Combining Speech and Gender Classification for Effective Emotion Recognition

Rahul Vivek Purohit
Asst. Prof. Ajay Kumar Garg Engg. College
27th Km Stone, Delhi-Hapur Bypass Road,
Adhyatmic Nagar, Ghaziabad,U.P(India)

Syed A Imam
Asst. Prof Faculty of Engineering & Technology,
Jamia Millia Islamia university New Delhi, India

ABSTRACT
The applications of emotion recognition in consumer electronics are increasing day by day. However the accuracy and stability of the decisions made by appliances largely depends on the efficient recognition of these emotions. The performance may degrade drastically due to interfering noise. This paper proposes a method which may improve the accuracy significantly. Results have confirmed that this system may help to improve the recognition results.

General Terms
Pattern Recognition, Gender Classification

Keywords
Speech Recognition, Gender classification, subharmonic-to-harmonic ratio

INTRODUCTION
The most important advantage of advancement in technologies must be in allowing peoples to live a comfortable life. Home appliances make one of the important tools in this regard. Improved technologies are making these appliances touch free day by day by the use of speech and emotions. However the stability and performance of these devices depends on the environment in which they are recorded. Any randomness in performance is caused because of the inadequate matching between the system database and actual speech or emotion. Person individuality also affects this performance severely under noisy background.

Many algorithms and methods have been proposed for speech and emotion recognition in the past. Most of these are based on the various combinations of feature extraction and classification. Mohammad soleymani et al. [1] Prepared and tested an extensive database based on face, head, speech, eye gaze, ECG etc. ASSESS (2000) [2] obtained an accuracy of 55% and MEXI of 60% for speaker independent system. Ververidis et al. [3] extracted 87 static features and achieved a recognition rate of 51.6% G. Zhou et al. [4] has used nonlinear Teager Energy Operator (TEO) feature for stressed/neutral classification and compared it with the traditional pitch and MFCC feature. Yacoub et al. [5] has differentiated between angry emotion and neutral emotion while achieving an accuracy of 94%. Dellaert et al. [6] has compared three classifiers: maximum likelihood Bayes classification, kernel regression and k-NN (Nearest Neighbor) with four emotion categories. Scherer [7] extracted 16 features and achieved overall accuracy 40% for fourteen emotional states. Other good works on speech emotional classification may be found in [8-10]. However the randomness in performance due to the noisy environment which results in inaccuracy of results are to be addressed as this makes the very backbone of the stability problem in home appliances.

The focus of this paper is primarily to try to resolve the mismatch problem between the system data base and the actual person`s speech which can make the decision taking ability of a home appliance more robust. This paper proposes the combination of speech and gender classification in a two level hierarchical system .The results shows a significant improvement on the traditional methods used for speech recognition under noisy environment.

1. TWO LEVEL HIERARCHICAL SYSTEM
This system may be represented in two steps. In first emotional features are collected using speech processing and in second feature extraction is used to classify the gender of the individual. Fig. 1 shows the required two level block diagram.

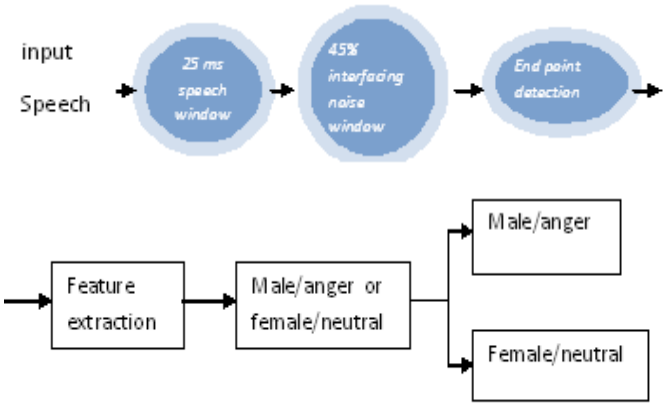


Fig.1 Two level categorization of emotions

The speech input to the system is categorized in four different groups to make the system robust namely male/anger, female/neutral, male/neutral and female/anger with the use of gender classification system and mean features.

In first step emotional features are collected using 25ms speech window, 45% interfering noise window at the ends of speech segment. With the use of an end- point detection method, the speech and non speech segments are differentiated for the purpose of extracting the emotional features. Three kinds of features have been extracted from

each speech segment, namely pitch, energy and 12th order MFCC.

