Emotion Recognition from Speech using Prosodic and Linguistic Features

Mahwish Pervaiz
Computer Sciences Department
Bahria University, Islamabad
Pakistan

Tamim Ahmed Khan
Department of Software Engineering
Bahria University, Islamabad
Pakistan

Abstract—**Speech signal can be used to extract emotions. However, it is pertinent to note that variability in speech signal can make emotion extraction a challenging task. There are a number of factors that indicate presence of emotions. Prosodic and temporal features have been used previously for the purpose of identifying emotions. Separately, prosodic/temporal and linguistic features of speech do not provide results with adequate accuracy. We can also find out emotions from linguistic features if we can identify contents. Therefore, We consider prosodic as well as temporal or linguistic features which help increasing accuracy of emotion recognition, which is our first contribution reported in this paper. We propose a two-step model for emotion recognition; we extract emotions based on prosodic features in the first step. We extract emotions from word segmentation combined with linguistic features in the second step. While performing our experiments, we prove that the classification mechanisms, if trained without considering age factor, do not help improving accuracy. We argue that the classifier should be based on the age group on which the actual emotion extraction be required, and this becomes our second contribution submitted in this paper.**

Keywords—Emotion Extraction; Prosodic Features; Temporal Features; Dynamic Time Wrapping; Segmentation

I. Introduction

User interfaces are becoming increasing complicated as the requirements, standards and de-facto standards are improving day by day. User interfaces are also providing speech processing systems to help users give commands without physically interacting through a keyboard. Speech is an important means of human to human (and machines) communications, and it carries additional information as well. We can find out the underlying emotional and psychological aspects as well. Speech processing provides a list of various features and characteristics of sound that can be further analyzed to reveal valuable information.

Although research work is done and various applications are developed but emotion recognition from speech is still a challenging task. Main reason for this is variability of expression even for the same emotion. According to representation of emotions in two-dimensional space, joy and anger both have common acoustic attributes like amplitude of voice, pitch, number of times their speech meets zero axis. In the same manner, fear and sad have some common attributes. Therefore, problem occurs between recognition of these two sets of emotions due to the fact that we extract emotion directly and only from speech signal or text and due to the feature set we use for recognition of emotion. Acoustic features of speech like pitch, energy or volume are also somewhat misleading if considered alone e.g., if a person is angry, he might speak in normal tone using harsh words to express his anger. Similarly, people may shout in anger while some don’t. Some people speak loudly when they are happy or excited while other may not. Therefore, we can say that people use speech signal’s features and speech semantics to present their emotions in our everyday life. This makes it clear that we need to extract emotions from acoustic as well as from semantic features to arrive at a conclusive result regarding the hidden emotions in a speech signal.

Therefore, we present an approach to consider prosodic and linguistic features to improve emotions extraction from speech. Our motivation for this research is an expected use in an e-learning system which does and offline emotions analysis of kindergarten students and reports their emotional changes throughout the day. The rest of the paper is organized as follows; we explain our emotions extraction approach in Section 2 and we introduce our two stages of our proposed model in Section 3 and Section 4 respectively. We discuss our experiments for evaluation in Section 5 and discuss our results in Section 6. We finally present related work in Section 7 and conclusion in Section 8.

II. Emotion Extraction from Speech

We implement a two-staged model where we extract emotions from speech signal using prosodic and temporal features in the first stage. We extract emotions from words with semantic orientation in the second stage. We present an overview of our proposal is given in Figure 1.

While considering prosodic features, we consider pitch, energy and Zero Crossing Rate (ZCR), where temporal features including Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coefficient (LPC) are considered in Stage-1. We use Support Vector Machine (SVM) classifier for classification of speech based on temporal and prosodic features. While extracting linguistic features in the second stage, we extract words by segmenting speech signal using pause between words. We then use dynamic time wrapping (DTW) with MFCC for matching signal for recognition of that word. Finally, we compare extracted words with our dictionary which contains words with class labels to see if a word or its synonym is used to express an emotion in word under consideration. Our reason to consider MFCC and LPC in stage-1 is the fact that we are extracting emotions at segment level and according to [21] spectral features specially MFCC and LPC are very beneficial to identify emotions of short length segment [21]. This helps us to get better performance of our system.

