International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Machine Learning for Classification of Emotion in Speech

Vivek Sharma, Chandrabhanu Mishra, Saroj Meher, Santosh Mishra

Abstract: ***The naturalness of the speech in any human being comes from his or her emotions. All human beings deliver and construe the messages with heavy use of emotions. So there is a need to develop a speech interface through which emotions embedded in the speech signal can be analyzed and processed. There are many speech translation systems developed with intent to interpret the inherent emotions in the speech signals but lack in processing the embedded emotions in the speech as because there is a lacuna in their modeling and depiction. The main objective of any speech processing system is to retrieve interesting information from speech like features, models so that the retrieved knowledge of interest can be further used in various speech processing applications. The scope of the present paper is to travel around the attributes of speech and its respective models with a goal to distinguish emotions by imprisoning precise information about emotion. This paper also studied various sources like source of excitement, vocal track system's silhouettes and its sequence, attributes of supra- segment to obtain a rich source of emotional information of a speech. The paper end with a final conclusion saying that source of excitation and its characteristics may be be single handedly enough for efficient acknowledgment of emotions.***

Keywords: Linear prediction (LP), Glottal Volume Velocity (GVV), Glottal Closure Instants (GCI)

I. INTRODUCTION

There are various sources like written text and video demonstrating emotions in a particular speech. In case of a video, the face and its expression and in a written text the punctuation marks can be a rich source of non-linguistic information of emotional expressions. A speech in a spoken form can have various interpretations based on the way how it has been spoken. As an example the word OKAY in English language has different meanings like approbation incredulity, approval, lackadaisical attitude or an allegation. The semantics of a spoken speech cannot be well interpreted by comprehending the spoken text.

Manuscript published on January 30, 2020.
* Correspondence Author
Vivek Sharma*, Biju Patnaik University of Technology, Rourkela, Odisha, India.
Chandrabhanu Mishra, Associate Professor, Electronics & instrumentation engineering, College of Engineering and Technology, Bhubaneswar, Odisha, India.
Saroj Meher, Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore, India.
Santosh Mishra, Biju Patnaik University of Technology, Rourkela, Odisha, India.

So it is necessary to develop a speech processing system which can analyze the semantics along with emotions embedded in the message as non linguistic information.
Speech plays an important role as natural medium to establish the communication between human and machine. A speech processing system will be able to reach to the standard of a human being if it can analyze and interpret efficiently the embedded information in the textual speech.
The main objective of ant speech processing software should not be only to process the textual messages in the speech but also to retrieve and interpret the emotional values and embedded purpose of the orator. The identification, Recognition as well as be familiar with inherent and implicit emotions in a speech signal is a budding research area in recent times.

1.1 Emotion Oriented Acknowledgement of Speech

A spoken speech when said with its own inherent way of expressing the emotions makes it more natural. In order to develop any speech processing software intending to recognize a speech, a speaker or language need to have a thorough knowledge of emotions. Capturing of adequate knowledge about emotion is done with a meticulous review of literature with an aim of retrieval of emotional features from a speech. The sources from which various attributes of the speech can be retrieved are source of excitation, system of vocal tract and prosodic features.

II. RELATED WORK

Vocal Track system and source of excitation are the dual pledge of dialog in a dialog generation machine [3]. So in case of a dialog the retrieval data like message, language, speaker and feelings are qualitatively available in the sources of excitation as well as vocal track system [4]. The proximity of feelings in explicit data present in dual source of excitation and in framework of vocal track system [2] is measured by visceral examination to strengthen the debate. A dataset of five sentences in hindi language spoken by a female working as radio craftsman has been considered in different styles expressing the feelings of the sentence. The Three passionate arrangement expressions are taken under a directed listening examination. The data sets are arranged in three different groups. Group one has 40 exclusive (5 sentences ×8 feelings) dialog enunciations looked over a Hindi fervent database. The passionate conversation records of group one verbalizes to the combined data of source of excitation and framework of vocal tract. Group two articulates to only source of excitation (LP lingering) data of group one words. Group three is about enclosing the word of expressions being spoken to data of vocal tract [1].

