A NEURAL NETWORK APPROACH
FOR HUMAN EMOTION RECOGNITION IN SPEECH

Muhammad Waqas Bhatti[1], Yongjin Wang[2] and Ling Guan[2]
[1]School of Electrical and Information Engineering, The University of Sydney.
Sydney, NSW 2006, Australia
[2]Department of Electrical and Computer Engineering, Ryerson University.
Toronto, Ontario, Canada M5B 2K3
bmwaqas@yahoo.com.au, {ywang, lguan}@ee.ryerson.ca

ABSTRACT

In this paper, we present a language-independent emotion recognition system for the identification of human affective state in the speech signal. A corpus of emotional speech from various subjects, speaking different languages is collected for developing and testing the feasibility of the system. The potential prosodic features are first identified and extracted from the speech data. Then we introduce a systematic feature selection approach which involves the application of Sequential Forward Selection (SFS) with a General Regression Neural Network (GRNN) in conjunction with a consistency-based selection method. The selected features are employed as the input to a Modular Neural Network (MNN) to realize the classification of emotions. The proposed system gives quite satisfactory emotion detection performance, yet demonstrates a significant increase in versatility through its propensity for language independence.

1. INTRODUCTION

As computers have become an integral part of our lives, the need has arisen for a more natural communication interface between humans and machines. To accomplish this goal, a computer would have to be able to perceive its present situation and respond differently depending on that perception. Part of this process involves understanding a user's emotional state. To make the human-computer interaction (HCI) more natural, it would be beneficial to give computers the ability to recognize situations the same way a human does.

In the field of HCI, speech is primary to the objectives of an emotion recognition system, as are facial expressions and gestures. It is considered a powerful mode to communicate intentions and emotions. This paper explores methods by which a computer can recognize human emotion in the speech signal. Such methods can contribute to applications such as learning environments, consumer relations, entertainment etc.

In the past few years, a great deal of research has been done to recognize human emotion using audio information. In the paper [1], the authors explored several classification methods including the Maximum Likelihood Bayes classifier, Kernel Regression and K-nearest Neighbors, and feature selection methods such as Majority Voting of Specialist. However, the system was speaker dependent, and the classification methods had to be validated on a completely held-out database. In [2], the authors proposed a speaker and context independent system for emotion recognition in speech using neural networks. The paper examined both prosodic features and phonetic features. Based on these features, one-class-in-one (OCON) and all-class-in-one (ACON) neural networks were employed to classify human emotions. However, no feature selection techniques were used to get the best feature set, and the recognition rate was only around 50%.

In this paper, we present an approach to language-independent machine recognition of human emotion in speech. The potential prosodic features are extracted from each utterance for the computational mapping between emotions and speech patterns. The discriminatory power of these features is then analyzed using a systematic approach, which involves the combination of SFS [3], GRNN [4] and consistency-based selection method. The selected features are then used for training and testing a modular neural network. Standard neural network and K-nearest Neighbors classifiers are also investigated for the purpose of comparative studies.

The remainder of this paper is organized as follows. In section 2, we give an overview of the emotion recognition system. Feature selection is described in section 3. In section 4, classification approaches and experimental results are presented. Finally, discussions and conclusions are given in section 5.

2. OVERVIEW OF THE EMOTION RECOGNITION SYSTEM

The functional components of the language-independent emotion recognition system are depicted in Figure 1. It consists of seven modules: speech input, preprocessing, spectral analysis, feature extraction, feature subset selection, modular neural network for classification, and the recognized emotion output.

