ISSN:0975-9646

Emotion Recognition System for English Dialects using Dimensional Theory

M. Chinna Rao M.Tech,(Ph.D)
Associate Professor, Kakinada Institute of Engg & Tech

Dr. A.V.S.N. Murthy
Professor of Mathematics, Srinivasa Institute Of Engg & Tech
Mummidivaram,

Abstract--In this paper Emotion recognition often refers to the science and technology of developing algorithms and implementing them on machines to recognize the Emotions. Research in the field of Emotion recognition has made a number of significant advances in the last two decades. emotion recognition results from speech signals, with particular focus on extracting emotion features from the short utterances typical of Interactive Voice Response (IVR) applications. This paper proposes a new approach to emotion recognition, making use of two of the emotional dimensions, *arousal* and *valence* and their relationship with different kind of features. The classification of various types of emotions such as sadness, boredom, happy, and cold anger. Emotion has used a dimensional theory proposed a dimensional model of emotion that is founded on a biological organization of appetitive and defensive motivational connections in the human cognitive system. We use a database from the Linguistic Data Consortium at University of Pennsylvania, which is recorded by 8 actors expressing 15 emotions. Results indicate that hot anger and neutral utterances can be distinguished with over 90% accuracy. We show results from recognizing other emotions. We also illustrate which emotions can be clustered together using the selected prosodic features.

Keywords: Emotion recognition, speech signal processing, arousal and valence.

I. Introduction

Emotion recognition in human speech has gained increasing attention in recent years due to the wide variety of applications that benefit from such technology. Although human emotions are hard to characterize and categorize [9], research on machine understanding of human emotions is rapidly advancing. Recognition of the emotional state of a person from the speech signal has been increasingly important, especially in human computer interaction. Emotion recognition depend on which emotions we want a machine to recognize and for what purpose. Emotion recognition has applications in human and machine interaction. In the area of speech technology during the last decade have worked on different aspects of emotions in speech. Linear prediction model features, cepstral features, mel-scale filter bank cepstral features, and log frequency power coefficients, which are used for various speech applications, have been tested for analysis and classification of emotion speech signals [1],[2],[3],[11],[12]. Although the emotion analysis of speech has gained some results, however, the practical experience to recognize emotions from English is extraordinarily deficient. In this paper, an analysis framework of speaking words and extraction of emotional features are constructed. In this paper we proposed a dimensional theory of emotion that is founded on a biological organization of appetitive and defensive motivational connections in the human cognitive system. Lang and colleagues believed that the biological organization underlying their model accounts for much of the variance in evaluative judgments behind emotional responses.

II. Emotion Recognition in Dimensional Model

Emotion recognition of speech can be viewed as a pattern recognition problem [2,10]. The results produced by different experiments is characterized by the *features* that are believed to be correlated with the speaker's emotional state, the type of *emotions* that we are interested in the *database* used for training and testing the dimensional and d) the type of dimensionas used in the experiments. To compare dimensions results, we must use the same dataset and agree on the set of emotions. The purpose of this section is not to compare results reported in earlier research but instead to review briefly techniques used in emotion recognition. Dimensional model of emotion that is founded on a biological organization of appetitive and defensive motivational connections in the human cognitive system. This model proposes three underlying dimensions to emotional response. The first dimension, valence, reflects the degree to which an emotional response is positive or negative. The second dimension, arousal, indicates the level of activation associated with the emotional response and ranges from very excited or energized. A third dimension of emotional response is dominance. Measures of dominance range from a feeling of being in control during an emotional experience to a feeling of being controlled by the emotion. Factor analyses of emotional data have consistently supported a strong two-factor solution consisting of arousal and valence as dimensions of emotional response. Dominance has been found to be a less stable dimension of emotional response and less reliably measured (Bradley&Lang, 1994). Early in the development of a three-dimensional model of emotion, it was recognized that dominance accounts for very little of the variance in emotional response (Mehrabian & Russell, 1974). Russell

(1980) argued that dominance should maintain some theoretical importance but relegated dominance to a secondary factor. It has been proposed that dominance may be a content feature of emotional stimuli but has few motivational or behavioral consequences in determining emotional response (Bradley&Lang, 2000).For this reason, a significant amount of research has theoretically acknowledged a three-dimensional model of emotional response but has primarily focused on manipulating and measuring arousal and valence (i.e., A. Lang, Dhillon,& Dong, 1995; P. J. Lang et al., 1993). The conceptualization of valence and arousal as the primary dimensions of emotional experience is quite common. This research suggests that human emotional experience can be mapped onto a two-dimensional space with valence and arousal as its axes, with a focus on emotion as it is being experienced by a person. Various techniques are used to elicit emotion in experimental participants, but the theory focuses on emotion. Tato *et.al.* [13] discussed techniques that exploit emotional dimension other than prosody. Their experiments showed how "quality features" (based on formant analysis) are used in addition to "prosody features" (pitch and energy) to improve the classification of multiple emotions. The quality features were mostly speaker-dependent and hence cannot be used in voice response s. Yu *et.al.* [15] used support vector mechanisms for emotion detection. They built classifiers for four emotions: anger, happy, sadness, and neutral. Since SVMs are binary classifiers, their recognizers worked on detecting one emotion versus the rest. An average accuracy of 73% was reported.

III. Emotion Characteristics in the LP Residual

Speech signals, as any other real world signals, are produced by exciting a system with source. A simple block diagram representation of the speech production mechanism Vibrations of the vocal folds, powered by air coming from the lungs during exhalation, are the sound source for speech. Hence, as can be from Fig. 1.12.the glottal excitation forms the source, and the vocal tract forms the system. One of the most powerful speech analysis technique is the method of linear predictive analysis. The philosophy of linear prediction is intimately related to the basic speech production model. The Linear Predictive Coding (LPC) analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint. Since speech can be modeled as the output of linear, time varying system excited by a source, LPC analysis captures the vocal tract system information in terms of coefficients of the filter representing the vocal tract mechanism. Hence, analysis of speech signal by LP results in two components, namely the synthesis filter on one hand and the residual on the other hand. In brief, the LP residual signal is generated as a byproduct of the LPC analysis, and the computation of the residual signal is given below.

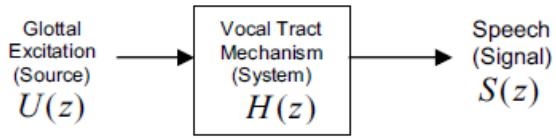

Fig.1.12. Source and System representation of Speech production

If the input signal is represented by Un and the output signal by Sn, then the transfer of the system can be expressed as,

$$H(z)=\frac{S(z)}{U(z)}$$

Where S(z) and U(z) are z-transforms of Sn and Un respectively.

Consider the case where we have output signal and the system and have to compute the input signal. The above equation can be expressed as S(z) = H(z)U(z)

$$U(z)=\frac{S(z)}{H(z)}$$

$$U(z)=\frac{1}{H(z)}S(z)$$

$$U(z)=A(z)S(z)$$

Where A(Z) = 1/H(Z) is the inverse filter representation of the vocal tract system.

Linear prediction models the output Sn as the linear function of past outputs and present and past inputs. Since prediction is done by a linear function, the name linear prediction. Assuming an all-pole for the vocal tract, the signal Sn can be expressed as linear combination of past values and some input Un as shown below.

$$Sn=-\sum_{k=1}^{p} a_k S_{n-k}+GU_n$$

Where G is a gain factor. Now assuming that the input Un is unknown, the signal Sn can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of

S_n be $\tilde{S}_n$,Where

$$\tilde{S}_n=-\sum_{k=1}^{p} a_k S_{n-k}$$

Then the error between the actual value S_n and predicted value $\tilde{S}_n$ is given by $e_n=S_n-\tilde{S}_n$. This error *en* is nothing but LP residual of signal is shown in below fig.

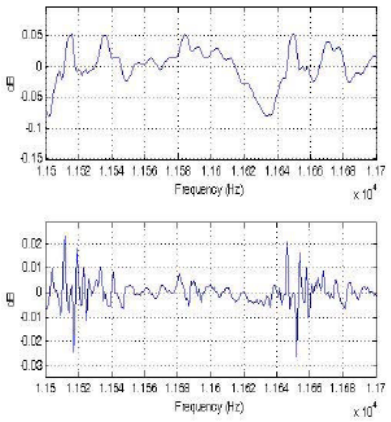

Fig.. Actual signal and its LP Residual

IV. Speech Database

The evolution of an emotion classifier relies heavily on the quality of the database used for training and testing and its similarity to real world samples (generalization). Speech data used for testing emotion recognition can be grouped under three categories depending on the way the speech signal was captured. The first method uses actors to record utterances, where each utterance is spoken with multiple feigned emotions. The actors are usually given the time to imagine themselves into a specific situation before speaking. The second method called Wizard-Of-Oz (WOZ) uses a program that interacts with the actor and drives him into a specific emotion situation and then records his responses. The third method, which is hard to obtain, is actual real-world recording of utterances that express emotions. In our experiments, we used a database from the Linguistic Data Consortium, University of Pennsylvania [14]. The original data set has 9 hours of English recordings in sphere format and their transcripts. The dataset is encoded in 2-channel interleaved 16-bit PCM. Each speech file is a continuous recording of several emotions from one speaker. We developed a splitter component that takes these recordings and the associated transcripts and emits separate utterances and their transcripts. Each utterance file represents one utterance by one actor expressing one emotion. As a result we obtain a set of 2433 utterances roughly distributed over fifteen emotions: neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, and contempt. These are short utterances of 3-4 words each 16-bit PCM, 22.05 KHz, and one channel. The utterances are spoken by 8 actors, mostly in the mid-20s with five females and three males.

V. Performance of Emotion Recognition

The system has been implemented in Matlab7 on WindowsXP platform. The result of the study has been presented in tables 1.1, 1.2. We have used LP order of 12 for all experiments. We have trained the model using dimensional model for different states and mixtures. Testing is performed using different test speech lengths such as 5 sec, 8 sec, 10 sec. Here, recognition rate is defined as the ratio of the number of Emotions identified to the total number of Emotions tested. For testing length of 10 sec is outperformed, where as for testing length of 8 sec is also on par with 10 sec testing length recognition rate. The table 1.1, 1.2 shows identification rate increases with different emotion states and mixtures.

