Emotional recognition from the speech signal for a virtual education agent

A Tickle[1], S Raghu[1] and M Elshaw[2]
[1]Department of Aviation, Aerospace, Electrical and Electronic Engineering, Faculty of Computing and Engineering, Coventry University, Priory Street, Coventry, UK
[2]Department of Computing, Faculty of Computing and Engineering, Coventry University, Priory Street, Coventry, UK
Contact Email: andrew.tickle@coventry.ac.uk

Abstract. This paper explores the extraction of features from the speech wave to perform intelligent emotion recognition. A feature extract tool (openSmile) was used to obtain a baseline set of 998 acoustic features from a set of emotional speech recordings from a microphone. The initial features were reduced to the most important ones so recognition of emotions using a supervised neural network could be performed. Given that the future use of virtual education agents lies with making the agents more interactive, developing agents with the capability to recognise and adapt to the emotional state of humans is an important step.

1. Introduction

This paper describes a neural network approach to speech emotion recognition from a set of acoustic features of a speech waveform. Emotions would enable a communication virtual agent to both maintain a user's positive emotional state and allow it to judge and refine its current dialogue strategy. One way to determine the emotional state of the user in this context is through an interpretation of the speech signal collected using a microphone. This approach outlined in this paper focuses on using auditory sensors rather than visual ones as they are not as intrusive and rely less on environmental conditions.

In order to use virtual agents in the role of educator, it is necessary to provide them with the capability to sense the emotional state of the human user [1, 2] so they can offer the appropriate and natural response. By recognising and expressing emotions, the interactive educational agent could act in an enthusiastic manner towards the teaching material and also show empathy towards student development. If the interactive agent has an appealing and believable personality, this would have the effect of making the interaction more enjoyable for the students [6]. It is felt that students will relate to virtual agents if they can connect with their personality and their emotional representation.

There has seen a growth in interest in emotional speech recognition using intelligent approaches to aid agent-person interaction. For example, Zhang et al. used back propagation neural network to recognise emotions such as anger, clam, happy, sad and surprise [9]. Traista and Elshaw (2012) [7] developed a self-organising neural network approach to perform emotional speech recognition that made use of a emergent unsupervised approach. Vogt et al. (2008) [8] developed EmoVoice a real-time emotion recognition system based on a state vector machine within a virtual agent named Greta. Despite the research carried out into emotional speech recognition, there is little agreement on how to perform feature selection, and so this paper will offer an indication of an approach to do so.

2. Methodology

As can be seen from Figure 1, the methodology for developing an emotional speech recognition system involves extracting features from the speech signal collected by a microphone, reducing the feature set to a more manageable level, and then recognising the emotions using an intelligent technique. In the remainder of this paper we will provide a more detailed description of how this was achieved.

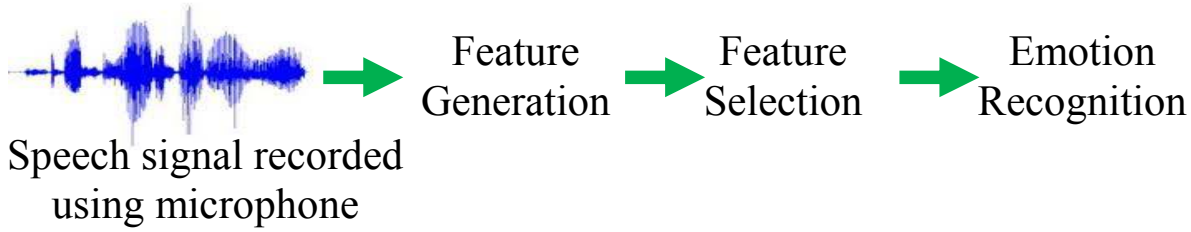

Figure 1. The components of the emotion recognition system.