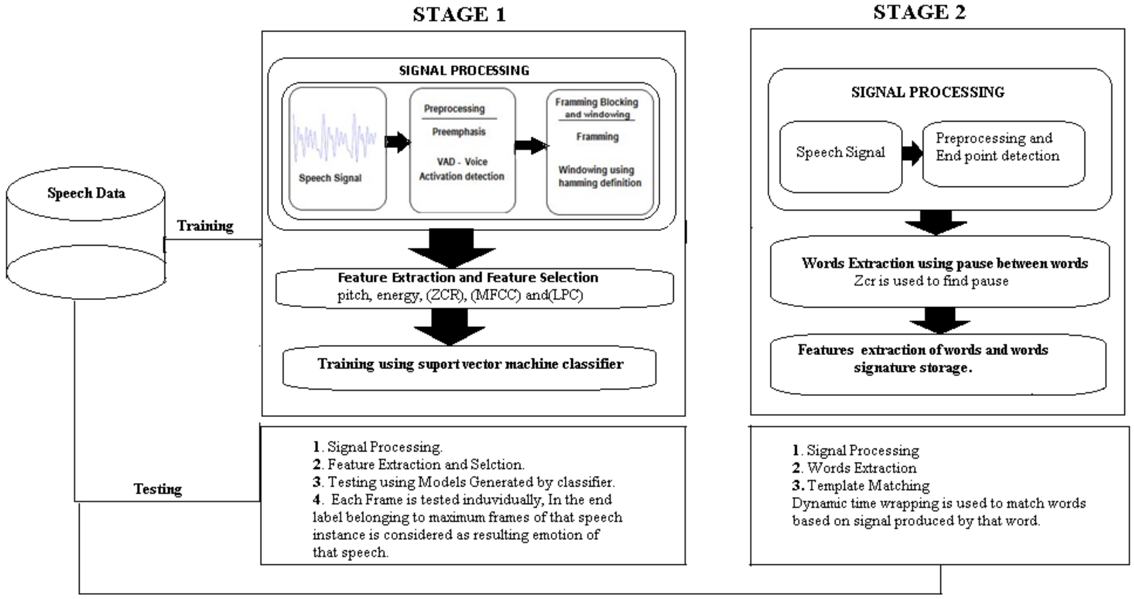

STAGE 1
SIGNAL PROCESSING
Speech Signal
Preprocessing
Preemphasis
VAD - Voice Activation detection
Framming Blocking and windowing
Framming
Windowing using hamming definition
Speech Data
Training
Feature Extraction and Feature Selection
pitch, energy, (ZCR), (MFCC) and(LPC)
Training using suport vector machine classifier
Testing
1. Signal Processing.
2. Feature Extraction and Selction.
3. Testing using Models Generated by classifier.
4. Each Frame is tested induvidually, In the end label belonging to maximum frames of that speech instance is considered as resulting emotion of that speech.
STAGE 2
SIGNAL PROCESSING
Speech Signal
Preprocessing and End point detection
Words Extraction using pause between words
Zcr is used to find pause
Features extraction of words and words signature storage.
1. Signal Processing
2. Words Extraction
3. Template Matching
Dynamic time wrapping is used to match words based on signal produced by that word.

Fig. 1. Emotion Recognition from Speech

III. Stage-I: Emotion Extraction from Speech with Prosodic & Temporal Features

In this stage, emotions are extracted from speech signal using prosodic and temporal features as discussed in above section. This stage comprises of three steps, Signal processing, features extraction and calculation and training of the classifier. We explain of these steps in detail.

A. Signal Processing

As speech signal contains noise that can increase error rate so it is important to minimize noise as possible. Signal processing is performed on speech signal to improve the correctness and efficiency of the extraction process. The following steps are involved in signal processing.

- Pre-Emphasize
- Voice Activity Detection
- Framing and Windowing

Speech signal suffers from additive noise because of its high dynamic range. We apply Pre-Emphasize to spectrally flatten speech signal and removal of low frequency noise. We first remove DC components by extracting mean value from all samples value and then filtering is applied on it. Pre-Emphasizer are often represented by first order high pass filter. The configuration of this filter in time domain is given by Eq. 1. Sample of original signal and pre-emphasize signal is shown in Figure 2.

$$\boldsymbol{h(n) = 1, -0.9375} \qquad (1)$$

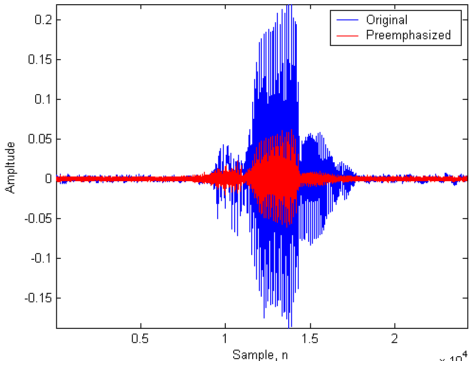


Fig. 2. Pre-emphasized signal

We use voice activity detection to detect start and end point of speech. It is pertinent to note that VAD could be applied in time domain as well as frequency domain. In Time domain volume and ZCR with high order difference are used and in frequency domain variance and entropy of spectrum is used for end point detection. We then select features threshold and then any frame with high values than threshold is considered as positive and negative otherwise [18]. There are four intuitive way to select threshold, as shown by the following four equations:

$$\begin{aligned} Vth &= vmax * \alpha, \qquad (2)\\ Vth &= vmedian * \beta,\\ Vth &= vmin * \gamma,\\ Vth &= v1 * \delta \end{aligned}$$

