Emotion Detection from Speech Signals and its Applications in Supporting Enhanced QoS in Emergency Response

R. M. Hegde, B. S. Manoj, B. D. Rao, and R. R. Rao
Department of Electrical and Computer Engineering
University of California San Diego
{rhegde, bsmanoj, brao, and rrao}@ucsd.edu

ABSTRACT

Networking in the event of disasters requires new hybrid wireless architectures such as Wireless Mesh Networks (WMNs). Provisioning Quality of Service (QoS) in such networks which are quickly deployed during emergencies demand radical solutions. In this paper, we provide a new QoS approach for voice calls over a wireless mesh networks during emergency situations. According to our scheme, the contention and back-off parameters are modified based on the emotion content in the voice streams. This paper also looks at methods for detecting emotion from an incoming voice call using the speech signal. The issues of interest in such situations are whether the caller is in a state of extreme panic, moderate panic, or in a normal state of behavior. The communication network behavior should be modified to provide differentiated QoS for calls based on the degree of emotion. We use several features extracted from the speech signal like the range of pitch variation, energy in the critical bark band, range of the first three formant variations, and speaking rate among others to discriminate between the three emotional states. At the back end the Gaussian mixture modeling techniques is used to model the three emotional states of the speaker. Since a large number of features increase the computational complexity and time, a feature selection technique is employed based on the Bhattacharya distance, to select the set of features that give maximum discrimination between the classes. These set of features are employed to simulate an emotion recognition system. The results indicate a promising emotion detection rate for the three emotions. We also present the early results on detecting the emotion content in the speech and using this in the MAC layer differentiated QoS provisioning scheme. Our scheme provides an end-to-end delay performance improvement for panicked calls as high as 60% compared to normal calls.

Keywords

Emotion detection, Speech processing, Feature extraction, VQ, GMM, Networking, QoS, MAC Layer, Throughput

INTRODUCTION

In this paper we explore methods by which the emotional state of the speaker making a call in a crisis scenario can be effectively detected from the speech signal. This decision about the emotional state of he speaker is further used to resolve some communication challenges in transmitting the voice data over a network. Conventional techniques to detect emotion from speech use features like pitch, loudness, speaking rate, short term energy derived from the speech signal (Cowie, Douglas-Cowie, Tsapatsoulis, Votsis, Kollias, Fellenz and Taylor, 2001; Yacoub, Simske, Lin, and Burns, 2003; Ang, Dhillon, Krupski, Shriberg, and Stolcke, 2002; Seppänen, Väyrynen, and Toivanen,2003). In this context we define a set of features that are most relevant to the detection of emotion, specifically with respect to whether a speaker is in a state of extreme panic, moderate panic or in a normal state of expression. The features used in this study include pitch information (Hess, 1983), short term energy, speaking rate, the first three fundamental frequencies and the individual frame autocorrelation measures. A Gaussian mixture model (GMM) (Huang, Ariki, and Jack, 1990; Rabiner and Juang, 1993), is computed for each of the three emotional states of the speakers using the aforementioned features derived from the speech signal. During the testing phase the same set of features are derived from speech signal and the emotional identity of the speaker is hypothesized by maximizing the probability of the feature set belonging to a particular GMM, representing a particular emotional state. We begin with the overall system architecture for providing better Quality of Service for the panic calls. There are two major components for this system, (a) the detection of panic calls and (b) the provisioning of QoS for the identified panic calls compared to other normal calls. We begin presenting the panic-detection method with an illustration showing the variation of certain features like spectrogram, F0 contour, energy contour and the formant structures extracted from speech signals collected from speakers in three different states of emotion (extreme panic, moderate panic, and normal emotional state). The preparation of a representative database for this particular task is then described (San Francisco virtual museum). A feature selection technique (Fukunaga,1990; Tooldiag) using the sequential forward search is then employed to select the features that contribute to maximum discriminability among the three different emotional states. The whole feature set and the reduced feature set selected from the feature selection technique are then used to perform emotion detection from the speech signal. The results for detecting the three emotions are promising. When emotion detection is restricted to two emotional states, namely extreme panic and normal emotional state the recognition improves considerably. We conclude with a discussion on its implications in communication challenges for emergency response.

