Cognitive Approach for Behavioral Analysis of Speech Data

[1]Dheeraj D, [2]Sushmitha K, [3]Sahana A.G,[4]Yashaswini R
[1]Assistant Professor, [2,3,4] UG Students, Dept. of Information Science &Engineering,
Global Academy of Technology, Bangalore, India.

Abstract : Recognition of emotion has been an important research area which may reveal some valuable input for different purposes. People expresses their emotions indirectly or directly through their speech, gestures, facial expressions or writings. Many different sources of information, such as speech, text and visual can be used to analyse emotions. Currently, writings take many forms of social media posts, micro-blogs, news articles, etc., and the content of these posts can be useful resource for text mining to discover and show various aspects, including emotions. Extracting emotions behind these posts is a complicated task. To solve this problem, researchers from different fields are trying to find a sufficient and good way to more precisely detect emotions of human from various sources which includes text and spoken words. In this sense, different techniques based on word and sentences, machine learning, natural language processing methods, etc., have been used to achieve better accuracy.

IndexTerms **– Emotion Recognition, Machine learning, Feature Extraction, Text Data Set**

I. INTRODUCTION

One of the most fundamental aspects of human behaviour is the ability to communicate. Humans may communicate with each other (speak and write) using natural languages. While speech refers to the vocalized form of human communication text represents the written format of human communication. Even though, the communication may be possible by natural language text or speech, the vocalized form of human communication or speech is considered to be the most convenient form of human communication. This leads to the development of speech recognition system [12]. However, understanding the meaning of human speech by a machine is relatively a complex task and is an active area of research.

Recent research makes more emphasis on the recognition of verbal and nonverbal information, especially on the topic of emotion detection. Scientists came to know that emotional skills can be an important component of intelligence, especially for human to human communication. In the past years, there are several works on emotion detection. Some research focused on multimodal emotion recognition and other research focused on only one kind of information. However, very few papers focused on emotion detection from text input. In this paper, an emotion recognition system is proposed to classify five basic emotions, including happy, sad, anger, hate and neutral. The proposed emotion recognition system is given an input in form of speech data where the speech is further converted into text and that text will undergo text mining to detect emotion present in input speech given minimum, median, and standard deviation of pitch, energy and Pitch vibration is not considered here instead words are given more prominence to classify the emotion.

For emotion recognition using text input module, we assume that the reaction of an input speech given is essentially represented by its word appearance in sentence. Two basic word types “emotional keywords” and “emotion modification words” are manually defined and used to extract emotions from the input. All of the extracted keywords and emotion modification words have their corresponding “intensity values” and “modification values.” which are defined manually. For each input sentence, the emotion intensity values are averaged and triggered by the modification values to give the current emotion output.

II. PROBLEM STATEMENT

Emotion recognition from the speaker’s speech is very difficult because of the following reasons: In differentiating between different emotions which specific speech features are more useful is not clear. Because of the existence of different sentences, speakers, style of speaking, speaking rates variability was introduced, because of which speech features get directly affected. The same utterance may show different emotions. Each emotion may correspond to the different portions of the utterance. Therefore it is very difficult to differentiate these portions of utterance. Another problem is that expressing the emotion is depending on the speaker and his/her environment and culture. As the environment and culture gets change the speaking style also gets change, which is another issue in front of the speech emotion detection system. There may be one or more types of emotions, long term emotion and transient one, so it is not specific which type of emotion the recognizer will recognize [2].

In the stream of human-computer interaction (HCI), emotion detection from the computer is still a challenging problem, especially when the detection is based solely on speech, which is the basic medium of human communication. In human-computer interaction systems, emotion detection could provide users with services that are personally improvised by being adaptive to their emotions. Therefore, detecting emotion from speech could have more potential applications in order to make the computer more adaptive to the user’s needs.

The work carried out is that we take speech data from standard datasets. The speech is converted into text and based on the text, the emotions are classified. It classifies emotions as happy, sad, hate, anger, neutral. When the statement undergoes 2 kinds of emotions it results in neutral statement.

III. PRE-PROCESSING

Speech can be defined as the expression of one’s thoughts and feelings through articulation of sounds and it is the most natural, and preferred means of communication among humans. But processing the speech data by a system to extract useful information is a bit complicated task as it involves analyzing intricate variations in speech in the form of language spoken, accent,

dialect which differ from person to person. In order to analyze the speech data and perform required operations on them, we need to first process the data to convert it to a suitable format. Here, we perform speech processing by using the features present in the speech and transforming it to text.