We use the algorithm presented in [16] to detect end points. After detection if end points we segment speech signal into blocks. Speech signal after voice activity detection is presented in Figure 3. Classification accuracy is proportion to the length of utterance [2] where spectral analysis of speech signal at sub utterance with small size of frame level gave good classification rate. Instead of taking whole speech utterance as a single block, we divide speech signals into small matrix with appropriate length of each frame. We use frame length of 40ms and sample size of 16000Hz and then we calculate total number of points in wav file by multiplying sample rate with frame size. After dividing into frame, we apply windowing on each frame. Hamming window reduces signal discontinuity on end of frames [11]. The coefficients of a Hamming window are computed from the following equation. The window length is L=N+1.

$$\boldsymbol{w\ (k) = 0.54 - 0.46cos(\ 2 \prod k\ K - 1\ )} \qquad (3)$$

B. Feature Calculartion and Feature Selection

After preprocessing is done framing and windowing is applied. Features are extracted from each frame for parameterization of feature vector. Features we have extracted are pitch, MFCC, and LPC. Pitch related statistics convey considerable information in speech about emotions status in a speech segment [1] and Yu et al [23]. We calculate pitch for each frame using correlation function.

Human ear hearing uses nonlinear frequency units [12]. Therefore for each utterance with original frequency another subjective pitch is measured on Mel scale. Mel scale offers linear frequency spacing below 1000h and logarithmic frequency spacing above 1000Hz. As given in literature spectral features gave more classification accuracy at frame level than prosodic features. Kim et al. argued that statistics relating to MFCCs also carry emotional information [10]. In this research we extracted First 13 coefficient of MFCC. Next, we convert signal from time domain to frequency domain using Fast Fourier Transform (FFT). Spectral Analysis showed that in speech signal different timbre correspond differently over frequency with different energy distribution. We apply FFT for converting signal to frequency domain in order to find out the magnitude frequency for each frame. We then measure step Mel frequency wrapping subjective pitch on Mel scale and spectrum is simulated with the use of filter bank. One filter is used for each component. Finally, we convert Log Mel spectrum back to time domain which results in MFCC as real numbers and we use discrete cosine transform to convert back into time domain. Basic idea of linear predictive coding is to predict current frame from linear sequences of past frames.

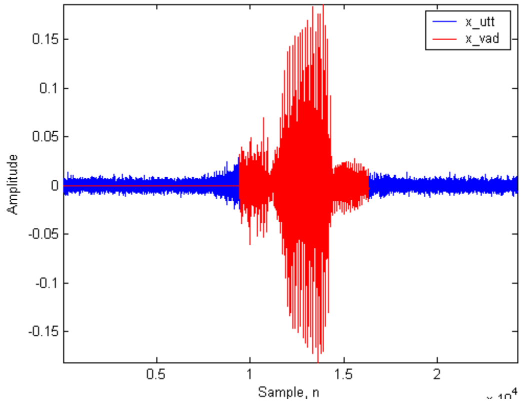


Fig. 3. Speech sample after voice activity detection

We use signal processing tool box for computing LPC coefficients of each frame. We use the following equation to compute LPC as:

$$\boldsymbol{lpc(\ x, p\ )} \qquad (4)$$

Where p is the pth order linear predictor (FIR Filter) that is used to predict current frame x.

$$x\ (\ n) = -a(2)x(\ n-1\ ) - a(3)(\ n-2\ ) \dots \dots \dots \dots \dots \dots \dots \\ - a(\ p+1\ )x(\ n-p\ )$$