SYSTEM ARCHITECTURE

The system architecture for enabling the emotion-ware (panic-aware) Quality of Service (QoS) in the response network is described here. In this work, we consider a Wireless Mesh Network (WMN) as the principal emergency response network. The emotion detection system runs either at the source node or at each intermediate node in the WMN. When the first node that handles the speech source, identifies the emotion level of the call, and that information is passed to the network layer and/or Medium Access Control (MAC) layer. The MAC layer support is essential to provision better QoS in multihop wireless networks. This can be achieved by modifying the contention resolution techniques (Murthy and Manoj, 2004). There can be three types of emotion detection processes in our system; (i) one-time emotion detection, (ii) run-time continuous emotion-detection, and (iii) run-time periodic emotion detection. In the first case, one-time emotion detection approach, the source node, either a WMN node or the client associated to WMN node, identifies the emotion content in the speech and thereby marks the voice packets indicating the emotion level. In this case, the emotion detection algorithm is applied only at the beginning of the call setup process and process of detection is carried out at the source node. The second approach, run-time continuous, attempts to detect and extract the emotion parameters from the speech packets during the call's duration. In this case, the packets are continuously marked or unmarked based on the outcome of the detection process. Here, one important aspect is the length of speech input for detecting the emotion content. The minimum duration of speech for detecting emotion parameters is three seconds. According to this approach, the emotion detection algorithm runs continuously to monitor the variations of the emotion content in the speech. One benefit of this approach is that the packet marking can be modified in run-time depending on the variation of emotion in the speech. For example, when a user's speech, during its course, changes from moderate to extreme panic level, our system can provide runtime improvement in the QoS. Finally, the run-time periodic algorithm provides a periodic detection of emotion algorithm and updates the packets with its latest observation on the emotion content in the speech stream. Compared to the run-time continuous emotion detection algorithm, this approach reduces processing overhead involved in the detection process. On the QoS provisioning, the following approaches can be used: (a) the differentiated services at the network layer, (b) the differentiated services at the MAC layer, and (c) hybrid differentiated services provisioning. In the first case, the QoS is provisioned by the network layer by providing multiple priority queues each for a particular QoS level. The second case provides a MAC layer QoS provisioning mechanism, by classifying the traffic into a number of categories based on the priority and thereby providing preference during contention process at the MAC level. Finally, in the third case, the QoS provisioning can be done together by the MAC and by the network layer. In this paper, for the emotion detection process, we use the one-time emotion detection method and for the QoS provisioning, we used the MAC layer solution as that forms the most effective component in achieving the QoS for WMNs. According to our system, the source node identifies the emotion content of the call and marks the packet either as a best-effort packet, packet belongs to moderately panicked voice source, or packet belongs to extremely panicked voice source. This marking is done at the Type of Service (ToS) field of the IP packet header. Once a packet reaches the MAC layer using IEEE 802.11 protocol, the ToS field is verified. When a packet belonging to a panicked call, marked to reflect the emotion content of the packet, reach the MAC layer, it is given appropriate differentiated treatment in the contention process.

THE PANIC-AWARE QOS PROVISIONING MECHANISM

The primary objective of our QoS provisioning scheme is to provide better delay and throughput performance for the calls originated by panicked sources compared to normal voice sources. Here, once the emotion-detection module detects the emotion content and marks the packets reflecting the emotion level. The medium access scheme used in IEEE 802.11 standard is the Distributed Coordination Function (DCF) and is described as follows. When a node has packets to be delivered to another node, it checks the channel status. If the channel is idle, it waits for DCF Interframe Space (DIFS) and transmits the packet if it still finds the channel free. Thus the channel access delay at light loads becomes only DIFS. Alternatively, if the channel turns busy after DIFS time, the node initiates the binary exponential back-off process. During this back-off process, the node defers its transmission attempt for a uniformly random amount of time within the Contention Window (CW). The value of CW varies between *CWmin* and *CWmax*. Once the CW value is chosen, the back-off counter is set and the counter is decremented for every time slot. During the back-off process, if the channel turns busy then the back-off counter is frozen until when the channel becomes idle for a period equal to DIFS. Once the back-off counter becomes zero, the node acquires the channel and starts transmitting the data packet or the control packets preceding the data packet. If there is a collision during the transmission, the node increases the contention window by a factor of two and begins the back-off process again. While the initial CW is set to a random value between (0, CWmin), for every collision, the contention window is doubled up to a maximum of CWmax. At high channel load, the CW acquires a larger value thus leading to higher access delay. We, in this paper, differentiate the back-off mechanism based on the emotion content in the packet. In order to provide a better delay performance, we use a different back-off factor, *BackoffFactor,* to use with the back-off process. Thus, for the packets containing emotional speech data, we provide, instead of a binary exponential scheme, a *BackoffFactor* based exponential scheme. Therefore, the CW value is not doubled upon the detection of a collision, instead by a factor designated by the *BackoffFactor* value. The *BackoffFactor* used for the packets belong to the moderately and extremely panicked voice sources are 1.8 and 1.5, respectively. The performance of the proposed scheme is presented later in this paper. The *BackoffFactor* is chosen to provide better QoS, interms of lower delay and higher throughput, for the calls belong to the extremely panicked sources compared to both the moderately panicked source and normal calls.