Speech creates specific vibrations in the air in the form of analog signals. Since the system cannot process analog signals directly, it is first converted into digital signals. To perform this conversion, the given analog signal is sampled by taking precise measurements of the signal at frequent intervals. This step is required as the temporal characteristics of speech signal vary with time, which makes it difficult to process. We can assume that the speech signal will be constant between shorter intervals, thereby making it easier to process. Now, the sampled signal is filtered using suitable techniques to remove unwanted noise, and is separated into different levels of frequency to measure the difference in the speed of different sound samples. Next, this signal is divided into small segments known as phonemes. The program then matches with these segments to known phonemes in the appropriate language (English, here) to make meaningful representation of sentences. Next, to figure out the information hidden in the speech signal, Hidden Markov model is used, where each phoneme acts like a link in the chain, and a completed chain forms a word. During this process, each phoneme is assigned a probability score, based on the training performed on the phonemes. The phoneme data is trained on lots of human-transcribed speech data to create acoustic model of words. Based on this model, the system makes a prediction of the sentence by considering the phonemes with higher probability score and combining them.

IV. METHODOLOGY

We consider the following three performance metrics to compare the four routing protocols.

In this paper, we propose a novel approach for Behavioral Analysis as shown in Figure 1 using speech as the main input and predicting the type of emotion associated with the given input.

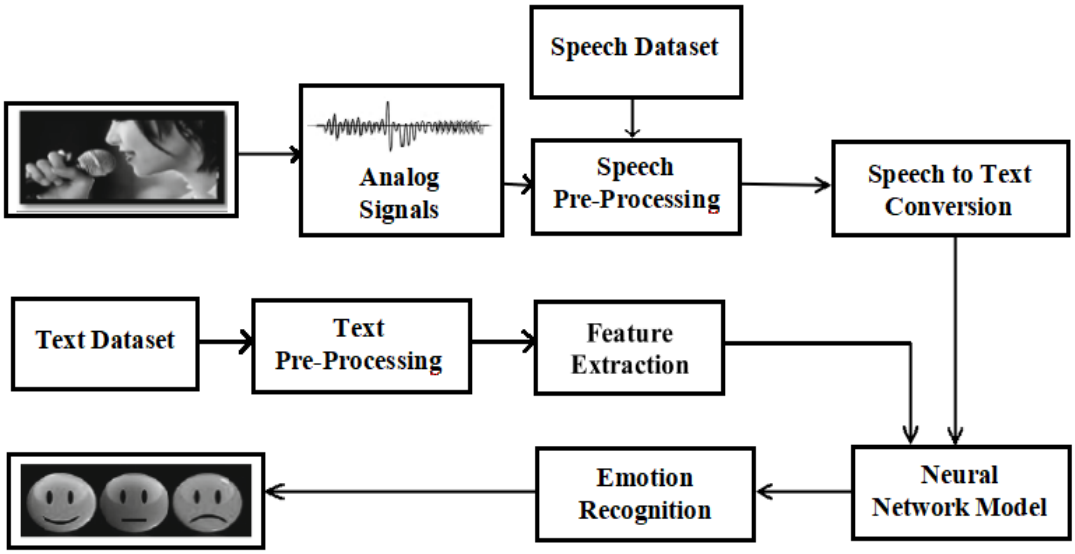


Fig 1 : Proposed Methodology

We first process the speech input, and convert it to text by making use of the intrinsic features present in the speech input. The resultant text is pre-processed before building a Model. The pre-processing steps is categorized into 2 steps :

Step 1 : Data Pre-Processing

This step involves cleaning the data before proceeding to Feature Extraction.

Here, we apply Natural Language Processing techniques to remove noise/unwanted characters from the text dataset. Some of the techniques include:

- Tokenization - Here, we convert individual sentences into words.
- Punctuation Removal - Here, we remove unwanted punctuations, tab spaces, and other unwanted characters from the text dataset.
- Stopwords Removal - Here, we remove some specific words that do not contain a specific semantic. Eg: Words such as ”the”, ”is”, etc. that occur frequently in sentence.
- Lemmatization - Here, we make use of English Vocabulary to determine the part of speech and remove unwanted words.

Step 2 : Feature Extraction

In this step, we transform the text to real valued vectors that can be used by the algorithms for building a Model. We use Word Embedding technique for Feature Extraction.

