2013 IEEE International Symposium on Multimedia

Emotion Recognition Modulating the Behavior of Intelligent Systems

Asim Smailagic, Daniel Siewiorek, Alex Rudnicky, Sandeep Nallan Chakravarthula,
Anshuman Kar, Nivedita Jagdale, Saksham Gautam, Rohit Vijayaraghavan, Shaurya Jagtap
Department of Electrical and Computer Engineering,
Carnegie Mellon University,
Pittsburgh, PA, 15213, USA
Email: asim@cs.cmu.edu

***Abstract*—The paper presents an audio-based emotion recognition system that is able to classify emotions as anger, fear, happy, neutral, sadness or disgust in real time. We use the virtual coach as an application example of how emotion recognition can be used to modulate intelligent systems' behavior. A novel minimum-error feature removal mechanism to reduce bandwidth and increase accuracy of our emotion recognition system has been introduced. A two-stage hierarchical classification approach along with a One-Against-All (OAA) framework are used. We obtained an average accuracy of 82.07% using the OAA approach, and 87.70% with a two-stage hierarchical approach, by pruning the feature set and using Support Vector Machines (SVMs) for classification.**

***Keywords*—emotion recognition; voice and speech analysis; interaction design; well-being**

I. Introduction

The goal for an intelligent Quality of Life system is to be able to adjust its response according to the user's emotions. Emotion recognition can detect a person's mood and hence allow the system to adapt to it, thereby providing an improved user experience. Emotions often drive human behavior and the detection of emotional state of a person is very important for system interaction in general and in particular in the design of intelligent systems such as Virtual Coaches [1]. A model of human behavior that can be instantiated for each individual includes emotional state as one of its primary components. Example emotional states that we are addressing are: anger, fear, happy, neutral, sadness and disgust.

Our proposed system has the following salient characteristics:

1) It uses short utterances as real-time speech from the user.
2) Prosodic and phonetic features, such as fundamental frequency, amplitude, and Mel-Frequency Cepstral Coefficients are used as the main set of features by which we can characterize the human speech samples. By doing so, we focus on the aspect of using only audio as input for emotion recognition without any additional facial or text features.
3) Our experiments use an OAA approach and a two-stage classification between different emotions.

Our emotion recognition system adjusts the behavior of the Virtual Coach for stroke rehabilitation exercises, depending on the user's emotion. For example, on detecting the emotion as angry, our system integrated with the Virtual Coach, advises the patient to 'take rest'. Our models can classify six emotions, and a subset of those emotions (anger, fear, happy and neutral) was chosen for the virtual coach application in consultations with clinicians and physical therapists from two rehabilitation hospitals in Pittsburgh, Pennsylvania.

II. Related Work

The task of emotion recognition is a challenging one and has received immense interest from researchers [2].

In [3], the authors use a supra-segmental Hidden Markov Model approach along with an emotion dependent acoustic model. They extracted prosodic and acoustic features from a corpus of word tokens, and used them to develop an emotion dependent model that assigned probabilities to the emotions – happy, afraid, sad and angry. The label of the emotion model with the highest generating probability was assigned to the test sentence. The paper reports human performance on detecting the emotional state of the speaker at 70% accuracy.

Paper [4] presents an analysis of fundamental frequency in emotion detection reporting an accuracy of 77.31% for a binary classification between 'expressive' or emotional speech including different emotions, and neutral speech. In this work, only pitch related features were considered. The overall emphasis of the paper was to analyze the discriminative power of pitch related features in contrasting neutral speech with emotional speech. The approach was tested with four acted emotional databases spanning different emotional categories, recording settings, speakers and languages. There is a reliance on neutral models for pitch features built using HMMs in their approach, otherwise the accuracy decreases by up to 17.9%.

Many automatic emotion classification systems use the information about speaker's emotion that is contained in utterance-level statistics over segmental spectral features [5]. Additionally, [6] uses class-level spectral features computed over consonant regions to improve accuracy. The authors compare performance on two publicly available datasets for six emotion labels - anger, fear, disgust, happy, sadness and neutral. Average accuracy for those six emotions using prosodic features on the LDC dataset [7] was 65.38%.

The Sensitive Artificial Listener (SAL) [8] is a spontaneous speech dataset that comprises of naturalistic interaction between the user and an agent. The emotions are annotated using a dimensional model.