For detection of pitch, many methods have been used in the noisy environment namely, harmonic sum spectrum, sub harmonic to harmonic ratio and average magnitude difference function. As the sub harmonic to harmonic ration uses fast Fourier transform and its performance is better than the others in noisy environment, it is used here to detect pitch. This method compares the maximum amplitude of the speech pitches to make a decision.

2. COMPARISON OF RESULTS

For the experimental purpose, two kind of data base has been used. Data-1[18] and our own data base, data-2. Table 1 shows the comparison of these databases.

Table I
Differences in recording environment
For Two data base

characteristics	Data base 1	Data base 2
Number of individuals	15 male/female	25 male/female
Number of sentences	45 sentences for four emotional features	80 sentences for four emotional features
Number of speech samples	15*45*4 = 2,700	25*80*4 = 8,000

With the use of pair wise coupling [19] classification algorithm the emotions are categorized for the different sets of recorded and actual speech. These results are shown in fig.2.

If the recorded and actual speech is from the same database, this algorithm may provide good accuracy as shown in fig 2(a). However the performance degradation is drastic if recorded and actual speech belongs to different sets as shown in fig 2(b). This occurs because of different recording environments for speech and may affect the decision capability of the device.

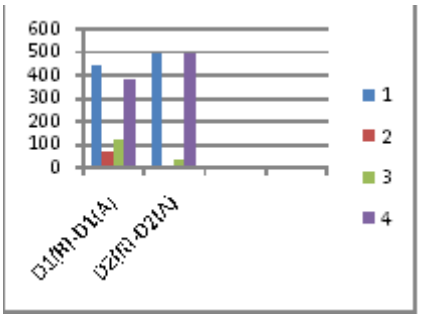


Fig.2 (a) Comparison of recorded and actual data base, Accuracy- 81.3% for D1, 96.2% for D2
1-Neutral-Neutral, 2-Neutral-Anger
3-Anger-Neutral, 4- Anger-Anger

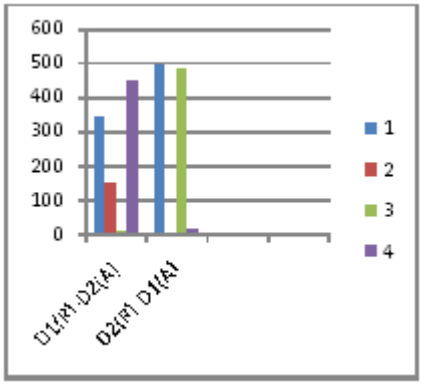
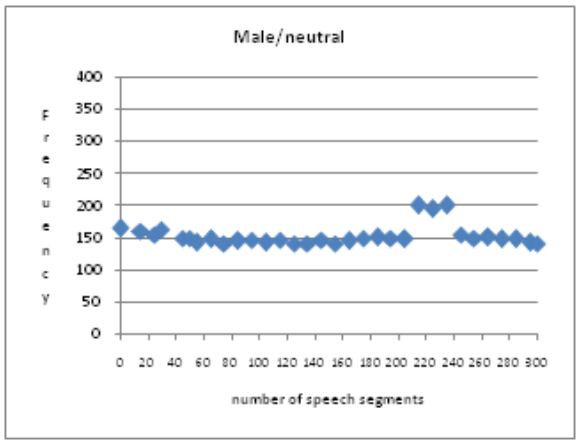
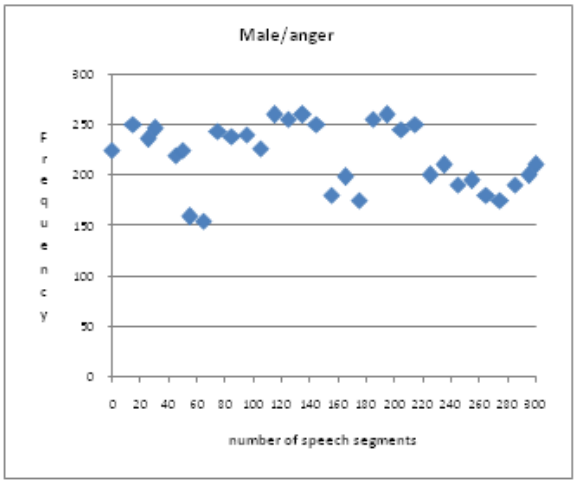


Fig.2 (b) Comparison of recorded and actual data base, Accuracy- 56% for D1(R), 79.4% for D2(R)
1-Neutral-Neutral, 2-Neutral-Anger
3-Anger-Neutral, 4- Anger-Anger

In the proposed algorithm, the pitch is calculated from 15 males and 10 females with 20 sentences from each database for each emotion. Fig 3 shows the results.





