International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 4, April 2017

ISSN: 2394-4404

An Improved Approach Of Emotion Recognition Combining Spectral And Prosodic Features With Reference To Assamese Language

Akalpita Das

Purnendu Acharjee

Assistant Professor

Pranhari Talukdar

Professor

Abstract: ***The Speaking rate feature of speech can be explored for discriminating robust emotions. In real life, it is found that certain emotions are used to be very active with high speaking rate while some are very passive with low speaking rate. Keeping this motivation, a Phase II emotion recognition system has been proposed where three broad groups (active, neutral and passive) are taken in Phase I and each broad group are further classified in Phase II. In each stage classification of emotions are done by exploring Spectral and prosodic features. The combination of both spectral and prosodic features found to be performed better.***

Keywords: ***Spectral, Prosodic, Excitation, Formant.***

I. INTRODUCTION

In the process of classification of emotions, it is seen that similar emotions always lead to misclassification. Such misclassification need to be reduced by taking an extra measure on performing early classification of those most confusing emotions into some different sub-groups. By grouping acoustically over lapping emotions into separate categories in stage one and classifying individual emotions in stage two. In the current study, "*speaking rate"* is chosen as the decisive factor for the sub-grouping of acoustically similar emotions. Later both Spectral and prosodic features in combination is used for further classification [1]. In stage one; by using spectral and prosodic features each emotion is categorized in 3 wider groups: 1) active 2) normal, and 3) passive emotions. Such wider groups are made based completely on speaking rate. In stage two, individual emotion sub grouping performed within each wider group.

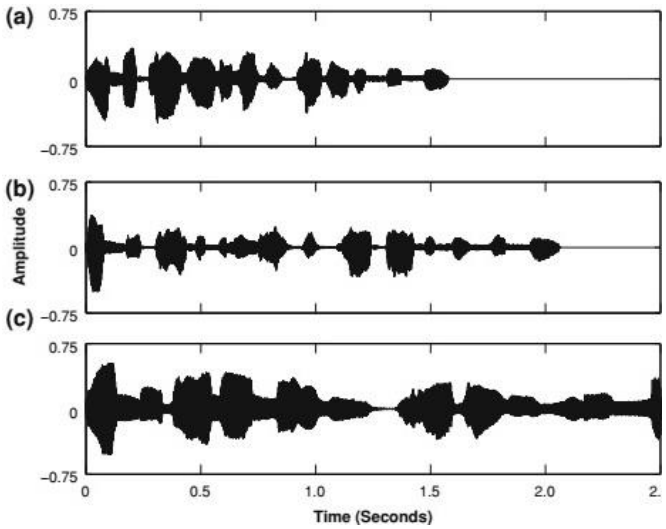

Figure 1.1: Duration of "কি কৰি আছা" for emotions a) disgust, b) neutral and c) sarcasm

Since it is observed that Excitation source feature has no appreciable influence in speech emotion recognition so it is avoided. We know that speaking rate is only a measure in utterance of number of syllables per unit amount time. It is accepted as most important characteristics for each speaker

[2]. While experimenting with different emotions, the variability of speaking rate is observed clearly along with other parameters like duration and frequency of pauses, vowel durations, gender and age of the speaker etc [3, 4]. We know speaking rate of a speaker greatly depends on psychological and physiological characteristics [5]. It has been observed that younger people contain faster speaking rate than the older people. Male speakers' speaking rate is comparatively faster than that of female speakers [5]. By conscious manipulation of source and system parameters, we can vary speaking rate by either insertion or deletion of different length pauses at different levels [6, 7]. It can be achieved due to variations change in the excitation source characteristics and in articulator movements [6]. The effect of change in speaking rate is nicely visible on prosodic features. Again we see that Pitch values for spontaneous speech are greatly varies by speaking rate variation [8]. Fast speech always characterized by the overall reduction of pitch range [7].

II. MOTIVATION

We know speaking rate depends upon speaker and gender to a large extent. The speech waveforms correspond to "কি কৰি আছা" /ki kori AsA/ taken from Assamese GU-EMO-SPEECH database. The same text is uttered by the speakers, here the speech durations are different, it's because of different embedded emotions. While expressing different emotions, speakers took minimum time for disgust and maximum for sarcasm. This observation clearly indicates that expression of speech emotions directly influenced by the speaking rate. Therefore characteristics of speaking rate need to be explored to classify emotions.

Emotion Type	Duration in Seconds
Surprise	2.10
Sarcasm	2.10
Neutral	2.18
Happiness	2.12
Fear	1.92
Disgust	1.65
Compassion	2.10
Anger	1.60

Table 1.1: Average duration of the speech utterances of GU-EMO-SPEECH for different Emotions

III. IMPLEMENTATION

The rate of vocal folds' vibration influences the behavior of the vocal tract when producing sound units. The opening of vocal folds of a human is observed to be normally same due to their muscular restriction. It's also observed that higher order formants (*F*2, *F*3 and *F*4) are used to be more distinctive with respect to emotions. To analyze classification of speech utterances emotional state based on speaking rate, an Assamese speech database with varying speaking rates is collected at GU-SPEECH-LAB. The database contains speech utterances of 5 different speaking rates. They are super slow, slow, normal, fast, super fast. GU-SPEECH-LAB has contributed to the recording of 10 Assamese sentences in 5 different speaking rates, in total 500 (10 *speaker* ×10 *sentences*×5 *speaking rates*) utterance.

EMOTION RECOGNITION PHASE I

Formant analysis of slow and fast utterances is performed on speech data. From figure 6.2 we can see the distribution formant frequencies for slow and fast utterances through graphs which indicate the distinctive properties of formant frequencies with reference to speaking rate. A classification system has been developed for analyzing the influence of speaking rate on spectral features. LPCC is calculated using the frame size of 20ms with the shift of 5ms. By using Gaussian mixture models, a new model for Assamese Emotional Speech is developed, where based on speaking rate using spectral features the classification of speech utterances is performed. Table 6.2 shows the classification performance. Average classification performance is found to be about 82%.

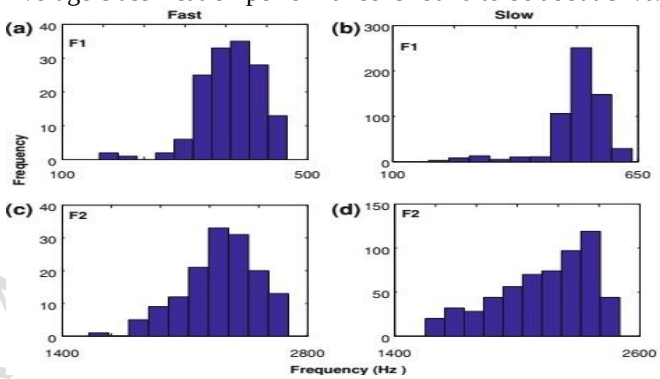

Figure 1.2: Distribution of frame wise F1 and F2 a) F1 for fast, b) F1 for slow c) F2 for fast d) F2 for slow

Rate	Super-Fast	Fast	Normal	Slow	Super-Slow
Super-Fast	0	0	0	5	100
Fast	0	2	7	95	0
Normal	0	10	95	5	0
Slow	25	60	12	9	0
Super-Slow	55	35	15	1	0

Table 1.2: Classification of speech utterances based on the speaking rate using spectral features

EMOTION RECOGNITION PHASE II

In Phase I, all emotions grouped into 3 main groups (active, normal and passive) corresponding to speaking rates as fast, normal, and slow. It is observed that active emotions are enthusiastically expressed with extra energy, but passive emotions are expressed with less energy. This way a gross-level-emotion-classification is done. In Phase II, individual emotion classification is performed.

This Phase II classification allows avoiding misclassification which normally happens in case of Phase I classification. As is shown in fig 1.2, the entire emotion classification systems based on proposed Phase II. In Phase I, depending on the speaking rate features, the unknown utterance is classified into one of the three categories (active

ISSN: 2394-4404

(fast), normal, and passive (slow)) emotions. In finer classification, each main category utterances are again classified into individual emotion.