978-0-7695-5140-1/13 $26.00 © 2013 IEEE
DOI 10.1109/ISM.2013.72

III. Scenario of Use

As one example of an intelligent system, we have chosen to discuss our Virtual Coach for Stroke Rehabilitation Exercises. We built a virtual coach to evaluate and offer corrections for rehabilitation of stroke survivors. The Virtual Coach for Stroke Rehabilitation Exercises is composed of a Microsoft Kinect sensor for monitoring motion, a machine learning model to evaluate the quality of the exercise, and a tablet for the clinician to configure parameters of exercise. A normalized Hidden Markov Model (HMM) was trained to recognize correct and erroneous postures and exercise movements.

Coaching feedback examples include encouragement, suggesting taking a rest, suggesting a different exercise, and stopping all together. For example, as shown in Fig. 1, if the user's emotion is classified as angry, our system advises the user to 'take a rest'. An interactive dialog was added to elicit responses from the user, as shown in Fig. 2. Based on these responses, the emotion is gauged by the audio emotion recognizer. The coaching dialog changes depending on performance, user response to questions, and user emotions.

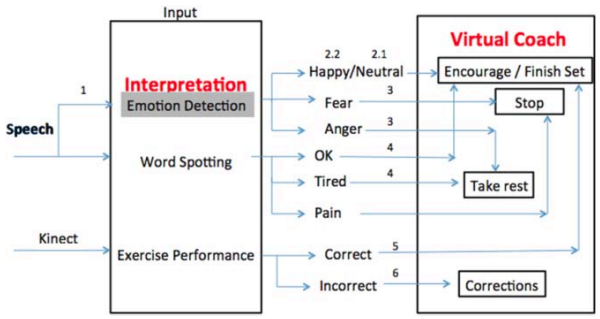


Fig. 1: Flow diagram of emotion recognition system integrated with Virtual Coach

Figure 3a shows a patient using the Virtual Coach. Figure 3b illustrates the situation when the system recognizes the user emotion as angry, and advises the user to 'take a rest'. The system was tested at two rehabilitation hospitals.

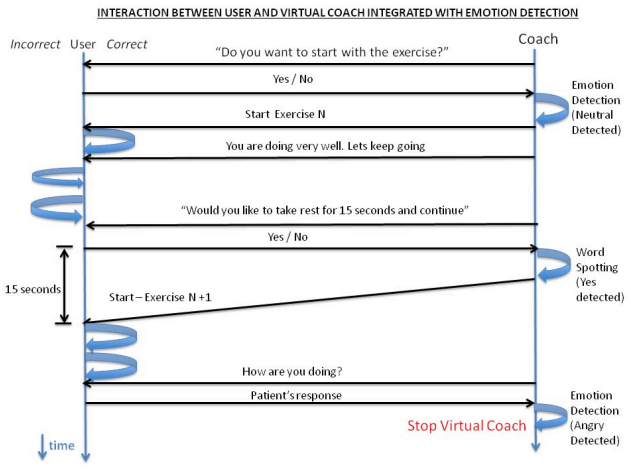


Fig. 2: Interaction dialog between user and Virtual Coach integrated with emotion recognition

IV. System Architecture

The emotion recognition system that is integrated with the Virtual Coach comprises of two main modules, namely feature extractor and classifier.

A. Feature Extractor

A total of 42 prosodic and phonetic features were used in the machine learning model that was used for Emotion Recognition. These include 10 prosodic features describing the fundamental frequency and amplitude [9]. These features are used for our task of emotion classification as they accurately reflect the state of emotion in an utterance.

In addition to the prosodic features, we are also using phonetic features such as Mel Frequency Cepstral Coefficients (MFCC) [10]. These are generated by binning the signal with triangular bins of increasing width as the frequency increases. Mel frequency coefficients are commonly used in both speech and emotion classification.

Using the prosodic and phonetic features together, as opposed to using only prosodic features, helps achieve higher classification accuracy. Our approach towards feature extraction focuses on the utterance-level statistical parameters such as mean, standard deviation, minimum, maximum and range. A Hamming window of length 25ms is shifted in steps of 10ms, and the first 16 Cepstral coefficients, along with the fundamental frequency and amplitude, are computed in each windowed segment. Statistical information is then captured for each of these attributes across all segments.