FEATURES FOR DETECTING EMOTION FROM THE SPEECH SIGNAL

In this work, we focus on the computation of only acoustic features such as pitch (Hess,1983), energy, formant structures, F0 contours, and related features, from the speech signal. To illustrate the significance of such features in detecting emotion from the speech signal we consider two sets of speech recordings one from a female speaker who is in an extremely panicking state and the other from a female speaker who is in a normal emotional state. These sentences are selected from the sample recordings from the 1989 earthquake in San Francisco available from the San Francisco Museum online. The utterance of a female speaker in a normal emotional state despite the 1989 San Francisco earthquake taking place, picked from the 911 recordings available at the San Francisco museum of art website is shown in Figure 1 (a). The corresponding energy contour, smoothed F0 contour, and the first five formant trajectories are shown in Figure 1 (b), (c), and (d), respectively. Similar plots for a female speaker in an extremely panicky state is shown in Figure 2. In Figure 1, the utterance corresponds to a female speaker asking the 911 despatcher the following question, “Hi, Can I fix … Can I start doing my own things after the earthquake ..?” In Figure 2, the speech signal corresponds to the following transcription, “Hello, We have an emergency here, where ?, In Geneva avenue, the water pipe is flooding, it broke I guess, its flooding the building…”.

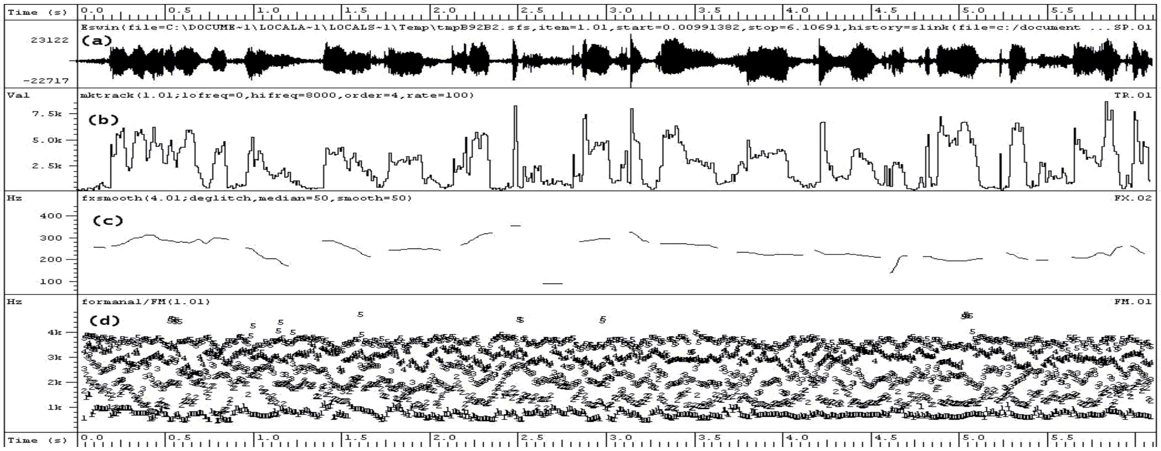

Figure 1: Illustration of the Energy, F0, and the formant contours for a speech signal recorded from a person in normal emotional state. (a) The speech utterance, (b) The energy contour, (c) The smoothed F0 contour, and (d) The trajectories of the first five formants.

It is worthwhile to note from Figures 1 and 2, that the range of energy for speakers in the two emotional states is considerably different. Note that Figure 2, displays a larger range of variation in energy indicating that speakers in panic produce speech signals with larger energy variations. The range of variation in the F0 contour (also called the fundamental frequency track) for the two emotional states is different. The range of F0 contour variation is quite large for a speaker in a panic state when compared to a speaker who is in a normal state of emotion. The speaking rate of the two speakers in the two emotional states is different which also reflects in the densities of the first five formant trajectories as shown in Figure 1 (d) and Figure 2 (d).

