Emotion Recognition and Classification in Speech using Artificial Neural Networks

Akash Shaw
National Institute of Technology
Warangal – 506004
Telangana, India

Rohan Kumar Vardhan
National Institute of Technology
Warangal – 506004
Telangana, India

Siddharth Saxena
National Institute of Technology
Warangal – 506004
Telangana, India

ABSTRACT
To date, little research has been done in emotion classification and recognition in speech. Therefore, there is a need to discuss why this topic is interesting and present a system for classifying and recognizing emotions through speech using neural networks through this article. The proposed system will be speaker independent since a database of speech samples will be used. Various classifiers will used to differentiate emotions such as neutral, anger, happy, sad, etc. The database will consist of emotional speech samples. Prosodic features like pitch, energy, formant frequencies and spectral features like mel frequency cepstral coefficients will be used in the system. Further the classifiers will be trained by using these features for classifying emotions accurately. Following classification, these features will be used to recognize the emotion of the speech sample. Thus, many components like pre-processing of speech, MFCC features, classifiers, prosodic features come together in the implementation of emotion recognition system using speech.

General Terms
Pattern Recognition, Speech.

Keywords
ANN, MFCC, prosodic features, emotion classification and recognition, pre-processing.

1. INTRODUCTION
The interaction between humans and computers has received a lot of attention off late. It is one of the most popular areas of research and has great potential. Teaching a computer the understanding of human emotions is an important aspect of this interaction. A lot of successful applications related to speech recognition are available in the market. People can use their voice to give commands to car, cell-phones, computer, television and many electrical devices. Thus, to make a computer understand human emotions and give a better interaction experience becomes a very interesting challenge.

The most common way to recognize any speech emotion is extracting important features that are related to various emotional states from the speech signal (i.e. energy is an important feature to distinguish happiness from sadness), feed these features to the input end of a classifier and obtain different emotions at the output end. This process is shown in the figure below.

In this paper, the aim is to classify a batch of recorded speech signal into four categories, namely: happy, sad, angry, natural. Before extraction, pre-processing is performed on the speech signals. Samples are taken from the speech and the analog signal is convereted to digital signal. Then each sentence is normalized to ensure that all the sentences are in the same volume range. At last, segmentation separates signal in frames so that speech signal can maintain its characteristics in short duration. Commonly used features are chosen for study and subsequently extracted. Energy is the most basic feature of speech signal. Pitch is frequently used in this topic and autocorrelation is used to detect the pitch in each frame. After autocorrelation, statistical values are calculated for speech signals. Formant is another important feature. Linear Predictive Coding (LPC) method is used to extract the first formant. Similar to pitch, statistical values are calculated for first formant. Mel frequency cepstral coefficient (MFCC) is a representation of short-term power spectrum on a human-like mel scale of frequency. First three coefficients of MFCCs are taken to derive means and variances. All features of the speech samples are put into Artificial Neural Network (ANN), which consists of an input matrix along with a target matrix, which indicate the emotion state for each sentence composed the input of neural network. Artificial Neural Network is used to train and test the data and perform the classification, in the end figures of mean square error and confusion will be given to show how good the performance is. [1][2][11][12]

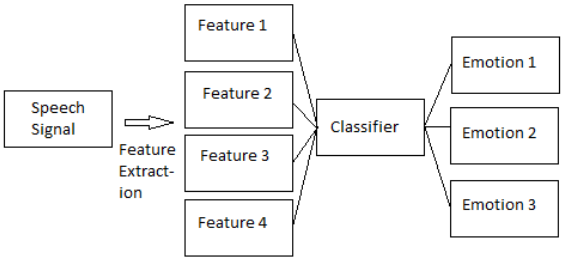

Figure 1: Flow of emotion recognition and classification

2. PRE-PROCESSING FOR EMOTION RECOGNITION
Prior to feature extraction, some necessary steps are taken to manipulate speech signal. Pre-processing mainly includes sampling, normalization and segmentation.

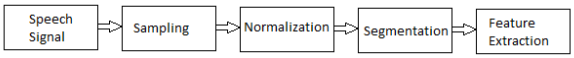

Figure 2: Pre-processing for emotion recognition

Speech signal is analog in form and it needs to be converted into digital form for processing. Analog signal is converted into discrete time signal with the help of sampling. Sampling ensures that the original characteristics of the signal are retained. According to sampling theorem, when the sampling frequency is greater than or equal to twice the maximum

analog signal frequency, the discrete time signal is able to reconstruct the original analog signal.