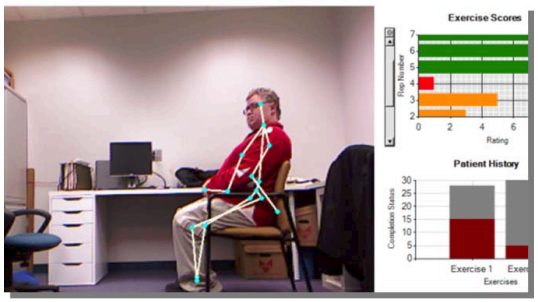
We calculated the mean and standard deviation for each of the 16 Cepstral coefficients giving us 32 features. In addition, the mean, standard deviation, minimum, maximum and range were calculated for fundamental frequency and amplitude, thus giving us the remaining 10 features. This results in 42 features for our dataset.

B. Classifier

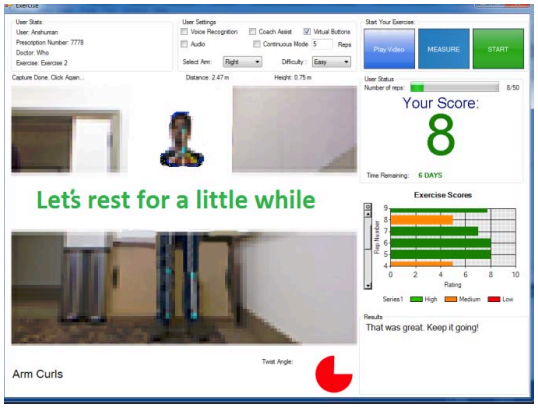
For the purpose of classification, we used Support Vector Machines with Linear, Quadratic and Radial Basis Function kernels, due to the property of SVMs to generate hyperplanes for optimal classification [11]. We ran experiments with different parameters for different kernels, and the best performing model, along with its parameters, was stored for each classification to be used later with the Virtual Coach.

V. Model Evaluation

Performance of three classification methodologies were evaluated on the syntactically annotated audio dataset produced by Linguistic Data Consortium (LDC) [7] and on a custom audio dataset. Additionally, the least discriminative features from the 42 phonetic and prosodic features were pruned to further improve the accuracy.



(a) Stroke patient using the Virtual Coach, with dashboard indicating his exercise performance



(b) When emotion is recognized as angry, the system advises the user to "take a rest"

Fig. 3: Virtual Coach user interface

TABLE I: Mapping of LDC emotions to six basic emotions

Basic Emotion	LDC Emotion	Number of utterances
Anger	Hot anger	139
Disgust	Disgust	179
Fear	Anxiety	183
Happy	Happy	179
Neutral	Neutral	112
Sadness	Sadness	155

A. Audio Datasets

1) LDC Audio Dataset: The primary dataset used in this study is the LDC audio dataset [7]. The corpus contains audio files along with the transcripts of the spoken words as well as the emotions with which those words were spoken by seven professional actors. The transcript files were used to extract short utterances and the corresponding emotion labels. The utterances contained short, four-syllable words representing dates and numbers, eg. 'August 16th'. The left channel of the audio files was used after sampling the signal down to 16 kHz, on which our classification algorithms were run.

As our OAA algorithm classifies six basic emotions anger, fear, happy, neutral, sadness and disgust, the emotion classes from the LDC corpus corresponding to these six emotions were selected. Table I shows this mapping along with the number of audio files from the dataset corresponding to each of the six emotions. A total of 947 utterances was used.

2) CMU/University of Pittsburgh Audio Dataset (Banana Oil): This dataset was expanded at the beginning of the project. We recorded 1,440 audio files from 18 subjects, with 20 short utterances for neutral, angry, happy and fear emotions in the context of the Virtual Coach application. Each audio file was 1-2 seconds long. The subjects were asked to speak the phrase "banana oil" exhibiting all four emotions. This phrase was selected because of its lack of association between the words and the emotions assayed in the study (i.e. anger or neutral), thereby allowing each actor to "act out" the emotion without any bias to the meaning of the phrase.

The subjects were given 15 minutes for the entire session, wherein they were made to listen to pre-recorded voices for two minutes, twice, after which they were given two minutes to rehearse and perform test recordings. In addition, for fear emotion, a video was shown as an attempt to incite that particular emotion. After recording the voice samples, subjects were asked if they felt the samples were satisfactory, and in case they weren't, the recording was performed again for the unsatisfactory ones.

Finally, after all samples had been recorded, they were renamed to conceal the corresponding emotion labels. For the purpose of emotional evaluation, seven 'evaluators' listened to the samples at the same time, and each one independently noted what she felt was the true emotion label for that particular file [11]. Throughout this process, the labels from one evaluator were not known to the rest. Finally, a consensus of labels was taken for each file, which was then decided as the ground truth label for that particular file. In addition, the consensus strength was also determined, based on which the ones with the strongest consensus were used for the final dataset of 464 files, 116 for each emotion. The evaluators were students from Carnegie Mellon University, who are fluent speakers of English language [12].