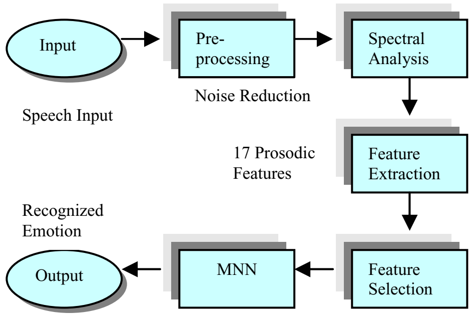

Figure 1: Structure of the Emotion Recognition System

2.1 Data Acquisition

In order to build an effective language-independent emotion recognition system and test its feasibility, a speech corpus containing utterances that are truly representative of an emotion was recorded. Our experimental subjects were provided with a list of emotional sentences, and were directed to express their emotions as naturally as possible by recalling the emotional happening, which they had experienced in their lives. The data were recorded for six classes: *happiness, sadness, anger, fear, surprise and disgust*. Since our aim is to develop a language-independent system, subjects from different language backgrounds are selected in this study. The speech utterances were recorded in English, Chinese, Urdu and Indonesian. Over 500 utterances, each delivered with one of six particular emotions, were recorded at a sampling rate of 22050 Hz, using a single channel 16-bit digitization.

2.2 Preprocessing

The preprocessing prepares the input speech for recognition by eliminating the leading and trailing edge. The volume is then normalized to improve detection by the spectrogram generator. Unvoiced sounds are cut if they appear dominant in the signal. A noise gate with a delay time of 150 ms, and a threshold of 0.05 is used to remove the small noise signal caused by digitization of the acoustic wave. The threshold is the amplitude level at which the noise gate starts to open and let sound pass. A value of 0.05 is selected empirically through the observation of background static "hiss" in the quiet parts of sections.

2.3 Spectral Analysis and Feature Extraction

Previous works have explored several features for classifying speaker affect: phoneme and silence duration, short-time energy, pitch statistics and so on. However, as prosody is believed to be the primary indicator of a speaker's emotional state, we choose prosodic features of speech for emotion analysis. A total number of 17 prosodic features are extracted by analyzing the speech spectrogram. These 17 possible candidates are listed in Table 1.

Feature	Description
1.	Pitch range (normalized)
2.	Pitch mean (normalized)
3.	Pitch standard deviation (normalized)
4.	Pitch median (normalized)
5.	Rising pitch slope maximum
6.	Rising pitch slope mean
7.	Falling pitch slope maximum
8.	Falling pitch slope mean
9.	Overall pitch slope mean
10.	Overall pitch slope standard deviation
11.	Overall pitch slope median
12.	Amplitude range (normalized)
13.	Amplitude mean (normalized)
14.	Amplitude standard deviation (normalized)
15.	Amplitude median (normalized)
16.	Mean pause length
17.	Speaking rate

Table 1: List of 17 Feature Vectors

Due to the nature of the investigation, we usually do not have a precise idea about the features needed in pattern classification. Therefore we consider all possible candidate features in feature extraction. However, some of the features may be redundant or even cause negative effects. Hence, we utilize a systematic feature selection approach to choose features to achieve maximum performance with the minimum measurement effort.

3. FEATURE SELECTION

The ultimate goal of feature selection is to choose a number of features from the extracted feature set that yields minimum classification error. In this study, we propose the adoption of an efficient one-pass selection procedure, the sequential forward selection (SFS) [3] approach that incrementally constructs a sequence of feature subsets by successively adding relevant features to those previously selected. To evaluate the relevancy of the subsets, we adopt the general regression neural network (GRNN) [4]. In this section, we first analyze the discrimination power of the 17 extracted features using the SFS method with GRNN. We then discuss some limitations of GRNN as the number of selected features grows, and introduce a consistency-based selection [5] as a complementary approach.

3.1 Sequential Forward Selection

The SFS is a bottom-up search procedure where one feature at a time is added to the current feature set. At each stage, the feature to be included in the feature set is selected among the remaining available features, which have not been added to the feature set. So the new enlarged feature set yields a minimum classification error compared to adding any other single feature.

