SPEECH EMOTION DETECTION BASED ON NEURAL NETWORKS

Kamran Soltani, Raja Noor Ainon
Faculty of Computer Science and Information Technology, University of Malaya
Kuala Lumpur, Malaysia
kamran@perdana.um.edu.my, ainon@um.edu.my

ABSTRACT

Emotion detection in spoken dialogues is an area that has traditionally been studied in psychology and linguistics but in recent years the engineering community has become increasingly active in this area, due largely to its importance in spoken language man-machine interfaces. Besides techniques in signal processing and analysis it also requires psychological and linguistic analysis. This paper reports an experimental study on six emotions, happiness, sadness, anger, fear, neutral and boredom. It uses speech fundamental frequency, formants, energy and voicing rate as extracted features. Features are selected manually for different experiments in order to get the best results. The selected features are included into a features vector with different sizes as input for different neural network classifiers. To carry out this experimental study a specific tool for language-independent emotion recognition tool has been designed and used. The database which is used for this experiment is the Berlin Database of Emotional Speech.

1. INTRODUCTION

Previous research on emotions, both in psychology and speech [1, 2, 3, 4 5, 6, 7, 8] tell us that we can find information associated with emotions from a combination of prosodic, tonal and spectral information. In addition, speaking rate and stress distribution also provide some clues about emotions. In this study the kinds of features that might carry more information about the emotional meaning of each utterance are considered. The features that contribute to emotions may be different for different spoken languages [9].The approach is to calculate which features carry more information, and to combine these features to get a better recognition rate. It also depends on which emotions we want a machine to recognize and its purpose. The motivation for our research is to explore the effect of selecting different features on automatic recognition of different emotions in speech. The application of emotion recognition can be seen in computer games, and call centers [4, 6]. As figure 1 illustrates emotion recognition in this study has three stages: Feature extraction, feature selection and classification. Basic features and statistics are computed in feature extraction followed by the analysis of features components in feature selection. Classification is then carried out by using a neural network classifier.

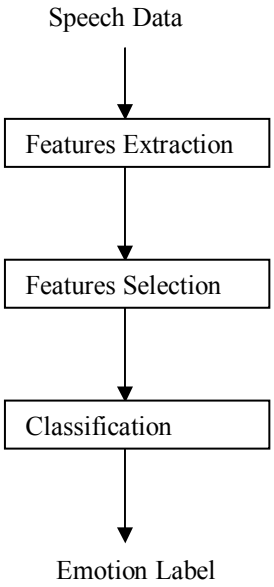

Figure 1. Emotion Recognition.

2. FEATURES EXTRACTION

Although in previous research different types of information, e.g. MFCC coefficients, have been used for feature selection [2, 6], in this research we have selected the fundamental frequency, energy, formant and voicing rate [11,12] as the basic features based on our preliminary study. .For example, the utterance level statistics for fundamental frequency (FO) used are FO maximum, FO standard deviation, FO range and FO mean.

3. FEATURES SELECTION

In selecting features for the classifier we attempt to identify those that will contribute more significantly to emotion speech recognition by using manual forward selection followed by backward elimination in order to rank the features. The higher ranked features will be selected for the classifier.

1-4244-0779-6/07/$20.00 ©2007 IEEE

4. CLASSIFIER

We use a two-layer back propagation neural network architecture [10] with a 8-, 12- or 16-element input vector,10 or 20 nodes in the hidden log-sigmoid layer and six nodes in the output linear layer. The number of inputs corresponds to the number of features and the number of outputs corresponds to the number of emotional categories.

5. DATABASE

The quality of the database will affect the performance of the classifier in the training and Testing. We used the Berlin Database of Emotional Speech [13] it contains about 500 utterances spoken by actors in a happy, angry, anxious, fearful and bored way as well as in a neutral version. The utterances are recorded from 10 different actors and actress (five Male and five Female) and ten different texts. To train and test our classifier we used 70% of this data for training and 30% for testing.

6. EXPERIMENTAL RESULTS

For our experiments we calculated some descriptive statistics, mean, standard deviation, minimum, maximum, and range, for the following acoustical variables: fundamental frequency F0, energy, voicing rate and first formants F1. Features vector with 8, 12 and 16 features are used in three experiments by using a two-layer backpropagation neural network as classifier.

In the first experiment, 16 features were selected as the top 16 features by using the features selection techniques described in section 2. These features are as follows: F0 maximum, F0 standard deviation, F0 range, F0 mean, energy maximum, energy standard deviation, energy range, energy mean F1 maximum, F1 standard deviation, F1 range, F1 mean, F1 Minimum voicing rate maximum, voicing rate standard deviation and voicing rate minimum. They were selected to investigate how sets of features influence the accuracy in different emotion categories.

The confusion matrix for the first experiment is shown in Table 1, with six intended emotions as column indices and their corresponding recognized emotions as the row indices. The recognition rate for each emotion is shown in bold and is underlined.

Table 1. Recognition of emotions in experiment 1.