B. Classification Methodology

While our focus was on classifying all six emotions correctly, we also wanted to concentrate on classifying positive (happy/neutral) against negative emotions (anger/fear) in the context of virtual coach for stroke rehabilitation. Therefore, we ran two experiments, namely One-Against-All and Two-Stage Hierarchical classification.

The samples were first split into 70% and 30% partitions for training and testing respectively. A 10-fold cross-validation approach was used on the training set for model, and files corresponding to each emotion were grouped randomly into 10 folds of equal size. Finally, the results were accumulated over all 10 folds, from which the confusion matrix was calculated.

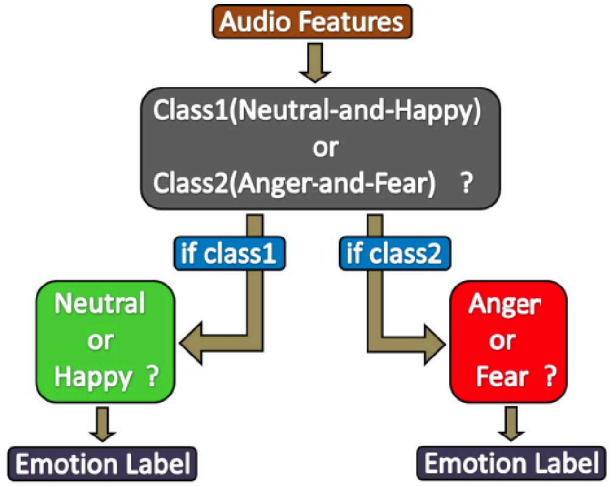


Fig. 4: Overview of two-stage hierarchical classification framework

The results over all passes were combined by summing the entries in the confusion matrices from each fold.

1) OAA Classifier: With One-Against-All approach, we trained our classifier to separate one class from the remaining classes, resulting in six such classifiers, one for each emotion. This results in an imbalance in the number of training examples for positive and negative classes. In order to remove any bias introduced by this class imbalance, the accuracy results from the binary classifier were normalized over the number of classes to compute balanced accuracy [6].

2) Two-Stage 4-Emotions Classifier: The confusion matrix obtained from the 4-emotion classification exhibited relatively less confusion in the emotion pairs Neutral-Happy and Angry-Fear, as compared to the four other pairs. In addition, thorough observation of feature histogram plots for all four emotions revealed that some features were able to sufficiently discriminate between certain emotions, while not being able to do so for the rest, and vice versa.

This led us to explore the possibility of developing a model which could achieve high classification accuracy across the emotions, by performing classification cascade between different sets of emotions, thereby resulting in the two-stage classifier. In this framework, the first stage determines if the emotion detected was a positive one (Class1), i.e. Neutral or Happy, or a negative emotion (Class2), i.e. Anger or Fear, as shown in Fig. 4. Depending on the result of the first stage, the emotion would then either be classified as Neutral or Happy, or as Anger or Fear.

C. Feature Pruning

We initially had 42 features, where 32 represented Cepstral, 5 pitch and 5 amplitude information. However, we found that some features did not add any information for the purpose of distinguishing between different emotion classes. Therefore, we ranked features based on their discriminative capability, with the aim of removing the low ranked ones. Histogram plots for each feature indicated that, for most cases, the distribution within each class could be approximated by a unimodal Gaussian. Figure 5 shows histograms of two features for Anger-versus-Fear classification, one with high (Fig. 5a) and low (Fig. 5b) discriminative ability, respectively.

In order to quantify the discriminative capability of each feature, a parameter M was defined for classes i and j, such that $M(i,j)$ is the percentage of files in class j that occupy values inside the range of values from class i with $i \neq j$.

For a feature having values distributed over k classes, we would have a matrix M of size $k \times (k-1)$, where each row contained the overlap values between a particular class and each of the $(k-1)$ remaining classes. The lesser the overlap a feature offered, the higher was its discriminative capability. Depending on the type of classification to be performed, the appropriate average overlap was calculated.