If we want to find the most discriminatory feature set, the algorithm will stop at a point when adding more features to the current feature set increases the classification error. For finding the order of the discriminatory power of all potential features, the algorithm will continue until all candidate features are added to the feature set. The order in which a feature is added is the rank of feature's discriminatory power.

3.2 General Regression Neural Network

The GRNN is then used to realize the feature selection criteria in measuring classification error. The GRNN is a memory based neural network based on the estimation of a probability density function. The main advantage of the GRNN over the conventional multilayer feed-forward neural network is that, unlike the multilayer feed-forward neural network which requires a large number of iterations in training to converge to a desired solution, GRNN needs only a single pass of learning to achieve optimal performance in classification. In mathematical terms, if we have a vector random variable x, a scalar random variable y, let X be a particular measured value of x, then the conditional mean of y given X can be represented as:

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y_i \exp(-\frac{D_i^2}{2\sigma^2})}{\sum_{i=1}^{n} \exp(-\frac{D_i^2}{2\sigma^2})} \qquad (1)$$

where D_i^2 is defined as:

$$D_i^2 = (X - X_i)^T (X - X_i) \qquad (2)$$

In the above equation (1), *n* denotes the number of samples. X_i and Y_i are the sample values of the random variable *x* and *y*. The only unknown parameter in the above equation is the width of the estimating kernel σ. However, because the underlying parent distribution is not known, it is impossible to compute an optimum value of σ for a given number of observations. So we have to find the σ value on an empirical basis. A leave-one-out cross validation method is used to determine the σ value that gives the minimum error.

3.3 Experimental Results Using SFS and GRNN

By applying SFS/GRNN, the discriminatory power of the 17 candidate features is determined in an order of {1,13,17,4,8,9,3,16,6,7,15,14,12,5,2,11,10}. The mean square error versus the feature index is plotted in Figure 2.

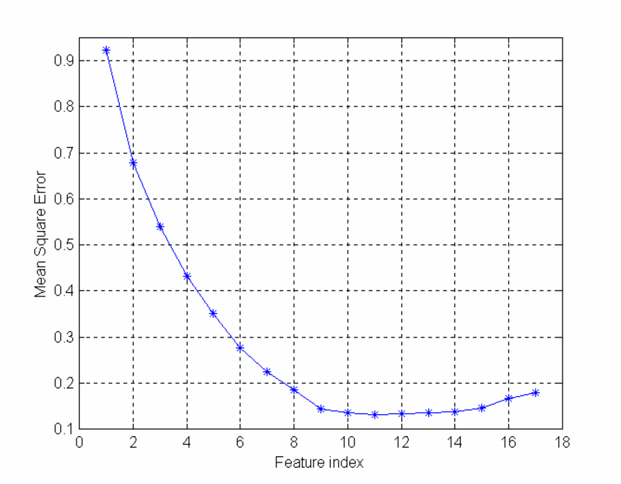

Figure 2: Error Plot for SFS/GRNN

The abscissa in Fig. 2 corresponds to feature order number. From the curve above, we can observe that the minimum mean square error occurs at the point where the top 11 features are included, which correspond to the feature numbers {1,13,17,4,8,9,3,16,6,7,15}. However, we can also observe that when the feature index number is greater than 9, the error curve is almost flat. A possible interpretation of this outcome is that due to the approximation nature of the GRNN modeling process which does not incorporate any explicit trainable parameters, it will be increasingly difficult for the network to characterize the underlying mapping beyond a certain number of features due to the limited number of training samples and their increasing sparseness in high-dimensional spaces. Thus, the order of features beyond the minimum point may not necessarily reflect their actual importance. We therefore need to consider more carefully about the relevance of those features around and beyond the minimum. An alternative approach is suggested in the next section to review the effectiveness of features from the point where the error curve begins to flatten.

