



# Emotion Recognition Using Prosodic Features

Hasir Mushtaq - 2018102049  
Dolton Fernandes - 2018111007

Mentor: Sparsh Garg

## Dataset:



- [A Database of German Emotional Speech](#)
- Ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and are interpretable in all applied emotions.
- Emotions Present in the recordings: neutral (neutral), anger (Ärger), fear (Angst), joy (Freude), sadness (Trauer), disgust (Ekel) and boredom (Langeweile).
- Recordings were taken with a sampling frequency of 48 kHz and later downsampled to 16 kHz.

## Dataset: Audio Labelling:

Text Utterance:

| code | text (german)  | try of an english translation   |
|------|--|---|
| a01  | Der Lappen liegt auf dem Eisschrank.   | The tablecloth is lying on the fridge.                                    |
| a02  | Das will sie am Mittwoch abgeben.  | She will hand it in on Wednesday.   |
| a04  | Heute abend könnte ich es ihm sagen.   | Tonight I could tell him.   |
| a05  | Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.               | The black sheet of paper is located up there besides the piece of timber. |
| a07  | In sieben Stunden wird es soweit sein.   | In seven hours it will be.  |
| b01  | Was sind denn das für Tüten, die da unter dem Tisch stehen?                        | What about the bags standing there under the table?                       |
| b02  | Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.                | They just carried it upstairs and now they are going down again.          |
| b03  | An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht. | Currently at the weekends I always went home and saw Agnes.               |
| b09  | Ich will das eben wegbringen und dann mit Karl was trinken gehen.                  | I will just discard this and then go for a drink with Karl.               |
| b10  | Die wird auf dem Platz sein, wo wir sie immer hinlegen.                            | It will be in the place where we always store it.                         |

Emotion:

| letter              | emotion (english) | letter | emotion (german) |
|---------------------|-------------------|--------|------------------|
| A                   | anger             | W      | Ärger (Wut)      |
| B                   | boredom           | L      | Langeweile       |
| D                   | disgust           | E      | Ekel             |
| F                   | anxiety/fear      | A      | Angst            |
| H                   | happiness         | F      | Freude           |
| S                   | sadness           | T      | Trauer           |
| N = neutral version |                   |        |                  |

## Dataset:



Percentage of files for each emotion:

The different emotions are not equally represented, so this is one thing we need to take care about.

| Emotion | % of Dataset    |
|---------|-----------------|
| Anger   | 28              |
| Disgust | 8               |
| Others  | Approx. 12 each |

## About the Task (Emotion Detection using speech signal):



- Emotion classification is one of the most challenging tasks in a speech signal processing domain. The problem of speaker or speech recognition becomes relatively an easier one when compared with recognizing emotion from speech.
- The basic principle behind emotion recognition lies with analysing the acoustic difference that occurs when uttering the same thing under different emotional situations.
- In addition to the features corresponding to the speaker and/or the speech, the sound signals do have some features that represents the emotional state of the speaker.
- We are interested in this part of the speech.

## Features: 1) Pitch

1. Pitch is the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords.
2. We choose the following pitch features for extraction:

| Feature                |  |
|------------------------|--|
| Minimum                | Maximum value obtained from the pitch contour.   |
| Maximum                | Maximum value obtained from the pitch contour.   |
| Mean                   | Average value obtained over the pitch contour.   |
| Median                 | Median value obtained over the pitch contour.  |
| Standard Deviation     | Standard Deviation extracted from the pitch contour.   |
| Time of Min. & Max.    | Times at which min. and max. occurred in the contour.  |
| First & Third Quartile | Median of the lower & upper half of the signal values sorted in ascending order, respectively. |
| Mean Slope             | Average absolute slope across all turning points in a pitch contour.                           |


## Features: 1) Pitch



On researching for the importance of pitch in emotion detection, we looked at a couple of papers like [this](#) and [this](#) paper. We concluded the following:

- Pitch and log energy are the most considered parameters of emotion and speaking styles evaluation. Experiments confirmed in the above mentioned paper conclude that using pitch information, 75% correct rate can be achieved when their pitch means and variance are observed.
- We notice that neutral, slow and soft styles involve low values of pitch.
- Pitch value increases significantly for loud and angry; average pitch reaches 253 Hz for angry and 216Hz for loud.
- Also mean pitch of anger, fear, and happiness is significantly higher than that of neutrality and sadness.

