Exploring Speech Features for Classifying Emotions along Valence Dimension

Shashidhar G. Koolagudi and K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India
koolagudi@yahoo.com, ksrao@iitkgp.ac.in

Abstract. Naturalness of human speech is mainly because of the embedded emotions. Today's speech systems lack the component of emotion processing within them. In this work, classification of emotions from the speech data is attempted. Here we have made an effort to search, emotion specific information from spectral features. Mel frequency cepstral coefficients are used as speech features. Telugu simulated emotion speech corpus (IITKGP-SESC) is used as a data source. The database contains 8 emotions. The experiments are conducted for studying the influence of speaker, gender and language related information on emotion classification. Gaussian mixture models are use to capture the emotion specific information by modeling the distribution. An average emotion detection rate of around 65% and 80% are achieved for gender independent and dependent cases respectively.

Keywords: Emotion; Emotion recognition; Gausssian mixture models; Telugu emotional speech database; Prosody; Spectral features; Valence.

1 Introduction

Most of state-of-the-art speech systems can efficiently process neutral speech, leading to incomplete and imperfect communication. Human beings always encapsulate the message, with in an envelop of desired emotion. This inbuilt emotion successfully conveys the intended meaning of the message, from the speaker, to the listener. So speech systems capable of processing emotional content of the signal along with proper message, are claimed to be more complete and meaningful. To bring in, the missing naturalness into the processed speech, one has to explore the mechanism of capturing emotions from the natural speech. Emotions through a nonverbal communication, play an important role, in the analysis of, telephonic conversations, call center dialogues and interactive voice response systems(IVRS) [1]. Medical doctors may use emotional content of the patient's voice as a diagnosing tool. Extracting emotions from the tapped telephone conversation of crime suspects, may help forensic department to nab the culprits. Robotic pets and humanoid partners may be more natural and enjoyable, if they can express and recognise emotions . So today's applications prefer the speech systems that can understand and produce emotions.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 537–542, 2009.

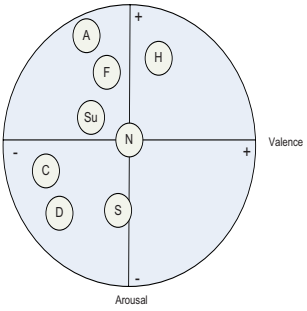

Fig. 1. The distribution of 8 emotions on a two dimensional emotional plane of Arousal and Valence. (A-Anger, C-Compassion/Sad, D-Disgust, F-Fear, H-Happy, N-Neutral, S-Sarcastic, Su-Surprise).

The emotional content of the speech can be visualized in a 3-dimensional space. The three dimensions are, arousal or activation, valence or pleasure and power or dominance [2]. Arousal is a loudness in expression. Valence is a perceptual pleasure, indicating positivism or negativity of the emotion. Power indicates dominance or weakness of an expression. Fig.1 shows the projected three dimensional emotional space, on the plane containing activation and valence as axes. One can observe here that the emotions anger and happy, are not easily distinguishable using arousal, but they can be using valence.

Normally human beings use long term speech features like: energy profile, intonation pattern, duration variations and so on, for detecting the emotions [3]. These are known as prosodic features. This is the reason that, most of the emotional speech related literature is shaped around the close vicinity of either prosodic features or their derivatives. But it is difficult to distinguish the emotions, that share common acoustic and prosodic features, using only these longterm features. It is shown in the Table 1, that quite frequently the emotions like anger and happy are inter miss-classified, with only prosodic features.

Table 1. Percentage of inter miss-classification of Anger and Happy emotions quoted by different researchers in the literature

Reference	Language	Percentage of anger utterances classified as happy	Percentage of happy utterances classified as anger
Serdar yeldirim, et.al. [4]	English	42	31
Dimitrios Virviridis, et.al. [5]	Scandinavian language (danish)	20	14
Felix Burkhardt, et.al. [6]	German (Berlin Emotion Database)	–	12
Oudeyer Pierre, et.al. [7]	Concatenated synthesis (English)	35	30
S G. Koolagudi, et. al.[8]	Telugu	–	34
S. G. Koolagudi	Berlin, German	27	20
Raquel tato [9]	German	24	25

2 The Collection of Speech Database and Selection of Features

The word 'emotion' is often open to quite large number of interpretations. In reality, emotions are basically pervasive in nature, but research models are mostly being built for their full blown versions. The emotional speech corpora may be collected in three major ways. The expert artists are asked to produce verbal emotions to get *simulated database*. The artists or normal people are made to produce different emotions by creating respective situations, without the knowledge of the speaker, giving *induced emotional database*. *Naturalistic data corpus* contains the recordings of natural emotional conversations. In this work, simulated speech corpus of Telugu language, is used for emotion analysis. The 8 emotions considered are anger, compassion, disgust, fear, happy, neutral, sarcastic and surprise . 10 speakers (5 male and 5 female) of varied age and experience from All India Radio (AIR) station, are recorded for preparing the corpus. 15 linguistically neutral sentences are uttered by each speaker in 10 different sessions. So the database contains 1500 utterances ($15sentences \times 10sessions \times 10speakers$) for each emotion and 12000 utterances ($1500 \times 8emotions$) in total. Proper care has been taken to include all phoneme classes, speaker, gender, text and session variations. The speech is recorded with a sampling frequency of 16kHz and each sample is stored as a 16 bit number [8].

Here we tried to identify the reliable speech features, that are least influenced by gender, speaker, language, cultural and contextual variations, while detecting emotions. Valence or perceptual pleasure is mainly observed in the speech due to conscious vocal effort through articulator activities. So appropriate spectral features - representing the vocal tract characteristics may be used as the features of interest. Therefore MFCC features are used to represent valence information of emotional utterances. The motivation to use these features, is that 'MFCC's' represent vocal tract characteristics of a speaker and pleasure (valence) of an emotional utterance. In fact MFCC's are derived using audio critical bands of human perceptive system and understanding of an emotion, by human being, is largely an individualistic perception. In this work the experiments are conducted to justify the hypothesis 'Spectral features alone are sufficient to classify the emotions along valence axis'.

3 Results and Discussion

The identification of different emotions would be reasonably possible, only when all the emotions under study, are properly discriminated in an emotional space using suitable features. Here the spectral features are explored for characterizing the emotions, with respect to the valence dimension. To capture the distribution of emotions, Gaussian mixture models are used. Single GMM is used for one emotion. So the model, in total contains 8 GMM's. 12 MFCC features extracted from a frame of a signal, along with one energy value, formed a feature vector of size 13. All the utterances of specific emotion are taken from the first 8 sessions

Table 2. Confusion Matrix for classification of emotions for the model trained using single female utterances. 13 MFCC's were used to construct feature vector, GMM contains 128 components and converged with 200 iterations.

	Anger	Compassion	Disgust	Fear	Happy	Neutral	Sarcastic	Surprise
Anger	60	0	27	0	0	3	10	0
Compassion	0	87	3	7	3	0	0	0
Disgust	23	0	70	0	0	0	7	0
Fear	10	3	0	84	0	0	0	3
Happy	0	0	0	0	84	10	3	3
Neutral	3	0	0	0	0	97	0	0
Sarcastic	3	0	3	0	0	0	94	0
Surprise	0	3	0	3	23	0	0	71

of the corpus and used for training the GMM's. 30 randomly selected utterances of each emotion, from the remaining 2 sessions are used for validating the trained models.

In this work the effort has been made to reduce the discrimination discrepancies caused due to prosodic features, by classifying the emotions on the basis of valence dimension. Table 2 is a confusion matrix obtained from the emotion classification results of the single female actor's utterances. GMM's with 128 components and 200 epochs towards convergence, yielded an average emotion detection rate of 80.42%. The spectral features (MFCC's) here, are able to clearly classify the emotions like anger and happy, which share similar acoustic characteristics along arousal or activation axis. It may be observed from Table 2, that none of either happy or anger utterances is inter miss-classified.

The summary of the classification results obtained using different, GMM configurations and text dependent training sets is briefed in Table 3. Column 2A represents the emotion classification performance of the models, built using single male speaker utterances. Similar results of female speaker are tabulated in column 2B. Performance of female emotion recognition is better due to clear expression of emotions by female artists. It is also supported by the MOS of subjective listening tests. Columns 2C and 2D represent the classification results, where speaker related information is generalised by training the models with the utterances of multiple speakers of the same gender. Column 2E represents the performance, where emotion recognition models are built with both the genders.

Table 4. consolidates the emotion recognition performance for text independent cases. Here the texts of training and testing utterances are different. This experiment is to verify the effect of phoneme related information on the classification results. Almost near to similar results are observed for text dependent and independent cases for female voices. Observe the columns 2A as well as 2B of tables 3 and 4 respectively . The higher performance in case of male emotion recognition, for text dependent case is obvious because of phonemic information playing role during classification (see the columns 2A of Tables 3 and 4) and male voices are less expressive compared to female voices. But the similar trend is not observed for female emotive utterances, the slight improvement in the emotion

Table 3. Average emotion classification performance for text dependent case. Emotions considered for analysis are anger,compassion, disgust, fear, happy neutral, sarcastic, surprise.

1 GMM's C-No.of components I-No.of epochs	2 Training Sets				
	2A Single Male	2B Single Female	2C 3 Males	2D 3 Females	2E 3Males + 3Females
64C-100I	73.33	80.63	66.25	73.25	59.25
64C-200I	74.58	79.17	65.75	74.75	63.38
128C-100I	75.83	80.42	73.25	76.37	61.63
128C-200I	77.92	80.42	73.37	76.75	63.75

Table 4. Average emotion classification performance for male and female voices in text independent case.Average emotion classification performance for text dependent case. Emotions considered for analysis are anger,compassion, disgust, fear, happy neutral, sarcastic, surprise.

1 Gaussian Mixture Models C-No.of components I-No.of epochs	2 Training Sets	
	2A Single Male	2B Single Female
64C-100I	63.33	84.12
64C-200I	63.75	80.00
128C-100I	66.25	85.42
128C-200I	65.83	83.33

recognition performance of text independent case justifies the very little role of phoneme based information towards classification. It reveals the fact that along with the speaker and phoneme specific information, spectral features also contain robust emotion specific information. The clear classification of emotions like, anger and happy motivates one to attribute this robust emotion specific information to MFCC features and hence the classification can be justified to be along valence or pleasure dimension. The above experiments are designed to show the minimal influence of speaker, gender and phoneme related information during the classification.

4 Summary and Conclusions

In this work, features representing the characteristics of vocal tract system (Spectral features) were proposed to discriminate the emotions. The hypothesis considered is that, 'the characteristics of VT system will follow the valence dimension of the emotions'. It has been shown here that, the emotions can be efficiently discriminated using the features contributing to emotional valence. Valence being pleasure in perception, is mainly contributed by the spectral features. So MFCC

features were used for classifying the emotions. These are found to be robust amongst other features, while classifying the emotions amidst speaker, gender and language variations. Maximum average emotion recognition performances for 8 given emotions, in case of single male and female speakers are around 78% and 80% respectively.

It is still a difficult challenge to classify the closely spaced emotions along pleasure and power axes. For example anger and disgust are very close in the negative region along pleasure axis, and they are on the either side of the origin along power axis. It is important to note that one cannot claim the spectral features, as the sole carriers of emotion specific information in their pleasure space, but results have shown that they have major contribution towards the valence. The combination of prosodic and spectral features may be more robust to represent the all 3 known dimensions of emotions.

References

1. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Trans. Speech and Audio Processing 13, 293–303 (2005)
2. Jin, X., Wang, Z.: An Emotion Space Model for Recognition of Emotions in Spoken Chinese. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 397–402. Springer, Heidelberg (2005)
3. Rao, K.S., Yegnanarayana, B.: Intonation modeling for indian languages. CSL (2008)
4. Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Busso, C., Deng., Z., Lee, S., Narayanan, S.: An acoustic study of emotions expressed in speech. In: Int'l Conf. on Spoken Language Processing (ICSLP 2004), Jeju island, Korean (October 2004)
5. Ververidis, D., Kotropoulos, C., Pitas, I.: Automatic emotional speech classifcation. In: ICASSP 2004, pp. I593–I596. IEEE, Los Alamitos (2004)
6. Burkhardt, F., Sendlmeier, W.F.: Verification of acousical correlates of emotional speech using formant-synthesis. In: ITRW on Speech and Emotion, Newcastle, Northern Ireland, UK, September 5-7, pp. 151–156 (2000)
7. Oudeyer, P.-Y.: The production and recognition of emotions in speech: features and algorithms. International Journal of Human Computer Studies 59, 157–183 (2003)
8. Koolagudi, S.G., Maity, S., Kumar, V.A., Chakrabarti, S., Rao, K.S.: IITKGP-SESC: Speech Database for Emotion Analysis. In: Communications in Computer and Information Science, JIIT University, Noida, India, August 17-19. Springer, Heidelberg (2009)
9. Tato, R., Santos, R., Pardo, R.K.J.: Emotional space improves emotion recognition. In: 7th International Conference on Spoken Language Processing, Denver, Colorado, USA, September 16-20 (2002)