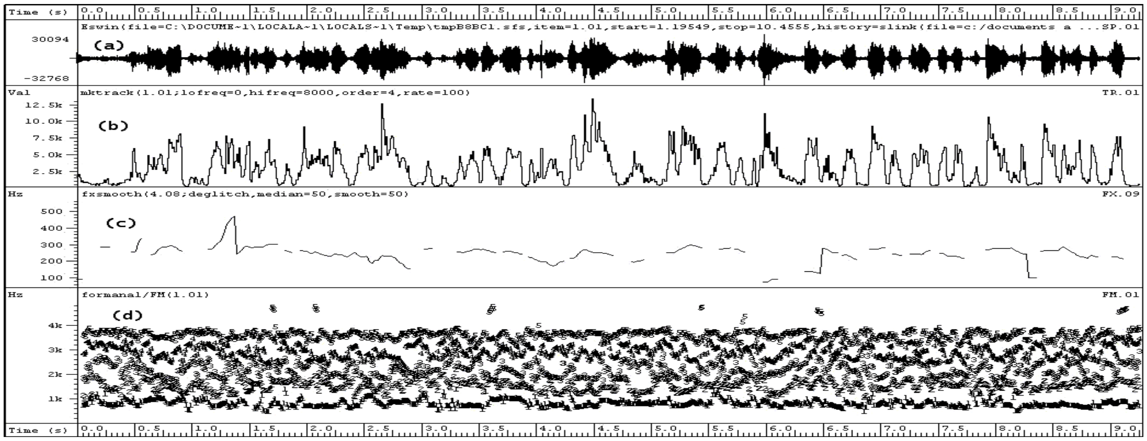

Figure 2: Illustration of the Energy, F0, and the formant contours for a speech signal recorded from a person in a extremely panicky emotional state. (a) The speech utterance, (b) The energy contour, (c) The smoothed F0 contour, and (d) The trajectories of the first five formants.

From the aforementioned observations and from our previous experimental results on emotion detection, we use the following features derived from the speech signal for the detection of the emotional state of the speaker.

1. Range of F0 variation 2
2. Mean of F0
3. Median of F0
4. Standard deviation of F0
5. Mean energy
6. Accumulated sum of energies computed from 64 linear filter banks
7. Range of Energy above 3 KHz
8. Range of energy below 3 KHz
9. Energy in the 1700 – 2000 Hz critical band
10. Duration of the voiced segment to account for the speaking rate
11. Maximum value of the autocorrelation of each frame
12. Range of the first formant F1
13. Range of the second formant F2
14. Range of the third formant F3

LABELLING AND DATABASE PREPARATION

Two sets of data were used to prepare the database for emotion detection used in this study. The first set of speech data came from the recordings of the actual 911 conversations available on the web from the Virtual museum of the city of San Francisco at http://www.sfmuseum.net/1989/sc911.html. The other set of data came from the speech recordings done in a quiet room environment where volunteers were asked to simulate three emotional states corresponding to extreme panic, moderate panic and a normal state of emotion. The transcripts were selected from the Santa Cruz 911 transcripts available at the Virtual museum of the city of San Francisco. Each message recorded was labeled manually by listening to them on high fidelity noise cancelling head phones. Tags were assigned to each speech segment that was clearly distinguishable as conveying information that a speaker was in extreme panic, moderate panic and a normal state of emotion. Three tags 0, 1, and 2 were used for moderate panic, extreme panic and a normal state of emotion respectively. We extracted a total of 250 such segments in each class of emotion and tags were assigned to them. 75% of this data is used as the training set and the other 25% of this tagged data is used as a testing set. An equitable distribution of the three emotional state tags is maintained in both the training and the test speech data.

MODELING THE EMOTIONS

Training the Gaussian mixture models

Prosody has been used by various researchers for detecting emotion (Seppanen et. al, 2003). Combining prosody with the content of the speech signal may be a promising way of performing effective emotion detection (Yacoub et. al,2003). In this work we simply concentrate on the acoustic features alone and temporarily ignore the automatic recognition of the speech signal and its content. The speech signal is first segmented into voiced and unvoiced regions. To perform the segmentation, we use the smoothed F0 contour as shown in Figure 3. In Figure 3 (a), is shown the speech signal corresponding to the utterance “fell on her , I got [pause].. I got an ambulance”. The thick lines denote the boundaries as detected from the regions of continuity of the smoothed F0 contours as shown in Figure 3 (c). Note that Figure 3 (c) illustrates the smoothed version of the raw F0 contour shown in Figure 3 (b).