## Features: 2) Energy



Energy of the speech signal can be calculated as follows (for discrete inputs) :


$$E = \sum X(m)^2$$

We selected the following features for energy feature extraction:

| Feature            |  |
|--------------------|--|
| Min. & Max Energy  | Min. and Max. Energy obtained from the energy function |
| Median             | Median value obtained over the energy function         |
| Mean               | Average value obtained over the energy function        |
| Standard Deviation | Standard deviation over the entire energy function     |
| Variance           | Variance over the entire energy function               |
| Range              | Max. - Min.  |



## Features: 2) Energy



Referring to the same papers as in the pitch slide and this paper, we conclude the following about the usage of energy for emotion recognition:


- Signal energy exhibits higher and sharper peaks in the case of Anger, when comparing this with other emotions. In all of these emotion states, Neutral almost always give the smoothest and lowest curves of the energy contour.
- In case of emotional speech, the signal energies are highest for Fear and lowest for Anger and signal energies for Happy are in between that for Anger and Fear.
- Loud and anger speeches tend to have higher amount of energy levels about 81dB.
- The neutral versions of speech possesses a speech energy of about 78 db.
- Happiness is also found to have significantly higher energy levels surprisingly comparable to that of anger.

## Features: 3) Intensity

- Features we chose to extract: minimum, time of minimum, maximum, time of maximum, first quartile, third quartile, mean, standard deviation.

| Feature:           |  |
|--------------------|--|
| Minimum            | Minimum value of signal.   |
| Time of minimum    | Time instant at which this minimum value is seen.                        |
| Maximum            | Maximum value of signal.   |
| Time of maximum    | Time instant at which this maximum value is seen.                        |
| First Quartile     | Median of the lower half of the signal values sorted in ascending order. |
| Third Quartile     | Median of the upper half of the signal values sorted in ascending order. |
| Mean               | Average of the signal values.  |
| Standard Deviation | Standard deviation of the signal values.                                 |
| Median             | Median of the signal values.   |

## Features: 3) Intensity




**Intensity** is perceived as the loudness of the sound. Variation with different emotions:

We went through this [paper](#) which tells about the importance of intensity on emotional speech.

- The portrayals of the same emotion with different intensity yields different patterns of acoustic cues, including higher voice intensity for the strong emotions than the weak emotions.
- Example: Sadness and Boredom are two similar sounding emotions but vary in the intensity.
- The function of sound intensity is pretty important, as indicated by the fact that people most often report voice cues such as loudness or talking speed to judge the emotional states of others in everyday life.
- Example: Loud voiced signal generally means Anger, Joy, and is hardly seen in the case of emotions like Sadness, Fear, etc.

## Features: 4) Speech Rate



- What is it? - It is a measure of how fast an interlocutor is talking. Popularly measure in words per minute (wpm) & syllables per minute (spm).
- Paper referred to: [The Performance of the Speaking Rate Parameter in Emotion Recognition from Speech](#)
- The speaking rate is the quotient of the number of phonemes (distinct units of sound) in the utterance and the length of the utterance in seconds. Unit - phonemes per second (pps). - We are using this, since phoneme level transcriptions are already provided.
- Relation with emotion ? - Generally speaking, emotions like Fear provide the highest speaking rates, followed by Joy and Anger. Disgust and Sadness show the lowest rates.

## Features:



### Methods to extract:

- Intensity and Pitch features have been extracted using Praat which is a free computer software package for speech analysis in phonetics.
- Energy features have been extracted using a python script.
- Phoneme level transcriptions are provided for each of the audio files in .lablout format. We wrote a python script to calculate number of phonemes and duration of the audio to determine speech rate as  $SR = \text{no\_of\_phonemes} / \text{duration (pps)}$ .

## Model:

|                         |         |             |
|-------------------------|---------|-------------|
| dense_input: InputLayer | input:  | $[(?, 26)]$ |
|                         | output: | $[(?, 26)]$ |

← Input layer of size  $N \times 26$ , where 26 is the size of feature vector.

|              |         |           |
|--------------|---------|-----------|
| dense: Dense | input:  | $(?, 26)$ |
|              | output: | $(?, 17)$ |

← A Dense layer input size 26 and output size 17.

|                |         |           |
|----------------|---------|-----------|
| dense_1: Dense | input:  | $(?, 17)$ |
|                | output: | $(?, 7)$  |

← A Dense layer input size 17 and output size 7 with softmax at the output layer. This is the classification layer which predicts the class. 7 is the number of emotions (classes).

## Training/Testing:



- We concatenated all the features extracted to make a final feature vector of size  $1 \times 26$  for each audio file.
- We split the data into training and testing in 80:20 ratio and stratify it so that data for each class is equally represented.
- We used [categorical cross entropy loss](#) as our loss function.
- We use adam optimizer.
- We trained our model with batch sizes of 5 and for 500 epochs.

## Results:



- After training for 500 epochs we achieved a training accuracy of 73.13%.

```
Epoch 500/500  
86/86 [=====] - 0s 1ms/step - loss: 0.7147 - accuracy: 0.7313
```

---

- After evaluating our model on the testing data, we got a testing accuracy of 67.29%.

```
22/22 [=====] - 0s 972us/step - loss: 0.9048 - accuracy: 0.6729
```



## Results:



### Confusion Matrix:

