Improving Fluency using Sentiment Analysis on Indian Languages

CONTENTS

- Problem Statement
- Literature Survey
- DataSet & Feature Preprocessing
- ML Methodology
- Performance Metrics
- Deployability

The Problem

Same Emotion; Different Expressions

Each language has its own unique way to express an emotion. Along with vocabulary, enunciation plays a key role. Can we use ML to bridge this enunciation and pronunciation gap?

Thus helping new speakers, aspiring artists, and bridging cross-cultural language gaps.

We aim to bridge this gap by developing an ML model to classify sentiments in Hindi speech and give feedback to correct any issues in the tonality or voice modulation etc.

Scherer, Klaus & Banse, Rainer & Wallbott, Harald. (2001). Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. Journal of Cross-cultural Psychology - J CROSS-CULT PSYCHOL. 32. 76-92. 10.1177/0022022101032001009.

Literature Review

1) IITKGP-SEHSC : Hindi speech corpus for emotion analysis

Objective:

- To develop a Hindi Speech Corpus for analyzing and recognizing emotions in speech signals and support emotion aware speech system such as speech recognition, human-computer interaction and speech synthesis
- Analyze how different speech features help in recognizing emotion.

Dataset Overview:

- 8 Emotion classes: Anger, Disgust, Fear, Happy , Neutral, Sad, Sarcastic, Surprise
- 10 speakers, 5 Male, 5 Female
- 12,000 total utterances worth 9 hours of recording

S. G. Koolagudi, R. Reddy, J. Yadav and K. S. Rao, "IITKGP-SEHSC : Hindi Speech Corpus for Emotion Analysis," 2011 International Conference on Devices and Communications (ICDeCom), Mesra, India, 2011, pp. 1-5, doi: 10.1109/ICDECOM.2011.5738540

Inference:

The authors used

I) SVM with an accuracy of ~75%

II)GMM with an accuracy of ~79%

Gap Identified: This paper just did sentiment analysis on the emotions and did not concern itself with giving feedback to the user

2)MNITJ-SEHSD : Hindi Emotional Speech Database

Objective:

- Improve Speech Emotion Recognition (SER) for Human-Computer Interaction
- Test different ML methodologies and different features to validate which work best for speech analysis
- Compilation of a Hindi dataset for future research

Dataset:

- 5 Emotion classes: Anger, Fear, Happy, Neutral, Sad
- 10 speakers, 5 Male, 5 Female
- 500 total utterances using neutral sentences

Overview:

Used three models with different parameters

- **Model I**
  - Prosidic Features
  - E.g energy, pitch,duration & zero-crossing rate
  - Used an SVM with RBF Kernel
- **Model II - Spectral Features**
  - Spectral Features
  - Extracted MFCC,Mel spectogram & chromogram
  - Total 180 dimensional vector fed to an SVM
- **Model III**
  - A five layer deep CNN
  - Log-Mel-Spectrogram given as input

K. Chauhan, K. K. Sharma and T. Varma, "MNITJ-SEHSD: A Hindi Emotional Speech Database," 2023 International Conference on Communication, Circuits, and Systems (IC3S), BHUBANESWAR, India, 2023, pp. 1-6, doi: 10.1109/IC3S57698.2023.10169497

Inference:

The authors used

I) SVM on prosidic features with an accuracy of ~57%

II)SVM on spectral features with an accuracy of ~90%

III) A 5 layer deep CNN with an accuracy of 87%

Gap Identification: No feedback mechanism was employed to help the user improve

The Gap

Currently, no research exists that gives feedback to users using sentiment analysis, especially in the case of regional languages like hindi.

Dataset Acquisition And Description

Speech Emotion Recognition Hindi

Nature of Dataset	• Contains 8 emotions (i.e. Happy , Sad,Angry,Disgust,Fear,Sarcastic,Surprise,Neutral) • Consists of 3200 total Utterances • Comes to 400 speech files per emotion	• Authored by Vishal Bhardwaj (Asst.Prof) • From Govt. Lahiri PG College Chirmiri • Falls under The Open Database License (ODbL) license
Quantitative Study	• Comprises of 4 males & 4 females actors giving a balanced datset • Each actors voiced 10 sentences in all the emotions	• Examples include : ◦ मुझे अच्छे अंक लाने हैं। ◦ सूरज एक अच्छा विद्यार्थी है।
Data Collection & Ethical Concerns	• All the actors were paid fairly for their work • The names & age are also given with the Dataset	• The collection was done in the Department of Computer Science And Application (Lahiri PG College) • No source for the funding is given

https://www.kaggle.com/datasets/vishlb/speech-emotion-recognition-hindi/data

Feature Extraction And Preprocessing

Extraction

SVM - Using Librossa(python library):-

Extracted mean and std of

- 40 MFCC Features
- 40 Mel-Spectogram Features
- 12 Chromograms
- 4 Formants (Pitch,Lip Rounding etc)
- 8 Prosidic Features including speaking rate

BiLSTM - Using Librossa(python library):-

Extracted time series data for

- 40 MFCC Features
- 40 Mel-Spectogram Features
- 12 Chromograms
- 4 Formants (Pitch,Lip Rounding etc)
- 8 Prosidic Features including speaking rate

Preprocessing

- Standardised the dataset using StandardScalar which converts everything into a range b/w [-1,1]
- Added padding on low duration samples to make the duration same for all (LSTMs)

Some Extracted Features (Before Pre-Processing)

mfcc_1_mean	mfcc_2_mean	mfcc_3_mean	mfcc_4_mean	mfcc_5_mean	mfcc_6_mean	mfcc_7_mean	mfcc_8_mean
-314.743439	99.861763	-4.827002	18.381813	-29.213797	-7.485608	-4.269964	-12.223748
-334.666046	112.437325	-10.156618	18.329895	-27.459194	-7.406189	-4.707757	-13.640003
-327.535522	103.649338	-6.033504	18.470753	-25.512444	-7.880752	0.860493	-12.586468
-307.994873	99.186653	-2.461936	20.922201	-29.544270	-9.691394	0.989803	-13.275464
-314.492249	101.228951	-1.076013	16.390169	-21.410463	-10.601329	1.424581	-12.769308

rms_mean	rms_std	zcr_mean	zcr_std	f0_mean	f0_std	F1	F2	F3	emotion
0.083147	0.089705	0.105895	0.056749	948.319153	758.237732	384.367279	725.533325	2990.557617	anger
0.072900	0.079170	0.088251	0.040122	865.732788	660.807251	0.000000	0.000000	463.995117	anger
0.093216	0.117069	0.089196	0.051633	938.010620	775.331055	317.746765	1053.976440	2941.628418	anger
0.125638	0.132192	0.089084	0.037741	1080.873047	939.845032	0.000000	0.000000	336.284821	anger
0.111746	0.135845	0.081196	0.032735	836.752197	572.179871	298.772339	1254.057739	3020.142578	anger

Methdology

Support Vector Machines (SVM)

- Simple SVM used to classify the emotion based on the distance of their features
- Kernel: “RBF”
- C: “10”
- gamma: “0.01”
- An accuracy of 80% was achieved
  (These values were reached through grid search)
- Intuition: Similar emotions would also have similar clusters for their extracted features
- Used one vs rest classification technique

BiLSTM From Scratch

- Trained a BiLSTM on Hindi Data
- 128 units, 0.3 dropout, 32 batch size, 0.001 LR
- Accuracy of 74% was achieved
- LSTM did not perform well due to lack of data

BiLSTM with Transfer Learning

- Trained a BiLSTM on english data
- Applied transfer learning by freezing LSTM layers and fine tune final layer
- An accuracy of ~60% was achieved
- Did not work well because emotion expression in English and Hindi is very different