3.4 Consistency-Based Feature Selection

In this paper, we use a consistency-based approach as a complementary approach to evaluate the relevance of features around and beyond the minimum error point. The consistency measure $\boldsymbol{C}$ of each feature is computed by:

$$\boldsymbol{C} = \frac{\text{mean inter-class distance}}{\text{mean intra-class distance}} \qquad (3)$$

where the distances are the space of the features under consideration. A given feature is said to have a large discrimination power if its intra-class distance is small and inter-class distance is large. Thus, a greater value of the consistency implies a better feature.

3.5 Experimental Results Using Consistency-Based Selection

By computing the consistency measure for the features around and beyond the minimum point using equation (3), we find that feature 2 has the highest consistency measure. The consistency ranking of the next top three features is the same as achieved by SFS/GRNN. These features are 6, 7, and 15. As a result, we choose four highly consistency features, namely {2,6,7,15}.

Using the combined SFS/GRNN and consistency-based method, we get a total of 12 features. These 12 features are used as input to the MNN which will be described in the next section.

4. EXPERIMENTAL RESULTS

In this section, we present a modular neural network (MNN) architecture, which effectively maps each set of input features to one of the six emotional categories. It should be noted that although we use a GRNN for feature selection, but GRNN has the disadvantage of high computational complexity [4], and is, therefore not suitable for evaluating new samples. Thus we apply a modular neural network based on back-propagation for classification which requires less computation. In the experiments, we compared the performance of MNN, Standard neural network and K-nearest Neighbors classifier.

4.1 Recognizing Emotions

In this study, the motivation for adopting a modular structure is based on the consideration that the complexity of recognizing emotions varies depending on the specific emotion. Thus, it would be appropriate to adopt a specific neural network for each emotion and tune each network depending on the characteristic of each emotion to be recognized. The MNN implementation is based on the principle of “divide and conquer”, where a complex computational task is solved by dividing it into a number of computationally simple subtasks, and then combining their individual solutions. Modular architecture offers several advantages over a single neural network in terms of learning speed, representation capability, and the ability to deal with hardware constraints [6]. The architecture of the MNN used in this study in shown in Figure 3. The proposed hierarchical architecture consists of six sub-networks, where each sub-network specializes in a particular emotion class. In the recognition stage, an arbitration process is applied to the outputs of the sub-networks to produce the final decision.

4.2 Recognition Results

The experiments we performed are based on speech samples from seven subjects, speaking four different languages. A total of 580 speech utterances, each delivered with one of six emotions were used for training and testing. The six different emotion labels used are happiness, sadness, anger, fear, surprise, and disgust. From these samples, 435 utterances were selected for training the networks and the rest were used for testing.

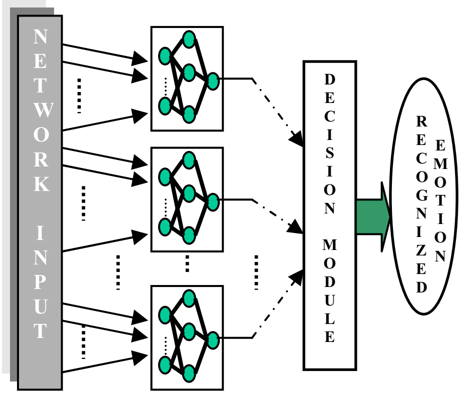

Figure 3: Modular Neural Network Architecture

We investigated several approaches to recognize emotions in speech. A standard neural network (NN) was first used to test all the 17 features and the 12 selected features. The number of input nodes equals to the number of features we used. Six output nodes, associated with each emotion and a single hidden layer with 10 nodes were used. A learning process was performed by the back-propagation algorithm. The system gives an overall correct recognition rate of 77.24% on 17 features and 80.69% on 12 selected features.

In the second experiment, we examined the K-nearest Neighbors classifier (KNN) on the 12 selected features. Leave-one-out cross validation was used to determine the appropriate k value. This classifier gives the overall recognition rate of 79.31%.