	Anger %	Boredom %	Fear %	Happy %	Sad %	Neutral %
Anger	**78.3**	0	0	17.6	0	4.1
Boredom	0	**69**	13.2	0	14.1	3.7
Fear	0	12.4	**67.2**	0	13.6	6.8
Happy	23.7	0	0	**64.8**	0	11.5
Sad	0	9.6	8..3	0	**71.3**	10.8
Neutral	0	1.6	1.2	12.8	8.8	**75.6**

In the second experiment, we chose 12 top features: F0 maximum, F0 standard deviation, F0 range, F0 mean, energy maximum, energy standard deviation, energy mean F1 maximum, F1 standard deviation, F1 range, voicing rate standard deviation and voicing rate minimum. Table 2 shows the confusion matrix for the second experiment.

Table 2. Recognition of emotions in experiment 2.

	Anger %	Boredom %	Fear %	Happy %	Sad %	Neutral %
Anger	**81.4**	0	0	12.9	0	5.7
Boredom	0	**73.2**	10.5	0	14.4	1.9
Fear	0	13.9	**66.5**	0	12.9	6.7
Happy	17.6	0	0	**72.9**	0	9.5
Sad	0	10.4	5.3	0	**72.4**	11.9
Neutral	0	1.3	1.9	9.9	8.3	**78.6**

In the third experiment we selected the best first 8 features as features vector. These features are: F0 maximum, F0 range, F0 mean, energy maximum, energy standard deviation, F1 maximum, F1 standard deviation, voicing rate standard deviation. Table 3 shows the confusion matrix for the third experiment.

Table 3. Recognition of emotions in experiment 3.

	Anger %	Boredom %	Fear %	Happy %	Sad %	Neutral %
Anger	**82.3**	0	0	10.2	0	6.6
Boredom	0	**76.5**	8.7	0	9.3	5.5
Fear	0	11.4	**70.1**	0	8.2	10.3
Happy	14.7	0	0	**75.4**	0	9.9
Sad	0	8.4	3.2	0	**74.1**	14.3
Neutral	0	1.5	2.1	8.2	4.8	**83.4**

It can be seen that *Neutral* was the emotion that was least difficult to recognize from speech as opposed to *Fear* which was the most difficult. In this experiment we reached the overall accuracy rate of 77.1%.

7. SPEECH EMOTION RECOGNITION TOOL

For doing this experimental work we developed a language-independent emotion recognition system that gives us this ability to change different parameters in features extraction and training steps as well as selecting different features for features vector and checking the accuracy rate of our classifier. This system is developed using MATLAB.

8. CONCLUSIONS

In this paper, we tried to find which features can contribute significantly to the speech emotion recognition. Using different features illustrates that in our research *anger* and *neutral* are two emotions that can easily be recognized whereas *fear* is the most difficult

one. Energy and pitch are two important features for speech emotion recognition according to this paper; however, for reducing the overlapping among the emotions linguistic features can be useful, for example using linguistic filter for recognizing between anger and happiness may increase the accuracy rate. In addition, although we found neural network a powerful classifier for the speech emotion recognition, using fuzzy logic in tandem with neural network may give the higher recognition rate due to fuzzy data in some cases.

REFERENCES

[1] Chul Min Lee, and Shrikanth S. Narayanan, "Toward detecting emotions in spoken dialogs", IEEE Transaction on Speech and Audio Processing, vol. 13, no. 2, pp. 293- 303, Mar. 2005.

[2] Dellaert, F. Polzin, and T. Waibel, "Recognizing emotions in speech", in Proc. ICSLP 96, vol. 3, pp. 1970-1973, USA, Oct. 1996.

[3] Kaxuhiko Takahashi, Ryohci Nakatsu and J. Nicholson,"Emotion Recognition in Speech Using Neural Network**"**, Neural Computing &Applications**,** vol. 9, no. 4, pp. 290-294, Dec. 2000.

[4] Sherif Yacoub, Steve Simske, Xiaofan Lin and John Burns," Recognition of Emotions in Interactive Voice Response Systems", in Eurospeech 2003, 8th European Conference on Speech Communication and Technology,Switzerland, Sep. 2003.

[5] Muhammad Waqas Bhatti, Yongjin Wang and Ling Guan," A neural network approach for human emotion recognition in speech", in Proc. ISCAS '04,vol 2,pp 181-4,Vancouver,Canada,2004.

[6] Valery A. Petrushin*,"*Emotion Recognition in Speech Signal: Experimental Study: Development and Application", in Proc. ICSLP 2000, USA, 2000.

[7] K.R. Scherer," Adding the affective dimension: A new look in speech analysis and synthesis", in Proc. ICSLP 96, USA, Oct. 1996.

[8] Ralf Kompe and JM Pardo,"Emotional space improves emotion recognition", in Proc. ICSLP 2002, USA, 2002.

[9] Scherer K. R., Banse R. and Wallbott H. G., "Emotion inferences from vocal expression correlate across languages and cultures", Journal of Cross-Cultural Psychology, vol. 32(1), pp. 76-92., 2001.

[10] Kevin Gurney, An introduction to Neural Network, Taylor & Francis, UK, 1997.

[11] UCL Department of Phonetics and Linguistics
http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html

[12] School of Computing, Canberra University.
http://www.ise.canberra.edu.au/un7190/Week04Part2.htm

[13] Berlin Emotional Speech Database.
http://pascal.kgw.tu-berlin.de/emodb/index-1024.html