Word Embedding : It is a text representation where words having the same meaning have similar representation. In other words, it can be defined as the representation of words in a coordinate system where closely-related words, based on a corpus of relationships, are placed together. Some of the popular Word embeddings are :

- Word2Vec embedding

Word2vec takes a large corpus of text as input and produces a vector space where each corresponding vector in the space has a unique word being assigned. Word vectors will be positioned in the vector space such that words that share similar contexts in the text corpus are located in closer to each other in the space. Word2Vec is quite popular for capturing the meaning of text and demonstrating it on tasks like calculating analogy questions. Eg: Something like a is to b as c is to?

- GloVe embedding

Glove stands for "The Global Vectors for Word Representation". It is can be viewed as an extension to the Word2vec embedding method for learning word vectors efficiently. GloVe explicitly constructs a word-context or word co-occurrence matrix across the whole text corpus by making use of statistics. This results in a learning model that may result in a much better word embedding.

We have used GloVe embedding in our work, due to its efficiency over the Word2Vec model, and the availability of a large quantity of pre-trained word vectors in different dimensions, which provide a better word embedding.

Now, after the text pre-processing is completed, we build a neural network model for text classification. The embedded text data is fed into the neural network, with multiple dense layers, for training. The trained model is then saved and is used for emotion recognition. The model is then run on the text sentence that is obtained as output from speech to text conversion. The model should be able to predict the emotion underneath the text sentence successfully.

DATASET

Our work makes use of 2 datasets namely : RAVDESS dataset for Speech input, and Twitter Dataset containing different classes of Emotions. The description of each of the dataset is as follows:

- RAVDESS dataset - RAVDESS (The Ryerson-Audio-Visual Database of Emotional Speech and Songs) is an open-source database that consists of 24 professional actors (out of which 12 are male actors ,and 12 are female actors), who vocalize 2 lexically matched statements in a Neutral North American accent.The RAVDESS database includes speech consisting of different emotions, namely calm, sad, happy, angry, fear, surprise and disgust expressions. It contains 7356 files in overall, and each file is rated 10 times on emotional validity, intensity, and genuineness by 247 research participants.

- Twitter Dataset - This is the dataset from Twitter public API, wherein the category of emotion was labelled at the end of each tweet. This dataset comprises of about 40,000 tweets that are classified into 13 emotion classes. But, most the emotion classes were highly similar, hence we combined many of those classes into 5 classes. These 5 classes are Happy, Sad, Angry, Hate, Neutral.

This dataset was preferred since the text is short, and informal, and the tweets covers a wide range of data, with different emotions, thereby justifying our work.

V. IMPLEMENTATION

- Convolutional Neural Network (CNN)

CNN is a class of deep, feed-forward artificial neural networks and use a variation of multilayer perceptron's designed to require minimal pre-processing. They are basically just several layers of convolutions with nonlinear activation functions like ReLU applied to the results. Instead of the traditional neural networks, CNN use convolutions over the input layer to compute the output. This results in local connections, where each region of the input is connected to a neuron in the output. Each layer applies different filters, typically hundreds or thousands, and combines their results.

- Long short-term memory (LSTM)

In LSTM there are several architectures of units. A common architecture is composed of a cell and three "regulators" usually called gates of the flow of information inside the LSTM unit: an input gate, an output gate and a forget gate. Some variations of the LSTM unit do not have one or more of these gates or maybe have other gates.

Intuitively, the cell is responsible for keeping track of the dependencies between the elements in the input sequence. The input gate controls the extent to which a new value flows into the cell, the forget gate controls the extent to which a value remains in the cell and the output gate controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit. The activation function of the LSTM gates is often the logistic function.

There are connections into and out of the LSTM gates, a few of which are recurrent. The weights of these connections, which need to be learned during training.

- C-LSTM network

The Convolutional Long Short-Term Memory Network is an LSTM architecture specifically designed for sequence prediction problems with spatial inputs. It is a combination of both CNN and LSTM. The architecture involves using Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs to support sequence prediction.

- ANN

ANN is the modeling of the human brain with the simplest definition and building blocks are neurons. There are about 100 billion neurons in the human brain. Each neuron has a connection point between 1,000 and 100,000. In the human brain, information is stored in such a way as to be distributed, and we can extract more than one piece of this information when necessary from our memory in parallel. We are not mistaken when we say that a human brain is made up of thousands of very, very powerful parallel processors.

In multi-layer artificial neural networks, there are also neurons placed in a similar manner to the human brain. Each neuron is connected to other neurons with certain coefficients. During training, information is distributed to these connection points so that the network is learned.

Disadvantages : Hardware dependence, Unexplained behaviour of the network, Determination of proper network structure, The duration of the network is unknown.

- RNN

Recurrent Neural Network are a type of Neural Network where the output from previous step are fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence.

Disadvantages of Recurrent Neural Network

1. Gradient vanishing and exploding problems.
2. Training an RNN is a very difficult task.
3. It cannot process very long sequences if using tanh or relu as an activation function.

LIMITATIONS

Choosing a machine learning algorithm is a complicated task and the choice can easily change depending on the problem and the structure of the data that we are working with.

Computational complexity is also an important point that has to be considered when we want to compare the resources that will be needed for a given algorithm. Furthermore, the number of dimensions of the dataset must be taken in account.

In our case, most of the data will be categorical. We won't have a huge amount of data and the model has to be able to constantly evolve and improve itself.

APPLICATION

Analyzing emotions can be helpful in many different domains. One such domain is human computer interaction. With the help of emotion recognition, computers can make better decisions to help users. With the increase in popularity of robotic research, emotion recognition will also help making human–robot interaction more natural.

The possible applications of speech recognition in HCI may include voice user interfaces such as voice dialing, data entry, preparation of structured documents, speech-to-text processing, and aircraft. The ASR technology may be useful for learning different languages by listening to the proper pronunciation, in addition to helping a person develop fluency with their speaking skills [1]. Students who are physically disabled or suffer from repetitive strain injury/other injuries to the upper extremities can be relieved from having to worry about handwriting, typing by using speech-to-text programs. They can also utilize speech recognition technology to search the Internet or use a computer at home without physically operating a mouse or keyboard. Speech recognition can allow students with learning disabilities to become better writers without concerning about spelling and other mechanics of writing.

The following pieces of software have been developed: Emotion Recognition Game (ERG), Emotion Recognition software for call centres (ER), and a dialog emotion recognition program (Speak Softly). The first program has been mostly developed to demonstrate the results of the above research. The second software system is a full-fledge prototype of an industrial solution for computerized call centres. The third program, which just adds a different user interface to the core of the ER system, demonstrates the real time emotion recognition.

i. Emotion Recognition Game : The program allows a user to compete against the computer or another person to see who can better recognize emotion in recorded speech. The program serves mostly as a demonstration of the computer's ability to recognize emotions, but one potential practical application of the game is to help autistic people in developing better emotional skills at recognizing emotion in speech.

ii. Emotion Recognition Software for Call Centres :

- Goal : The goal of the development of this software was to create an emotion recognition agent that can process telephone quality voice messages (8 kHz/8 bit) in real-time and can be used as a part of a decision support system for prioritizing voice messages and assigning a proper human agent to respond the message.
- Agent : It was not a surprise that anger was identified as the most important emotion for call centres. Considering the importance of anger and scarcity of data for some other emotions we decided to create an agent that can distinguish between two states: "agitation" which includes anger, happiness and fear, and "calm" which includes normal state and sadness. To create the agent, we used a corpus of 56 telephone messages of varying length (from 15 to 90 seconds) expressing mostly normal and angry emotions that were recorded by eighteen nonprofessional actors. These utterances were automatically split into 1-3 second chunks, which were then evaluated and labeled by people. They were used for creating recognizers using the methodology described above.
- System Structure : The ER system is a part of a new generation computerized call centre that integrates databases, decision support systems, and different media such as voice messages, email messages and a WWW server into one

information space. The system consists of three processes: the wave file monitor agent, the message prioritize agent, and the voice mail centre. The wave file monitor reads every 10 seconds the contents of voice message directory, compares it to the list of processed messages, and, if a new message is detected, it calls the emotion recognition agent that processes the message and creates emotion content files, which describe the distribution of emotions in the message. The prioritize is an agent that reads the emotion content files, sorts messages considering their emotional content, length and some other criteria, and suggests an assignment of a human agent to return the calls. Finally, it generates a web page, which lists all current assignments. The voice mail centre is an additional tool that helps operators to visualize emotional content of voice messages; sort them by name, date and time, length, and emotional content; and playback the whole message or a part of it.

COMPARISION

Initially CNN algorithm is used for classification of input given where cnn is like neural networks that are made up of neurons with learnable weights and biases. Each neuron receives many inputs takes a summation off weights over them, pass it through a function and responds with an output. But because of its drawbacks: high computational cost, If one don't have a good GPU they are quite slow to train (for complex tasks), since they use a lot of training data, there may be memory losses.

So, LSTM was chosen over CNN for its advantages: LSTM has the capability of bridging long time lags between inputs. In other words, it is able to remember inputs from up to 1000 time steps in the past. This capability makes LSTM a better for learning long sequences with long time lags.

However, LSTM has disadvantage: it is slower than other activation functions, such as sigmoid, tan (h) or rectified linear unit then comes a problem which Lstm to select Uni-directional or Bi-directional but because of some of valid reasons bi-directional lstm stands high which is explained as follows:
Comparing Uni-directional with Bi-directional

Unidirectional LSTM only preserves information of the past because the only inputs it has seen are from the past.

Bidirectional will run your inputs in two ways, one from future to past and another from past to future and what differs this approach from unidirectional is that in the LSTM that runs in backwards direction you preserve information from the future and using the two hidden states that are combined you are able in any point in time to preserve information from both directions.

What they are suited for is a very complicated question but Bi-LSTMs show very good output as they can understand better context, I will try to explain through an example. Let's say we try to predict the next word in a sentence, on a high level what a unidirectional LSTM will see is:

The boys went to…

And will try to predict the next word only by this context of sentence, bi-lstm information further down the road for example

Forward LSTM: The boys went to…

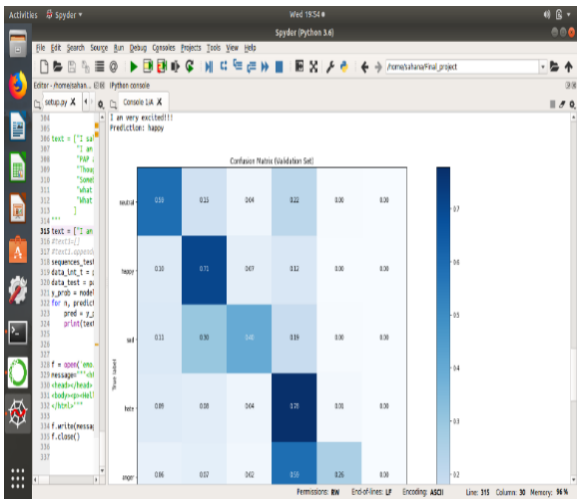
Backward LSTM: …and then they got out of the pool

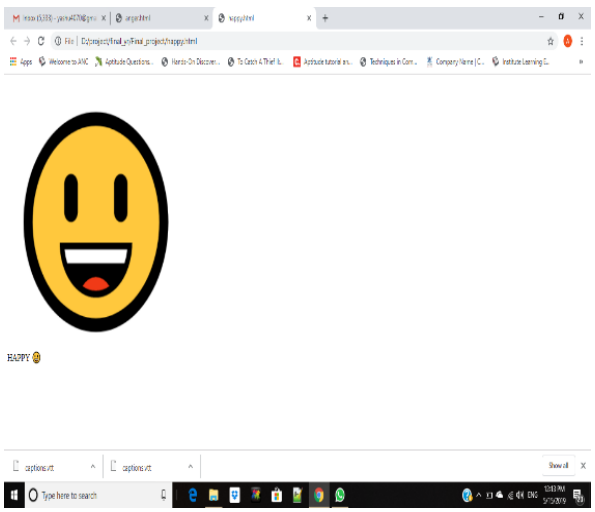
You can see that using the information from the future it could be easier for the network to understand what the next word is. So bi-lstm is better then uni-lstm.

CNN with LSTM indeed has achieved impressive success on a wide range of sequence modelling task.

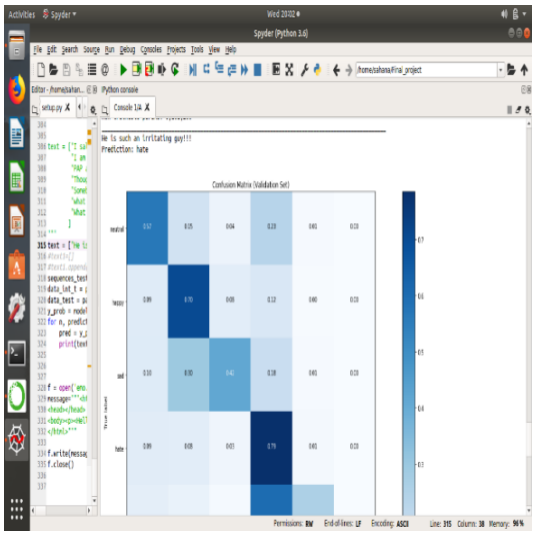
SNAPSHOTS

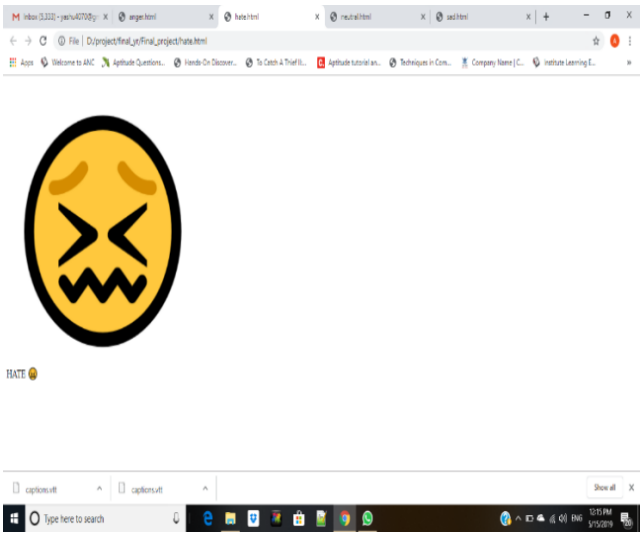
- For emotion classified as Happy :





➢ For emotion classified as Hate :





VI. CONCLUSION

In this paper, an emotion recognition system with speech input is proposed. The emotional state of speech input is converted into text and fed into the textual emotion recognition module made up of CNN-LSTM (bi-directional). The emotion recognition of converted text information is done based on the pre-defined values for emotion modification and emotion descriptors. The final emotional state obtained is further smoothed by the previous emotion history. The final experimental result shows that the strategy provides a promising result for emotion recognition. Automatic emotion recognition from human speech are increasing now a day because it results in the better interactions between human and machine.

VII. FUTURE ENHANCEMENT

There are a number of ways that this project could be extended. Perhaps one of the most common tool in emotion detection system is the neural networks and Hidden Markov model for a automatic, recognition. Also the integration of other modalities such as video based or manual interaction will be investigated further.

VIII. ACKNOWLEDGMENT

With great regards, gratitude and reverence to my Guide Dr. S Vagdevi, Prof & Head, Dept. of EEE, GSSSIETW, Mysore for providing all the support and guidance. I wove great respect and gratitude to all my family members for the blessings and encouragement showered on me.

REFERENCES

[1] P. Sanderson, “Cognitive work analysis and the analysis, design, and evaluation of human-computer interactive systems, in proc: Computer Human Interaction, pp.220-22, 1998.
[2] M. E. Ayadi, M. S. Kamel, F. Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases”, Pattern Recognition 44, PP.572-587, 2011.
[3] I. Chiriacescu, “Automatic Emotion Analysis Based On Speech”, M.Sc. THESIS Delft University of Technology, 2009.
[4] T. Vogt, E. Andre and J. Wagner, “Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization”, LNCS 4868, PP.75-91, 2008.
[5] S. Emerich, E. Lupu, A. Apatean, “Emotions Recognitions by Speech and Facial Expressions Analysis”, 17th European Signal Processing Conference, 2009.
[6] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, “Speech Emotion Recognition Using Hidden Markov Model”, Eurospeech, 2001.
[7] P.Shen, Z. Changjun, X. Chen, “Automatic Speech Emotion Recognition Using Support Vector Machine”, International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.
[8] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features and Methods", Elsevier Speech communication, vol. 48, no. 9, pp. 1162-1181, September, 2006.
[9] Z. Ciota, “Feature Extraction of Spoken Dialogs for Emotion Detection”, ICSP, 2006.
[10] E. Bozkurt, E, Erzin, C. E. Erdem, A. Tanju Erdem, “Formant Position Based Weighted Spectral Features for Emotion Recognition”, Science Direct Speech Communication, 2011.
[11] C. M. Lee, S. S. Narayanan, “Towards detecting emotions in spoken dialogs”, IEEE transactions on speech and audio processing, Vol. 13, No. 2, March 2005.
[12] T. Sakai, S. Doshita, “The Automatic Speech Recognition System for Conversational Sound”, IEEE Transactions on Electronic Computers, Vol.12, No.6, 1963.