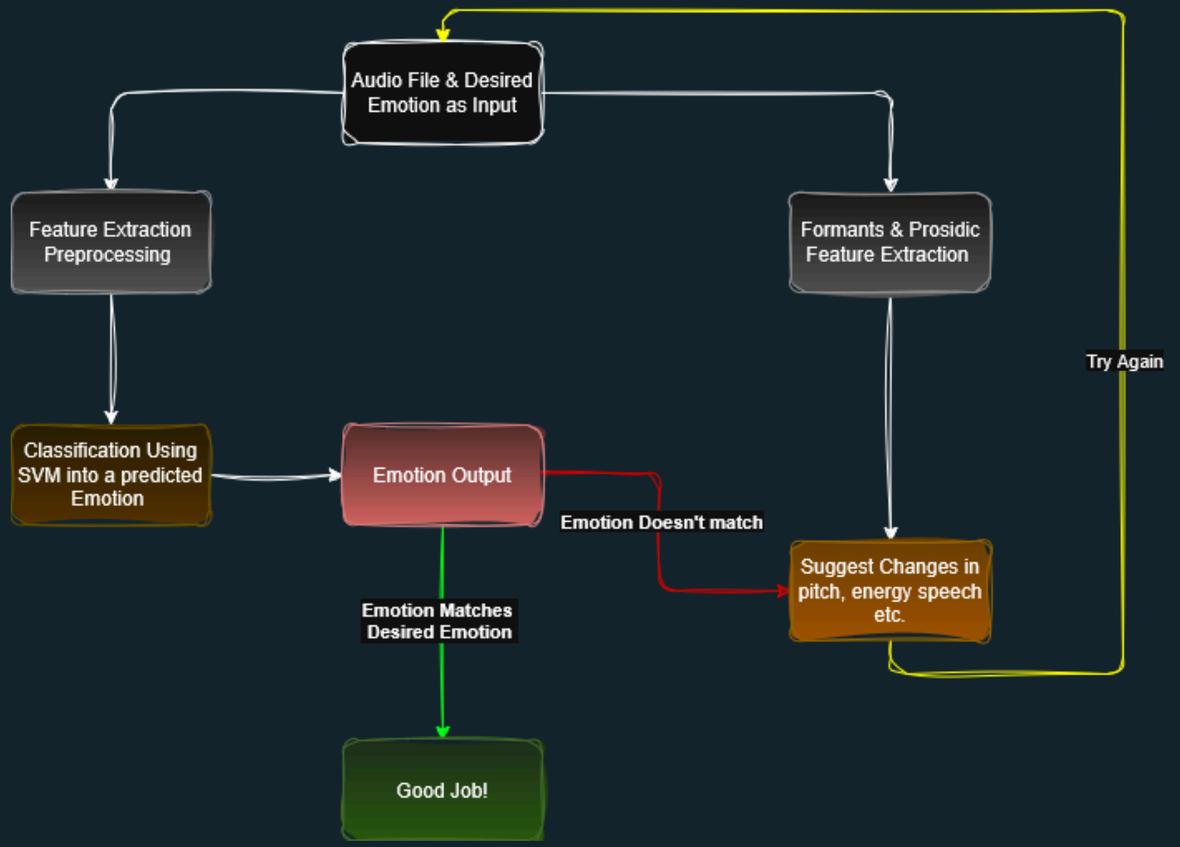
Boulmaiz, T., Guermoui, M. & Boutaghane, H. Impact of training data size on the LSTM performances for rainfall–runoff modeling. Model. Earth Syst. Environ. 6, 2153–2164 (2020). https://doi.org/10.1007/s40808-020-00830-w

Voting Classifier

- Trained an SVM , BiLSTM & random forest classifier
- An accuracy of 75% was achieved with this classifier
- Interpretation
  - The accuracy comes out as the average of all the classifiers which could be due to not enough data being present to feed every model

Feedback System

- Statistics based feedback
- Take the delta between input audio features and avg features of target emotion
- Suggest changes based on delta



```
Target Emotion    : happy
Predicted Emotion : sad
[ENERGY] Good
- Your vocal energy matches the target emotion.

[F1 - VOCAL OPENNESS] Too Closed
- Your articulation sounds too closed.
- Open your mouth more while speaking.
- Use stronger vowel projection.

To sound more 'happy', focus on:
- Lower your overall pitch
- Stabilize your pitch variation
- Open your mouth more for clearer vowels
- Use a brighter, more forward tongue position
- Relax facial tension and vocal strain
```

The current feedback comments mainly aim to correct the sound of the vowel via the shape of the mouth

We will need a specialists(speech therapist) input for more accurate and helpful feedback while this serves as a proof of concept

Challenges

Dataset Availability

- The biggest hurdle encountered was Dataset Availability
- The professors from IITKGP could not be contacted even after countless attempts
- A delayed reply from the authors of HindiSER and unforeseen circumstances prevented us from obtaining the dataset in time

Model/Software Issues

Due to the previous mentioned constraints on the data availability, the ability of Neural Nets & LSTM were heavily hampered leading to poor performance