Finally, we tested the proposed modular neural network (MNN) by using the 12 selected features. The architecture of the network was the same as depicted in figure 3. Each subnet consisted of a 3-layered feed-forward neural network with one 12 element input vector. Each sub-network was trained in parallel. It is also noted that the modular network was able to learn faster than the standard neural network. Furthermore, the classification performance was improved with the added benefit of computational simplicity. This approach achieves the best overall classification accuracy of 83.31%.

5. DISCUSSIONS AND CONCLUSIONS

The comparison of the recognition results using different approaches is shown in Figure 4. The number that follows each classifier corresponds to the dimension of the input vector. The results show that, by applying SFS/GRNN in conjunction with a consistency-based selection method, the performance of the system was greatly improved. It also demonstrates that the proposed modular neural network produces a noticeable improvement over a standard neural network and a K-nearest Neighbors classifier which has been adopted in some of the other literatures [1, 7]. The time for training a set of sub-networks in MNN was also much less than for a large standard NN. This leads to efficient computation and better generalization.

In this paper, we have presented an approach to language-independent machine recognition of human emotion in speech. We have investigated the universal nature emotion and its vocal expression. Although language and cultural background have some influence on the way in which people express their emotions, our proposed system has demonstrated that the emotional expression in speech can be identified beyond the language boundaries.

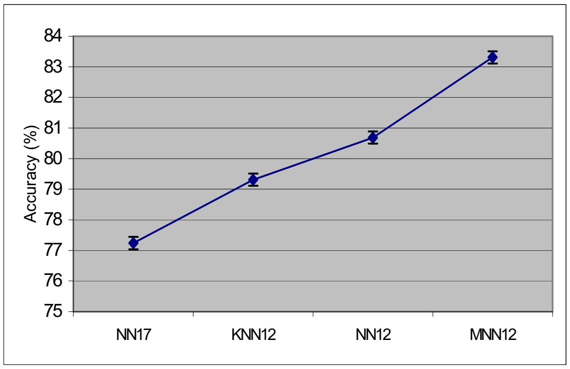

Figure 4: Comparison of the Recognition Results

The results of these experiments are promising. Our study shows that prosodic cues are very powerful signals of human vocal emotions. In the future, we intend to extend our system by integrating audio data with video data. As shown in [8], some of the emotions are video dominant while others are audio dominant. The combination of acoustic speech with the video of the human face conveys more information about the human emotional state than the acoustic speech alone. This complementary relationship will help in obtaining higher emotion recognition accuracy.

6. REFERENCES

[1] F. Dellaert, T. Polzin and A. Waibel, “Recognizing Emotion in Speech”, *Proceedings of the ICSLP '96*, October, 1996.

[2] J. Nicholson, K. Takabashi and R. Nakatsu, “Emotion Recognition in Speech Using Neural Network”, *Neural Information Processing,* 1999.

[3] *J. Kittler,* “Feature Set Search Algorithms”, *In C. H. Chen, editor, Pattern Recognition and Signal Processing, pages 41-60. Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, 1978.*

[4] D. F. Specht, “A general regression neural network,” *IEEE Trans. Neural Networks, vol. 2, no. 6, pp 568-576,* 1991.

[5] M. Wu et al., “Dynamic resource allocation via video content and short-term statistics,” *IEEE Int. Conf. on Image Processing, vol. III, pp. 58-61,* 2000

[6] R. A. Jacobs, M. I. Jordon, “A Competitive Modular Connectionist Architecture,” *in Advances in Neural Information Processing systems 3, pp. 767-773,* 1991

[7] C. M. Lee, S. Narayanan and R. Pieraccini, “Recognition of Negative Emotions from the Speech Signal”, *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on ,* 9-13 Dec. 2001

[8] L. C. De Silva, T. Miyasato and R. Nakatsu, “Facial Emotion Recognition Using Multi-modal Information”, *International Conference on Information, Communications and Signal Processing, ICICS ’97,* Singapore, 1997.