Recognition of Negative Emotions from the Speech Signal

C. M. Lee[1], S. Narayanan[1], R. 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 methods for automatic classification of 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 recorded from a commercial application deployed by SpeechWorks. Linear discriminant classification with Gaussian class-conditional probability distribution and k-nearest neighborhood methods are used to classify utterances into two basic emotion states, *negative* and *non-negative*. The features used by the classifiers are utterance-level statistics of the fundamental frequency and energy of the speech signal. To improve classification performance, two specific feature selection methods are used; namely, promising first selection and forward feature selection. Principal component analysis is used to reduce the dimensionality of the features while maximizing classification accuracy. Improvements obtained by feature selection and PCA are reported in this paper. We reported the results.

1. INTRODUCTION

The 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, and for a variety of automatic training and educational applications. For this reason automatic emotion recognition in speech has recently received wide attention [1]. The goal of an automatic emotion recognizer is to assign category labels that identify emotional states. While cognitive theory argues against such categorical labeling [2], it nevertheless provides a pragmatic intermediate choice, especially from an engineering standpoint. Primary reasons include (1) A lack of a definite description and agreement on a set of basic emotions (2) A lack of consistency in description: the same emotional category tends to be described in the literature in diverse manner [1]. 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 obtaining 66% of overall accuracy [3].

We favor the notion of domain-dependent emotions, and focus on a reduced space as opposed the entire space of emotions, for the development of 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.

Several pattern recognition methods have been explored for automatic emotion recognition [4, 5]. For example, Dellaert, et al. used maximum likelihood Bayes classification, Kernel regression, and k-nearest neighbor methods [6], whereas Roy and Pentland used Fisher linear discrimination method [7]. In the study proposed in this paper we used linear discrimination (LDC) and k-nearest neighborhood (k-NN) classifiers.

Most of the reported studies have used speech recorded from actors that were asked to express (feign) pre-defined emotions. One exception is the study by Batliner et al [5]. In this work, a 'Wizard-of-Oz' scenario was used to collect data. Subjects were asked to communicate with a real computer, and the study reported classification of the utterances into two emotions: 'emotional' and 'neutral'. Another relevant study is by Petrushin who developed a real-time emotion recognizer using neural networks for call center applications [4]. He achieved ~77 % classification accuracy in two emotion states, 'agitation' and 'calm' for 8 features chosen by a feature selection. For the work reported in this paper, we used a corpus of sentences from a human-machine spoken dialog application deployed by SpeechWorks used by real customers.

The most common acoustic features used for emotion recognition are pitch-related and energy-related [3-8]. For example, McGilloway, et al. studied 32 different features for the classification of 5 emotion states [8]. The features are concerned with F0 (usually regarded as a pitch), energy, duration and tune (segments of the pitch contour bounded at either end by a pause of 180 ms or more). Benchmark testing was done on the features and highly contributing features were found using linear discriminant analysis (LDA). In our work we used ten utterance-level statistics derived from F0 and energy as the acoustic features for emotion recognition. A description of those features is provided in Sec.3.

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 signal and evaluated them with respect to the resulting classification accuracy/error [4, 6]. Some of the features may be highly correlated, with redundant information 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.

The rest of the paper is organized as follows: Section 2 describes the speech data used in the experiments. Section 3 discusses the features extracted from the speech data, and describes the PCA feature reduction method and two other feature selection methods that have been used in the experiments. Section 4 presents the results of classification experiments and Section 5 provides the conclusions.

2. SPEECH DATA PREPARATION

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. The speech database contained 1187 calls, each having average of 6 utterances; therefore, the total number of utterances was approximately 7200. The first step in data processing was to mine this data to select portions that would enable us investigate emotion recognition. In order to facilitate the labeling process, we first used objective measures such as ASR accuracy, total number of dialog turns, and rejection rate to narrow down the inventory to potentially useful dialogs for our experiments. This was followed by subjective tagging of the data into one of two possible emotion categories -- negative and non-negative by two different human listeners. In our study, 'negative' emotions represent anger or frustration in human speech, whereas 'non-negative' emotions are the complement of that, i.e., they represent neutral or positive emotions such as happiness or delight. The order of utterances was randomly chosen in order for listeners not to be influenced in guessing the emotions by the situation in the dialog (minimizing thus the effect of discourse context). After the human listening test, it turned out that most of the 'non-negative' utterances were neutral, i.e., had no apparent display of emotions. For this work, we selected those utterances for which both listeners had complete agreement in their tagging. Overall percentages of agreements were 70.5 % for male speech data and 65.5 % for female speech. That resulted in 40 'negative' and 182 'non-negative' emotion-tagged utterances from 45 calls of female speakers and 31 'negative' emotion and 122 'non-negative' emotion-tagged data from 29 calls of male speakers. To ensure a proper balance, 80 utterances from female speakers (40 for each category) and 62 utterances from male speakers (31 each) were chosen for our experiments.

Data sparsity is a critical challenge and a reality in the study of emotion recognition. Display of negative emotions is relatively infrequent in realistic human-machine interactions but nevertheless important to be detected. Hence, algorithm development for classification should attempt to cope with this issue.

3. FEATURE EXTRACTION

In our experiments, we computed only acoustic features such as pitch and energy related features from the speech signal. Other features would be useful for the emotion recognition: for instance, linguistic information, e.g., the use of swear words 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 [5]. Here, we focus only on acoustic features.

Base Features

The acoustic features chosen 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. These are referred to as base features since they provide the starting point for this study. 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.

Feature Reduction by PCA

To improve classification performance, we reduced the dimension of the features by principal component analysis (PCA) [9]. PCA involves computation of the sample $d \times d$ covariance matrix Σ of the full feature set with d-dimensions, calculation of the eigenvalues and eigenvectors of Σ, and finally sorting it according to decreasing eigenvalues. Then, the largest k eigenvectors are chosen to form a $k \times k$ matrix A whose columns consist of the k eigenvectors. We can obtain new features set by preprocessing features according to:

$$\mathbf{x}' = A^T(\mathbf{x} - \boldsymbol{\mu}) \tag{1}$$

where $\boldsymbol{\mu}$ is the mean vector for $\mathbf{x}$.

Obviously the feature set after PCA is different from the base feature set since it is located in the projected feature space, and the dimension of the features can be usually reduced.

Feature Selection

All of the base acoustic features we proposed are not equally useful for emotion recognition thereby motivating the need for feature selection. The rationale for feature selection is that new or reduced features might perform better than the base features since we can eliminate irrelevant features from the base feature set. This can also reduce the dimensionality, which can otherwise hurt the performance of the pattern classifiers. For feature selection, we used promising first selection (PFS) and forward selection (FS) methods [6].

In PFS, we used k-NN classification on each feature dimension separately, and ordered them according to increasing leave-one-out cross validation error, thus adding new feature dimension successively each time. We start by including in the feature set the feature that shows the best performance, and proceeded by adding the 2nd best feature, calculating the classification error, and continuing until all the feature components are exhausted.

Forward Selection chooses features and adds the best performing features in conjunction with the dimensions already selected, i.e., it tries to find out all the possible combinations unlike the PFS method where new features are added in isolation.

4. EXPERIMENTAL RESULTS

We used two 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. In the following experiments all the error rates shown were calculated by leave-one-out cross validation for the maximal use of available data. In this procedure, we use all but one data for training, and reserve one remaining piece of data for testing. This procedure was repeated for each

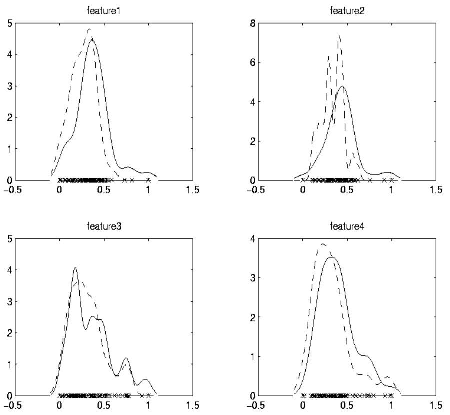

Figure 1: Probability density function estimation by Parzen Window of 4 feature dimensions -- mean, median, STD, and maximum of F0 -- in Base Features (Female Speakers). Straight lines represent 'negative' and dashed lines represent 'non-negative' emotion, respectively (not normalized).

sentence in the selected corpus. The performance of the classifiers was assessed by computing the average classification error rate.

Comparison of Base Features and Preprocessed Features by PCA

Classifier Method	Error Base Features, %	Error PCA (7-dim), %
LDC	22.5	21.25
k-NN (k = 3)	30	26.25

Table 1a: Results of classification error for base features and features preprocessed by PCA (Female Speakers). For PCA, the best results were for dimension of 7.

Classifier Method	Error Base Features, %	Error PCA (6-dim), %
LDC	43.35	25.81
k-NN (k = 3)	35.48	33.87

Table 1b: Results of classification error for base features and features preprocessed by PCA (Male Speakers). For PCA, the best results were for dimension of 6.

First, we tested the classification performance using 10 base features with LDC and k-NN classifiers and the results are shown in Tables 1a and 1b. Results for female and male speakers are given separately due to significant gender dependency in the results. Using PCA, we reduced the dimension of features down to 7 (females) and 6 (males), which was obtained by cross validation test for each reduced dimension. Results indicate that PCA yields improved performance for both LDC and k-NN classifiers. Note that random guessing error is 50%.

For the base features, k-NN performed more consistently than LDC, and it can be due to the fact that Gaussian assumption may be not valid. The estimated pdf of 'negative' and 'non-negative' emotion classes for 4 base features is shown in Figure 1. The figure shows that the distributions of the class-conditional probabilities of the base features are not strictly Gaussian.

Feature Selection

In order to assess the performance of each feature in isolation, we calculated the classification error rates for each feature and each classifier; the results for female speakers are given in Table 2 (and are similar for males). These results show that the discriminative ability of each feature in isolation for each classifier is different with respect to the classification error. We have boldfaced the 3 most discriminative features for each classifier. For k-NN classifier, F0_median, Energy_mean, and Energy_STD have the most discriminative power in terms of minimum classification error.

Feature	Error LDC, (%)	Error k-NN (%) (k = 3)
F0_mean	41.25	48.75
F0_median	42.5	**32.5**
F0_STD	50	50
F0_Max	47.5	63.75
F0_Min	43.75	57.5
Energy_mean	**26.25**	**35**
Energy_median	33.75	45
Energy_STD	37.5	**38.75**
Energy_Max	**26.25**	41.25
Energy_range	**32.5**	42.5

Table 2: Classification error results of each feature in isolation in terms of error rate (female speakers).

Then we performed PFS and FS (Section 3.3) and calculated the classification error for each newly added feature dimension in increasing order of error. The classification results for female and male speakers are shown in Tables 3a and 3b, respectively. The number of features selected by each method is given in parenthesis. For females, the features selected by PFS were F0_median, Energy_mean, Energy_STD, Energy_max, and Energy_range whereas FS selected F0_median, Energy_max, and F0_mean as the best feature combination.

For males, the features selected in the base features for PFS were all but Energy_range, while FS selected Energy_mean, F0_median, F0_mean, F0_max as the best feature combination.

Method	Error Base Feature (%)
PFS	20 (5)
FS	22.5 (3)

Table 3a: Classification error results for promising feature selection and forward feature selection for base features in female speakers using k-NN (k = 3).

Method	Error Base Feature (%)
PFS	24.19 (9)
FS	24.19 (4)

Table 3b: Classification error results for promising feature selection and forward feature selection for base features in male speakers using k-NN (k = 3).

Next we performed the experiment for PFS using LDC for both female and male speech data. For female speech, the classification error was 20% with 3 features, which were Energy_mean, Energy_max, and Energy_range. For male speech data, the classification error was 29.03 % with 4 features, which were F0_STD, Energy_mean, F0_max, and F0_min.

5. CONCLUSIONS

In this paper, we explored automatic recognition of negative emotions in speech signals from a real-world application using pattern recognition techniques, such as LDC and k-NN, in conjunction with feature selection and reduction methods (PFS, FS and PCA). Such techniques provided improved performance compared to base features: overall classification error rates in female speech data were 20% for PFS, 22.5% for FS, and 26.25% for PCA with k-NN classifier, respectively. And the results for male speech were 24.19% for PFS, 24.19% for FS, and 33.87% for PCA. The results are comparable to those by Petrushin, which had ~77% classification accuracy for 8 features (after feature selection) in two emotion states using neural network methods [4].

The reason for gender-specific experiments are due to the fact that pitch-related features are very different between female and male, especially mean, max, and min of F0. Further note that k-NN showed more consistent results since the differences in classification error between females and males were smaller than for LDC, i.e., k-NN depended less on chosen data. That result may be due to the non-Gaussianity of class-conditional pdfs. Although the results are promising, several important issues remain to be addressed.

First we should consider that the data used was from a real-world application, so no pre-assigned emotion categories were available (in contrast to data obtained by explicitly elicited emotions). In this respect, more human listening tests should be performed to ensure the emotion states of utterances.

Second, research on the emotion recognition has focused just on the features obtained directly from the signal or through some feature selection technique such as the forward feature selection method where the final features are a subset of the original features. However, sometimes it is more appropriate to perform PCA directly on the original data to discover the underlying dimensions of features in the sense of pattern recognition. The improvements obtained by PCA in our study suggest the promise of such an approach for the emotion recognition problem. There are many methods to perform preprocessing of the data such as factor analysis, independent component analysis and so on. Such studies on feature selection/reduction may also reveal the inherent characteristics of human emotions.

Finally, better pattern recognition techniques should be investigated since there is no clear-cut definition of emotion states/categories or their acoustic correlates [1]. Therefore, pattern classification methods that can deal with that uncertainty in the emotion states should be studied and developed. Such studies should also adopt a principled way for incorporating linguistic and dialog information in emotion recognition.

6. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, “Emotion Recoginition in Human-Computer Interaction,” *IEEE Signal Proc. Mag.*, 18(1), pp. 32-80, Jan 2001.

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

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

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

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

[6] Dellaert, F., Polzin, T., and Waibel, A., “Recognizing Emotion in Speech,” *ICSLP'96 Conference Proceedings*, 1996.

[7] 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.

[8] 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.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd Ed. (Prepublication Ed.), 1999.