Volume is an important fact when calculating speech energy and other features. Normalization process uses the signal sequence to ensure each sentence has a comparable volume level.

Speech is a random signal and its characteristics change with time, but this change is not instantaneous. Segmentation process divides the signal sequence into multiple frames with overlap. Overlapping is done to avoid loss of data due to aliasing. The signal s(n) becomes $s_i(n)$ once framed, where i indicates the number of frames. After pre-processing, characteristics of the whole speech signal can be studied from statistical values. [3]

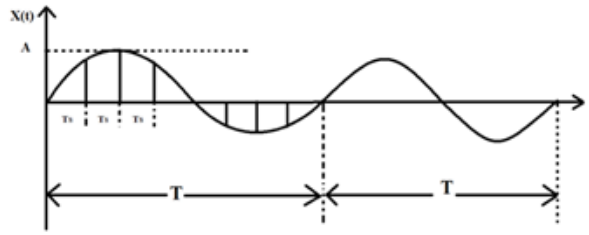

Figure 3: Sampling process

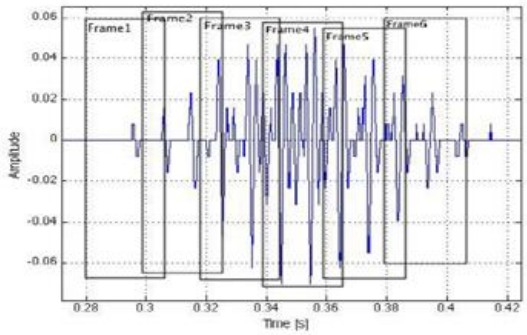

Figure 4: Segmentation process

3. FEATURES FOR EMOTION CLASSIFICATION AND RECOGNITION

3.1 Energy

Energy is the most basic feature in speech signal processing. It plays a vital role in emotion recognition, e.g. speech signals corresponding to happiness and anger have much higher energy than those of sadness. [6]

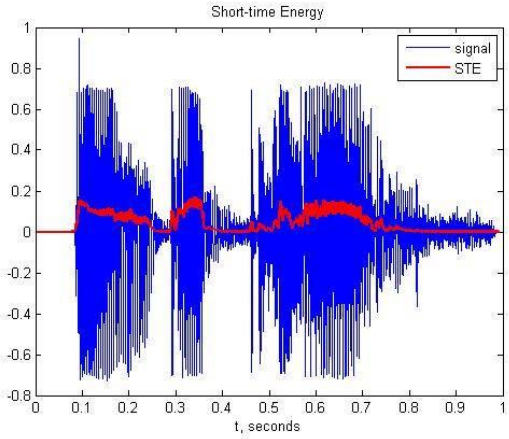

Figure 5: Short term enegry

3.2 Pitch

Pitch is known as the perceived rising and falling of voice tone. It is the perceptual form of fundamental frequency because it sets the periodic baseline for all higher frequency harmonics contributed by oral resonance cavities. It represents the vibration frequency of vocal folds during speaking.

There are many ways to estimate pitch from a speech signal. Auto-correlation method is used because it is a commonly used method and is easy to practice. This method uses short-term analysis technique to maintain characteristic for each frame, which means pre-processing should be fully applied before pitch extraction. Since autocorrelation can decide the period of a periodic signal, autocorrelation is applied for each frame. [6]

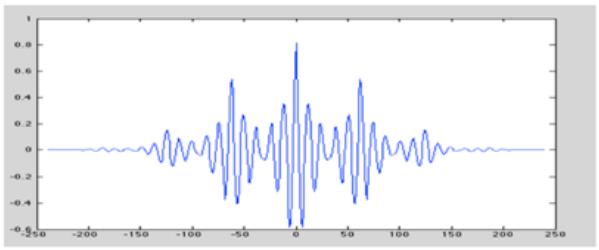

Figure 6: Autocorrelation

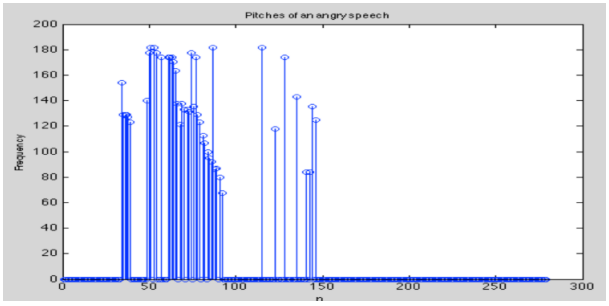

Figure 7: Pitch of an angry signal

3.3 Formant Frequencies

Formant frequencies are defined as resonances in vocal tract and they determine characteristic timbre of vowel. It is also a very useful feature for speech recognition and could be found in many speech emotion studies. The peaks of the frequency response from a linear prediction filter are the formants. [6][14]

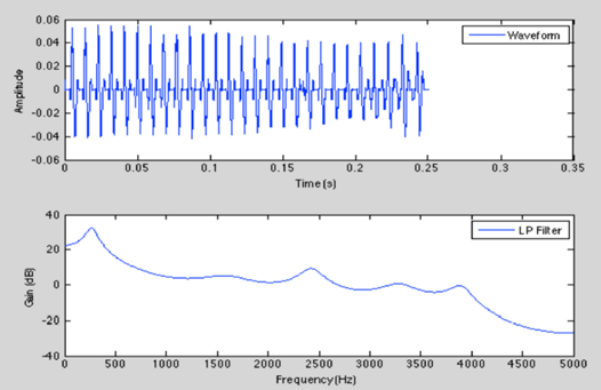

Figure 8: Frequency response of a linear predictive filter

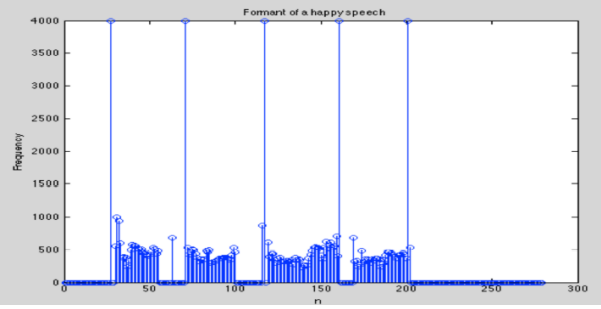

Figure 9: Formant frequencies of a happy speech

3.4 Mel Frequency Cepstral Coefficients (MFCC)

The Mel-Frequency Cepstrum Coefficients (MFCC) is an accurate representation of short time power spectrum of a sound. The advantage of MFCC is that it imitates the reaction of human ear to sounds using a mel scale instead of linearly spaced frequency bands. [2]

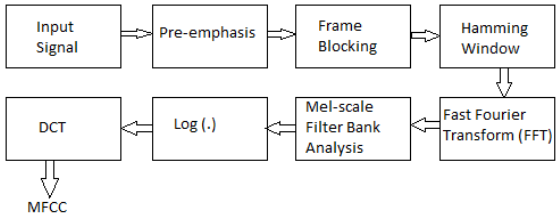

Figure 10: MFCC block diagram

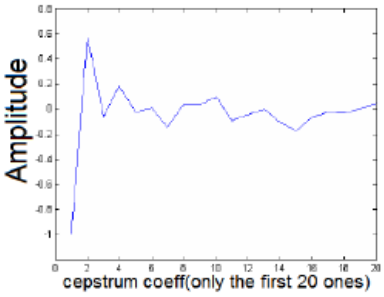

Figure 11: MFCC of a speech signal

4. RESULTS

4.1 Classification with Neural Network

The general working flow is shown in figure. Input data and target data are loaded. Input data here is a matrix of the features extracted from the speech inputs. Target data indicates the emotional states of these inputs. Next, the percentage of input data into 3 categories namely training, validation and testing is chosen randomly. The training set fits the parameters of the classifier i.e. finds the optimal weights for each feature. Validation set tunes the parameters of a classifier that is it determines a stop point for training set. [13]

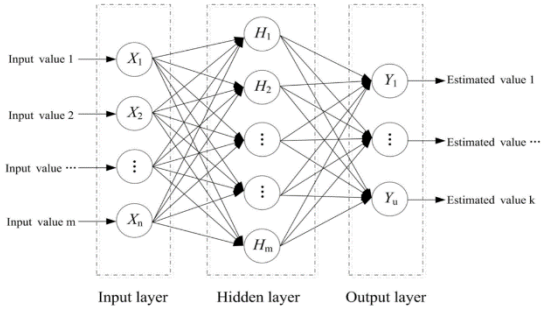

Figure 12: Artificial Neural Network

Finally test set tests the final model and estimates the error rate. The default value sets training in 70 percent and 15 percent each for the rest. Initially the default values are used. Next, the number of hidden layers is chosen. As discussed previously, more the number of hidden layers, more complicated the system, better the result. Lastly the network is trained several times. The mean square together with error rate will indicate how good the results are. [4][5]

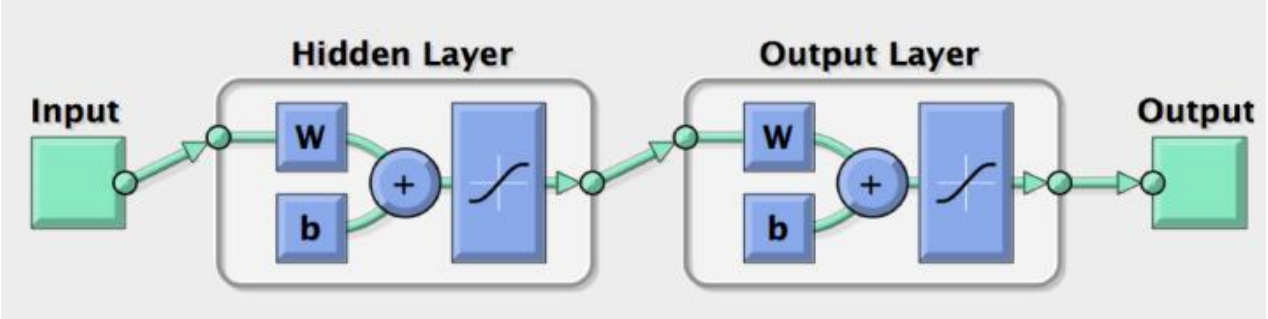

Fig 13: Flow process of speech emotion classification

Table 1: Average values of energy and pitch for different emotions

Emotion	Average energy	Average pitch
Neutral	0.0295	233.033
Happy	0.04829	295.247
Angry	0.04939	217.535
Sad	0.0254	214.789

4.2 Other Remarks

The average log normalized values of energy and pitch for each emotion are shown in the table. Angry emotion has the highest energy, happy has the second highest and sad has the lowest energy. Linear prediction co-efficients are obtained using a high pass all pole filter and consequently first three formant frequencies are obtained. [14] To obtain mel frequency cepstrum coefficients, pre-processing is applied again for a short time as shown in the figure. [6] [7][8]

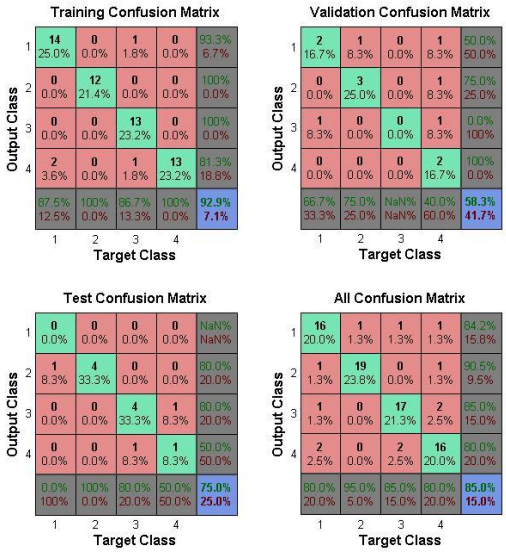

Figure 14: Confusion Matrix for emotion classification with classification accuracy of 85 percent

51	happy	happy	Hit
53	happy	happy	Hit
55	happy	happy	Hit
57	happy	happy	Hit
59	happy	happy	Hit
61	angry	sad	Miss
63	angry	angry	Hit
65	angry	neutral	Miss
67	angry	angry	Hit
69	angry	angry	Hit
71	sad	sad	Hit
73	sad	sad	Hit
75	sad	sad	Hit
77	sad	angry	Miss
79	sad	sad	Hit
2	neutral	angry	Miss
4	neutral	neutral	Hit
6	neutral	neutral	Hit
8	neutral	happy	Miss
10	neutral	sad	Miss
12	happy	happy	Hit
14	happy	neutral	Miss
16	happy	angry	Miss
18	happy	happy	Hit
20	happy	happy	Hit
22	angry	angry	Hit
24	angry	angry	Hit
26	angry	angry	Hit
28	angry	angry	Hit
30	angry	angry	Hit
32	sad	angry	Miss

Figure 15: Recognition of emotion of each speech signal

5. CONCLUSIONS

Firstly, all the speech corps were recorded. It is comparably easier for the classifier to decide the category of each speech e.g. sad speeches are all slow, depressing and powerless.