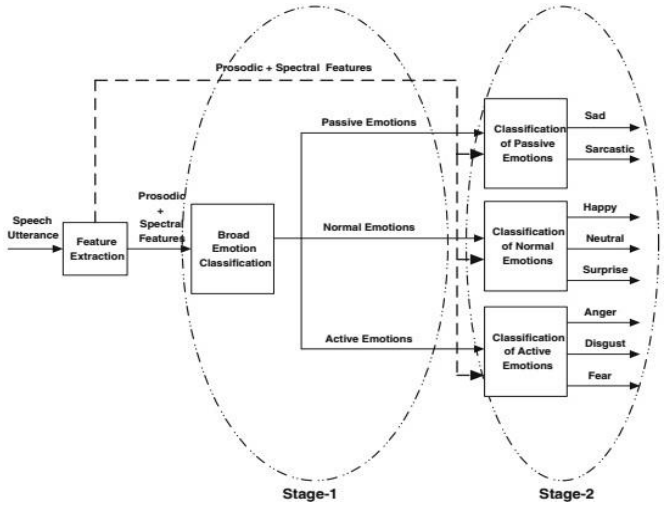

Figure 1.3: Block diagram of Phase II emotion recognition system using speaking rate

IV. EMOTION RECOGNITION AT GROSS LEVEL

Based on duration analysis, eight emotions broadly categorized, into three main groups as active (*fast*) [anger, disgust, fear], normal [neutral, surprise] and passive (*slow*) [sadness, Sarcasm].

Emotion Status	Features Set along with percentage of recognition		
	Spectral+Prosodic	Prosodic	Spectral
Active	97	97	90
Neutral	90	82	95
Passive	95	80	80
Average	97	85	90

Table 1.2: Gross-level emotion classification performance combining of prosodic and system features

Gross-level-emotion-classification is designed by using three feature sets: namely, 1) spectral, 2) prosodic and 3) combination of both. To capture the characteristics of 3 emotional categories, three GMM's set are exclusively trained. To develop speech-emotion-recognition-models, 1800 (10 *sentences* × 10 *sessions* × 3 *emotions* × 6 *speakers*) utterances are used for training *active* and *normal* emotion models, and 1200 (10 *sentences* × 10 *sessions* × 2 *emotions* × 6 *speakers*) utterances are used for training *passive* emotion models. The best classification performance attained by 64 Gaussian component models, using converged after 30 iterations. The broad emotion model is tested using 1200 (5 *sentences*×10 *sessions*×3 *emotions*×4 *speakers*) utterances of active and normal emotions, 400 (5 *sentences* × 10 *sessions* × 2 *emotions* ×4 *speakers*) utterances of passive emotions. The performance of emotion recognition on broad emotion is shown in Table 1.3. It is evident that the emotion recognition performance for active (fast), normal emotion is much better than passive (slow) emotions. This happens because of the dominance in energy and pitch features in active and normal speech emotions.

In this study, the new technique called (weighted) score level fusion technique, has been applied to combine the scores from both spectral and prosodic features. From each speech utterance, features (spectral and prosodic) are extracted and emotion recognition models are developed. From each utterance feature vector a probability score is calculated from each model. Corresponding to that model, the total of vector wise probability score results effective score for each utterance. The highest effective score model for each speech utterance is a hypothesized emotion. To take the decision related to the emotion category normally the weighted sum of two effective scores is used. The combination of weights 0.6 for spectral and 0.4 prosodic respectively found to perform better. The combined measures can recognize the emotions namely active or fast emotions almost without any error. It is found that the average recognition for three broad categories is 89%. It is promising performance to continue the study further.

Emotions(Active)	Recognition Anger	Recognition Disgust	Recognition Fear
Anger	75	25	0
Disgust	20	80	0
Fear	5	0	95

Table 1.3: Finer level classification of active emotions using spectral and prosodic features

Emotions (Normal)	Recognition Anger	Recognition Disgust	Recognition Fear
Anger	85	15	0
Disgust	00	97	0
Fear	15	0	95

Table 1.4: Finer level classification of normal emotions using spectral and prosodic features

Emotions (Passive)	Recognition Sadness	Recognition Sarcasm
Sadness	95	5
Sarcasm	00	97

Table 1.5: Finer level classification of passive emotions using spectral and prosodic features

CONCLUSION

From above tables it may be concluded that the emotions like anger, disgust and surprise emotions are not recognized well compared to other emotions. From the subjective evaluation it is observed that the quality of expression of anger, disgust and surprise is not as discriminative as that of other emotions. Table 1.7 shows the comparison of emotion recognition performance using Phase I and Phase II classifications. From Table, it is seen that almost all emotions are correctly recognized after the Phase I. In Phase II, around 85% of average recognition performance is found. Both Phase II and Phase I are developed using the combination of spectral with 21 LPCCs and local prosodic features. When the Phase II approach is employed it is seen that there is a major improvement in the recognition performance.