VI. Result Analysis

In this paper the testing has performed by recording 50 Emotions voices and found the results in the form of percentage of recognition summarized below. The Emotion's voices are recorded for both training and testing. The training data is collected in the form of english Voice utterances and the testing data is collected in the form of telugu Voice utterances. The training data is collected for different time durations like 10 Sec, 15 Sec and 20 Sec, correspondingly the testing should be performed with different test durations like 5 Sec, 8 Sec, 10 Sec. With this following data the testing should be performed by changing the state space as 2 states and 3 states and finally found that the percentage of recognition is increased with the increased test and train time durations.

Results of Percentage of recognition with 3 states and diff. train and test durations:

		Testing Duration		
		5 Sec	8 Sec	10 Sec
Training Duration	10	72%	78%	85%
	15	79%	82%	89%
	20	83%	86%	93%

Table 1.1. Percentage of Recognition Results

Results of Percentage of Recognition with 2 states and diff. train and test durations:

		Testing Duration		
		5 Sec	8 Sec	10 Sec
Training Duration	10 Sec	68%	74%	78%
	15 Sec	75%	79%	81%
	20 Sec	78%	82%	88%

Table.1.2. Percentage of Recognition Results

VII. Conclusion

In this paper we have demonstrated the importance of information in the excitation component of speech for Emotion recognition task. Linear prediction residual was used to represent the excitation information. Performance of the recognition experiments shows that dimensional model can capture some Emotion-specific excitation information from the LP residual. Performance of the system for different HMM states shows that it could capture the Emotion-specific excitation information. Larger the training length, the better is the performance, although smaller number reduces computational complexity. The objective in this project was mainly to demonstrate the significance of the Emotion-specific excitation information present in the linear prediction residual for Emotion recognition. We have not made any attempt to optimize the parameters of the model used for feature extraction, and also the decision making stage. Therefore the performance of Emotion recognition may be improved by optimizing the various design parameters.

References

[1] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic Emotion identification and verification" J. Acoust. Soc. Ameri., vol. 55, pp.1304-1312, Jun. 1974.
[2] Ang, J., Dhillon, R., Krupski, A., Shriberg, E. and Stolcke A., "Prosody-based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog", in *Proc. of ICSLP-2002*, Denver, Colorado, Sept. 2002.
[3] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Noth, E., "Desperately Seeking Emotions: Actors, Wizards, and Human Beings", in *Proc. of the ISCA Workshop on Speech and Emotion*, the Queen's university of Belfast, Northern Ireland, Sept. 5-7, 2000.
[4] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J.G., "Emotion Recognition in Human-Computer Interaction", *IEEE Signal Processing Magazine*, 18(1), pp. 32-80, Jan 2001.
[5] Dellaert, F., Polzin, T., and Waibel, A., "Recognizing Emotion in Speech", in *Proc. of ICSLP 1996*, Philadelphia, PA, pp. 1970 -1973, 1996 [6] Hess, W.: *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin 1983.
[6]Bradley,M.M. (1994). Emotional memory: A dimensional analysis. In S. Van Goozen, N. E. Van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 97-134). Hillsdale, NJ: Lawrence Erlbaum.
[7]Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology, 50*, 260-268.
[8] Levit, M., Huber, R., Batliner, A., and Noeth, E., "Use of Prosodic Speech Characteristics for Automated Detection of Alcohol Intoxication", *ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. Red Bank, NJ October 22-24, 2001
[9] McGilloway *et.al.*, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark", *ISCA Workshop on Speech and Emotion*, Belfast 2000.
[10] Ortony, A., Clore, G.L., and Collins, A.: *The Cognitive Structure of Emotions*, Cambridge Univ. Press, 1988.
[11] Petrushin, V., "Emotion in Speech: Recognition and Application to Call Centers", in *Proc. of Artificial Neural Networks in Engineering*, pp. 7-10, Nov. 1999.
[12] Petrushin, V., "Emotion Recognition in Speech Signal: Experimental Study, Development, and Application", in *Proc. of International Conference on Spoken Language Processing*, ICSLP 2000.
[13] Quinlan, J.R.: *C4.5: Programs for Machine Learning*, Morgan Kauffman, 1993.
[14] Tato, R., Santos, R., Kompe, R., Pardo, J.M., "Emotional Space Improves Emotion Recognition", in *Proc. Of ICSLP-2002*, Denver, Colorado, September 2002

Mortha Chinna Rao received the B.Tech degree in Computer Science and Engineering from JNTU, Hyderabad, Andhra Pradesh, India, in 2004. And also, received M.Tech, in Software Engineering from JNTU, Hyderabad, Andhra Pradesh, India, in 2008. Presently He is pursuing Ph.D in JNTUK, Kakinada, Andhra Pradesh, India. He is working as Associate Professor in KIET, Kakinada. His research interests including Data mining and Emotion recognition and speech recognition in Image processing. He is member of Computer society of India and Indian Society for Technical Education.