We compute 13 LPC coefficient and extract zero crossing rate of the signal by calculating how many time a signal meet with zero axes. Next, we compute energy for each frame using log energy equation defined in [3] as:

$$\boldsymbol{E = 10 * log10\ (e +\ s\ 2n)} \qquad (5)$$

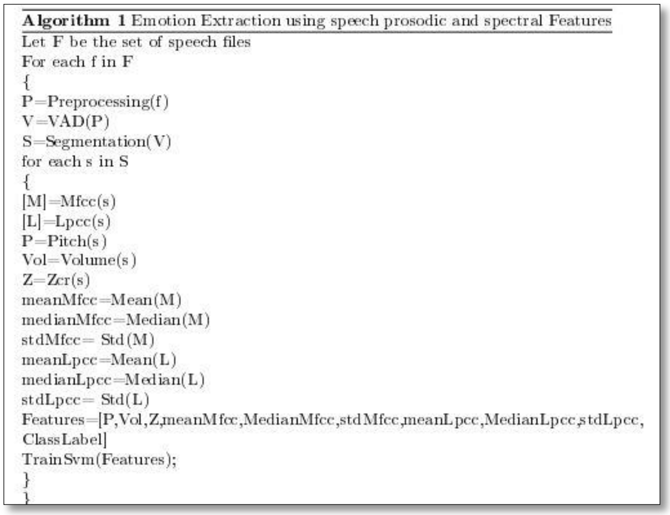
Here, "e" is a small positive constant added to prevent the computing of log of zero. It is important to note that e is much less than the mean-squared value of the speech samples. We use configuration for Energy as given by Eq. 6 below.

$$\boldsymbol{E1 = 10 * log10(\ eps \\ + sum(\ FRAME, 2\ ) \\ /size(\ FRAME, 2\ )\ )} \qquad (6)$$

We then use forward selection algorithm for selection of features after extraction of features. We begin with basic values of energy, volume, MFCC and LPCC and in each iteration, we add one feature and their statistics like mean, median, standard deviation and variance. We compare performance accuracy with the previous iteration at the end of each iteration.

C. Classification through SVM

We use a classifier to predict emotional labels using selected set of features. We use static classifier Support Vector Machines (SVM) that is inherently two class problem for linearly separable data. In our case, we have multidimensional data that is not linearly separable because some classes share same features that could not be linearly separable. We use SVM with One Versus All (OVA) for classification and we perform classification for every single frame of the speech signal and then for whole speech signal emotion with the maximum number of frames is finalized as an emotional label of that speech signal. We train SVM classifier with three data sets. We use 70% of data for training, and we use 30 percent data for testing. The number of models generated by SVM is the same number of emotion classes they have for each data set. Algorithms of emotion extraction with speech spectral and prosodic features is given in Algorithm 1.



```
Algorithm 1 Emotion Extraction using speech prosodic and spectral Features
Let F be the set of speech files
For each f in F
{
P=Preprocessing(f)
V=VAD(P)
S=Segmentation(V)
for each s in S
{
[M]=Mfcc(s)
[L]=Lpcc(s)
P=Pitch(s)
Vol=Volume(s)
Z=Zcr(s)
meanMfcc=Mean(M)
medianMfcc=Median(M)
stdMfcc= Std(M)
meanLpcc=Mean(L)
medianLpcc=Median(L)
stdLpcc= Std(L)
Features=[P,Vol,Z,meanMfcc,MedianMfcc,stdMfcc,meanLpcc,MedianLpcc,stdLpcc,
ClassLabel]
TrainSvm(Features);
}
}
```

IV. Stage - II: Speech Content Processing

In the second stage, we extract emotions based on linguistic features (words and semantics). We use same data sets for training and testing of this system. We perform signal processing as we do in Stage 1 but we do segmentation based on words using ZCR. ZCR is helpful in depicting pause between words and we extract features of each segment (word). We then store these words and features in a file which we treat as the dictionary with labels. We match words with dictionary we created earlier in testing phase, and extract emotions. We extract words by parsing speech signal and then detect segments based on pause between words. We compute Zero-crossing rate to segment speech and consider each segment as a single word.

We extract Mel Frequency Cepstrum Coefficient (MfCC) and Zero Crossing Rate (ZCR) of each word. We store and use MFCC and ZCR for word recognition later and we generate signal for each word uttered and we store in dictionary as discussed earlier. In case we recognize a word, we use corresponding label of that words as their corresponding emotion. We use Dynamic Time Wrapping (DTW) to recognize words uttered by speaker with words stored in our dictionary. We recognize corresponding label of those words and consider it as resulting label of speech.

V. Experimental Setup

We use three datasets for training and testing our model. These are Surrey Audio-Visual Expressed Emotions (SAVEE), Polish Emotional speech database (POLISH) and a locally developed dataset of Sky School Kindergarten students' dataset (KSD). Researchers of Technical University of Lodz, Poland develop Polish Emotional speech database and it contains recordings from four male and four female actors in a single session. Each actor utters five different sentences in 6 different emotions. Emotions are joy, boredom, fear, anger, sadness, and neutral. Sentences do not contain any emotional semantic. There are a total of 240 utterance in the database. The database contains 16 bit, 44 kHz recordings under studio noise environment.

University of Surrey develops SAVEE database [5] where the database comprises audio and visual data of four actors. Age of actors is between 27 to 31 years with average age 30 years. The database comprises of seven emotions i.e., anger, disgust, fear, happiness, sadness, surprise and neutral. Each participant utters fifteen sentences in seven emotions. There are four actors used where two possessing English accent; one possessing southern and one possessing Scottish accent.

We develop a dataset considering students between ages six to eight years. We record speech data for speech recognition and emotion recognition. For speech recognition, we select different emotional keywords mostly used by children and tag them with emotion classes like if there is word “disappointed” we tag it with class sad. If the word is “Wow”, we tag it with class happy or excited. In case, a word belongs to two classes, we tag it to class that contain most probability of having that word. We given students different words and sentences containing emotional keywords and ask them to speak that sentences three to four times for speech recognition module. We record student voices depicting different emotions for emotion extraction from speech signal. We then manipulate speech data at word level, we use five labelers for listening the data in successive order and annotate each sound independently as belonging to one of five classes. We label data at word level and sentence level. We give an emotional keyword and sentence where we find a label with maximum vote by labeler. We give label based on the prosodic and temporal features at sentence level but we do word level labeling based that word solely at the word level. We call our database Kindergarten Students Database (KSD).

VI. Results and Discussion

We present our results considering three speech data sets such that we run both stages separately as well as in conjunction to report overall improvements. We consider SAVEE dataset, Polish dataset and KSD datasets as discussed earlier. Our results for Stage 1 and Stage 2 for SAVEE dataset are displayed in Figure 4 and in Figure 5. Similarly results of polish emotional speech datasets with prosodic and temporal features (stage 1) are demonstrated in Figure 6 and with linguistic features (stage 2) are demonstrated in Figure 7. Figure 8 present results of stage 1 with our own developed dataset (KSD) and stage 2 results for the same dataset are presented in Figure 9. We report our combined results in Figures 10, 11 and 12. It is straightforward to conclude that results of prosodic and temporal features are better than linguistic features with SAVEE and POLISH data sets where linguistic features are proven to be better parameters than prosodic in case of KSD dataset.

The reason behind these result is the variability of speaker's accent, tone, age, fluency and ability to express emotions. SAVEE and POLISH data contains speech of adults while our data set contains voices of students from KG level. Adults can better express their emotions than KG students and their speech's tone varies with their emotion. Another important aspect of SAVEE and Polish datasets is that the recordings are from actors who can better differentiate emotions with their expression and style.

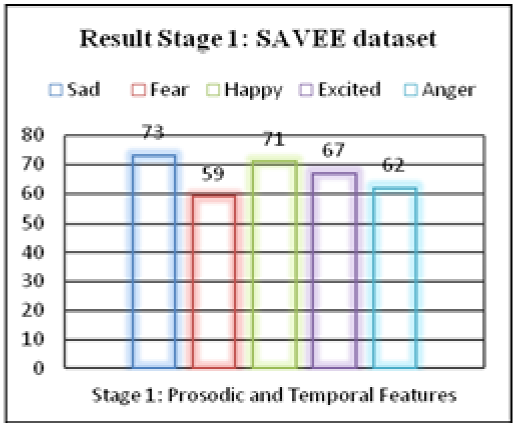


Fig. 4. Result Stage-1, SAVEE dataset

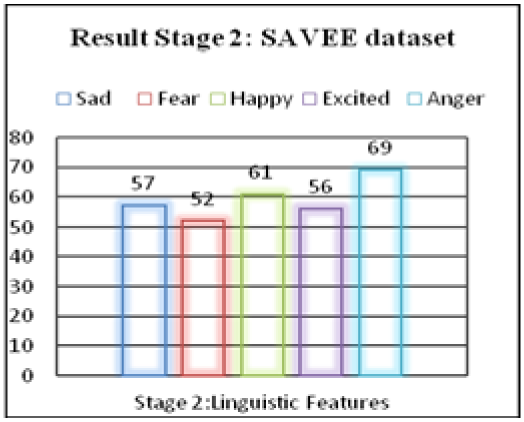


Fig. 5. Result Stage 2, SAVEE dataset

However, adults speak more fluently, and their speech has no clear pause between words and hence words segmentation is difficult for adults in SAVEE and POLISH datasets. Therefore, prosodic and temporal features extraction gives a better result for these datasets. Second reason for our first result because KG students are not able to express emotions more clearly with their voice. We noted that most KG students speak in the same volume in a happy and angry mood while they have the same tone in fear and in sad emotion. However, they (KG students) speak slower than adults in case of SAVEE and POLISH datasets and hence, word segmentation with KG students' speech gives more accurate results as compared to young people's speech. After concluding reasons for the fact that stage 1 is displaying an overall better performance than the outcome of stage-2, we focus on the overall results presented in Figures 10 - 12 where we present combination of both stages and extract emotion based on combine features. Although we could not segment all words of a speaker but with words that we could extract we used their semantic plus prosodic and temporal features of speech and then concluded emotion of speaker. Here we clearly find out an overall gain in almost each case.

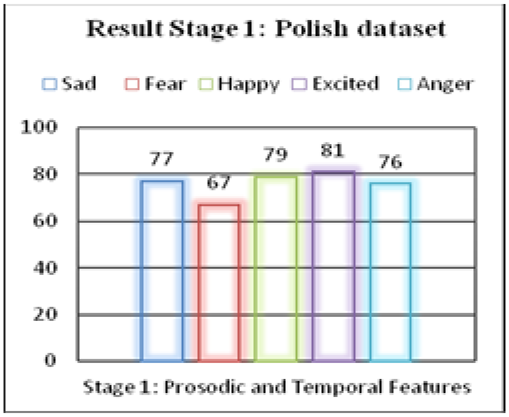


Fig. 6. Result stage 1 with Polish dataset

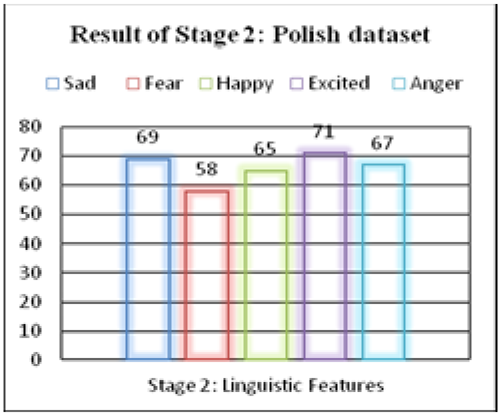


Fig. 7. Result stage 2 with Polish dataset

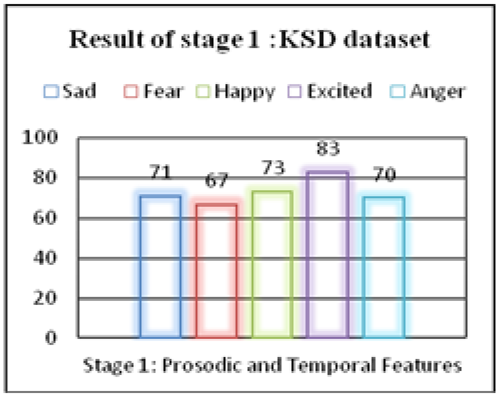


Fig. 8. Result stage-1 with KSD dataset

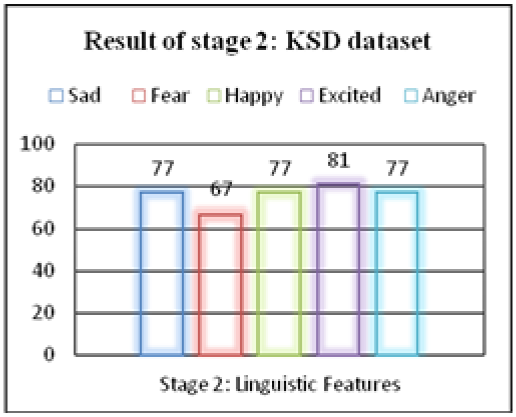


Fig. 9. Result stage-2 with KSD dataset

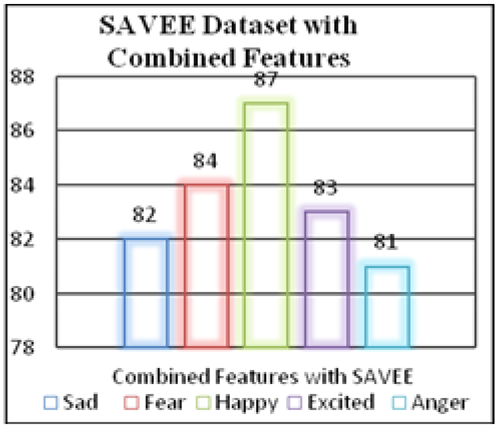


Fig. 10. Result of SAVEE dataset with combined features

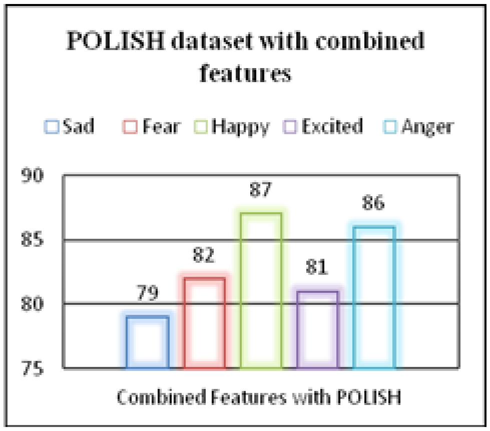


Fig. 11. Result of POLISH dataset with combined features

We also report overall complexity of the algorithm, and we calculate the overall time taken in case of running Stage-1, Stage-2 and both the stages combined. We find out that the time taken for all cases remains minutely less than the total time taken for each stage separately, and none of the cases is posing any serious constraints neither the data nor computation time is expanding drastically. This is mainly because we have performed same steps twice extracted - once during stage - 1 where we perform segmentation to divide speech signals into small segments with the help of ZCR and calculated spectral feature (MFCC) of speech segments to get information's of human's emotion and second time in stage - 2 when we performed segmentation using ZCR to segment words uttered by speaker and extracted MFCC to recognize words.

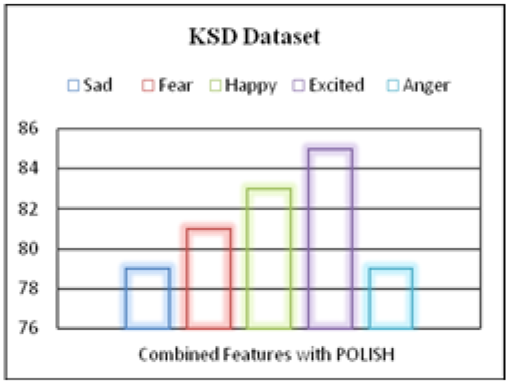


Fig. 12. Result of KSD dataset with combined features

VII. Related Work

Researcher have implemented various models for recognition of emotions from speech using different sets of features. Yildrim [22] conducts a study aimed to analyze how speech is modulated with the change in speaker emotional state. They measure vowel articulation, spectral energy and acoustic parameters of speech to find acoustic similarity and difference between different states of emotions. They also perform discriminant analysis to check the acoustic separability of emotion at utterance level. The author conclude that some emotions have similar acoustic properties and they mentioned happiness/anger and sadness/neutral in this category. Author of [3] present speech signal driven approach to deal with two class similarity problem. They use HMM based features with combination of prosodic and spectral features and achieved accuracy of 47.83% with two class similarity problem [3]. Alexander [6] present a new method to identify emotions using parameters of glottal airflow signal. The effectiveness of their system is tested