Emotion Detection in Speech

Mohammad Saad Khan (20162043)
Damodar Pai Dukle (20162055)

Abstract

In this project we explore the problem of Emotion recognition using a combination of Mel-frequency cepstral coefficients (MFCC) , Linear Prediction Cepstral Coefficients (LPCC) and frame level Prosodic features. We use various models for this classification task like Gaussian mixture models (GMM) , Deep Neural Network (DNN) and a joint model of conventional features and neural embeddings. We use two datasets, namely RAVDESS and EMODB; and report results for these. In the experiments, we observe that DNN outperforms GMM . The best result on RAVDESS dataset is obtained as 68.40% using a DNN and on EMODB dataset as 68.76% using a DNN.

Datasets

Berlin Database of Emotional Speech (EMODB)

- **Language**: German
- **Speakers**: 10 (5 Male, 5 Female)
- **Sentences**: 10 per speaker
- **Emotions**: anger, boredom, disgust, fear, happiness, sadness and neutral
- **Training data**: 8 speakers(4 male,4 female) per emotion
- **Testing data**: 2 speakers(1 male,1 female) per emotion

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

- **Language**: English
- **Speakers**: 24 (12 male, 12 female)
- **Sentences**: 2 per speaker with 2 repetitions
- **Emotions**: neutral, calm, happy, sad, angry, fearful, surprise, and disgust (with varying intensity)
- **Training**: 18 speakers (9 male, 9 female) per emotion
- **Testing**: 6 speakers (3 male, 3 female) per emotion

Preprocessing

Following steps describe preprocessing done to all speech files used for the experiments

1. Convert speech file to .wav extension
2. Use Audio Mixer software(SOX) to change frame rate to 8KHz
3. Convert file to single channel mono
4. Make Frames from speech file using window_size = 20ms with overlap =10ms using hamming window
5. For each frame generate a one hot representation of the emotion label

Feature Extraction

MFCC features

1. Mel Frequency Cepstral Coefficients :

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum

MFCCs are commonly derived as follows:

- Take the Fourier transform of (a windowed excerpt of) a signal.
- Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
- Take the logs of the powers at each of the mel frequencies.
- Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.

2. Delta and Delta-Delta Coefficients : Also known as differential and acceleration coefficients. The MFCC feature vector describes only the power spectral envelope of a single frame, but speech also has information in the dynamics i.e. what are the trajectories of the MFCC coefficients over time, these are the delta coefficients
   To calculate the delta coefficients, the following formula is used:

$$\boxed{d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}}$$

where d_t is a delta coefficient, from frame t computed in terms of the static coefficients c_{t+N} to c_{t-N}. A typical value for N is 2.

3. Delta-Delta (Acceleration) coefficients are calculated in the same way, but they are calculated from the deltas, not the static coefficients.

LPCC features

LPCCs are used to capture emotion-specific information manifested through vocal tract features. We carry out the 10th order LP analysis on the speech signal, to obtain 10 LPCCs per speech frame of 20 ms using a frame shift of 10 ms. The human way of emotion recognition depends equally on two factors, namely: its expression by the speaker as well as its perception by a listener. The purpose of using LPCCs is to consider vocal tract characteristics of the speaker, while performing automatic emotion recognition.

Prosodic features

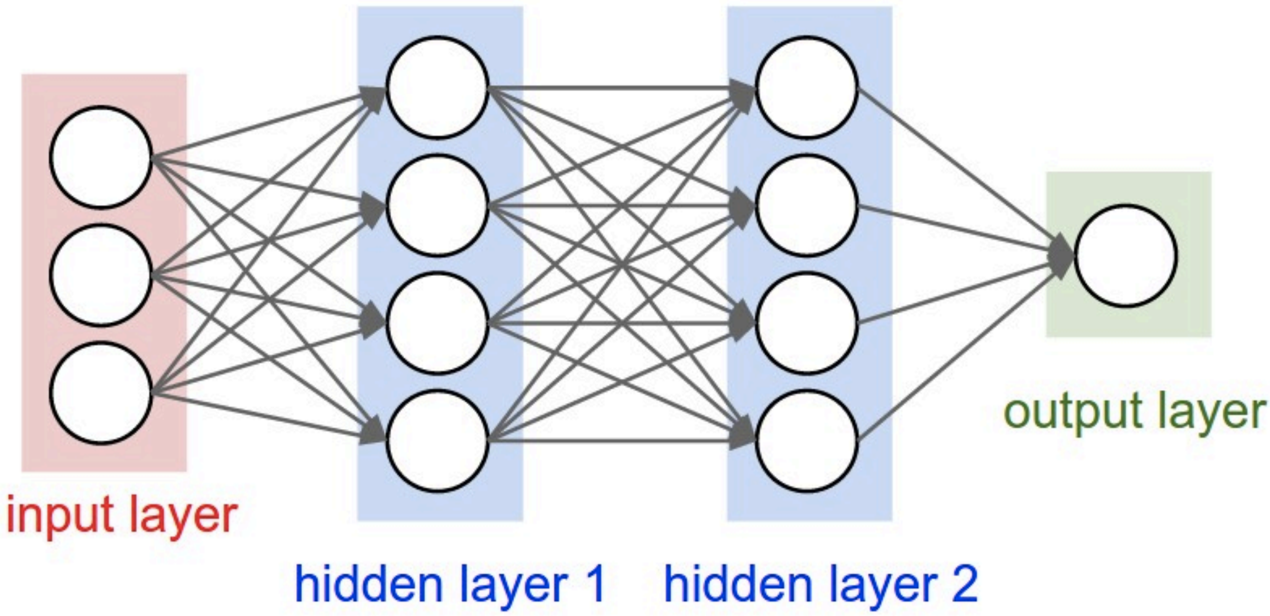
We have used the following five frame level prosodic features:

1. **Pitch Contour**:
   - Pitch contour captures the characteristics that are pertaining to articulation
   - From each frame, the pitch value is calculated using the Autocorrelation method. So, for a given speech sample we get a vector of dimension equal to number of frames.
2. **Energy Contour**:
   - Energy contour captures stress patterns in speech. The energy valleys in speech serve as delimiters for phonemes or vowels in speech.
   - The energy contour is obtained by tracking the variation in amplitude of the signal over time. We take the Root Mean Square (RMS) value of amplitudes in a frame. Doing so yields one value per frame which form a vector.
3. **Zero-crossing rate contour** :
   - Zero crossing rate (ZCR) is the rate of sign changes along a signal. It involves finding the number of sign changes within a frame. So, we get one value per frame which contributes to another vector.
4. **Number of Epochs per frame**:
   - Epochs are instants of significant excitation of the vocal-tract system.
   - We extract epoch locations using DYPSA (dynamic programming projected phase-slope algorithm) algorithm. Then we count the number of epochs in a given frame and construct a vector.
5. **Mean amplitude of epoch locations** :
   - After getting the epoch locations in a frame, we calculate the amplitude and take the mean. Since we get one value per frame, this feature also gives a number of frames sized vector.

After extraction of 39 MFCC, 10 LPCC and 5 Prosodic features, we combine them to get a matrix of **(n x 54)** where n is the number of frames in the signal.

DNN Model Description

Architecture



- We use a Feed forward network with 2 hidden layers.
- The input layer will have number of nodes equal to number of features which is 54 in our case.
- The output layer will have number of nodes equal to the number of emotions.
- The number of nodes in the hidden layers is a hyper parameter and is varied to get the best result.

- Metrics :

  - In our problem, we label all the frames for a single file with the same label as the emotion label of the file. Since emotion labels signify the emotion of the full file and not the individual frames, we often separate the Accuracy into 2 metrics

    - Frame Accuracy = Correct Frames / Total Frames
      - ❖ It tells how many frames were correctly classified by the model

    - Test Accuracy = Correct Tests / Total Tests
      - ❖ It tells how many labels in test data were predicted correctly
      - ❖ A test is Correct if the majority of frames in that file are predicted correctly

Observations and Results

RAVDESS DATASET

- **GMM with MFCC features (8 emotions)**

Expt 1 : Increasing Number of Mixtures

n_mixtures	Accuracy	Other Parameters	
8	39.84375	window	0.03
16	41.9270833333	window_overlap	0.015
32	**45.3125**	voiced_threshold_mul	0.05
64	45.3125	voiced_threshold_range	100
128	44.53125	max_iterations	100
256	41.6666666667	calc_deltas	FALSE

Expt 2: Calculating Delta features

calc_deltas	n_mixtures	Accuracy	Other Paramet
TRUE	16	35.67 (did not converge)	window
TRUE	32	43.75	window_overlap
TRUE	64	45.3125	voiced_threshold_mul
TRUE	**128**	**46.09375**	voiced_threshold_range
TRUE	256	43.2291666667	max_iterations
TRUE	512	40.3645833333	

Expt 3: Increasing Iterations

max_iterations	Accuracy	Other Parameters	
200	41.9270833333	window	0.03
400	**45.3125**	window_overlap	0.015
1000	42.96875	voiced_threshold_mul	0.05
2000	45.3125	voiced_threshold_range	100
		n_mixtures	128
		calc_deltas	TRUE

- **GMM with MFCC, LPCC and 5 Prosodic features (4 emotions)**
    - Emotions considered: Angry, Sad, Happy, Neutral