Groups	Emotions	Recog. Phase-I	Recog. Phase-II
Active	Anger	82	80
	Disgust	80	87

	Fear	75	95
Normal	Happy	88	92
	Neutral	85	99
	Surprise	72	85
Passive	Sadness	82	96
	Sarcasm	88	98

Table 1.6: Performance using single and Phase II classification approaches on GU-EMO-SPEECH

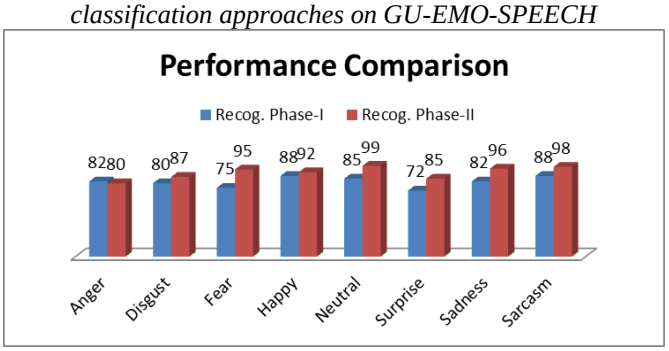

Figure 1.4: Performance of individual emotions using single and Phase II emotion classification

V. CONCLUSION

To resolve the classification ambiguity of highly confusing emotions, a two-stage classification approach has been proposed to enhance the emotion recognition performance. In this two-stage approach, the combination of spectral and prosodic features has been employed. In the Phase I, eight emotions are classified into three broader categories namely active, normal and passive based on the speaking rate. In the Phase II, within a broad group, emotions are classified into the individual category. It has been observed that, after the Phase I, emotion classification performance is very high. The proposed Phase II classification has considerably improved the emotion recognition performance. This method demonstrated the multi-stage emotion classification approach with feature combination.

REFERENCES

[1] S.G.Koolagudi, K.S. Rao, Phase II emotion recognition based on speaking rate. Int. J. Speech Technol. 14, 35–48 (2011)

[2] S.G. Koolagudi, S. Ray, K.S. Rao, Emotion classification based on speaking rate, in *Communications in Computer and Information Science*, ed. by S. Ranka, A. Banerjee, K.K. Biswas, S. Dua, P. Mishra, R. Moona, S.-H. Poon, C.-L. Wang. International Conference on Contemporary Computing, vol. 94, pp. 316–327, Springer, USA, 6–8 Aug 2010

[3] K.S. Rao, B. Yegnanarayana, Modeling durations of syllables using neural networks. Comput. Speech Lang. 21, 282–295 (2007)

[4] A.L. Francis, H.C. Nusbaum, Paying attention to speaking rate, in *Fourth International Conference on Spoken Language, 1996 ICSLP 96*, (Philadelphia, PA, USA), pp. 1537–1540 (V3), IEEE, October 1996. Center for Computational Psychology, Department of Psychology, The University of Chicago

[5] J. Yuan, M. Liberman, C. Cieri, Towards an integrated understanding of speaking rate in conversation, in *Interspeech 2006*, (Pittsburgh, PA, 2006), pp. 541–544

[6] M.S.H. Reddy, K.S. Kumar, S. Guruprasad, B. Yegnanarayana, Subsegmental features for analysis of speech at different speaking rates, in *International Conference on Natural Language Processing*, (Macmillan, India, 2009), pp. 75–80

[7] A. LI, Y. ZU, Speaking rate effects on discourse prosody in standard chinese, in *Fourth International Conference on Speech Prosody*, (Campinas, Brazil, 2008), pp. 449–452, 6–9 May 2008

[8] H. Yang, W. Guo, Q. Liang, A speaking rate adjustable digital speech repeater for listening comprehension in second-language learning, in *International Conference on Computer Science and, Software Engineering*, vol. 5, pp. 893–896, 12–14 Dec 2008

[9] S.G. Koolagudi, S. Maity, V.A. Kumar, S. Chakrabarti, K.S. Rao, GU-EMO-SPEECH: speech database for emotion analysis. Communications in Computer and Information Science, JIIT University, Noida, India: Springer, ISSN: 1865–0929 ed., 17–19 Aug 2009

[10] E.F. Lussier, N. Morgan, Effects of speaking rate and word frequency on pronunciations in convertional speech. Speech Commun. 29, 137–158 (1999)

[11] M. Richardson, M.Y. Hwang, A. Acero, X. Huang, Improvements on speech recognition for fast talkers, in *Eurospeech Conference*, Sept 1999