2.1 Speech corpus

The "Berlin Database of Emotional Speech" is used to train and test the algorithm in this paper [3]. The Berlin Database consist of 535 speech sample, which contain German utterances related to emotions such as anger, disgust, fear, joy, sadness, surprise and neutral, acted by five males and five females. The Berlin database was chosen to be used because of the quality of its recording and its popular use in emotion recognition based research work. Some of the utterance translated into English in the Berlin Corpus include: 'The tablecloth is lying on the fridge', 'She will hand it in on Wednesday' and ' In seven hours it will be'. Of the 535 speech sample, features extraction was carried out on 105 samples, consisting of 15 occurrences of each emotion. To consider how the trained model performed on samples from a person not in the data set, two samples of each emotion were recorded in English by a male volunteer.

2.2 Feature generation

To extract the features from the speech samples the data mining tools OpenSMILE was employed. OpenSMILE toolkit is a flexible feature extractor for speech processing and machine learning applications, which is fully written in C++ [4]. In this research, OpenSMILE was configured to extract a set of 988 acoustic features. The feature set that was extracted from the microphone speech signal is specified by the configuration file "emobase.conf" in the toolbox. The features included the following low-level descriptors (LLD): Intensity, Loudness, 12 MFCC, Pitch (F0), Probability of voicing, F0 envelope, 8 LSF (Line Spectral Frequencies) and Zero-Crossing Rate. Delta regression coefficients were computed from these LLD, and the following functionals were applied to the LLD and the delta coefficients: Max/Min. value and respective relative position within input, range, arithmetic. mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges. Figure 2 displays the output from OpenSmile when creating the features from the emotional speech signals.

2.3 Feature selection

The performance of the learning algorithm increases according to the quality of the input data. Among the 988 acoustic features extracted using the OpenSMILE tool, several of these features had overlapping information and a few features contained irrelevant information. In order to improve the processing speed of the system to a practical standard, it was necessary to severally reduce the features set to the most relevant ones. A view of the features in the 'explore' tool of WEKA

(http://www.cs.waikato.ac.nz/ml/weka/) provides a basic understanding of how each feature effects the emotion the samples describe. Graphs of each feature can be viewed under the pre-process tab.

For the process of feature selection, filtering and wrapping algorithms were employed. Using the WEKA machine learning tool, "info-gain attribute" algorithm was used to rank the features in the order of importance according to their respective weights. Similarly, using the "classify Subset Eval" algorithms the best-first list of features was generated. A hybrid set of features selected by considering both these algorithms was then used as the final training set for the multilayer perceptron.

The results of "classifySubset Eval" algorithm provided 11 features as the ones with highest importance. However, the performance of these 11 features alone was very poor, hence a bigger list of features were needed and so these features were combined with the 63 features available by rank by using the "info-gain attribute" ranker algorithm of WEKA. The combined feature set contained a total of 71 features.

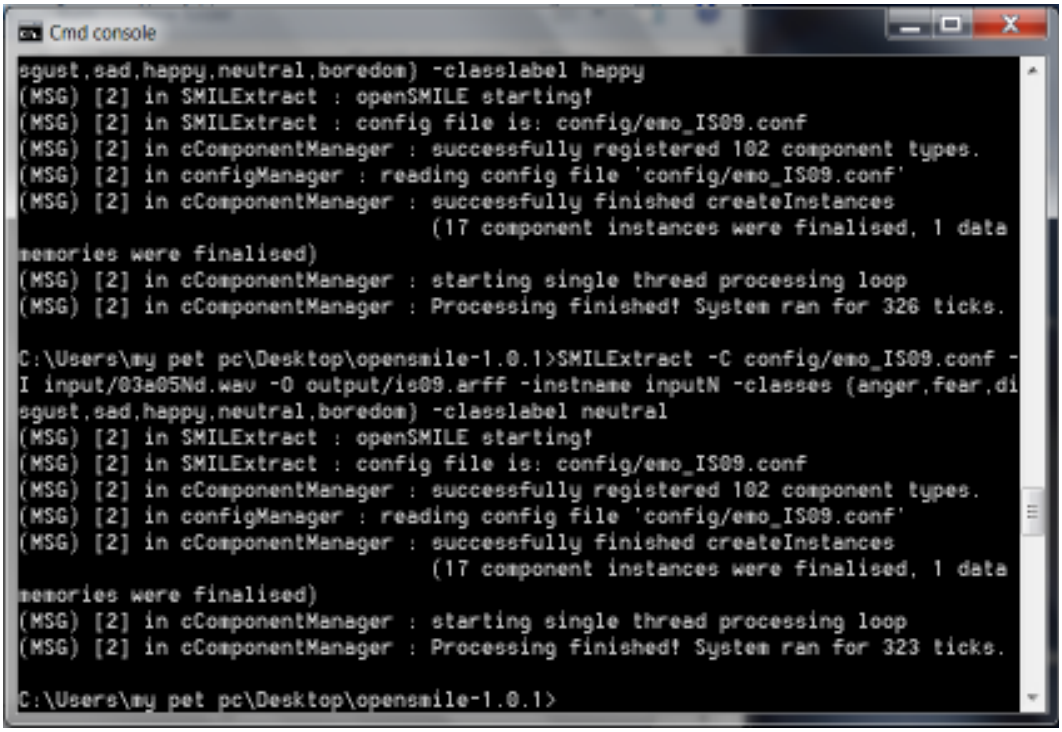

Figure 2. Screenshot of the OpenSMILE command console during feature extraction.

2.4 Multi-layer perception neural network

It was decided to use a neural network for emotion recognition as this learning technique has proved successful in the past. Multilayer perceptron is an important class of neural networks. As shown by Figure 3, it commonly consists of neurons that constitute the input layer, one or more hidden layers and an output layer of computational nodes. The learning rule typically used for the multilayer neural network is the back-propagation rule that allows the network to learn to classify. This rule creates the output of the network compares this with the required output and by propagating the error back through the network alters the weights to reduce the error [5]. During the forward pass the synaptic weights are combined with the input vector to produce on the output layer the classification. During the backward pass, the classification produced by the system is subtracted from the actual classification to get the error signal. The error signal is passed backwards through the network to adjust the weights.

3. Results

For the samples from the Berlin database after initial training was performed, a 10-fold cross validation evaluation was performed. This system's performance is measured based on the number and percentage of correctly recognised emotions. The number of instances correctly classified for the Berlin Database is 88 out of 105, producing an average accuracy of 83.8%. Considering the confusion matrix for samples from the Berlin Emotional Database in Table 1, accuracies differ significantly between the emotions, with some being easier to identify than others. For example, the multilayer neural network seems to struggle to differentiate between emotions anger and disgust. A few inaccuracies also occur in the neutral emotion class as well.

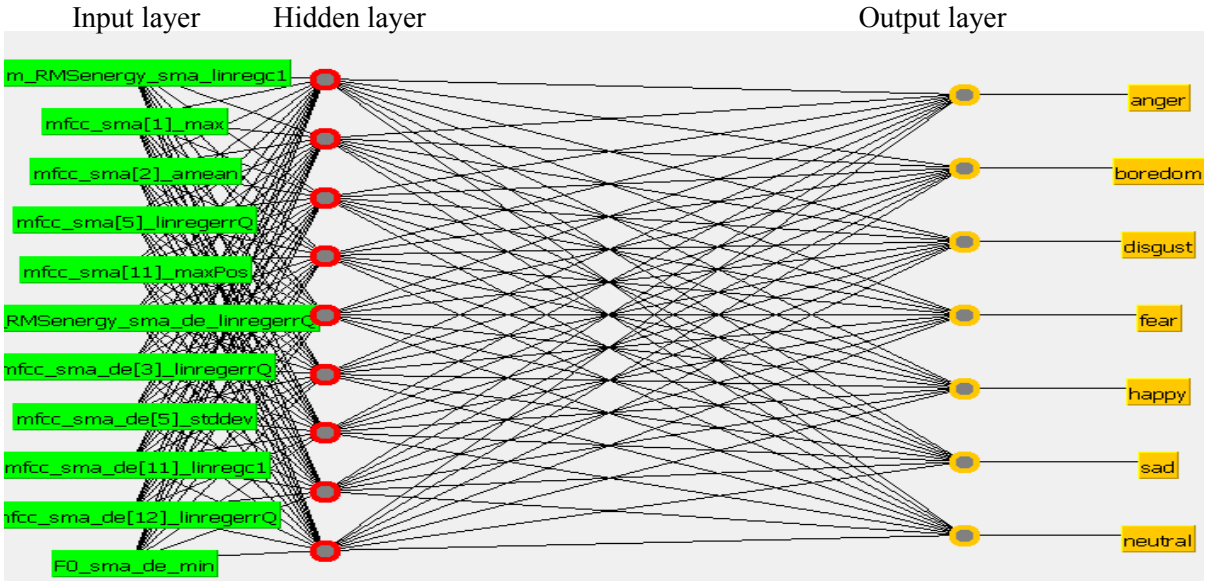

Figure 3. Example multi-layer perceptron model from WEKA showing the 11 features provided from the wrapping algorithms

The performance of the system was also tested by using the speech samples recorded from an English speaking male volunteer, as a preliminary study. These samples had 2 occurrences of each emotion. The average accuracy in this test was much poorer at 42.8%. The fall in accuracy can be explained by the lack of microphone quality, as a generic laptop microphone was used. The noise interference is a major reason for the fall in accuracy. As seen in Table 2, from the confusion matrix it can be observed that the error occurrence is random and has no conclusive patterns.

Anger	Boredom	Disgust	Fear	Happy	Sad	Neutral
12	0	2	0	0	0	1
0	14	0	0	0	1	0
3	1	10	1	0	0	0
0	0	0	13	0	1	1
0	0	0	0	14	0	1
0	1	0	0	0	14	0
0	1	0	0	2	1	11

Table 1. Confusion matrix – performance Results of Berlin Database

Anger	Boredom	Disgust	Fear	Happy	Sad	Neutral
1	0	0	0	1	0	0
0	0	1	0	0	1	0
0	1	0	1	0	0	0
0	0	0	1	0	1	0
0	0	0	0	1	0	1
0	1	0	0	0	1	0
0	0	0	0	0	0	2

Table 2. Confusion matrix – performance on samples taken from an English male volunteer

4. Discussion

Implementing an emotion recognition system for a virtual educational agent would be a huge step forward. Although, the approach presented in this report achieves good performance, further work would help reduce the feature set even further. The number of features used is a primary factor concerning the processing time taken by the system to complete its operations. In order to be able to implement this approach in hardware components, it would require a huge amount of processing memory.

Although part of the reduction in performance on the speech from the male volunteer collected as a preliminary study might be the attribute to the training set being in German and this data being recorded in English, it is not the real reason as both these languages represent emotion in speech in the same manner. Considering the closeness of the origin of these languages, it is anticipated that this emotion recognition system is language independent in this scenario. The poor performance is likely because of the low quality of the microphone and the noise level in the environment during the recording. If virtual educational agents are to be commonplace, there is a need to develop approaches to overcome noise in the environment without making the microphone sensor too costly.

5. Conclusions

The project presents a system for emotion recognition from speech whose accuracy is similar to present works using the Berlin Emotional Database. The report provides descriptions of feature extraction from the Berlin Database samples, various ways the features have been selected for the final set of 80 features and the multilayer perceptron model using which the results have been produced. The approach by which the features have been reduced offers new opportunities for emotion recognition. The system offers a real indication that intelligent autonomous emotion recognition system could be incorporated into social robots to enable them to interact with human users in a natural way.

Although the performance of the multi-layer perceptron neural network proved successful, there is an opportunity to achieve more reliable performance. This will be achieved by performing classification of the emotion using multi-classifiers and then combine the outcomes with a voting based system. There is also the opportunity to incorporate other sensor reading to give the emotional state of the student using the virtual education system. For example, the system could use a camera to explore the facial expressions of the student and a heart rate sensor to establish if emotions influence heart rate. As shown in Figure 4, it would be possible to combine top-down features such as the words and facial expression recognised with the lower level feature such as those gained from the speech signal. By being able to recognise emotions using the auditory input from the users, there are many areas as well as virtual educational agents, such as social assistive robots and telephone call centres that would benefit from the research outlined in this paper.

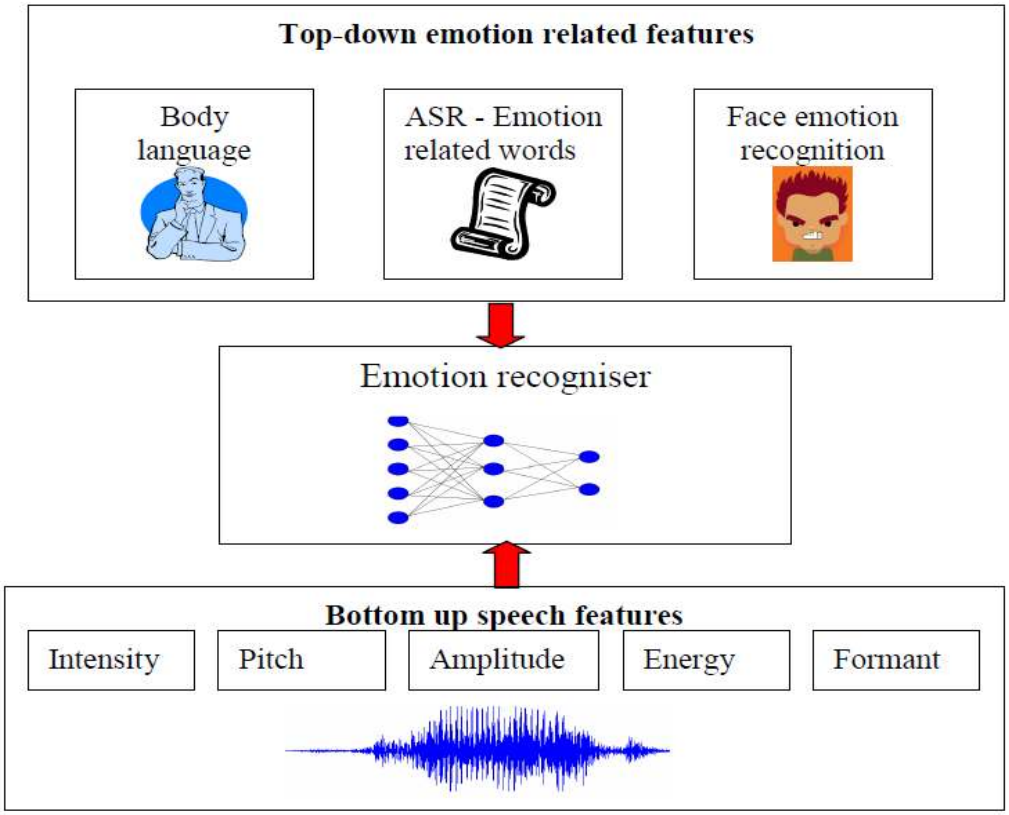

Figure 4. Top-down and bottom-up emotion recognition from multimodal sensors.

6. References

[1] Creed, C and Beal, R. 2005 Using emotion simulation to influence user attitudes and behaviour. *Proceedings of the 2005 Workshop on the role of emotion in HCI*

[2] Cerezo, E., Baldassarri, S., and Seron, F. 2007 Interactive agents for multimodal emotional user interation, *IADIS International Conference Interfaces and Human Computer Interaction*, pp. 35-42.

[3] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. 2005 A Database of German Emotional Speech. *Proceedings Interspeech 2005, Lisbon, Portugal*

[4] Eyden, F., Wollmer, M., and Schuller, B. 2010 Opensmile: the munich versatile and fast open-source audio feature extractor, *MM '10 Proceedings of the international conference on Multimedia*, pp. 1459-1462.

[5] Haykin, S 2005 Neural Networks: A comprehensive foundation, 2nd edition. Essex,Pearson Education.

[6] Hertzum, M., Andersen, H., Andersen, V. And Hansen, C. 2002 Trust in information sources: seeking information from people, documents, and virtual agents, *Interacting with Computers*, 14(5), pp. 575-599.

[7] Traista, A., and Elshaw, M., 2012 A hybrid neural emotion recogniser for human-robotic agent interaction, *Proceeding of 13th Engineering Applications for Neural Networks*..

[8] Vogt, T., André, E., and Bee., N. 2008 EmoVoice - A framework for online recognition of emotions from voice. *In Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer, Kloster Irsee*.