With Delta features

n_mixtures	Accuracy
8	54.27083333334
16	57.11805555556666
32	60.416666666666664
64	61.80555555556666
128	**63.19444444446666**

Without Delta features

n_mixtures	Accuracy
8	63.47031963466666
16	58.561643835575
32	63.0136986301
64	64.38356164380001
128	**64.3835616438**

- **DNN with MFCC, LPCC and 5 Prosodic features (4 emotions)**
    - Prosodic features used : Pitch , Energy, ZCR and Epoch number
    - Nh1 - Number of nodes in Hidden layer 1
    - Nh2 - Number of nodes in Hidden layer 2
    - b_size - Batch size
    - n_epochs - number of epochs for training
    - Optimiser - Keras optimiser used

Nh1	Nh2	b_size	n_epochs	Optimiser	Test Accuracy	Frame Accuracy
20	10	20	50	Adamax	51.0416666667	25.9713440
20	10	50	50	Adamax	66.40625	28.971344
10	10	50	50	Adamax	67.968	26.9297331
30	15	50	50	Adamax	64.01736	43.79181
30	15	50	30	Adamax	67.838541	30.778849
35	**12**	**50**	**50**	**Adam**	**68.40277728**	**32.69590**
40	20	50	50	Adamax	64.3835616	43.86560

EMODB DATASET

- **GMM with MFCC, LPCC and 3 Prosodic features (5 emotions)**

With Delta features

n_mixtures	Accuracy
8	54.1666666667
16	57.2916666667
24	**59.375**
32	55.2083333333
64	58.3333333333
128	58.3333333333

Without Delta features

n_mixtures	Accuracy
8	52.0833333333
16	**57.2916666667**
24	50.0
32	55.2083333333
64	46.875
128	52.0833333333

- **GMM with MFCC, LPCC and 3 Prosodic features (5 emotions)**
  - Prosodic features used : Pitch , Energy and ZCR

With Delta features

n_mixtures	Accuracy
8	49.1666666668
16	55.83333333332
24	57.70833333332
32	56.25
64	57.91666666666
128	**60.625**

Without Delta features

n_mixtures	Accuracy
8	53.33333333332
16	52.91666666668
24	53.95833333334
32	55.0
64	53.749999999980005
128	**56.041666666640005**

- **GMM with MFCC, LPCC and 5 Prosodic features (4 emotions)**
  - Prosodic features used : Pitch , Energy, ZCR, Epoch number, Mean epoch amplitude

With Delta features

n_mixtures	Accuracy
8	59.45205479448
16	65.75342465749999
32	68.4931506849
64	**69.58904109586**
128	68.4931506849

Without Delta features

n_mixtures	Accuracy
8	63.47031963466666
16	58.561643835575
32	63.0136986301
64	64.38356164380001
128	**64.3835616438**

- **DNN with MFCC, LPCC and 4 Prosodic features (4 emotions)**
  - Prosodic features used : Pitch , Energy, ZCR and Epoch number
  - Nh1 - Number of nodes in Hidden layer 1
  - Nh2 - Number of nodes in Hidden layer 2
  - b_size - Batch size
  - Optimiser - Keras optimiser used

Nh1	Nh2	b_size	Optimiser	Test Accuracy	Frame Accuracy
30	12	40	Adam	60.2739726027/	42.1635182999
35	**12**	**50**	**Adam**	**68.76712328764**	**42.4881936246**
40	12	40	Adam	65.7534246575	43.9443132625
35	12	40	Adamax	63.01369863010001	42.9878000787
20	10	50	Adamax	64.15525114151666	42.7292404565
20	10	50	Adamax	60.958904109	40.0875639512
35	15	50	Adam	61.6438356164	43.9787485242
10	10	50	Adamax	60.2739726027	39.1774891775

Software and System requirements

Tools

- MATLAB - extraction of Prosodic features
- SCILAB - extraction of LPCC
- Python 2.7 - MFCC extraction and model implementation

Libraries

- VoiceBox - DYPSA for epoch features
- pyAudioAnalysis - MFCC features
- Keras - DNN implementation
- Sklearn - GMM impelentation
- Numpy

References

- Shashidhar G. Koolagudi1 · Akash Bharadwaj1 ·Y. V. Srinivasa Murthy1 · Nishaanth Reddy1 · Priya Rao1(2017). Dravidian language classification from speech signal using spectral and prosodic features. Springer
- K.S. Rao et al., Language Identification Using Spectral and Prosodic Features, SpringerBriefs in Speech Technology, DOI 10.1007/978-3-319-17163-0

Links

- http://iitg.vlab.co.in/?sub=59&brch=164
- https://github.com/tyiannak/pyAudioAnalysis