with Artificial Neural Network (ANN), SVM, Hidden Markov Model (HMM), K Nearest neighbour (KNN), Bayesian classifier, Gaussian Mixture Model (GMM), decision tree and a new optimum path classifier (OPF). They find best results of glottal features with SVM and OPF. MFCC are extracted from data for classification of emotion [6], [9]. Spectral and prosodic features are identified in[13] and experiment are conducted after implementing three classifiers to identify emotions. The authors build Gaussian Mixture Model with prosodic features, SVM with prosodic features and Gaussian Mixture Model with Prosodic features. The authors calculate features and then apply algorithm to select most relevant features and show that prosodic features are more helpful in detection of emotion than spectral features. Another framework in which they extracted 133 speech features and aimed to identify feature set that would be appropriate to descriminate between seven emotions based on speech processing [7]. They use Neural network classifier with 35 input vectors and tested their model using Berlin dataset that include speaker dependant and speaker independent instances.

Zhu [25] identified emotions deficiencies in an E-Learning environment and proposed emotion recognition system for them. They build speech corpus from various subjects belongs to different languages and extracted prosodic features. They use Sequential Forward Selection (SFS) approach to select most appropriate feature set and used to classify emotions with General Regression Neural Network (GRNN). Considering emotion recognition a challenging task, Zhang performs an experiment with SVM is used for classification of 4 emotions and feature set of prosodic and speech quality features. He proposes that combination of prosodic and speech quality features increase 10% recognition rate [24]. The researchers in [14] extract quality Quality and prosodic features that overlap and complement each other in identification of emotions. Wang [20] combines these two stages and used optimal searching properties of Genetic algorithm considering personal characters of emotions. They achieve 86% recognition accuracy with this searching algorithm [20]. The author of [2] extend their research of 2009 and extract spectral features at stressed vowels, unstressed vowels and at consonant. They find that these levels contain more rich information's about emotions than utterance level. Their results show higher accuracy at segment level than at utterance level [2].

We present a new model where we extract emotions at two stages and then combine both stages result. Our results are 83.4 % with SAVEE 83% with polish 68% with KSD dataset. The results are better in the sense that when we have speech that has no clear words our stage1 with prosodic and temporal features help in identification of emotions. In case these features don't help to identify emotion correctly, words semantics help in identifying emotions of speaker. So this hybrid system of prosodic, spectral and words semantic features perform better than other system.

VIII. Conclusion

We have considered prosodic, temporal as well as linguistic features which has been helpful in increasing accuracy of emotion recognition. For the purpose of elaborating our results, we have used a two-staged model for emotion recognition. We have used the results by each stage separately and in a combined manner to present that our approach works better and produced significant results. We conclude from our results that considering prosodic as well as linguistic features together helps improving overall results without degrading performance. We propose algorithm where we extract emotions based on prosodic and temporal features in the first step and we extract emotions from word segmentation combined with linguistic features in the second step. While performing our experiments, we prove that the classification mechanisms, if trained without considering age factor, do not help improving accuracy. We provide our argument that the classifier should be based on the age group on which the actual emotion extraction be required and this becomes our second contribution submitted in this paper

References

[1] Baenziger, T. a. (2005). The role of intonation in emotional expressions. Speech communication, 3.

[2] Bitouk, D. a. (2012). Class-level spectral features for emotion recognition. Speech communication.

[3] Bozkurt, E. a. (2009). Improving automatic emotion recognition from speech signals. In NTERSPEECH (pp. 324-327).

[4] Dellaert, F. a. (1996). Recognizing emotion in speech. In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on (pp. 1970-1973).

[5] Hoque, M. E. (2006). Robust recognition of emotion from speech. In Intelligent Virtual Agents (pp. 42-53). Springer.

[6] Iliev, A. I. (2010). Spoken emotion recognition through optimum-path forest classification using glottal features. Computer Speech & Language, 3.

[7] Iliou, T. a.-N. (2009). Statistical evaluation of speech features for emotion recognition. In Digital Telecommunications, 2009. ICDT'09. Fourth International Conference on (pp. 121-126).

[8] K Sreenivasa and Koolagudi, S. G. (2012). Emotion Recognition Using Speech Features. Springer Science & Business Media.

[9] Kao, Y.-h. a.-s. (2006). Feature analysis for emotion recognition from Mandarin speech considering. In InterSpeech.

[10] Kim, D.-S. a.-Y. (1999). Auditory processing of speech signals for robust speech recognition in real-world noisy environments. Speech and Audio Processing, IEEE Transactions on, 55-69.

[11] Koolagudi, S. G. (2012). Emotion recognition from speech: a review. International journal of speech technology, 99-117.

[12] Lopez-Poveda, E. A. (2001). A human nonlinear cochlear filterbank. The Journal of the Acoustical Society of America, 3107-3118.

[13] Luengo, I. a. (2005). Automatic emotion recognition using prosodic parameters. In Interspeech (pp. 493-496).

[14] Lugger, M. a. (2007). The relevance of voice quality features in speaker independent emotion . In Acoustics, Speech and Signal Processing, 2007. (pp. 13-17).

[15] Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. In Proceedings of Artificial Neural Networks in Engineering (pp. 7-10).

[16] Rabiner, L. R. (1975). An algorithm for determining the endpoints of isolated utterances. Bell System Technical Journal, 297-315.

[17] Rabiner, L. R. (1975). Theory and application of digital signal processing. Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. .

[18] Shen, J.-l. a.-w.-s. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. In ICSLP (pp. 232-235).

[19] Ververidis, D. a. (2006). Emotional speech recognition: Resources, features, and methods. Speech communication, 1162-1181.

[20] Wang, Y. a. (2008). Adaptive and optimal classification of speech emotion recognition. In Natural Computation, 2008. ICNC'08. Fourth International Conference on (pp. 407-411).

[21] Wu, S. a.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. Speech Communication, 768-785.

[22] Yildirim, S. a. (2004). An acoustic study of emotions expressed in speech. In INTERSPEECH.

[23] Yu, F. a.-Y. (2001). Emotion detection from speech to enrich multimedia content. In Proceedings of the second IEEE pacific rim conference on multimedia: (pp. 550-557).

[24] Zhang, S. (2008). Emotion recognition in Chinese natural speech by combining prosody and voice quality features. In Advances in Neural Networks-ISNN 2008 (pp. 457-464).

[25] Zhu, A. a. (2007). Study on speech emotion recognition system in E-learning. In Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments (pp. 544-55