The VT parameters (LP Coefficients) of words of Group one serves as the energizing factors for the wave
proof of Group three articulations. The VT scaffold yield including the vocal tract qualities got keyed up with uneven tumult as the domain of white arbitrary clamor is level.
A randomized arrangement if followed prior to the conductance of subjective listening examination.
The emotions inherent in the statements are set to be identified and recognized by a total of 25 research candidates from which 16 were male and 9 were female candidates with everyone's mother tongue happens to be Hindi. A group of 8 emotions were demonstrated to them and they were asked to distinguish the utterances being played before them sequentially [5]. Five minutes interval is provided prior to the examination and evaluation of the subsequent set and a classification of 120 word sounds (40 word sounds from each set) has to be identified by each subject. The table below represents the emotion identification result and performance evaluation for each three groups in the form of confusion matrices [7]. In the 8 x 8 confusion matrix the diagonal elements demonstrates the percentage of accurately identified word sounds. The percentage of subjective listening examination has an average emotion identification performance of 47%, 53% and 60% by means of *source*, *system* and *source + system* respectively. The remarkable concluding consequences show the occurrence of emotion-precise information in both source of excitation system of vocal tract.

Table 1 : Result of Subjective Listening Examination showing the Performance evaluation of Emotion classification. Files of Emotional speech are produced using (i) only source of excitation signal, (ii) system of vocal tract attributes (iii) both excitational source and vocal tract system attributes.

Emotions	**Performance of Emotion acknowledgment**							
	Annoyance	Abhorrence	Panic	Pleasure	Impartial	Depression	Derision	Shock
Annoyance	40	17	7	10	10	0	6	10
Abhorrence	13	43	0	0	20	7	10	7
Panic	0	0	53	13	7	13	7	7
Pleasure	10	0	17	37	17	6	0	13
Impartial	0	0	20	0	57	23	7	0
Depression	3	10	20	7	10	50	0	0
Derision	7	17	10	3	0	0	43	20
Shock	3	10	10	7	0	0	20	50
Consequences of Subjective listening merely with system attributes (Group Two)								
Annoyance	43	17	13	10	7	0	0	10
Abhorrence	20	47	0	0	17	10	6	7
Panic	0	0	63	10	7	10	0	10
Pleasure	3	0	17	50	20	3	0	7
Impartial	3	0	13	10	57	10	7	0
Depression	3	0	13	10	57	10	7	0
Derision	0	3	10	7	13	60	0	7
Shock	10	7	13	7	0	0	13	50
Consequences of Subjective listening merely with normal speech word sounds (Group Three)								
Annoyance	70	13	3	3	7	0	0	4
Abhorrence	20	43	7	10	7	0	6	7
Panic	0	17	37	7	6	17	6	10
Pleasure	3	3	7	67	13	0	0	7
Impartial	0	7	3	0	87	3	0	0
Depression	0	13	20	0	13	54	0	0
Derision	0	7	3	0	10	0	73	7
Shock	0	3	23	17	10	0	0	47

LP residual signs and GVV signs of the vowel segment /a/ extracted from word sounds justifies the potentiality of excitation around the glottal closure instants (GCI) are precise to each emotion and highlights the perception of occurrence of emotion-specific information in the source of excitation. Figure 1 clears that each diversified emotions shows a change in phase at the GC regions and variation in residual signal amplitude with immediate frequency (epoch intervals or pitch periods).

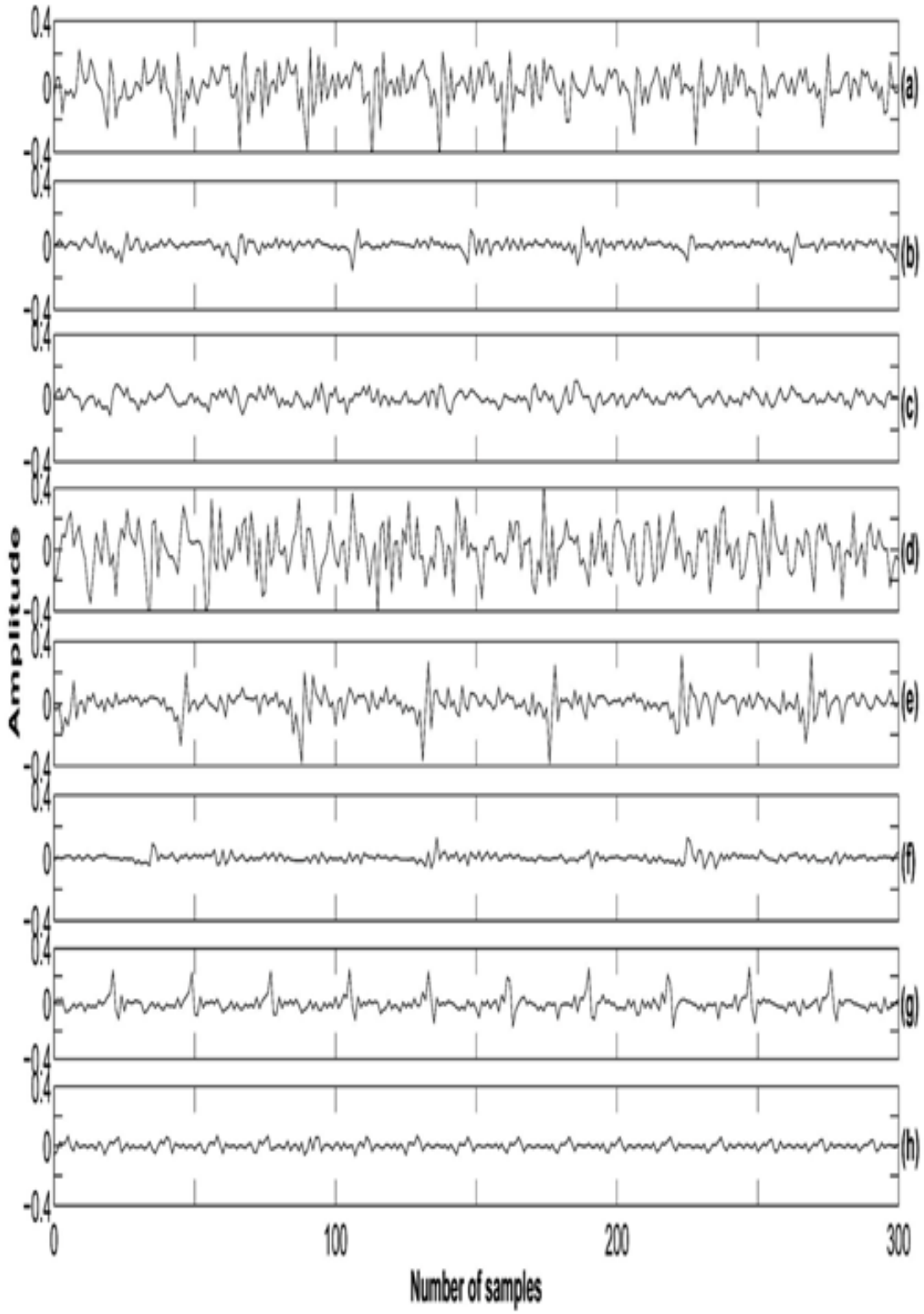

Figure 1: Variations of eight emotions found in LP residual signals with vowel segment /a/ for identified attributes of Table 1.

Figure 2 demonstrates the single exclusive glottal pulse shapes (durations of glottal opening and glottal closure, their ratio) disparity. As different scale of measurements are taken for plotting notable variations in amplitudes of GVV signal are penned down. A Research study motivation for carrying out the exploration of source of excitation for emotional speech acknowledgemnt is obtained from the plotting of LP residual and GVV signals as well as the outcomes of subjective listening examination narrated in Table 1.

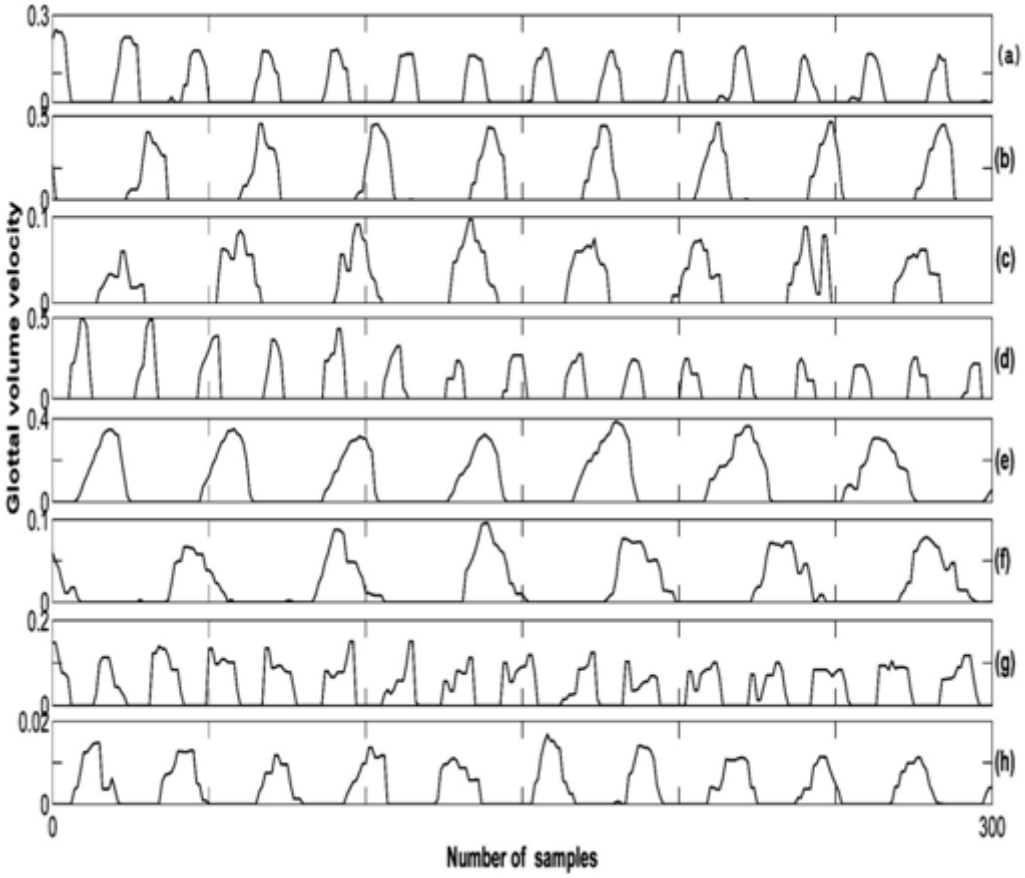

Figure 2 : Signs of Glottal Volume Velocity (GVV) for vowel segment /a/ , in eight different emotional attributes specified in Table 1

III. ACKNOWLEDGEABLE EMOTION ORIENTED ATTRIBUTES FROM SOURCE OF EXCITATION

A discourse sign making use of tenth request of LP investigation gives a lingering signal. Among various classifications of excitation, voiced excitation contains critical feeling explicit data, on the grounds that the comparing glottal vibration example might be distinctive for various feelings. The characteristic of overt feeling is acclaimed by the swiftness of glottis vibration, variations in GVV signals, excellence and raggedness of remarkable excitation at glottal finale. We have projected a relationship exist among the parameters like LP lingering signal, LP leftover stage, age, GVV signal. These parameters can be considered as main highlighted features for the recognition of feelings from dialogue. The explanation of these highlighted parameters is further extended in the next subsections.

3.1 Privileged Association between LP outstanding samples

Each exemplary data is considered as a linear weighted aggregation of past p sample data where p specifies the prediction order in a linear prediction analysis of speech [74]. Let $s(n)$ is the current sample, next for past p sample data the prediction can be stated as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k)$$

The error of prediction can be stated as the difference between the actual and predicted sample data and mathematically can be stated as

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k)$$

$a_k s$ is the coefficients of linear prediction and obtained from minimization of mean squared error.
Autocorrelation methods is adopted to resolve a group of p normal equations to obtain coefficients.

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \qquad n = 1,2 \ldots \ldots p$$

Where

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k) \qquad k = 0,1,2, \ldots \ldots n$$

$R(k)$ is described as autocorrelation function.
$(e(n))$ is presented as the outstanding signal and achieved when the sign of the speech is passed through inverse filter $A(z)$. The inverse filter can be given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$$

$|H(w)|$ represents the LP spectrum 2 and can be mathematically written as

$$|H(w)|^2 = \left|\frac{G}{1+\sum_{k=1}^{p} a_k e^{-jwk}}\right|^2$$

G in the above equation presents gain parameter obtained from the minimization of mean squared error.

An exemplary sample data of speech signal at 8 kHz is taken for which an LP order of 8-14 appears to be the appropriate sate for the derivation of LP residual. In our present work an exemplary sample data of speech signal at 8 kHz and LP order 10 is used to obtain LP residual. The source of excitation information is supposed to be present in the LP Residual. The associations of first and second order in the neighboring sample data are the result of analysis of LP investigation with autocorrelation coefficients whereas the investigation also shows the absence of associations with respect to shape and size of vocal tract. With the effect of above findings a justification is made for low correlation values for nonzero time lags for the autocorrelation function of the LP residual signal. So our study is further extended by the exploration of privileged association between the data samples of LP residual signal for potentiality of emotion-specific information. A nonlinear process is to be adopted for arresting the specific emotional information from privileged associations. As Neural Network has proved itself in detaining the potential nonlinear information from data samples, this research study has adopted the neural network based models to arrest the same. The potential emotion precise information is retrieved from residual data samples with the help of auto-associative neural network (AANN). The existence of privileged association between the data samples of LP residual are justified when the errors of trained AANN is plotted with the subsequent training of neural network model with the data sample of LP Residual including noise. The plotted Figure in Figure 3 clearly evidences exponential downfall of training error with the amplification of trained iterations. Such exponential downfall of error is an indication of privileged association and correspondence between the data samples. The noisy training data samples seem to be non declining; thus shows absence of association between noisy data samples.

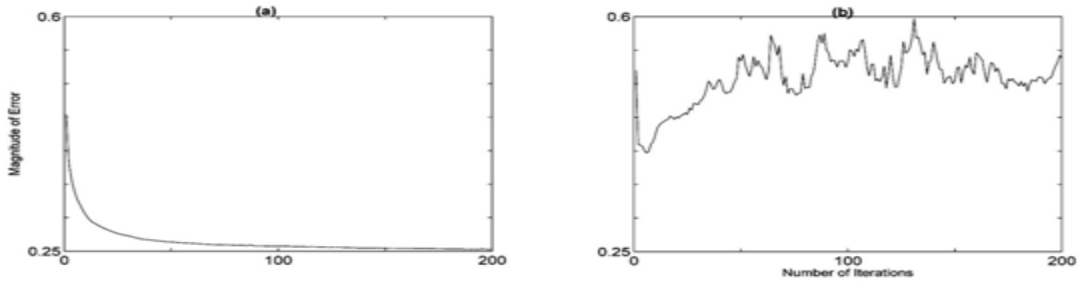

Figure 3 : Blueprints of Training error for AANN models. A) data sample of LP residual and B) data sample of LP residual with random noise.

IV. MODEL TAXONOMY

Both AANN and SVM models serve as the classification models to arrest precise emotional attribute information from the source of excitation attributes. SVM serves as the classification model on the basis of discriminative information present in the feature vectors where as AANN arrests the nonlinear associations between the feature vectors.

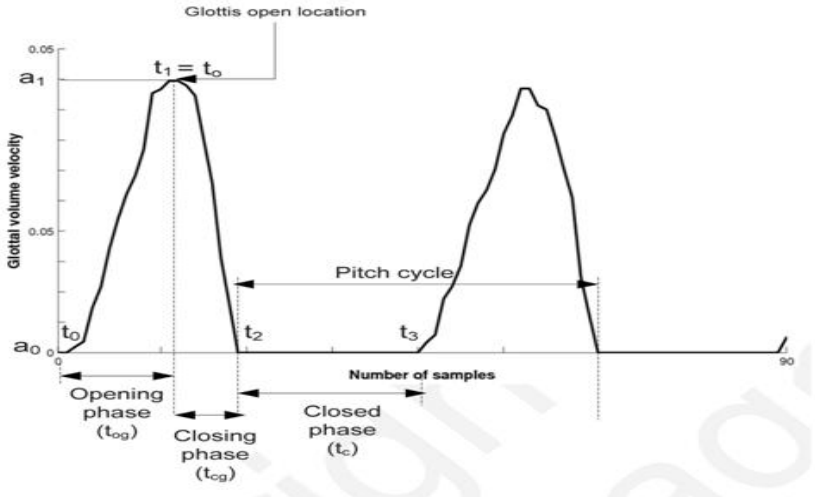

Figure 4: Glottal volume velocity signs demonstrating the two cycles

4.1 Auto-Associative neural networks

AANN models are the basic representation of feed-forward neural network (FFNN), that makes a mapping function of input vector onto itself so it is also called as auto-association or identity mapping [146, 147]. The AANN composed of an input layer, an output layer and one or more hidden layers.
The facets of input feature vector stay similar to the number of nodes in the input and output layers. The hidden layer nodes stay less in number in comparison to the nodes in either input or output layers. The alternative name of the hidden layer is also taken as dimension compression layer. The nature of the activation function is linear for the nodes of input and output layer where as the activation function shows either linearity or nonlinearity for the nodes of hidden layer. This research study has adopted a five layer AANN model as shown in Figure 5.

The structure of neural network model is expected to arrest the privileged association from the succession of LP residual samples i.e. 40L-60N-12N-60N-40L, where L represents linear nodalunits and N represents nonlinear nodal units.

The numerical value in Figure 5 points to the number of nodal units at hand in the respective layer. The size of feature vectors used to build up the model is designated by the presence of number of linear elements at the input layer. The *tanh*(*s*) as the activation function with *s* indicating the net input value of that unit is used by non linear nodal units. The structure of the network model is supposed to be settled on pragmatically.

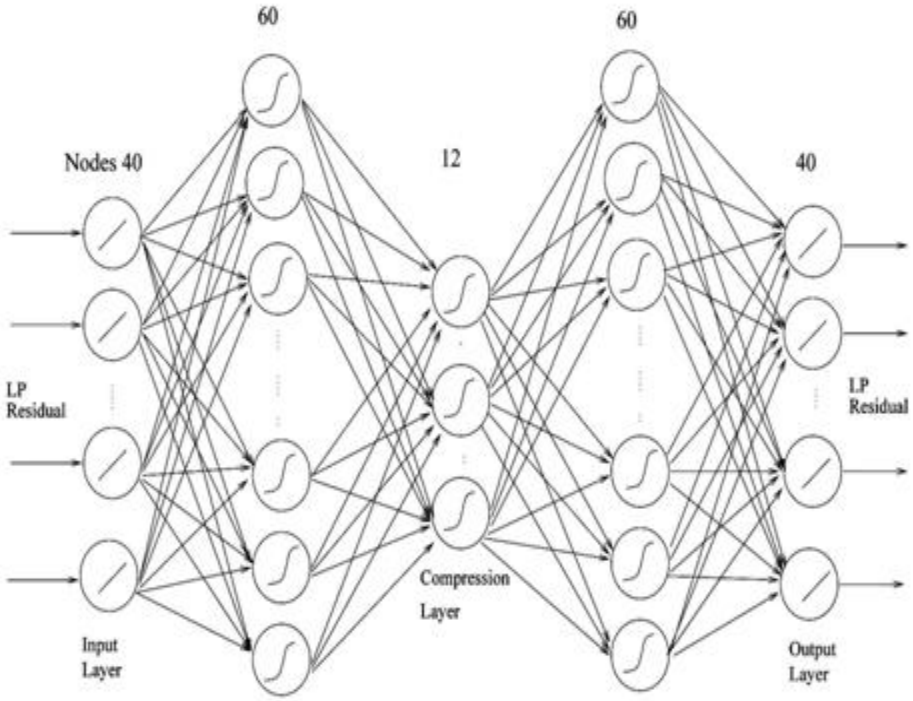

Figure 5: AANN model showing five layers.

V. CONCLUSION

The classification, categorization as well as characterization of the precise emotional features is done from the attributes or characteristic features retrieved from the source of excitation. The existence of precise emotion specific information in each and every constituent of speech is justified with the subjective listening examination conducted on speech word sounds from variety of sources like source of excitation, system for vocal tract as well a mixture of both. The LP residual data samples, segments of the LP residual data samples, epoch constraints, dynamics of the epoch attributes at syllable and statement level, as well as data samples of the GVV signal are taken as precise extracted attributes from the source of excitation for demonstration of emotions.

The remarkable concluding result proved that source of excitation single handedly may not be enough for effective acknowledgement of emotions. But the combined effort of both source of excitation attributes as well as system of vocal track attributes showed remarkable performance advancement in emotional acknowledgement. The precise attributes obtained from consonant-vowel transition area on its own has shown the emotional performance appreciation in comparison to spectral attributes obtained from the total speech. The recognition performance of pitch synchronously extracted attributes is further amplified in the midst of variety of spectral attributes. Local prosodic attributes (dynamics of prosody) has shown better emotion recognition in comparison to global attributes.
So it is concluded that since attributes of source, system, and prosodic show signs of inimitableness for precise emotion specific information thus a combination of above three justifies boosted emotion acknowledgment performance.

REFERENCES

1. Emotion Recognition Using Excitation Source Information."Sreenivasa Rao Krothapalli,Shashidhar G. Koolagudi" Springer 2012
2. D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," Speech Communication, vol. 49, no. 2, pp. 98–112, 2007.
3. M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," Speech Communication, vol. 49, no. 10-11, pp. 787–800, 2007.
4. D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am., vol. 90, no. 5, pp. 2394–2410, 1991.
5. J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," IEEE Trans. Affective Computing, vol. 2, no. 3, pp. 162–174, 2011.
6. R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," J. Personality and Social Psychology, vol. 70, no. 3, pp. 614–636, 1996. [6] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," in Speech Evaluation in Psychiatry, J. K. Darby, Ed. Grune and Stratton, New York, 1981.
7. M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in Proc. Int. Conf. Multimedia and Expo, Hannover, Germany, 2008, pp. 865– 868