Continuous Wavelet Transform based Speech Emotion Recognition

Pankaj Shegokar and Pradip Sircar
Department of Electrical Engineering,
Indian Institute of Technology Kanpur,
Kanpur 208016, India
E-mail: (pkjshgkr, sircar)@iitk.ac.in

***Abstract*—Emotion recognition from speech is one of the most interesting topics in research community and has developed to a great extent in the last few years. The real challenge in speech emotion recognition (ER) lies in the extraction of features that efficiently encapsulate the emotional information in speech and also do not depend on the speaker. This paper deals with the challenging task of speaker independent ER based on feature selection and classification algorithms. Features are selected based on continuous wavelet transform (CWT) and prosodic coefficients, and are classified and compared using support vector machine (SVM).**

***Index Terms*—Continuous wavelet transform, Morlet wavelet, Emotion recognition, Speech processing, Pattern recognition, Support vector machine**

I. Introduction

Emotion is a conscious experience which involves various biological and psychological activities. It is one of the key characteristics of humans and is reflected mainly in facial expression, body language and speech. We have focused here in capturing the emotions in various features from the speech signals. The problem of emotion recognition in speech signals is unique in its own sense. Since speech signals are non-stationary signals therefore it is necessary to use a tool which retains such information and hence captures the essence of various emotions. Speaker independent emotion recognition is of utmost importance, as it makes more practical sense. Useful applications of speech emotion recognition are human-machine communication, counselling, rehabilitation, automatic feedback in call-center systems.

Some notable works in this area includes emotion speech analysis and synthesis [1], [2]. Mel-frequency cepstral coefficients (MFCC) and its dynamic parameters based recognition is analyzed in [3]. The impact of the classification method and features selection for the speech emotion recognition accuracy is discussed in [4]. In [5], speech classification algorithm based on multi-level extraction techniques is presented.

Wavelet transform is an efficient signal processing method for multi-resolution analysis (MRA) and local feature extraction of non-stationary signals as in [6]–[8]. It involves decomposing a given signal into its scale and time components and thus gives frequency/scale versus time description. Wavelet analysis is localized both in space and scale. Analysis of multicomponent speech like signals by continuous wavelet transform (CWT) based technique can be found in [9]. Wavelet transform (WT) along with genetic algorithm based on hidden Markov model (HMM) is discussed in [10]. Entropy of discrete WT coefficients based ER is analyzed in [11]. Wavelet based feature extraction for phoneme recognition is discussed in [12]. Emotion recognition from speech under environmental noise conditions using Wavelet decomposition is discussed in [13].

In our work, we have demonstrated how the well known concepts of principal component analysis, non-negative matrix factorization methods on Morlet wavelet transformed matrix, and useful prosodic features can be used as features for emotion recognition.

II. Emotion Recognition System

The technique used for emotion recognition from speech signals is based on pattern recognition system. Fig. 1 explains the flow diagram of the method. Every step is explained in detail in the subsections to follow.

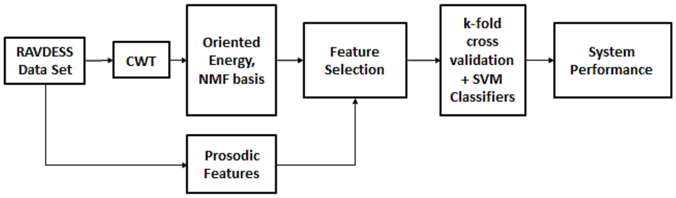

Fig. 1: Emotion recognition system

A. RAVDESS Data Set

The data set used is The Ryerson audio-visual database of emotional speech and song (RAVDESS) which consists of 24 actors (12 male, 12 female) speaking and singing with various emotions, in a north American English accent. It is freely available for scientific-research and general use under a non-commercial creative commons license [14]. The speech set consists of the 8 emotional expressions: Neutral, Calm, Happy, Sad, Angry, Fearful, Surprise, and Disgust. All emotions except neutral are expressed at two levels of emotional intensity: normal and strong. There are 2,452 unique vocalizations, all of which is available in three modality formats: full audio-video (720p, H.264), video-only, and audio-only (wave). We have used only male speech signals (720) in our work.

978-1-5090-0941-1/16/$31.00 ©2016 IEEE

To reduce the complexity, we have removed the initial silence part from each speech signal and changed the original sampling rate of 48 kHz to 16 kHz.

B. Continuous Wavelet Transform

The CWT of a function $f(t)$ at a scale $a \in \mathbb{R}^+$ and translational value $b \in \mathbb{R}$ is defined as:

$$W(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{-\infty} f(t)\overline{\psi}\left(\frac{t-b}{a}\right) dt \tag{1}$$

where $\psi(t)$ is a continuous function both in time domain and frequency domain called the *mother wavelet*, and $\overline{\psi}(t)$ is its complex conjugate. For various values of pair (a,b), mother wavelet produces different *daughter wavelets*. There are a lot of continuous wavelet functions available [8]. But Morlet wavelet is found to be closely related to human perception both hearing and vision. This work analyzes the Morlet wavelet in extracting emotions from speech.