In the feature extraction process, the accuracy of features plays an important role. The method used in this paper defined a reasonable domain for each detected word sample. For pitch and formant, it is not likely to examine the accuracy for each frame, so the reasonable ranges for pitches and formants are defined to filter other error values out. This is a quick and simple implementation as well, but it might also cause the edge effect. One way to improve this is by applying a start and end point detection algorithm to indicate the start and end sample point of real speech part.

For the classification process, the error rate is highly dependent on the training times. However, the suitable number of training times is neither fixed nor predictable in neural network system. For the random process of choosing training, validation and test data, the weights assign for each feature changes for different choice. To enable to have a desirable result, it is necessary to test the training process several times. After suitable times of training process, extra test signals are load into the system for emotion recognition. With an accuracy of 85 percent classification rate, the selected features (energy, pitch, formant and MFCC) prove to be good representations of emotion for speech signal.

The correctness of recognition of few emotions is shown in the figure. Also, the efficiency of classification of each emotion as a function of the number of hidden layers of ANN is summarised in the table as shown.

No. times ANN was trained	8	10	12	20	Overall Accuracy
Neutral	100%	61.26%	95.20%	91.96%	87.10%
Happy	81.80%	87.46%	94.10%	98.40%	90.44%
Angry	85.66%	95.66%	91.80%	87.30%	90.10%
Sad	56.96%	79.13%	88%	95.46%	79.88%
Overall Accuracy	81.10%	80.88%	92.23%	93.28%	86.87%

6. FUTURE WORK

For the further work the system could be improved by increase the accuracy of extracted features to classify more complicate speech samples i.e. for multiple speakers and more emotions. Also, the system can be designed to recognize emotions like hot and cold anger, panic, surprise, fear, etc. This emotion recognition system can be extended to images as

well. Image-based speech processing to recognize emotions can be incorporated in future works. Also, study shows that people suffering from autism often have difficulty expressing their emotions explicitly. The sole purpose of this project is to recognize emotions. This project can prove to be of assistance to them which will help them to socially interact with their peers. [9][10]

7. REFERENCES

[1] L. Rabiner and B. H. Juang, 1993, Fundamentals of Speech Recognition.

[2] Lawrence Rabiner, Ronald Schafer, Introduction to digital speech processing.

[3] Nicholson J., Takahashi K., Nakatsu R., "Emotion Recognition in Speech using Neural Networks", IEEE Trans. Neural Information Proc., 1999, Vol. 2, 495-501.

[4] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, "Neural Networks used for Speech Recognition", JOURNAL OF AUTOMATIC CONTROL, 2010, University of Belgrade.

[5] S. Haykin, 1999, Neural Networks: a comprehensive foundation.

[6] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, A. Stolcke, "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog", Proc. ICSLP, Denver, Colorado, USA, 2002, 2037-2040.

[7] S. Yacoub, S. Simske, X. Lin, J. Burns, "Recognition of Emotions in Interactive Voice Response Systems", Proc. European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003, 729-732.

[8] J. Liscombe, "Detecting Emotion in Speech: Experiments in Three Domains". Proc. HLT/NAACL, New York, NY, USA, 2006, 231-234.

[9] R.P. Hobson, "The autistic child's appraisal of expressions of emotion. Journal of Child Psychology and Psychiatry", 27, 1986, 321-342.

[10] K.A. Loveland, B. TUNALI-KOTOSKI, Y.R. Chen, J. Ortegon, D.A. Pearson, K.A. Brelsford, M.C. Gibbs, "Emotion recognition in autism: Verbal and nonverbal information", Development and Psychopathology, 9(3), 1997, 579-593.

[11] T. Vogt, E. André, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation", Proc. Language Resources and Evaluation Conference, Genoa, Italy, 2006, 1123-1126.

[12] F. Dellaert, T. Polzin, A. Waibel, "Recognizing Emotion in Speech", Proc. ICSLP, Philadelphia, PA, USA, 1996, 1970-1973.

[13] M. W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech", IEEE ISCAS, Vancouver, May 2004, 23-26

[14] Snell. R. "Formant location from LPC analysis data", IEEE Transactions on Speech and Audio Processing, 1(2),1993, pp. 129–134.