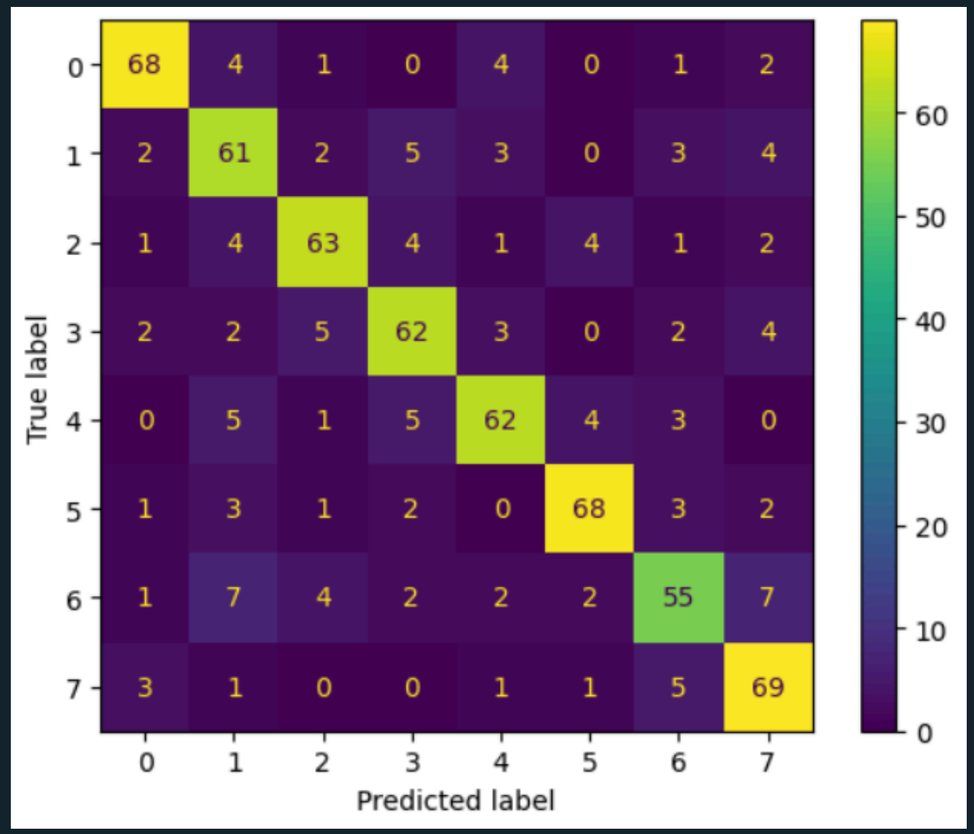
Better Feedback suggestions

Although we have suggested some common changes;For more personalized and accurate feedback we would need to consult with a professional speech therapist

Performance Metrics

SVM Classification Scores

	precision	recall	f1-score	support
0	0.87	0.85	0.86	80
1	0.70	0.76	0.73	80
2	0.82	0.79	0.80	80
3	0.78	0.78	0.78	80
4	0.82	0.78	0.79	80
5	0.86	0.85	0.86	80
6	0.75	0.69	0.72	80
7	0.77	0.86	0.81	80
accuracy			0.79	640
macro avg	0.80	0.79	0.79	640
weighted avg	0.80	0.79	0.79	640



Note:- After removing the emotion “Sarcasm” the accuracy of the model jumped to around ~85% from 80%

Feedback System Metric

- While using the feedback to classify the voices present in the dataset, The accuracy came out as expected
- When a real world voice input was given the accuracy dropped to ~30%

Classification Report:

	precision	recall	f1-score	support
anger	0.52	0.24	0.33	50
disgust	0.22	0.36	0.27	50
fear	0.21	0.26	0.23	50
happy	0.26	0.36	0.31	50
neutral	0.30	0.22	0.25	50
sad	0.26	0.20	0.23	50
sarcastic	0.27	0.16	0.20	50
surprise	0.51	0.62	0.56	50
accuracy			0.30	400
macro avg	0.32	0.30	0.30	400
weighted avg	0.32	0.30	0.30	400

Interpretation

- We strongly believe that this can be for two major reasons
  - The Dataset is too small
  - Our recording environment was not properly tuned to the enviorment present in the Dataset recordings.

Deployability

- In its current state, the model is not robust enough for deployment
- We hypothesize that, given a better dataset, this model can be made fit for deployment
- To make this a better dataset, it would need to be developed in-house
- The main challenge with scalability for this kind of project is the various regional accents present in India
- Other challenges include avoiding regional bias, gender bias, and quality of the recorded sound.

Thank You!