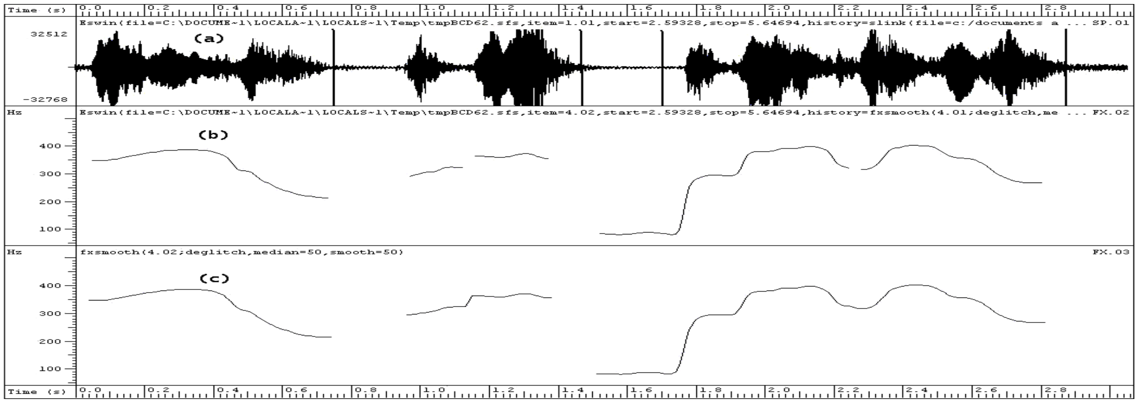

Figure 3: Segmentation of the speech signal for emotion tagging and feature extraction with thick lines denoting segment boundaries selected for emotion tagging and feature extraction. (a) The speech signal, (b) The F0 contour, and (c) The smoothed F0 contour.

The fourteen features listed in Section II are extracted from each voiced segment representing the different classes of emotion. We consider only the first ten seconds of each speech segment representing each class of emotion for the segmentation. Subsequently we build Gaussian mixture models (GMMs), using the training data discussed in Section III. This results in three GMMs representing each class of emotion. It should be noted that a single feature vector is formed per each tagged speech segment instead of a sequence of feature vectors. A model that consists of the component mixture densities, means, and co-variances derived rom the feature vector, is called the Gaussian mixture model (GMM) of that emotional class. The GMM λ, corresponding to a particular class of emotion is given by

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \quad \text{---- (1)}$$

where p_i are the component mixture densities, μ_i the means, and Σ^{-1} the co-variances. The training procedure is similar to the procedure followed in vector quantization. Clusters are formed within the training data. Each cluster is then represented with multiple Gaussian pdfs. The union of many such Gaussian pdfs is a GMM. The most common approach to estimate the GMM parameters is the maximum likelihood estimation where p(X|λ) is maximized with respect to λ. p(X|λ) is the conditional probability and vector $X = \{x_0, x_1, ..., x_{(M-1)}\}$, is the set of all feature vectors belonging to a particular class of emotion. Since there is no closed form solution to the maximum likelihood estimation, convergence is guaranteed only when large enough data is available. An iterative approach using expectation maximization (EM) algorithm is followed. The EM algorithm improves on the GMM parameter estimates by iteratively checking for the condition

$$p(X|\lambda^{k+1}) > p(X|\lambda^{k}) \qquad \text{---- (2)}$$

where k is the number of iterations.

Classification of the three emotions

During the testing phase, the test speech signal is first segmented into voiced and unvoiced regions as discussed in Section III (B). The fourteen features discussed in the preceding Section III (A), are extracted from the segmented speech signal. The sequence of feature vectors is fed through the three emotion models (GMMs). This results in log likelihoods for each emotion. The model that maximizes this log likelihood is declared as the correct match. Let the number of models representing different emotional classes be S. Hence λ_j, where j = {1, 2, 3, …., S}, is the set of GMMs under consideration. For each test utterance, feature vectors x_n at time n are extracted. The probability of each model given the feature vectors x_n is given by

$$P(\lambda_j|x_n) = \frac{p(x_n|\lambda_j)P(\lambda_j)}{P(x_n)} \qquad \text{---- (3)}$$

Since $P(x_n)$ is a constant, and $P(\lambda_j)$ the apriori probabilities are assumed to be equal, the problem is reduced to finding the λ_j that maximizes $p(x_n \mid \lambda_j)$. But $p(x_n \mid \lambda_j)$ is given by

$$P(x_n|\lambda_j) = p(\{x_0, x_1, ..., x_{M-1}\}|\lambda_j) \qquad \text{---- (4)}$$

where M is the number of feature vectors for each frame of the speech signal belonging to a particular class of emotion. Assuming that each frame is statistically independent, Equation 4 can now be written as

$$p(\{x_0, x_1, ..., x_{M-1}\}|\lambda_j) = \prod_{m=0}^{M-1} p(x_m|\lambda_j) \qquad \text{---- (5)}$$

Applying logarithm on Equation 5, and simplifying for S, we have

$$S_r = \max_{1 \leq j \leq s} \sum_{m=0}^{M-1} log[p(x_m|\lambda_j)] \qquad \text{---- (6)}$$

where S_r is declared as the emotional class to which the feature vectors belong. Note that $\{S_r, r = \{1,2,3\}\}$ is the set of all emotional classes. This approach lets the emotion detection to be done in near real time and suits applications that use such a detection scheme to prioritize transmission of emergency information.

FEATURE SELECTION

Feature selection techniques have been widely used in pattern recognition primarily to prune features that do not contribute to discrimination among classes. A few of them suit speech recognition applications, where recognition rate must be used as an internal criterion for selecting features. The sequential forward search (SFS) is one such technique. In this Section, we use the SFS technique, to evaluate the full set of fourteen features using the Bhattacharya distance metric. The Bhattacharya distance measure is used to investigate class separability criteria in this context since the classes to be recognized are few. The Bhattacharya distance is primarily defined for a two class problem and extension to a multi-class case is a combination of pairwise bounds as illustrated in the Equations 7 and 8. The Bhattacharya distance, for a two class case is given by

$$B_{pair}(X) = \frac{1}{2} \int [p(X|\omega_1)p(X|\omega_2)]^{1/2} dX \qquad \text{---- (7)}$$

while the Bhattacharya distance for an M class case is given by

$$B_M(X) = \frac{1}{2} \sum_{i>j}^{M} \sum_{j=1}^{M} \int [p(X|\omega_i)p(X|\omega_j)]^{1/2} dX \qquad \text{---- (8)}$$

We use the following algorithm for the calculation of the Bhattacharya distance using sequential forward search (SFS).

```
1. Start with an empty set P = {ø}, as the current set of selected features
2. Let Q be the full set of 14 dimensional features for emotion detection
3. While the size of P is less than 16

(a) for each  v Є Q
i. set P' ← {v} U P
ii. compute the Bhattacharya distance v* with P'
(b) set  P' ← {v} U P
(c) set  Q ← Q\{v*}
(d) save the Bhattacharya distance calculated with the current P
4. Return the feature number and the corresponding separability criterion (Bhattacharya distance)
5. Rank the features based on the separability criterion
```

The features ranked according to the algorithm mentioned above are listed in Table 1. The first ten features ranked according to the Bhattacharya distance are also used in the two sets of emotion detection tests whose results are listed in Table 2 and Table 3.

Feature Dimension	Feature Name	Rank assigned based on SFS feature selection
1	Range of F0 variation	4
2	Mean of F0	1
3	Median of F0	2
4	Standard deviation of F0	3
5	Mean energy	5
6	Accumulated sum of energies computed from	6
7	Range of Energy above 3 KHz	7
8	Range of energy below 3 KHz	8
9	Energy in the 1700 – 2000 Hz critical band	9
10	Duration of the voiced segment to account for	10
11	Maximum value of the autocorrelation of each	11
12	Range of the first formant F1	12
13	Range of the second formant F2	13
14	Range of the third formant F3	14

Table 1: Ranking of the fourteen features used for emotion detection based on the Bhattacharya distance and the SFS feature selection algorithm

PERFORMANCE EVALUATION OF THE EMOTION DETECTION MECHANISM

To evaluate the performance of the emotion detection system we performed two sets of experiments. The first set of experiments considered three emotional states namely extreme panic, moderate panic and normal state of emotion. The second set of experiments considered only two emotional states extreme panic and normal state of emotion. Table 1 gives a summary of the results of emotion detection for the three classes of emotion defined by us namely, extreme panic, moderate panic and normal state of emotion. In Table 2 is listed the results of performance evaluation of the emotion detection system for the second set of experiments considering only two states of emotion.

Feature Set	Number of Mixtures in the GMM	% Recognition - Extreme Panic	% Recognition - Moderate Panic	% Recognition - Normal state
Full Set (14 features)	32	75	55	62
Full Set (14 features)	64	80	65	68
Full Set (14 features)	128	87	70	75
Full Set (14 features)	256	87	72	75
Sub set using SFS (10 features)	32	65	45	52
Sub set (10 features)	64	72	55	62
Sub set (10 features)	128	86	73	78
Sub set (10 features)	256	86	75	82

Table 1: Performance evaluation of the emotion detection system for classification of three emotional states

Feature Set	Number of Mixtures in the GMM	% Recognition - Extreme Panic	% Recognition - Normal state
Full Set (14 features)	32	86	85
Full Set (14 features)	64	95	94
Full Set (14 features)	128	97	96
Full Set (14 features)	256	97	96
Sub set using SFS (10 features)	32	65	52
Sub set using SFS (10 features)	64	72	62
Sub set using SFS (10 features)	128	86	78
Sub set using SFS (10 features)	256	86	82

Table 2: Performance evaluation of the emotion detection system for classification of two emotional states

It is significant to note that as the number of mixtures is increased in the Gaussian mixture model the recognition performance improves. But an increase from 128 mixtures to 256 mixtures does not yield any improvement which is expected given the nature of the speech data collected. The recognition of speakers in extreme panic state is found to be quite good while it can be seen that the recognition of the other two emotional states is not very encouraging. This is primarily because of the lack of variation in the data collected which leads to a high degree of confusion between moderate panic and normal panic. It could also be a result of errors in manual emotion tagging performed for preparing the database. But restricting the classification to just two emotional states namely the extreme panic case and the normal state of emotion gives a very promising recognition performance for emotion detection. The result of such an experiment where emotion detection is performed for only two different states is listed in Table 2.

PERFORMANCE EVALUATION OF THE QUALITY OF SERVICE MECHANISM

We simulated the medium access control layer based Quality of Service (QoS) provisioning system using GlomoSim simulator. The Simulation parameters are described in Table 3. The results obtained through simulations are described below.

Parameter	**Value**

MAC protocol	IEEE 802.11
Propagation model	Two ray propagation model
Mobility	None
Number of nodes and transmission range	30 and 360 meters, respectively
Bandwidth per call	68Kbps
Maximum number of calls in the system	27
Average duration of calls	200seconds
Fraction of normal, moderate, and extreme calls	33% of the total calls per category
Terrain dimensions	2000mx2000m and 1000mx1000m

Table 3. Simulation parameters used in our study

Among the several results, we present two important performance observations here. Figure 1 presents the average end-to-end delay experienced by each call with respect to the number of calls in the network. The size of the terrain area (1000mx1000m) and the number of nodes (90 nodes) chosen are such that the network experience very high load. Our system is expected to provide the best performance when the network experience high traffic and contention. In this experiment, the number of calls in the network is increased from 3 to 27 and the source and destination of each call are chosen according to a uniformly distributed random variable. Maximum number of calls per node is limited to a maximum of two. Each class of calls, normal, moderately panicked, or extremely panicked, has the same percentage share of the total number of calls in the network. From Figure 1, we noticed that, the end-to-end delay is very minimal when the network load is less than 10 calls and the end-to-end delay is found to be increasing exponentially with the network load. The differentiated contention management mechanism employed in our scheme appeared to provide a better performance for both the moderately panicked calls and the extremely panicked calls compared to the normal calls. In addition, at high loads, we noticed that the extremely panicked calls are performing better. Therefore, once the emotion content of a call is detected and the packets are marked, the network is able to provide a better performance. In our experiments, we found delay performance improvement of 50% and 60%, compared to normal calls, for moderately panicked calls and extremely panicked calls, respectively.

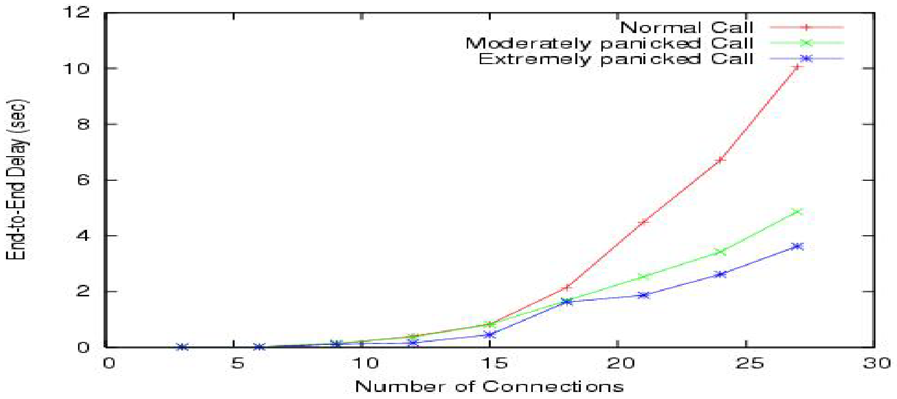

Figure 1 End-to-end delay performance.