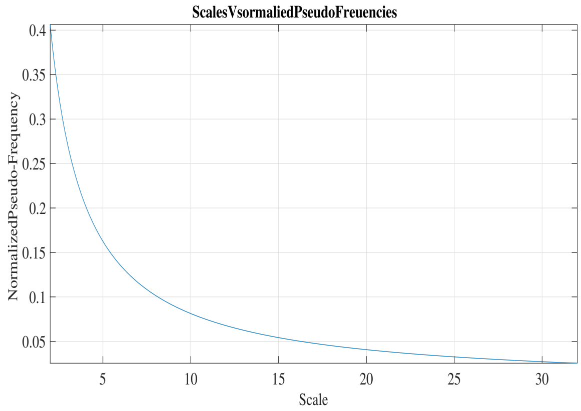

Fig. 2: Scale frequency relationship for Morlet wavelet

A complex-valued Morlet wavelet is defined as:

$$\psi(t) = \exp\left(-\frac{t^2}{2\sigma^2}\right)\exp(i\xi t) \tag{2}$$

Typically, $\sigma^2 = 1$ and parameter ξ allows trade off between time and frequency resolutions and generally $\xi = 5$. The relationship between scale and frequency for Morlet wavelet is inverse, as depicted in Fig. 2. The octave scales are defined [8] as:

$$scales = 2^{\frac{1}{nvoices}[nvoices:(1/nvoices):(nvoices*noctaves)]} \tag{3}$$

where $nocatves$ is the number of octaves used and $nvoices$ is the number of voices used. The idea here is to use intelligently those features which can capture the information regarding emotions in the wavelet matrix efficiently.

In our work we have used features based on the following two techniques:

1) *Principal Component Analysis*
The CWT matrix, $\mathbf{W}^{m\times n}$, is very large (number of scales × length of speech signal). Obviously, we need some compact representation without loosing any information regarding emotion. Principal component analysis (PCA) is suited for such problems where the aim is to determine a subspace of dimension $l \leq m$, such that the variance of the data after projection on this subspace is maximized. The l mutually orthogonal bases are called as *principal directions*. The oriented energy along the $\mathbf{u}_i$ principal direction is given by [15]

$$E(\mathbf{u}_i) = \sum_{k=1}^{n} (\mathbf{u}_i^T \mathbf{x}_k)^2. \tag{4}$$

and is used as the feature.

2) *Non-Negative Matrix Factorization*
It is an approximate factorization of a matrix into non-negative matrices [16]. Consider the CWT matrix, $\mathbf{W}^{m\times n}$, the task of NMF is to perform the following approximate factorization:

$$\mathbf{W} \approx \mathbf{AB} \tag{5}$$

where $\mathbf{A}^{m\times r}$ and $\mathbf{B}^{r\times n}$, $r < min(m,n)$ and all the matrix elements are non-negative. This non-negative constraint is again very practical, as there is no use of representing negative values in emotions. The most commonly used cost function is Frobenius norm of error matrix:

$$\begin{aligned} &\underset{\mathbf{A},\mathbf{B}}{\text{minimize}} && \|\mathbf{W}-\mathbf{AB}\|_F^2 \\ &\text{s.t.} && \mathbf{A}(i,k) \geq 0,\ \mathbf{B}(k,j) \geq 0,\ \forall\, i,j,k \end{aligned} \tag{6}$$

Factorizing the CWT matrix $\mathbf{W}$ such that $r = 1$, the vector $\mathbf{A}^{m\times 1}$ is called as NMF basis vector and is used as a feature vector.

C. Prosodic features of Speech

To capture information about prosody [3], following features are used directly from speech signals:

1) *Linear predictive coding (LPC) coefficients* of a p^{th}-order linear predictor (FIR filter) that predicts the current value of the real valued time series s based on past samples [17].
2) *Root-Mean-Square (RMS) energy* is defined as

$$RMSE\{s[n]\} = \sqrt{\frac{1}{N}\sum_n s^2[n]} \tag{7}$$

3) *Zero crossing rate (ZCR)* is defined as

$$ZCR = \frac{1}{T}\sum_{i=1}^{T-1} \mathbb{I}\{s_i s_{i-1} < 0\} \tag{8}$$

where s is a signal of duration T and $\mathbb{I}\{A\}$ is an indicator function which is 1, if its argument A is 1 and vice-versa.

4) *Entropy* is defined as

$$E\{s\} = \sum_i log(s_i^2) \tag{9}$$

where s is the signal and s_i are the coefficients of s in an orthonormal basis.

5) *Maximum, Minimum, Mean.*

Appending all the features, we form a single feature vector and the same process is repeated for all the speech signals.

D. Validation and Classification

Cross-validation is a model validation technique for estimating how the results of a statistical analysis will generalize to an independent data set. It is used to examine the predictive accuracy of the fitted model [18]. In k-fold cross validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model and the remaining $k-1$ subsamples are used as training data. This validation process is repeated k times(folds) with each of the k subsamples used exactly once as the validation data. The k results from the folds are averaged to produce a single estimation, called as predictive accuracy of the fitted model. There are various types of classifiers such as Bayesian classifiers, linear classifiers, non-linear classifiers etc, each suitable for certain types of problems [19], [20]. In our work, we have used a very popular type of non-linear classifier known as support vector machine (SVM).

III. Experiments and Results

All the results are shown for various emotions, actor 21 (male) and common sentence “Dogs are sitting by the door” in strong intensity.

A. CWT for various emotions

Continuous wavelet transform is calculated using octave scale with $nvoices = 8$, $noctaves = 5$, and shown below:

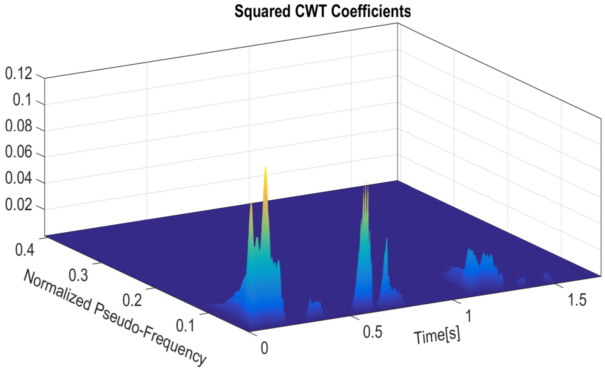

Fig. 3: CWT for Neutral emotion

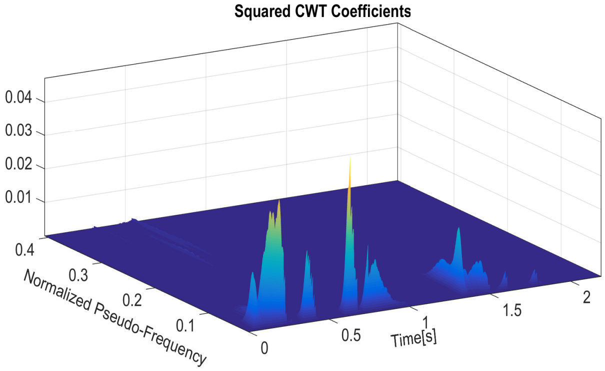

Fig. 4: CWT for Calm emotion

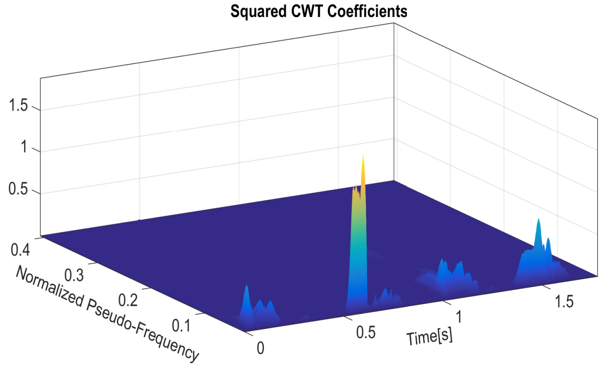

Fig. 5: CWT for Happy emotion

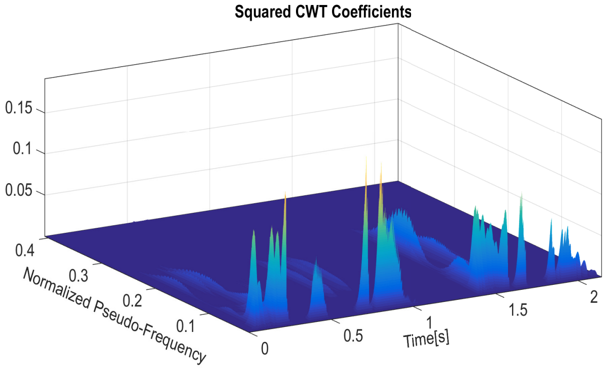

Fig. 6: CWT for Sad emotion

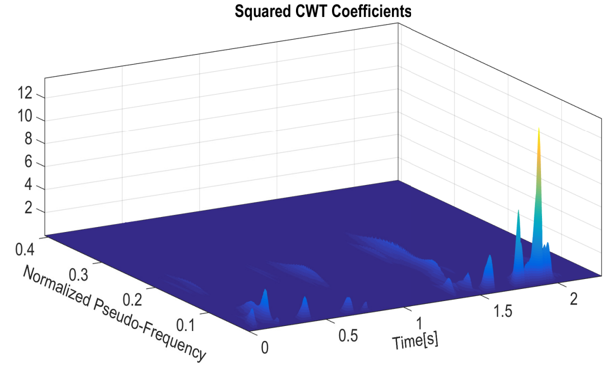

Fig. 7: CWT for Angry emotion

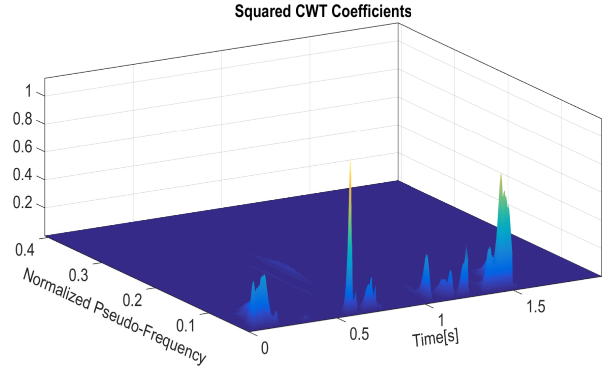

Fig. 8: CWT for Fear emotion

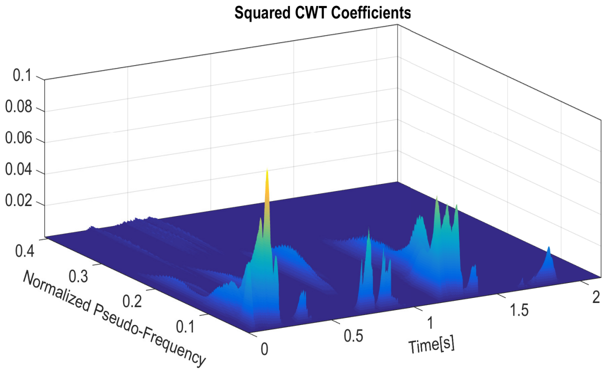

Fig. 9: CWT for Disgust emotion

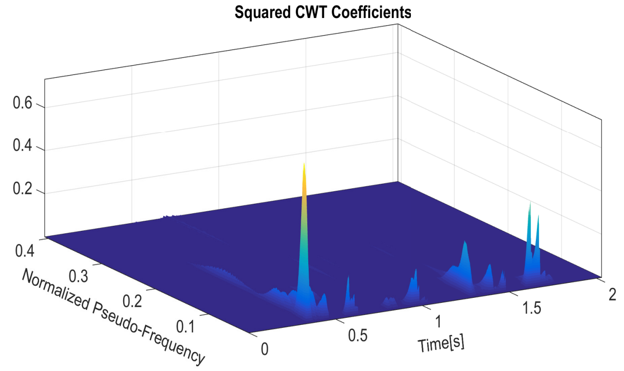

Fig. 10: CWT for Surprised emotion

Two main observations from these plots are that the ranges of energy for various emotions vary considerably and that most of the informations lie in the lower frequency range for majority of the emotions.

1) Oriented Energy: Applying the principal component analysis (PCA) on the CWT matrix and calculating the oriented energy, we find that almost 95% information lies in the first three principal directions for almost all emotions.

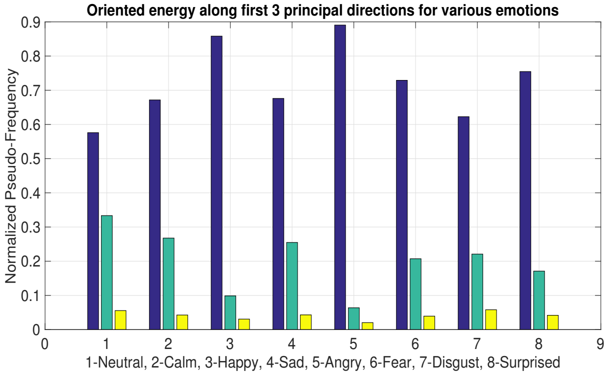

Fig. 11: Oriented Energy Distribution in various emotions

2) Matrix Factorization: Applying the non-negative matrix factorization (NMF) on the wavelet matrix we obtain the NMF basis vectors as depicted below:

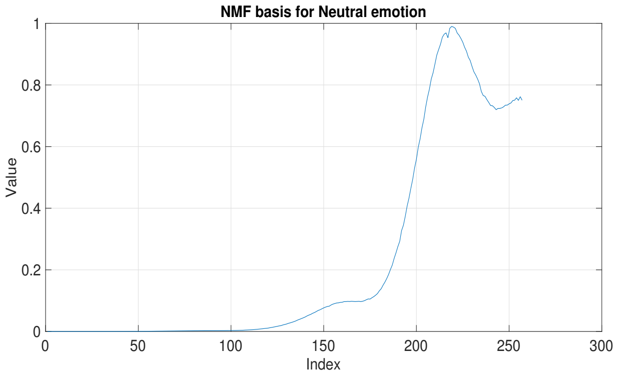

Fig. 12: NMF-1 basis vector for Neutral emotion

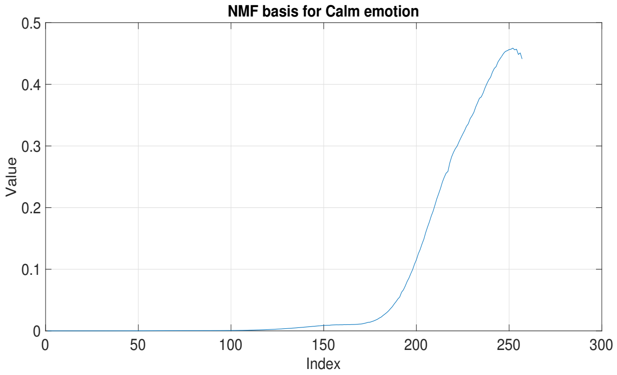

Fig. 13: NMF-1 basis vector for Calm emotion

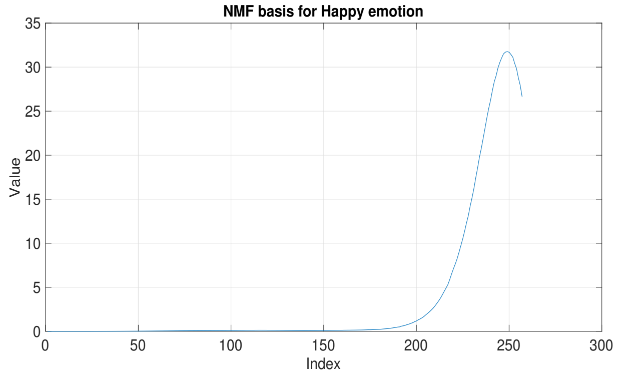

Fig. 14: NMF-1 basis vector for Happy emotion

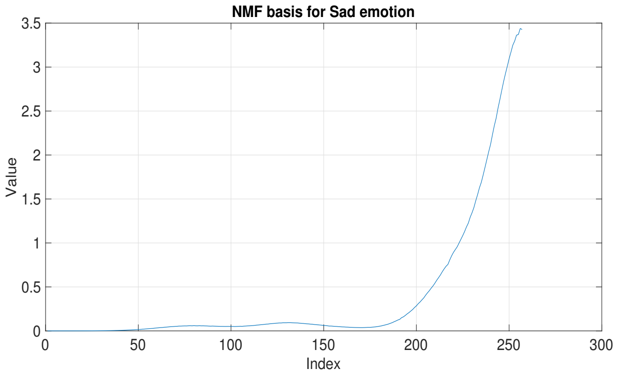

Fig. 15: NMF-1 basis vector for Sad emotion

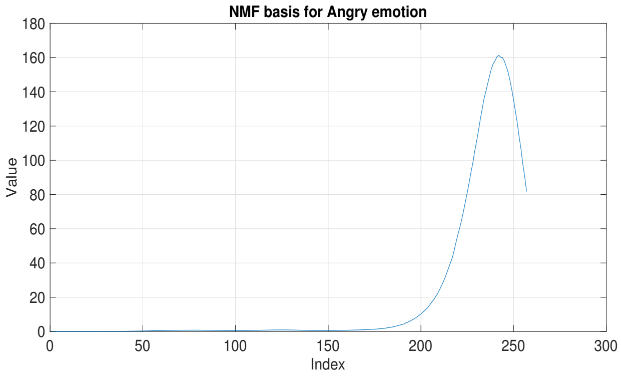

Fig. 16: NMF-1 basis vector for Angry emotion

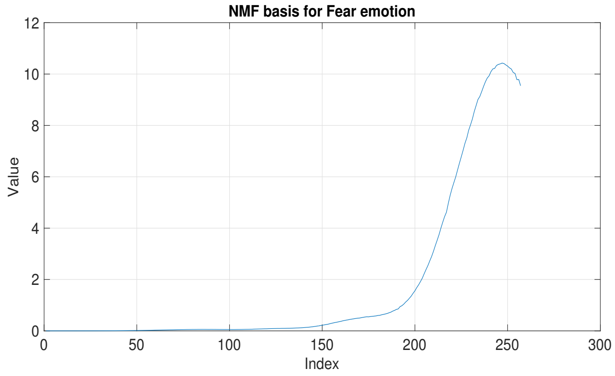

Fig. 17: NMF-1 basis vector for Fear emotion

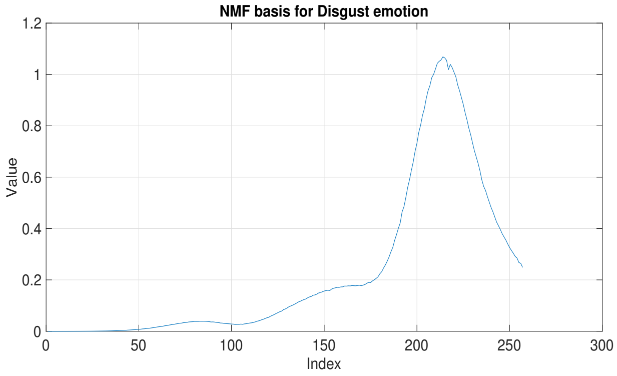

Fig. 18: NMF-1 basis vector for Disgust emotion

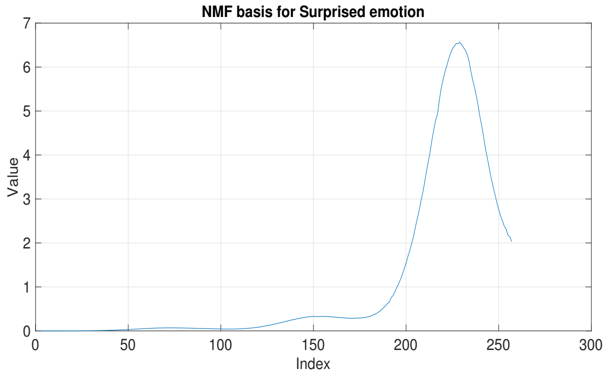

Fig. 19: NMF-1 basis vector for Surprised emotion

B. *Prosodic features*

1) *LPC coefficients:* The 12 LPC coefficients for various emotions are shown below:

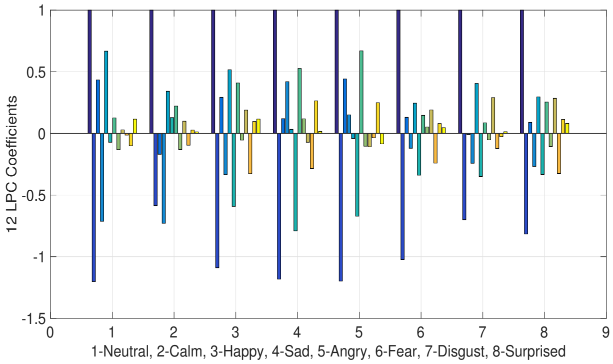

Fig. 20: LPC coefficients for various emotions

2) *RMS Energy:* The RMS energy for various emotions is shown below:

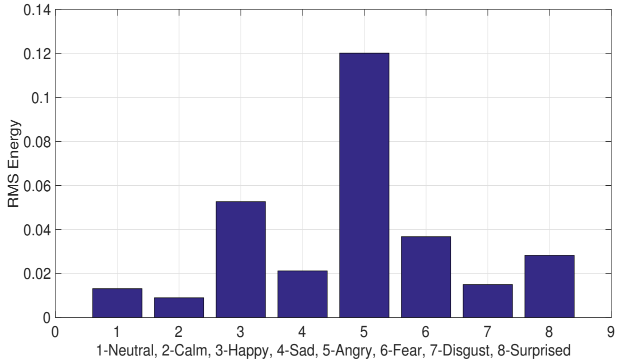

Fig. 21: RMS Energy in various emotions

3) *Maxima, Minima, ZCR:* The Maxima, Minima, ZCR for various emotions is shown below:

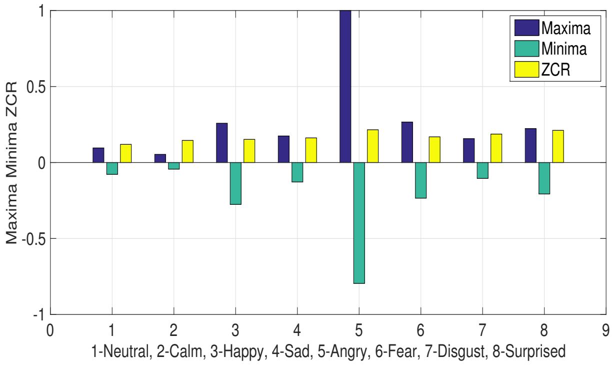

Fig. 22: Maxima, Minima, ZCR in various emotions

4) *Mean:* The Mean for various emotions is shown below:

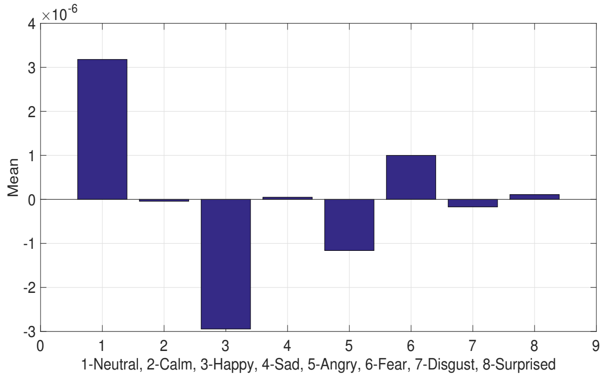

Fig. 23: Mean in various emotions

5) *Entropy:* The Entropy for various emotions is shown below:

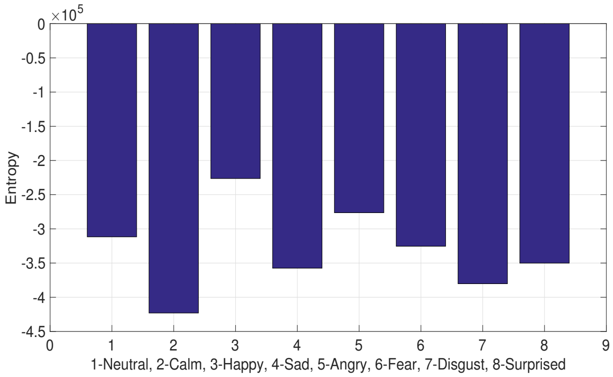

Fig. 24: Entropy in various emotions

IV. Performance

A. Validation Accuracy

Using 5-fold cross validation, we obtain the overall accuracy of the system for the various SVMs, as shown in the table below:

TABLE I: Accuracy for various SVM algorithms

Features	**Linear SVM**	**Quadratic SVM**	**Cubic SVM**	**Gaussian SVM**
3 PCA directions, NMF-1 basis, Prosodic (278) features	51.3%	60.1%	56.3%	52.3%
10 PCA directions, NMF-1 basis, Prosodic (285) features	52.6%	56.8%	55.2%	50.3%

B. ROC Curves

The receiver operating characteristic (ROC) curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR, also known as sensitivity, measures the proportion of positives that are correctly identified. Similarly, the FPR, also known as specificity, measures the proportion of negatives that are correctly identified. Performance of each emotion can be measured by the area under the ROC curve which is an indication of how each emotion is distinctively classified compared with others.

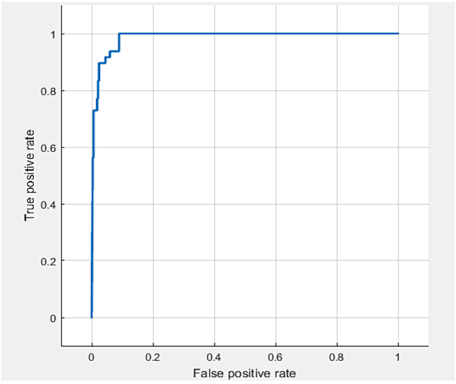

Fig. 25: ROC for Angry, Area=0.986731

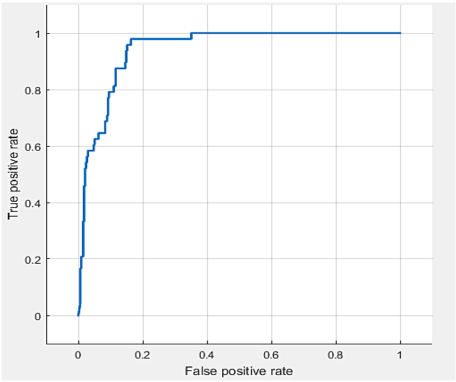

Fig. 26: ROC for Calm, Area=0.942584

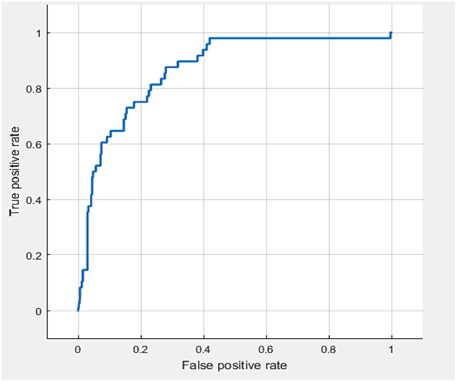

Fig. 27: ROC for Neutral, Area=0.866629

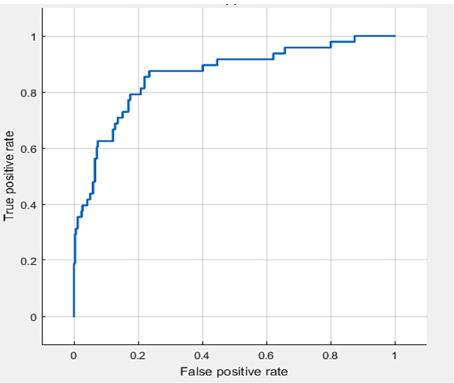

Fig. 28: ROC for Happy, Area=0.861483

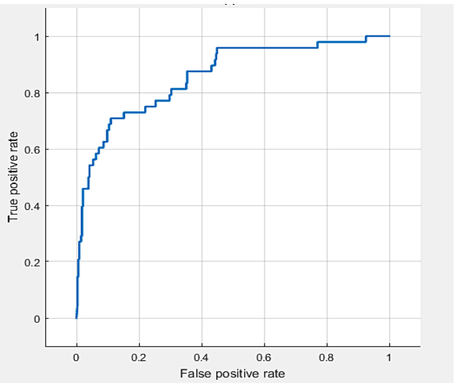

Fig. 29: ROC for Disgust, Area=0.857763

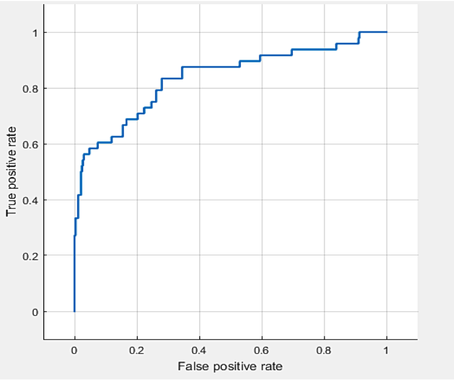

Fig. 30: ROC for Fear, Area=0.836062

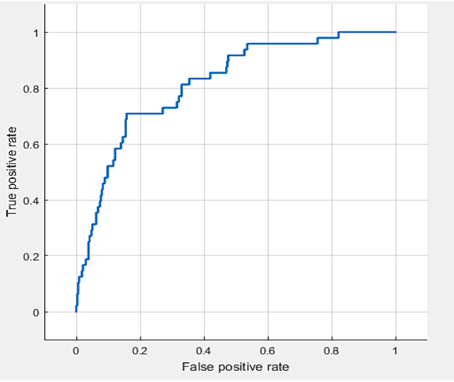

Fig. 31: ROC for Sad, Area=0.816344

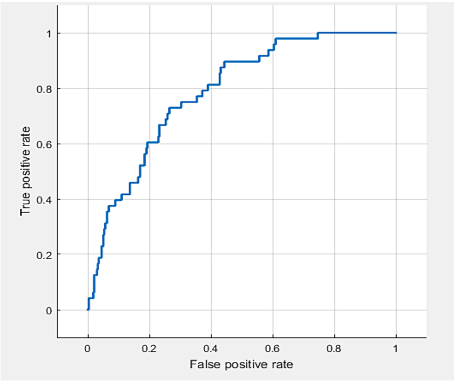

Fig. 32: ROC for Surprised, Area=0.788876

V. Conclusions

The results obtained are highly promising and shows that the classification based on the continuous wavelet transform features can be highly applicable in emotion recognition and classification system. Angry and Calm emotions are recognized very accurately. Performance of Neutral, Happy, Disgust and Fear emotions is also good. But for Sad and Surprised emotions accuracy is slightly compromised. The proposed method is very general since it is speaker independent, and tested with two versions of speech with different durations and intensities (strong and normal). The proposed method is a learning system as it learns from the given data set.

References

[1] J. S. Park, J. H. Kim and Y. H. Oh, “Feature vector classification based speech emotion recognition for service robots”, *IEEE Trans. on Consumer Electronics*, 55(3), 1590-1596, 2009.

[2] E. H. Kim, K. H. Hyun, S. H. Kim and Y. K. Kwak , “Improved emotion recognition with a novel speaker-independent feature”, *IEEE/ASME Trans. on Mechatronics*, 14(3), 317-325, 2009.

[3] P. Partila, M. Voznak and J. Tovarek, “Pattern recognition methods and features selection for speech emotion recognition system”. *Scientific World J.*, 2015.

[4] L. Chen, X. Mao, Y. Xue and L. L.Cheng, “Speech emotion recognition: Features and classification models”, *Digital Signal Processing*, 22(6), 1154-1160, 2012.

[5] A. Zelenik, B. Kotnik, Z. Kačič and A. Chowdhury, “Novel expressive speech classification algorithm based on multi-level extraction techniques”, Proc. 5th Intl. Conf. In Pervasive Computing and Applications (ICPCA), pp. 410-415, 2010.

