotions in spoken dialogs because they have learned how some words were related to the corresponding emotions. In this regards, for instance, psychologists have tried to recognize the language / linguistic of emotions by asking people to list the English words that describe specific emotions [7]. Such results would be useful for identifying emotional keywords; our interest was in associating emotions to words in spoken language / linguistic and it was highly domain dependent. We focus on categorizing negative emotions using data obtained from callers communicating with computerized dialog techniques. We obtained the emotional 'keywords' in this database by calculating the emotional salience of the words in the data corpus. The salience of a word in emotion detection could be defined as mutual information between a specific word and emotion category. Similar ideas have been used in natural language / linguistic acquisition [8]. In other words, salience of a word was 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 detection. This could be cast as a data fusion problem. Here, the acoustic and linguistic information flows were assumed independent and that each independent decision rule was known. Because we have two emotion classes, the problem was posed as a binary hypothesis test. The rest of the paper was organized as follows: Section 2 describes the data corpus used, In Section 3 we explain how to recognize the emotionally salient words in the data corpus and make a decision; the decision combination scheme based on acoustic and linguistic information was described in Section 4. Section 5 presents the experimental results, and discussion of the results was 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 function deployed by Speech Works in the Dept. of INSTRUMENTATION & USIC, Gauhati University. To provide reference data for computerized classification experiments, 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 1000 utterances for female speakers with 400 non-negative and 100 negative utterances and male (400 non-negative and 100 negative emotion-tagged utterances).

3. Emotional Salience:

The strategy here was to "spot keywords" for improving the detection of emotions. To recognize the keywords in the utterances, we adopted the information-theoretic concept of salience; a salient word with respect to a category was one which appears more often in that category than at in other parts of the corpus and was considered as a distance measure from the null words of which the relative frequency in each class was the same. We used a salience measure to find the keywords that were 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 work certain words more frequently in expressing their emotions because they have learned the connection between the certain words and the related emotions. This was 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 was given by [11]:

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

where $P(e_k / w_n)$ was the posterior probability that an utterance contain word w_n implies emotion class e_k, and $P(e_kn)$ denotes the prior probability of that emotion. We could 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 $i (w_n, e_k)$ was positive. Whereas, if the word w_n makes a class e_k less likely, $i (w_n, e_k)$ was negative. If there was 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 was defined as mutual information between a specific word and emotion class,

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

Emotional salience was 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 were 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 could, however, be extended to include a word pair or a word triplet. For instance, the word "Damn" would be followed by "It" rather than "Damn" alone, 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: Partial list of salient words in the database.

Here "Emotion" represents maximally correlated emotion class given words, i.e., the emotion class that maximizes the posterior probability of emotion in a given a word.

4. Decision Methods on Acoustic and Language / Linguistic Information:

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

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

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

4.1. Combination of Acoustic and Language / linguistic Information:

Let E_0 and E_1 denote non-negative and negative emotions, respectively. We consider the problem of combining acoustic and language / linguistic information at the decision level [11], and assume they were statistically independent to each other. The decision rule was given by:

$$\frac{P(E_1|A,W)}{P(E_0|A,W)} > 1, \quad \text{decide } E_1$$
$$\text{otherwise}, \quad \text{decide } E_0 \qquad (4)$$

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

$$P(E|A,W) = P(E|W)\frac{P(A|E,W)}{P(A,W)} \qquad (5)$$
$$\propto P(E|W)P(A|E) \qquad (6)$$

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

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

Classification Method		Error,%
Acoustic Only	LDC	39.43
	kNN(k=3)	32.25
Linguistic Only		27.35
Combination	LDC	18.25
kNN(k=3)	kNN(k=3)	26.76

Classification Method		Error,%
Acoustic Only	LDC	27.51
	kNN(k=3)	26.21
Linguistic Only		38.65
Combination	LDC	27.95
kNN(k=3)	kNN(k=3)	27.52

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 were 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, where i=0,1.

The decision of combined features could be implemented as a logical function [12] and we adopted an "OR" logical combiner, i.e., if either acoustic or language / linguistic features declared its emotional class to be E1, then the combined decision was also declared E1. The combined decision rule, therefore, was given by:

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

5. Experimental Results:

For acoustic information, we used two pattern classification methods to classify the emotion states conveyed by the utterances: one was LDC and the other was a k-NN classifier. Acoustic features comprise utterance-level statistics obtained from pitch (F_0) and energy of the speech data. These include mean, median, standard deviation, maximum, and minimum for F_0, 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 way. 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 / linguistic information.

Classification Method		Error,%
Acoustic Only	LDC	38.12
	kNN(k=3)	35.75
Linguistic Only		16.92
Combination	LDC	13.75
kNN(k=3)	kNN(k=3)	18.25

Classification Method		Error,%
Acoustic Only	LDC	28.52
	kNN(k=3)	27.90
Linguistic Only		32.02
Combination	LDC	18.95
kNN(k=3)	kNN(k=3)	18.99

Table 3. Classification error results for acoustic, linguistic features and the combination of acoustic and linguistic features.

We work all the data including female and male data to obtain the salient words in language / linguistic 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.

Finally, the combined decision for test data was made using Eq. 8. Experiment results were 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 were 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 could get better the performance of emotion recognizer significantly by combining acoustic and language / linguistic information. When we partition the data into training and test for language / linguistic information, the results for language / linguistic information only case were 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 / linguistic level was rather sparse and has significant consequences for

detection. At the same time, using the training data for testing has the danger of over fitting. This was, 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 / linguistic information in this case indicates over fitting.

6. Discussion

In this paper, we explored computerized detection of negative emotions in speech signals using data obtained from a real-world function. Both acoustic and language / linguistic information were used for the emotion detection. The results show that significant improvement could be made combining acoustic and language / linguistic 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 / linguistic information. There were several issues that need to be further explored in the future. First of all, data sparsity was even more a stringent problem for linguistic modeling than at the acoustic level since acoustic and linguistic data were at 2 different scales. In the test phase using language / linguistic 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 / linguistic 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 could incorporate word sequences to judge emotion states. This should be possible, again, with a larger corpus. The third issue was that there was previous research on collecting words related to emotion states, the so called 'language / linguistic of emotion' [8, 13]. If we could combine those word lists as the emotional language / linguistic lexicon, we may build a more general 'emotional language / linguistic model'. This was also related to the first issue of the data sparsity since if we could generate a general model of emotional lexicon of a language / linguistic, we could easily combine it with the domain data in estimating the salience of words. The problem was, however, that most of the words in the lists were 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 function domain). The fourth issue that should be further explored was how to best

combine acoustic and language / linguistic information. In this paper, we proposed it as a data fusion problem and combined information at the decision level using a logical "OR" function. However, there were several other possible combination schemes, e.g., feature level combination or giving different weights to acoustic and language / linguistic information in Eq. 6. The weights would be determined by confidence score of the acoustic and language / linguistic decision or relative effects on the decisions, and the formula could 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 significance in the decision made by language / linguistic and acoustic information only, and Th was a threshold. The last issue was 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 get better emotion detection.

7. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion detection 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, "Detection of negative emotions from the speech signal," in *Proc. Computerized Speech Detection and Understanding*, Dec 2001.

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

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

[5] S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S.S. Narayanan, "Politeness and frustration language / linguistic in childmachine interactions," in *Proc. Eurospeech*, Aalborg, Denmark,
2001.

[6] 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.

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

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

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

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

[11] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection techniques," *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,".