For Anger-versus-Rest classification, the average overlap was calculated as shown in (1).

$$\text{Overlap} = \frac{1}{l}\sum_{j} M\left(\text{anger}, j\right) \tag{1}$$
$$\text{where } j \in \{\text{neutral}, \text{happy}, \text{fear}\}, \; l = |j|$$

For a Class1-versus-Class2 classification, where Class1 consists of Neutral and Happy, and Class2 consists of Angry and Fear, the overlap was calculated as shown in (2)

$$\text{Overlap} = \frac{1}{k \times l}\sum_{i}\sum_{j} M\left(i,j\right), \tag{2}$$
$$\text{where } i \in \{\text{neutral}, \text{happy}\}, j \in \{\text{anger}, \text{fear}\}, k = |i|, l = |j|$$

Thus, for a given classification problem, we first rank features in decreasing order of discriminative ability, and successively remove the ones with the worst discriminative power, running the classification experiment with a reduced set each time.

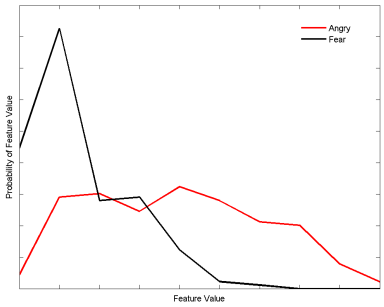
While our method is conceptually similar to feature selection methods such as Minimum-redundancy-maximum-relevance (mRMR) [13], which makes use of mutual information from a feature set for a target class, it is significantly different in the following ways.

1) Our focus is on feature removal, not on feature selection. This means that we concentrate on discarding features that do not contribute enough towards classification, rather than finding the set of features that contributes best to classification.
2) Mutual information is symmetric and averaged over all classes, while Overlap M is asymmetric and specific to a pair of classes, i.e. $M(i,j) \neq M(j,i)$. Thus, we can find a feature's discriminative power for classification between any set of classes.

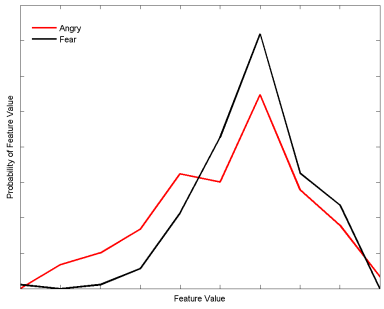
VI. Results

A. Most Discriminative Features

It was observed that each binary classification had its highest accuracy associated with a unique set of features. The complete set consisted of the mean of the first 16 Cepstral coefficients followed by the standard deviation of those coefficients, and the mean, maximum, minimum, standard deviation



(a) Histogram of feature 33, which exhibits good discriminative power



(b) Histogram of feature 7 which has bad discriminative power

Fig. 5: Histograms of features for Anger vs. Fear classification.

and range of the fundamental frequency and the amplitude, respectively. Analysis of the best feature set for each classifier suggests two important things.

1) Highest cross-validation accuracy for all emotions except fear emotion was obtained when the least discriminative features were pruned. OAA classifier for fear vs. rest used all 42 features.
2) Amplitude features, except the mean value, are not discriminative enough for problems involving neutral and disgust emotions, particularly for OAA classification.

The classification accuracy and the associated feature set for each experiment are summarized in Figure 6, where the presence of a blue bar indicates that the particular feature was used, while the absence indicates otherwise. The table shows that, for most of the cases, the best accuracy is achieved when the number of least discriminative features is removed for the LDC dataset.

B. Accuracy

In the One-Against-All experiments, the average classifier accuracy was found to be 82.07%, while in the two-stage classification framework, the average accuracy was 87.70%. For Anger vs. Fear and Class 1 vs. Class 2 classification tasks, SVM with quadratic kernels gave the best results, whereas RBF kernels performed best for the rest of the experiments. Table II shows our accuracy results for OAA classification and those of Bitouk's [6] using OAA classification for a six-class recognition task.

A comparison of our results for OAA classification, with those of Waibel's [3] shows that we achieved higher average accuracy, as shown in Table III. The CMU dataset was used in this experiment.

VII. Conclusion and Future Work

Our audio-based emotion recognition system was integrated with an intelligent system, the Virtual Coach for stroke rehabilitation exercises, that can adjust to the user's emotions.

The emotion recognition results show that the use of prosodic and phonetic features such as fundamental frequency, amplitude and Mel-Frequency Cepstral Coefficients perform well in the task of classifying emotions. We employed feature pruning to use the best performing features, based on their discriminative power to classify emotions, resulting in improved accuracy.

Our current user tests include stroke survivors interacting with the Virtual Coach application, integrated with emotion recognition, and using spontaneous speech.

As the next step, we plan to investigate more sophisticated methods to identify the best feature space or subspace for classification. We also intend to use Hidden Markov models to investigate the change in emotions of patients over the course of various exercises. In the future, we will integrate emotion recognition with other virtual coaches and intelligent systems.

We are also running larger user tests with stroke patients using our Virtual Coach for stroke rehabilitation exercises and are planning to report those results soon.

VIII. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. EEEC-0540865 and in part by the Pennsylvania Infrastructure Technology Alliance, a partnership of Carnegie Mellon, Lehigh University and the Commonwealth of Pennsylvanias Department of Community and Economic Development (DCED). The authors would like to thank the graduate students Sanuj Basu, Soham Chokshi and Ashish Samant for their valuable contributions to the project.

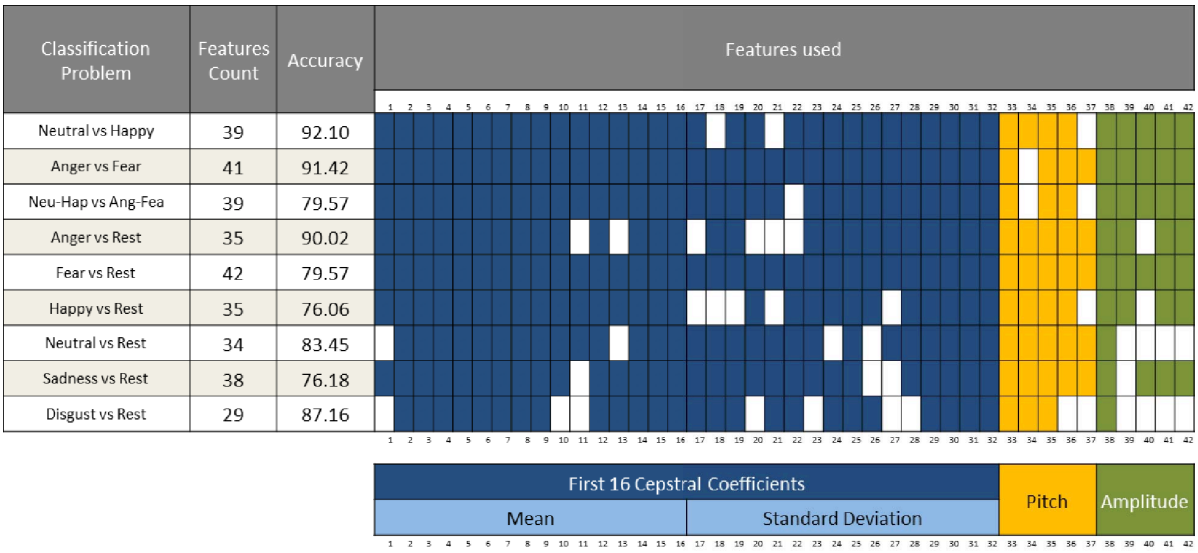


Fig. 6: Classification methodologies with highest accuracy and corresponding set of most discriminative features for LDC dataset

Emotion	Bitouk [6] (%)	Our Work (%)
Anger	71.9	90.02
Fear	60.9	79.57
Happy	61.4	76.06
Neutral	83.8	83.45
Sadness	60.4	76.18
Disgust	53.9	87.16

TABLE II: Emotion recognition accuracies on the LDC dataset

Emotion	Waibel [3] (%)	Our Work (%)
Anger	77.9	87.90
Fear	60.0	86.13
Happy	93.8	87.70
Neutral	-	91.40

TABLE III: Emotion recognition accuracies on the CMU datasets

REFERENCES

[1] D. Siewiorek, A. Smailagic, and A. Dey, "Architecture and applications of virtual coaches," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2472–2488, 2012.

[2] A. Ortony, *The cognitive structure of emotions*. Cambridge university press, 1990.

[3] T. S. Polzin and A. Waibel, "Detecting emotions in speech," in *Proceedings of the CMC*, vol. 16. Citeseer, 1998.

[4] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 582–596, 2009.

[5] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3. IEEE, 1996, pp. 1970–1973.

[6] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7, pp. 613–625, 2010.

[7] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," *Linguistic Data Consortium, Philadelphia*, 2002.

[8] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5–17, 2012. [Online]. Available: http://doc.utwente.nl/62670/

[9] R. Huang and C. Ma, "Toward a speaker-independent real-time affect detection system," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 1204–1207.

[10] R. Tato, R. Santos, R. Kompe, and J. M. Pardo, "Emotional space improves emotion recognition." in *INTERSPEECH*, 2002.

[11] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1057–1070, 2011.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[13] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.