[6] O. Rioul and M. Vetterli, “Wavelets and signal processing”. *IEEE Signal Processing Magazine*, 8, LCAV-ARTICLE-1991-005, 14-38, 1991.

[7] A. M. Grigoryan,“Fourier transform representation by frequency-time wavelets”, *IEEE Trans. on Signal Processing*, 53(7), 2489-2497, 2005.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[9] P. Sircar, K. Prasad and B. Harshavardhan, “Analysis of multicomponent speech-like signals by continuous wavelet transform-based technique”, Proc. 14th European Signal Processing Conf. (EUSIPCO), pp. 1-5, 2006.

[10] H. Zhiyan and W. Jian, “Speech emotion recognition based on wavelet transform and improved HMM”, 25th Chinese Control and Decision Conference (CCDC), pp. 3156-3159, 2013.

[11] S. Sultana, C. Shahnaz, S. A. Fattah, I. Ahmmed, W. P. Zhu and M. O. Ahmad, “Speech emotion recognition based on entropy of enhanced wavelet coefficients”, Proc. Intl. Symp. on Circuits and Systems (ISCAS), pp. 137-140, 2014

[12] C. J. Long and S. Datta, “Wavelet based feature extraction for phoneme recognition”, Proc. 4th Intl. Conf. In Spoken Language (ICSLP), vol. 1, pp. 264-267, 1996.

[13] J. C. Vásquez-Correa, N. Garciá, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla and E. Nöth, “Emotion recognition from speech under environmental noise conditions using wavelet decomposition”, Proc. Intl. Carnahan Conf. in Security Technology (ICCST), pp. 247-252, 2015.

[14] S. R. Livingstone, K. Peck and F. A. Russo, “Ravdess: The Ryerson audio-visual database of emotional speech and song”, Proc. Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science, 2012.

[15] B. De Moor, J. Vandewalle and J. Staar, “Oriented energy and oriented signal-to-signal ratio concepts in the analysis of vector sequences and time series”, In E.F. Deprettere (Ed.), *SVD and Signal Processing*, 209-232, 1988.

[16] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization”, *Nature*, 401(6755), 788-791, 1999.

[17] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Prentice Hall, 1978.

[18] F. Mosteller, “A k-sample slippage test for an extreme population”, In *Selected Papers of Frederick Mosteller*, pp. 101-109, Springer, 2006.

[19] S. Theodoridis and K. Koutroumbas, *Pattern Recognition* Elsevier, 2003.

[20] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.