CLASSIFYING EMOTIONS IN HUMAN-MACHINE SPOKEN DIALOGS

Chul Min Lee[1], *Shrikanth S. Narayanan*[1], *Roberto Pieraccini*[2]

[1]Dept. of Electrical Engineering and IMSC, University of Southern California, Los Angeles, CA
[2]SpeechWorks International, New York, NY
cml@usc.edu, shri@sipi.usc.edu, roberto@speechworks.com

ABSTRACT

This paper reports on the comparison between various acoustic feature sets and classification algorithms for classifying spoken utterances based on the emotional state of the speaker. The data set used for the analysis comes from a corpus of human-machine dialogs obtained from a commercial application. Emotion recognition is posed as a pattern recognition problem. We used three different techniques - linear discriminant classifier (LDC), k-nearest neighborhood (k-NN) classifier, and support vector machine classifier (SVC) - for classifying utterances into 2 emotion classes: *negative* and *non-negative*. In this study, two feature sets were used; the base feature set obtained from the utterance-level statistics of the pitch and energy of the speech, and the feature set analyzed by PCA. PCA showed that it can show the performance comparable to base feature sets. Overall, LDC achieved the best performance with error rates 27.54% on female data and 25.46% on males with the base feature set. SVC, however, showed better performance in the problem of data sparsity.

1. INTRODUCTION

Emotion recognition in human speech has obtained increasing attention in recent years as the need for machines to detect/recognize human emotions in human-machine interaction environments has grown [1]. Our motivation for the recognition of negative emotions in speech comes from the increased role spoken dialog systems are playing in human-machine interaction, especially for deployment of services associated with call centers such as customer care. Other applications include a variety of automatic training and education scenarios. Automatic emotion recognizer is a system for assigning category labels that identify emotional states. While cognitive theory in psychology argues against such categorical labeling [3], it provides a pragmatic choice, especially from an 'engineering standpoint'. Focusing on the archetypal emotions --happiness, sadness, fear, anger, surprise, and disgust -- is typically justified as a way into arriving at finer distinctions. For example, Scherer explored the existence of a universal psychobiological mechanism of emotion in speech across languages and cultures by studying the recognition of 5 emotions in nine languages, and obtained 66% overall accuracy [4].

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 the acoustic speech signal. The detection of negative emotions can be used as a strategy to improve the quality of the service in call center applications. While semantic and discourse information also contribute toward emotion recognition, the focus of this paper is on classification based on acoustic information only. In a previous study [2], we presented preliminary results on classification of negative and non-negative emotions. This paper expands upon those ideas including new results using a support vector machine classifier (SVC). Our study showed that SVC would overcome the problem of data sparsity, which is a critical challenge in real-world applications.

Various pattern recognition approaches have been explored for automatic emotion recognition [5][6]. Dellaert et al., for instance, used maximum likelihood Bayes classification, Kernel regression, and k-nearest neighbor methods [7], whereas Roy and Pentland used Fisher linear discrimination method [8]. In this paper we used linear discrimination classifier (LDC), k-nearest neighborhood (k-NN) classifiers and support vector machine classifier (SVC) as the classification methods. Petrushin developed a real-time emotion recognizer using neural networks for call center applications, and achieved ~77% classification accuracy in two emotion states, 'agitation' and 'calm' for 8 features chosen by a feature selection [5].

In this study, we used a corpus of sentences from a human-machine spoken dialog application deployed by SpeechWorks used by real customers [9]. So far, most of the reported studies have used speech recorded from actors that were asked to express pre-defined emotions. A notable exception is the study by Batliner et al [6]. They adopted a 'Wizard-of-Oz' scenario to collect data, which assumed that naive subjects were asked to communicate with a real computer, in two emotion categories such as 'emotional' and 'neutral'.

As to the acoustic features for emotion recognition, prosodic information is most common. McGilloway et al. studied 32 different features for the classification of 5 emotion states [10]. The features are concerned with F0 (pitch), energy, duration and tune (segments of the pitch contour bounded at either end by a pause of 180 ms or more). In our work, we used ten utterance-level statistics derived from F0 and energy as the acoustic features for emotion recognition.

By far, most previous research on front-end signal processing for emotion recognition has focused on a variety of feature sets obtained directly from the speech [5][7]. However, some of the features may be highly correlated and hence may not be optimal. Since we pose emotion recognition in human

speech as a pattern recognition problem, we can apply component analyses such as principal component analysis (PCA) to discover, and reduce, the underlying dimensions of the feature space. Another advantage of using PCA for dimensionality reduction is that the large dimensionality of the feature space can hurt the performance of the pattern classification if the size of the training data is small. Thus, in this work, we adopted PCA for feature reduction and used the resulting features as new feature set in the experiments.

2. SPEECH DATA CORPUS

The speech data (8kHz, mu-law) used in the experiments was obtained from real users engaged in a spoken dialog with a machine agent over the telephone for a call center application deployed by SpeechWorks. To provide reference data for automatic classification experiments, the data were independently tagged by 2 human listeners. Only those data that had complete agreement between the taggers (about 65% of the set) were chosen for the experiments reported in this paper. After the database preparation, we obtained 706 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). In doing experiments, we assumed that the prior probabilities of each class are equal to 0.5 since it is our hope of classifying the emotions in speech with just the given data.

Data sparsity is a critical challenge and a reality in the study of emotion recognition. In collecting emotional speech, it was found that explicit display of negative emotions is relatively infrequent in realistic human-machine interactions. But even in such cases, detecting negative emotions is important, e.g., from the viewpoint of customer service in automated caller systems. Hence, algorithm development for classification should attempt to cope with this issue of data sparsity.

3. FEATURE EXTRACTION

In our experiments, only acoustic features such as pitch and energy related features were calculated from the speech signal. Additional features would be useful for the emotion recognition: such as linguistic information, e.g., the use of swear words [11] and discourse information, e.g., repetition of the same sub-dialog. A scheme to combine those 'content-related' features with acoustic features was proposed in [6], in which they used details about topic repetition as their 'language' information. Here, we focus only on acoustic features.

3.1. Base Feature set and Feature set by PCA

We used two feature sets for the experiments; one is the base feature set and the other is the feature set by PCA. Base acoustic feature set for emotion recognition comprised utterance-level statistics obtained from the pitch (F0) and energy information of the speech signal. These included the mean, median, standard deviation, maximum, and minimum for pitch, and mean, median, standard deviation, maximum, and range (maximum – minimum) for energy. For pitch calculations, only voiced regions were taken into account. To compute the energy of the speech signals, we used a 30 ms Hamming window with 10 ms overlap. Further all the samples were normalized, i.e., the origin was shifted to the means of the features and the variances of all features were scaled to 1.

Another feature set we used was those after preprocessed by PCA. The dimensionality of the feature set with PCA can be reduced from that of base feature set [12].

4. CLASSIFICATION

In our experiments, three classification methods were applied to the data; linear discriminant classifier with Gaussian distribution (LDC), k-nearest neighborhood (k-NN) classifier, and support vector machine classifier (SVC). First two methods were adopted in a previous study [2]. LDC is a parametric method with Gaussian probability distributions as its class-conditional distribution of each class. After estimating the parameters such as means and variances, LDC classifies the class based on the maximum posterior probability using Bayes rule. The k-NN classifier first approximates the local posterior probability of each class by averaging over the *k* nearest neighbors. Classifying the class by k-NN is performed by voting the majority vote over those neighbors. In the experiments, it was shown that k-NN classifiers provided more reliable performance due to the fact that class-conditional probability distribution is not Gaussian. In SVC, the input vectors, $\mathbf{x}$, are mapped onto a high dimensional feature space *F* through some nonlinear mapping. The detailed description on SVC is in the following section.

4.1 Classification by Support Vector Machines

By preprocessing the data to represent patterns in a higher dimension than input space, we can separate them from two classes by a hyperplane in new feature space. Let Φ be the nonlinear mapping from original feature space to new high dimension feature space [12-14].

$$\Phi : R^N \rightarrow F \tag{1}$$

where *N* is the dimension of input feature space and *F* is high dimensional space. After transforming each input feature to $\mathbf{y}_k = \Phi(\mathbf{x}_k)$, k = 1, … , n (number of training patterns), we can find a hyperplane discriminating the new feature space, such as

$$g(\mathbf{y}) = \mathbf{w} \cdot \mathbf{y} + b \tag{2}$$

where b is margin and $\cdot$ denote inner product of two vectors. For two class problem, let $z_k = \pm 1$ denoting which class the *k*th pattern belongs. Thus a separating hyperplace satisfies

$$z_k g(\mathbf{y}) = z_k (\mathbf{w} \cdot \mathbf{y} + b) - 1 \geq 0 , \quad \forall k \tag{3}$$

Now consider the points for which the equality in Eq. (3) holds. These points lie on the hyperplane Φ, and are called support vectors. Support vectors are the training data that define optimal separating hyperplane and the most difficult patterns to be classified. Accordingly, these points are most informative patterns for the classification task.

To construct the optimal separating hyperplane in the feature space F, one only needs to calculate the inner product between support vectors and the vectors of the feature space by "kernel function", $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. The common kernel functions are polynomial functions with various

degrees and radial basis functions (RBFs). The decision function that is nonlinear in the input space is:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{sv} z_i \alpha_i K(x_i, x) + b \right) \quad (4)$$

where α_i's are Lagrange multiplier coefficients. To find α_i, it is sufficient to find the maximum of the objective functional

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_k - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (5)$$

The advantage of SVC approach is that the complexity of the resulting classifier is characterized by the number of support vectors rather than the dimensionality of the transformed feature space. As a result, SVC tends to be less prone to overfitting problem than other classification methods [13].

5. EXPERIMENTAL RESULTS

We used three pattern recognition techniques to classify the emotion states conveyed by the speech utterances: (1) linear discriminant classifier (LDC) that assumes each class has Gaussian probability distribution. (2) k-nearest neighborhood (k-NN) classifier and (3) support vector machine classifier (SVC). The first two classifiers were explored in detail in [2] and are used in this study, comparing with the performance of SVC. The training data set was selected 10 times in a random manner from the whole data set. Each training data set had 260 samples (130 data from each emotion class) for female and 240 (120 samples from each emotion class) for male. All the error rates in the experiments were calculated by 10-fold cross validation. That is to randomly divide the training data set into 10 disjoint sets of equal size, and classifiers are trained 10 times, each time with a different set held out as a validation set [12]. The estimated error rate is the mean of these 10 errors. The final error rates were the mean of 10 training dataset randomly chosen from the data pool.

5.1. Comparison between Classifiers

Classifier Method	Averaged Error, %	Standard Dev., %
LDC	27.54	1.86
k-NN (k=3)	28.35	2.33
SVC	27.69	2.11

(a)

Classifier Method	Averaged Error, %	Standard Dev., %
LDC	26.85	1.3
k-NN (k=3)	29.5	2.18
SVC	25.58	1.72

(b)

Table 1: Comparison of 3 classifiers in terms of averaged 10-fold cross-validated classification error in female data. (a) base feature, and (b) feature by PCA (dimension = 6). For SVC, polynomial kernels with degree of 1 were used.

Classifier Method	Averaged Error, %	Standard Dev., %
LDC	25.46	2.24
k-NN (k=7)	26.42	2.14
SVC	25.75	1.66

(a)

Classifier Method	Averaged Error, %	Standard Dev., %
LDC	23.62	1.21
k-NN (k=7)	26.33	2.1
SVC	25.04	1.32

(b)

Table 2: Comparison of 3 classifiers in terms of averaged 10-fold cross-validated classification error in male data. (a) base feature, and (b) feature by PCA (dimensioin = 6). For SVC, polynomial kernels with degree of 1 were used.

Table 1 and 2 show the performance comparison between LDC, k-NN, and SVC classifiers for female and male data. When determining the parameters (k for k-NN and kernel for SVC) of k-NN classifiers and SVC, we used 10-fold cross-validation, and set the parameters with the lowest misclassification error rate.

In the experiments, LDC performed better than k-NN classifiers and SVC, but within the range of standard deviation. We can conclude that 3 classifiers perform the same. Note that the feature sets by PCA, which reduced the feature dimensions, showed the comparable performance with the base feature sets. By reducing the feature dimension, we usually lose the information for classification, but PCA preserve enough information to perform comparably.

5.2. Learning Curves Along Training Sample Sizes

It is known that the generalization in SVC is high because the optimal hyperplane can be constructed from a small number of support vectors; therefore, it is less prone to the size of the training set than LDC even in higher dimensions [14]. In Figure 1, the error evaluation curve for female data is shown for base feature sets. The error is calculated by training each classifier in randomly chosen patterns out of the whole data each learning size in each class, repeating 10 times computing the classification error on the remaining unused patterns, and averaging the results of error rate. Both SVC and k-NN classifiers performs better than LDC in the sense that even in small number of training set, the generalization ability is better than LDC.

6. CONCLUSIONS

In this paper, we explored automatic recognition of negative emotions in speech signals using data obtained from a real-world application with pattern recognition techniques, such SVC, LDC and k-NN, in conjunction with feature reduction method of PCA. We also notice that SVC outperforms LDC in the light of generalization. Due to the data sparsity we

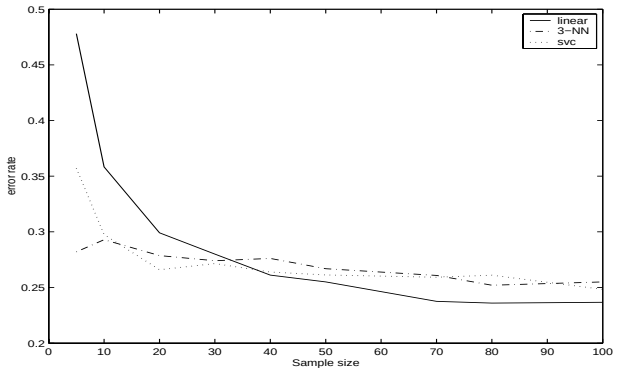

Figure 1: Test error evaluation curve for 3 classifiers (female data). For each learning size, training data were chosen at random out of the whole dataset for each classifier and tested against an unused data in the dataset. This process was repeated 10 times and the errors were averaged. For SVC, polynomial of order 1 (linear) kernel was used.

encounter in the real world applications, this is a very promising result and needs further research. Note also that PCA in these data sets showed comparable performance for all the classifiers even in the reduced feature dimensions. There are several issues to be further explored in the future.

First of all, we should consider more refined tagging methods for emotion states of the data used because it comes from a real-world application i.e., no pre-assigned emotion categories were available. In this respect, more human listening tests should be performed to ensure the emotion states of utterances.

Secondly, research on the emotion recognition has focused just on the acoustic features. However, sometimes it is more appropriate to perform PCA or other component analyses on the original signal data to remove redundancies. The comparable performances obtained by PCA in all the classifiers suggest the promise of such an approach for the emotion recognition problem in real-world data. In this respect, we should further explore other techniques such as factor analysis for emotion recognition.

The last issue is about classification methods. Since emotion states/categories do not have clear-cut boundaries, we need to explore and develop the classification methods to deal with that problem. Such studies should also have the ability of integrating other information such as linguistic and 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 Signal Proc. Mag.*, 18(1), pp. 32-80, Jan 2.

[2] C.M. Lee, S. Narayanan, R. Pieraccini, "Recognition of Negative Emotions from the Speech Signal," *Proc. Automatic Speech Recognition and Understanding*, (Trento, Italy), Dec. 2001.

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

[4] K. Scherer, "A Cross-Cultural Investigation of Emotion Inferences from Voice and Speech: Implications for Speech Technology," *ICSLP 2000*, Beijing, China, Oct. 2000.

[5] V. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers," *Artificial Neu. Net. In Engr. (ANNIE '99)*, pp. 7-10, Nov. 1999.

[6] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, "Desperately Seeking Emotions: Actors, Wizards, and Human Beings," *Proceedings of the ISCA Workshop on Speech and Emotion*, (to appear).

[7] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech," *ICSLP'96 Conference Proceedings*, Philadelphia, PA., pp. 1970 -1973, 1996.

[8] D. Roy and A. Pentland, "Automatic Spoken Affect Analysis and Classification". *In the Proceedings of the International Conference on Automatic Face and Gesture Recognition*, Killington, VT. 1996.

[9] http://www.speechworks.com/indexFlash.cfm

[10] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk and S. Stroeve, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark," *ISCA Workshop on Speech and Emotion*, Belfast 2000.

[11] Sudha Arunachalam, Dylan Gould, Elaine Andersen, Dani Byrd and Shrikanth S. Narayanan, ``Politeness and frustration language in child-machine interactions'', in *Proc. Eurospeech*, (Aalborg, Denmark), pp. 2675-2678, 2001.

[12] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd Ed., John Wiley & Sons, New York, NY, 2001.

[13] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, 1995.

[14] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Know. Disc.*, vol. 2(2), pp. 1-47, 1998.