We also carried out experiments to measure the throughput obtained per call among the three classes of calls under study in the network. Figure 2 shows the average throughput achieved per call for each of the three classes of calls. For the same terrain dimensions and the number of nodes as in the above experiment, we conducted the throughput measurement experiment and we obtained a peak throughput of about 68Kbps per call when the network load is very light. As the network load is increasing, we found the throughput achieved per call started decreasing. With high load in the network, the per call throughput achieved came down to a minimum of 37.5Kbps for the normal calls. While the per call throughput achieved by the moderately panicked calls and extremely panicked calls came down to a minimum of about 45Kbps. While, there are only minor deviations between the moderately panicked and extremely panicked classes, with further increase in the network load, we noticed the extremely panicked calls gain better throughput. The throughput gain achieved in the case of the panicked calls is about 20% in comparison to the normal calls.

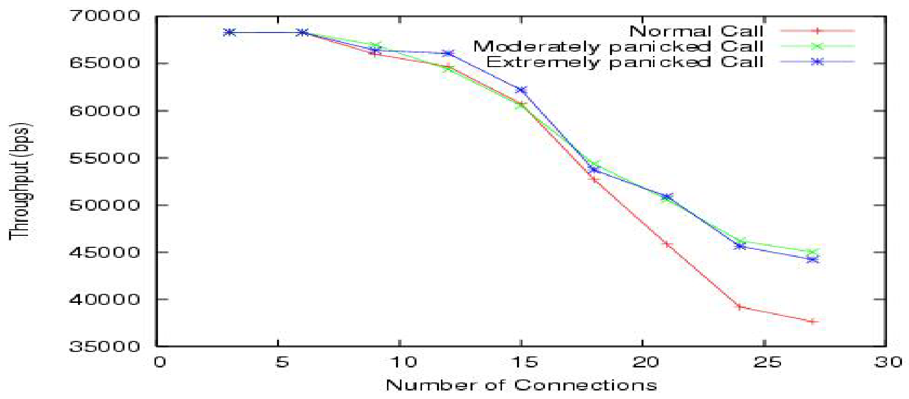

Figure 2. Average throughput achieved per call.

CONCLUSION

We proposed a system where the network QoS is provisioned based on the emotion content in the voice stream. It is also illustrated in this work that detection of emotion from a speech signal with specific reference to the three emotional states which we call extreme panic, moderate panic and normal state is possible without any manual intervention. Speech features for implementing an automated emotion detection system are identified. Feature selection is also implemented on the full feature set to derive a smaller feature set based on the SFS technique with Bhattacharya distance as a discriminability measure. The system seems to give promising results when the emotional classes are reduced to two states namely extreme panic and normal. Such an automated system can be used to feed an emergency response network where a decision on the emotion of the speaker may be of significant interest to implement more efficient and state of art emergency response systems. We have seen average end-to-end delay reduction to the tune of more than 60% for the panicked calls compared to the normal calls.

ACKNOWLEDGMENTS

Work described in this paper was funded by the RESCUE project at UCSD, NSF award #0331690, and the Responsphere project, NSF award #0403433.

REFERENCES

1. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G.Taylor, (2001) Emotion Recognition in Human-Computer Interaction, *IEEE Signal Proc. Mag.*, 18(1).
2. Sherif Yacoub, Steve Simske, Xiaofan Lin, John Burns, (2003) Recognition of Emotions in Interactive Voice Response Systems**,** HPL technical report.
3. J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, (2002) Prosody-based automatic detection of annoyance and frustration in human-computer dialog, ICSLP.
4. Tapio Seppänen, Eero Väyrynen and Juhani Toivanen, (2003) Prosody-based classification of emotions in spoken Finnish. Proc. 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, pp. 717 - 720.
5. Hess, W., (1983) Pitch Determination of Speech Signals. Springer-Verlag, Berlin.
6. X.D. Huang, Y.Ariki, and M.A. Jack, (1990) Hidden Markov Models for Speech Recognition. Edinburgh Univ. Press.
7. L. R. Rabiner and B. H. Juang, (1993) Fundamentals of Speech Recognition. New Jersey: Prentice Hall.
8. Virtual museum of the city of San Francisco at http://www.sfmuseum.net/1989/sc911.html
9. K. Fukunaga, (1990) Introduction to Statistical Pattern Recognition. Boston: Academic Press.
10. TOOLDIAG : A pattern recognition toolbox. http://documents.cfar.umd.edu/resources/source/tooldiag.html/
11. Murthy, C. S. R., Manoj, B. S., (2004) Ad hoc Wireless Networks: Architectures and Protocols. New Jersey: Prentice Hall PTR.