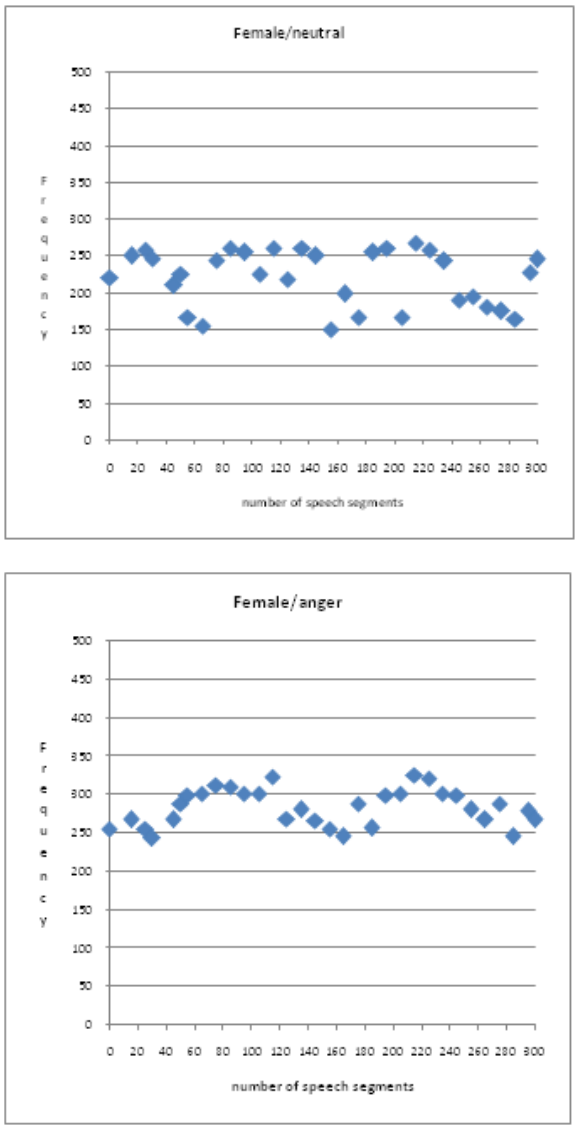


Fig.3 Pitch variation of emotions between male and female

The necessity of the proposed system can be clearly recognized from fig.3. As can be seen male/anger and female/neutral pitch varies almost in the same range of 150Hz to 260 Hz. Male/neutral pitch is almost always below 150Hz and have least variation while female/anger has largest pitch variation with a range of above 240 Hz. Because of the similarity of pitch ranges in male/anger and female/neutral, it is difficult to differentiate them and hence system performance may degrade.

The two level system proposed here may offer the solution to this problem. Here all the four sets are categorized on the basis of pitch first and then in the second level, energy and MFCC is used to categorized male/anger and female/neutral further. Table 2 shows the results with the proposed system.

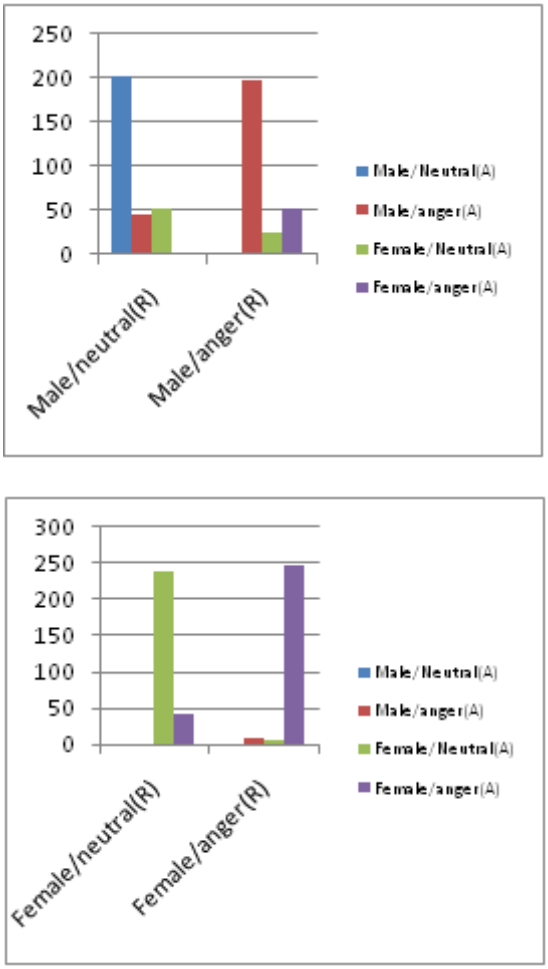


Fig.3 Comparison of D1(R)-D2 (A), Accuracy 86%

Comparison with fig 2 shows an improvement of around 30% with this system.

3. CONCLUSION

Since the emotions are not only affects by the environment bet also by the gender of the individual, this paper has proposed a two level system which can help improving the recognition results. As an improved system it`s applications may also extend to robotics and medicine therapy.

In future a speech recognition system that involves facial expression and gestures will be investigated.

4. REFERENCES

[1] Mohammad soleymani, J. Liichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging" in IEEE transaction on affective computing, Vol 3, No. 1, January-March 2012.

[2] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," in Proc. ISCA Workshop Speech Emotion, Belfast, U.K., 2000, pp. 207–212.

[3] D. Ververidis and C. Kotropolos, "Automatic emotional speech classification," in Proc. IEEE int. Conf. Acoust., Speech Signal Process., May 2004, vol. 1, pp. 593–596.

[4] Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Nonlinear Feature Based Classification of Speech under Stress. IEEE Transactions on speech and audio processing, vol. 9(3), IEEE Computer Society Press, Los Alamitos (2001)

[5] Yacoub, S., Simske, S., Lin, X., Burns, J.: Recognition of emotions in interactive voice response system. Eurospeech 2003 Proc. (2003)

[6] Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in Speech. Proc. International Conf. on Spoken Language Processing, pp. 1970– 1973 (1996)

[7] Scherer, K.R.: Adding the affective dimension: A new look in speech analysis and synthesis. In: Proc. International Conf. on Spoken Language Processing, pp. 1808–1811 (1996)

[8] T. Kostoulas, I. Mporas , T. Ganchev , N. Fakotakis, The Effect of Emotional Speech on a Smart-Home Application, Lecture Notes in Computer Science Vol. 5027, pp. 305-310, 2008

[9] J.S Park, J.H. Kim, Y.H. O, "Feature Vector Classification based Speech Emotion Recognition for Service Robots", IEEE Transactions on Consumer Electronics, Vol. 55, No. 3, August 2009

[10] Hua A., D. Litman, "Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs", ICSLP2006, Pittsburgh, Pennsylvania, USA, 2006.

[11] Kostov, V., Fukuda, S.: Emotion in user interface, Voice Interaction system. IEEE Intl. Conf. on systems, Man, Cybernetics Representation, vol. 2, pp. 798–803 (2000)

[12] Oriyama, T.M., Oazwa,: Emotion recognition and synthesis system on speech. IEEE Intl. Conference on Multimedia Computing and Systems, pp. 840–844 (1999)

[13] Lee, C.M., Narayanan, S., Pieraccini, R.: Classifying emotions in human-machine spoken dialogs. ICME'02 (2002)

[14] Gu, Li., Zahorian, S.A.: A new robust algorithm for isolated word endpoint detection. ICASSP2002, Orlando, USA (2002)

[15] Noll, M.: Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. Proceedings of the Symposium on Computer Processing Communications, pp. 779–797 (1969)

[16] Ross, M.J., Shaer, H.L., Cohen, A., Freudberg, R., Manley, H.J.: Average magnitude difference function pitch extractor. ASSP-22, 353– 362 (1974)

[17] Xuejing Sun, "A pitch determination algorithm based on subharmonic-to harmonic ratio", ICSLP, pp. 676-679, 2000

[18] B. S. Kang, "A text-independent emotion recognition algorithm using speech signal", MS Thesis, Yonsei University, 2000.

[19] T.-F. Wu, C.J. Lin and R. C. Weng Probability Estimates for Multiclass Classification by Pair wise Coupling", Journal of Machine Learning Research, 2004

[20] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[21] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[22] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[23] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[24] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender