Emotion state detection via speech in spoken Hindi

Group 15: Prakhar Kulshreshtha (13485), Soumya Gayen (13708)
MENTOR: Prof Piyush Rai
CS365: Introduction to AI Programming, IIT Kanpur

Introduction

- Emotion recognition from human speech is an important field of Digital Signal Processing as well as AI. Emotion content from speech can be extracted using phonetic features or prosodic features. Since same phrases can have different emotions when spoken differently, we discard the phonetic features and try to build a classifier solely based on prosodic features.
- We explore:
  - Various classifiers to classify human speech into 8 categories anger, fear, disgust, happiness, surprise, neutral, sadness and sarcastic
  - different features like MFCC, SSC, spectral energies, etc.

Dataset and Classifiers

- Dataset - Subset of IITKGP-SEHSC: Simulated Emotion Hindi Speech Corpus
  - 2 speakers - 1 male, 1 female
  - 15 sentences X 8 emotions X 10 separate sessions = 1200 utterances
- Classifiers employed
  - KNearestNeighbours
  - SVMs(Linear, Polynomial, RBF kernels)
  - Random Forests
  - Adaptive Boosting with decision tree stumps
  - Neural Network(MLP, Deep Neural Net)

Features

- MFCC and SSC as features
- MFCC - Mel Frequency Cepstrum Coefficients
  - Power Cepstrum:

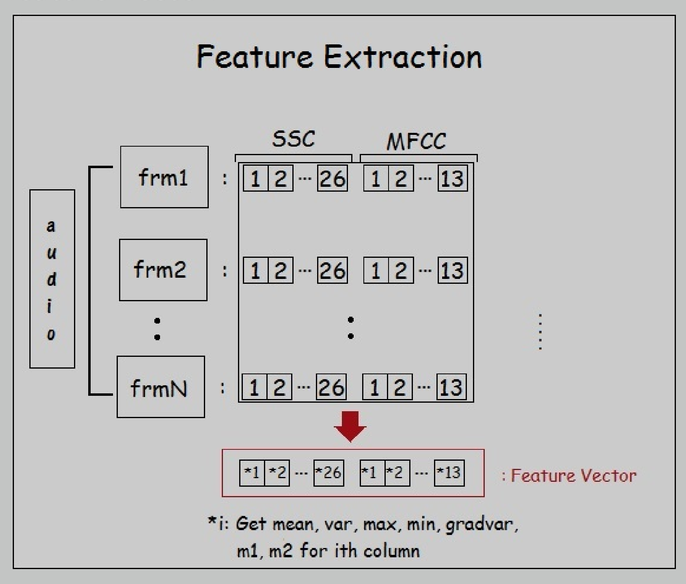
$$Cepstrum = \|F^{-1}\{log(\|F\{f(t)\}\|^2)\}\|^2$$

  - MFC (Mel Frequency Cepstrum) -Cepstrums spaced on the mel scale of frequency, which approximates the human auditory frequency response.
  - MFCC (MFC Coefficient) - 13 Coefficients to characterise MFC in a finite length audio frame.
- SSC - Spectral Subband Centroid

$$SSC(i) = \frac{\sum_{k=1}^{n} f_i(k)x_i(k)}{\sum_{k=1}^{n} x_i(k)}, i = 1 to 26$$

where i is the subband number, $f_i(k)$ are frequencies corr. to that band, and $x_i(k)$ are the power coefficient of that frequency.

- Calculate the two Coefficients for each frame, and then obtain the final representative feature vector:

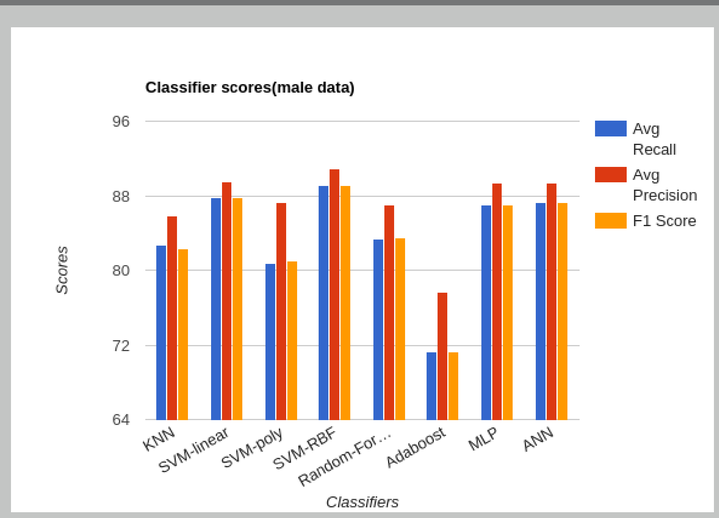


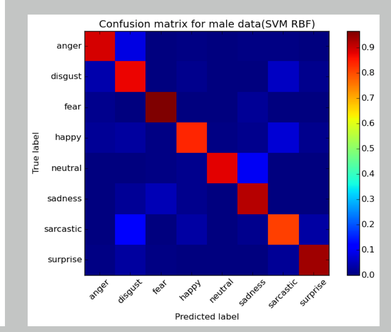
Hence we have 7X(13+26) = **273** features per utterance.

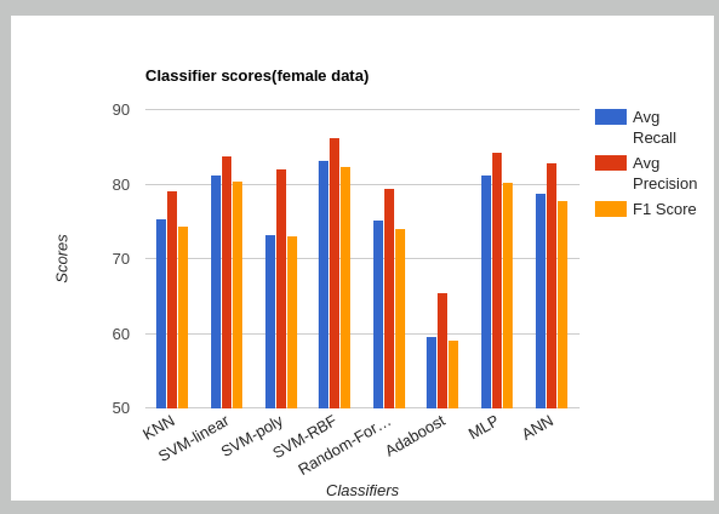
Evaluation Method:

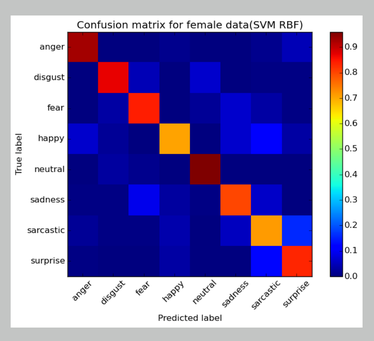
- 'unrandomized' K-Fold Cross Validation.
- For one speaker, 15 sentences, and per sentence 8 emotions in 10 sessions
- Divide dataset into 15 folds, each fold containing $\mathbf{8 \times 10 = 80}$ utterances corresponding to one sentence.
- Then simply test the Classifier for each fold (after training it on the remaining 14 folds).

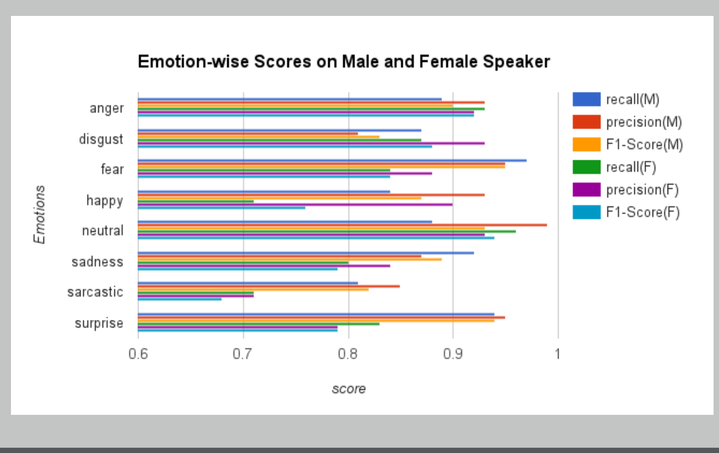
Results

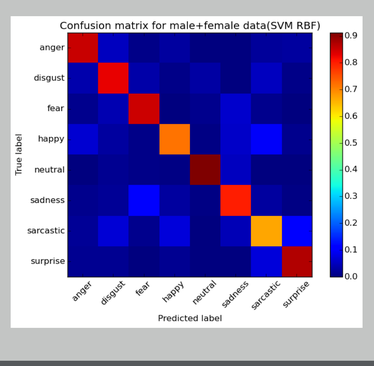












Conclusions

1. It is indeed possible for an AI system to recognize emotion from a spoken utterance, even if the system doesn't understand the meaning of the utterance at all.
2. Even basic classifiers are performing far better than the random guess (100/8 = 12.5%).
3. Out of all the classifiers we tested the best performance was given by SVM with RBF Kernel.

Classifier	Accuracy [%]	
	male	female
IIT-KGP SEHSC, Spectral Features, (overall avg)	77.38	80.75
IIT-KGP SEHSC, 32 Centered GMM, Session-wise Cross Validation, Speaker-7	87.22	-
Ours, SVM, Session-wise CV, speaker 3,4	84.42	82.58
Ours, SVM, Utterance-wise CV, speaker 3,4	**89.08**	**83.16**

4. When we fused male and female data into one dataset, 15 fold CV gives avg accuracy of **85.50%**. So if given a large no. of speakers, our system should be able to become speaker independent.

References

1. Advances in Multimedia Information Processing â PCM 2002: Third IEEE - Google Books
2. IITKGP-SEHSC : Hindi speech corpus for emotion analysis Shashidhar G. Koolagudi, Ramu Reddy, Jainath Yadav, K. Sreenivasa Rao, IIT-KGP
3. Python Speech features, explained and implemented, by James Lyons
4. SPECTRAL SUBBAND CENTROID FEATURES FOR SPEECH RECOGNITION Kuldip K. Paliwal School of Microelectronic Engineering Griffith University Brisbane, QLD 4111, Australia
5. Emotion Recognition in Speech Using MFCC and Wavelet Features, K.V.Krishna Kishore, P. Krishna Satish, Computer Science and Engineering Vignan University
6. Emotion and Gender Recognition of Speech Signals Using SVM S.Sravan Kumar, T.RangaBabu
7. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011 Christos-Nikolaos Anagnostopoulos Â· Theodoros Iliou Â· Ioannis Giannoukos
8. Poster Template from - **http://www-i6.informatik.rwth-aachen.de/ dreuw/latexbeamerposter.php**