

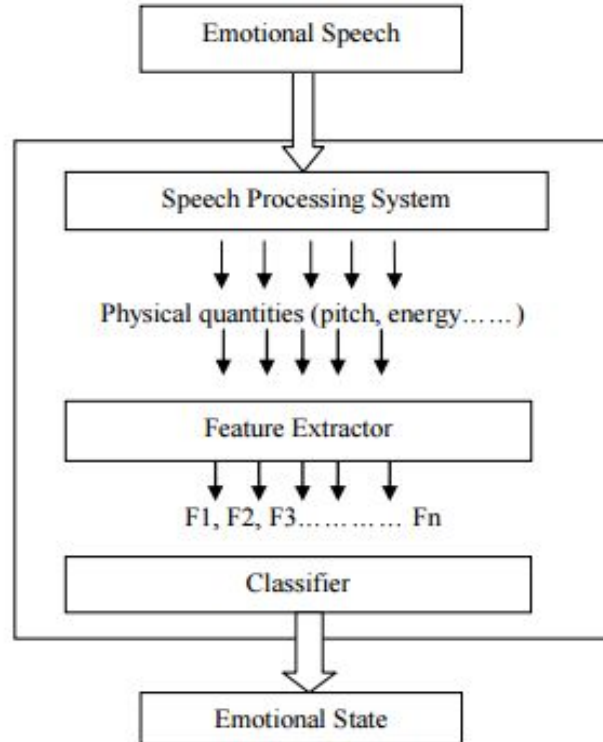
Emotion Orientation from Speech Audio Signal

Navjot Singh, Yash Bhalgat, Kalpesh Patil, Meet Shah

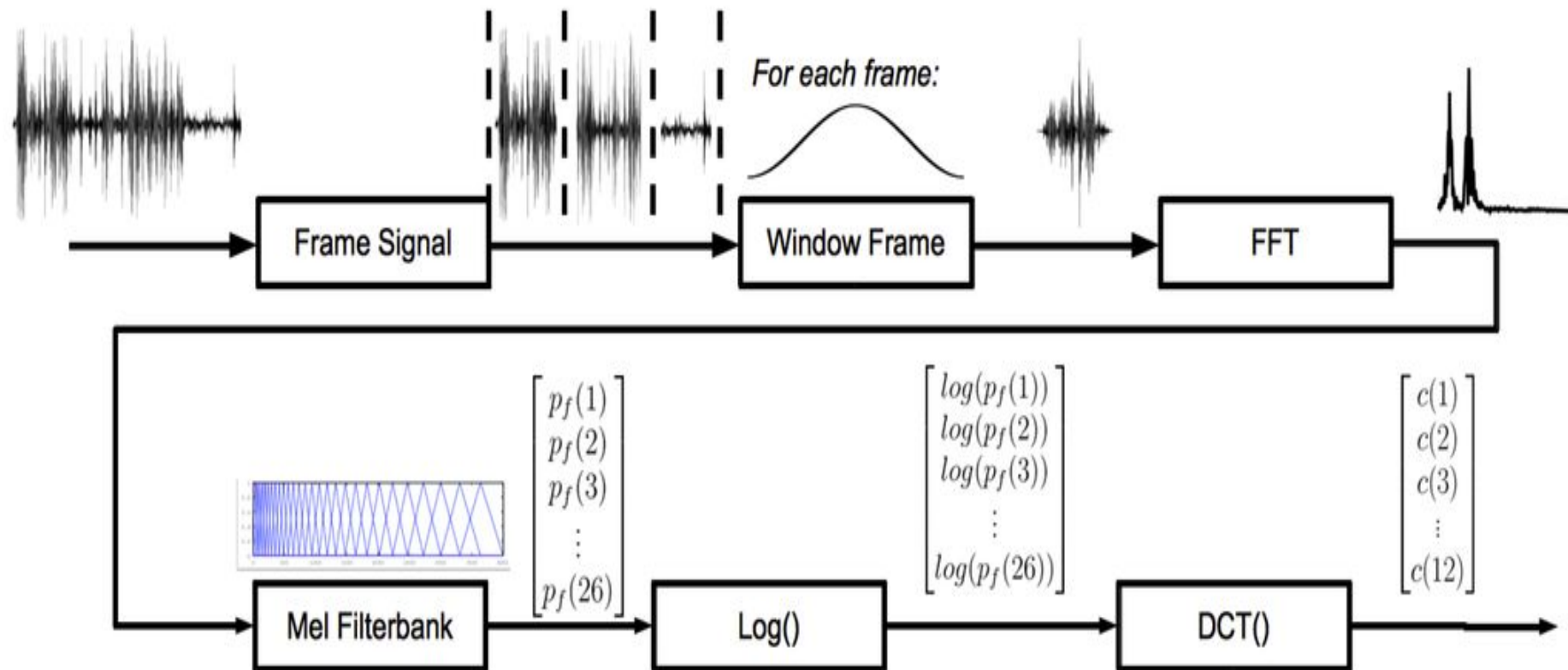
Guide : Prof. Vikram Gadre

GROUP 4

Block Diagram



Block Diagram



Framing the Signal

Signal framed into 20-40ms frames (Standard \rightarrow 25ms)

{Chosen so that the Audio signal does not change much}

\therefore 16 KHz signal $\rightarrow 25 \times 16 = 400$ samples

Frame Step = 10ms (Causes Overlapping of frames)

$s(n) \longrightarrow s_i(n)$ (n : ranges on the number of samples in frame (e.g 400) ;

(time domain signal)

i : denotes the frame number)



Discrete Fourier Transform (DFT)

Calculation of complex DFT for each frame $s_i(n)$ (e.g 400 sample frame)

$$S_i(k) = \sum_{n=1}^N s_i(n)h(n)e^{-j2\pi kn/N} \quad 1 \leq k \leq K$$

N -> length of analysis window ; K -> length of DFT



Periodogram estimate of PSD

- Motivated by the Human Cochlea, which vibrates on different spots depending upon incoming frequency.
- Identification of what frequencies are present in the frame.

$$P_i(k) = \frac{1}{N} |S_i(k)|^2$$

- We would generally perform a 512 point FFT and keep only the **first 257 coefficients**.

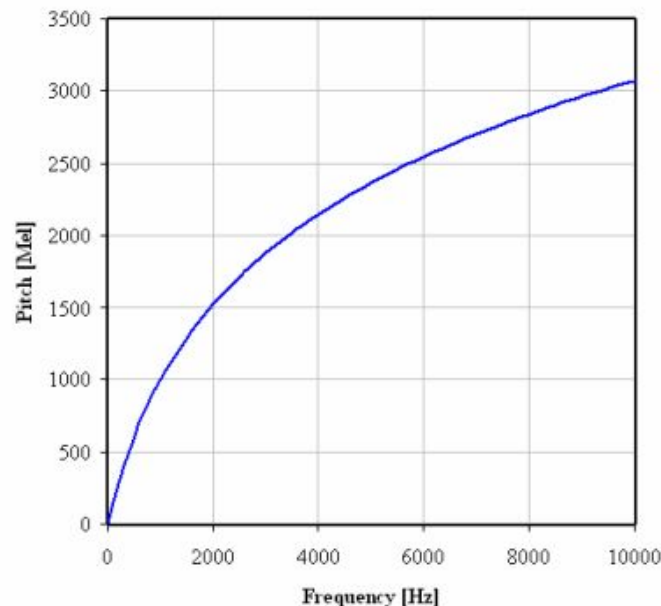


The Mel Scale

- The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency.
- Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies.

$$M(f) = 1125 \ln(1 + f/700) \quad (1)$$

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \quad (2)$$

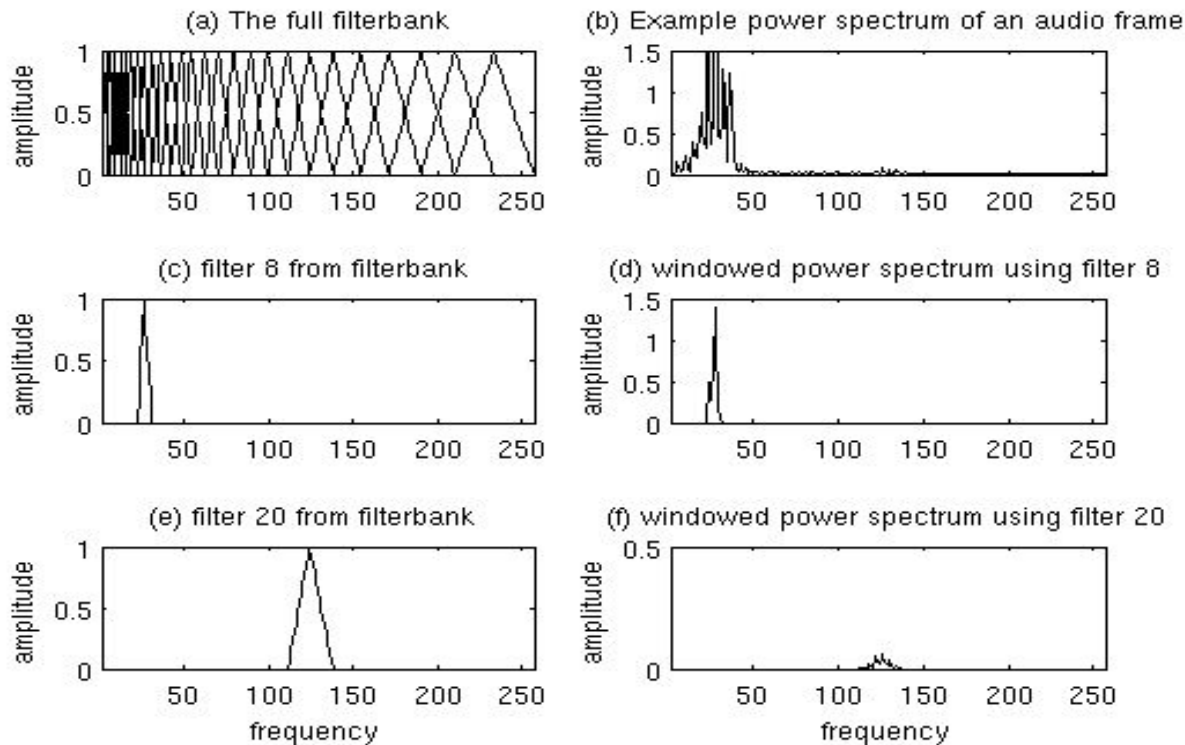


Mel Filter Bank

- Periodogram estimate contains lot of unnecessary information.
- Set of 26 triangular filters applied to periodogram estimate.
- Multiply each filterbank with the power spectrum (size 257), add up the coefficients for each bank separately.
Gives an indication of how much energy was present in each filterbank.
- Vector of size 26 (1 element for each filter bank) is obtained.



Example of Mel Filtering



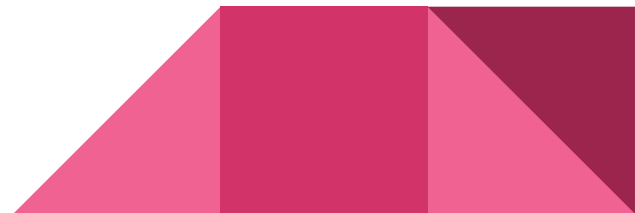
Logarithm and DCT

- Take log of each 26 filterbank energies.
- Take Discrete Cosine Transform (DCT) of this new vector.

The DCT **decorrelates the energies** which means diagonal covariance matrices can be used to model the features (useful in classifier techniques such as HMMs)

Calculation of DCT of N samples.

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right] \quad k = 0, \dots, N-1.$$



Some Refinement

- Higher DCT coefficients represent fast changes in filter bank energies. Hence, we take only the lower 12 coefficients of the 26 size vector. These 12 numbers for each frame are called **Mel Frequency Cepstrum Coefficients (MFCC)**
- A speech signal also contains information in the dynamics (apart from the 12 coefficients collected from each frame).

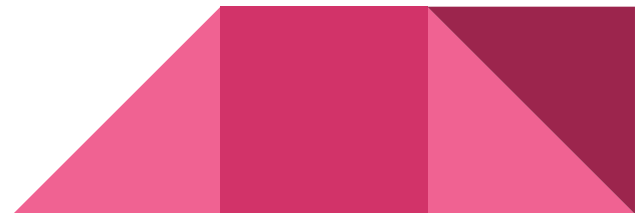
$$d_t = \frac{\sum_{n=1}^N n(c_{t+n} - c_{t-n})}{2 \sum_{n=1}^N n^2}$$

$t \rightarrow$ frame number.

$N \rightarrow$ typically chosen = 2

$C_t \rightarrow$ MFCC vectors of frame t

$d_t \rightarrow$ Delta Coeffs. for frame t



Classifier Techniques

Various classifier models and techniques can be used to assign emotion labels to various characteristic representation obtained from the audio speech signal.

Results from recent proceedings using such classifiers to solve the problem of emotion detection will now be presented.



Hidden Markov Model (HMM)

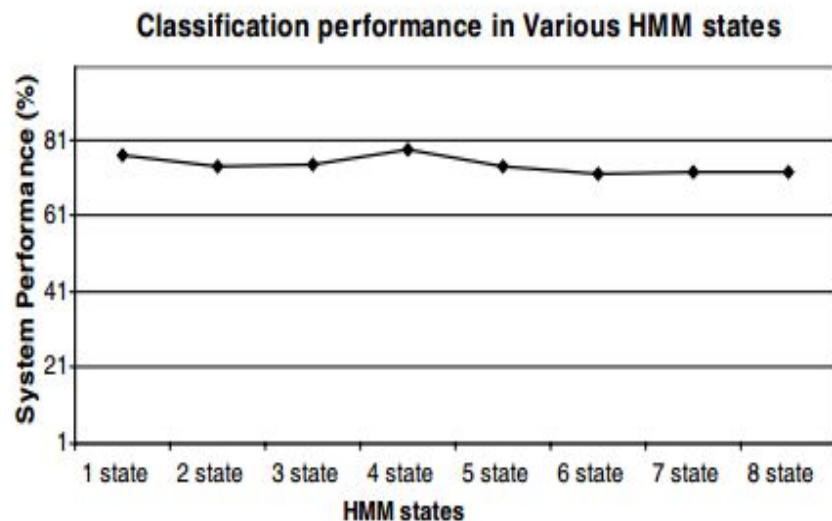
- Hidden Markov models (HMMs) are popular for speech recognition (Lee and Hon, 1989).
- According to Deller et al. (1993), the states in the HMM frequently represent identifiable classes in speech recognition. The number of states is often chosen to roughly correspond to the expected number of phonemes in the utterances.

The state transition probabilities and the output symbol probabilities are uniformly initialized.

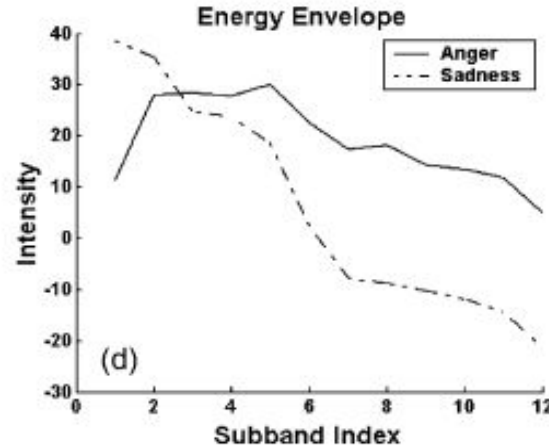
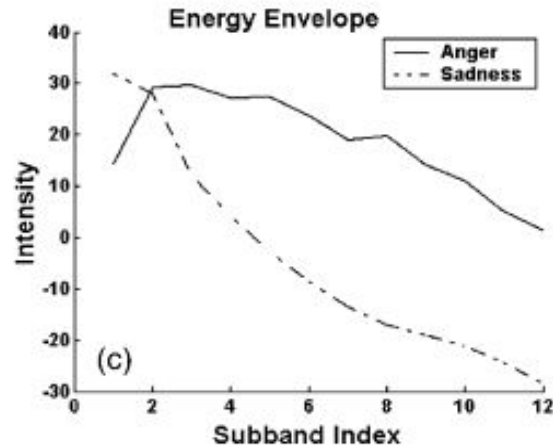
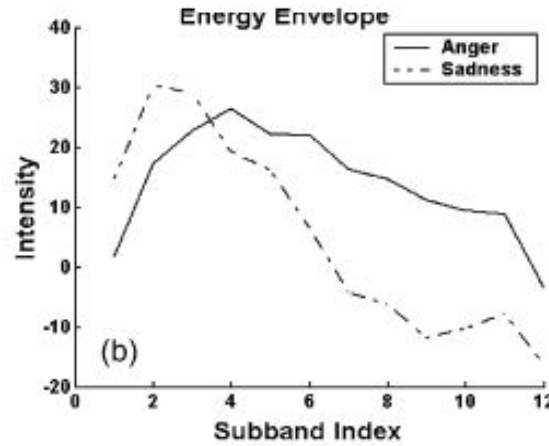
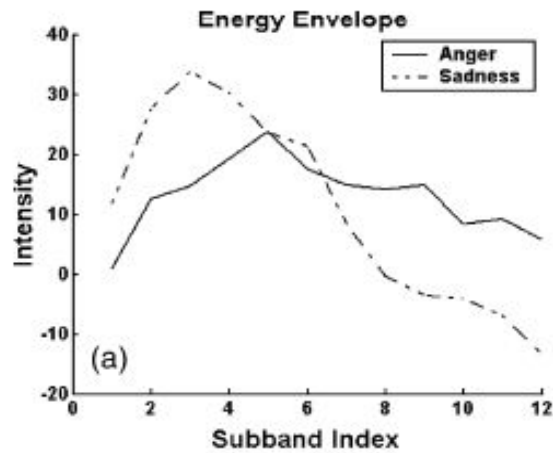
HMM Performance

To assess the effect of number of states for the HMM model, experiments were carried out using HMM models with one to eight states.^[1]

An accuracy of ~75% was obtained in a study of Burmese and Mandarin Speakers



[1] "Speech Emotion Recognition Using Hidden Markov Models" : Tin Lay New, S.W. Foo



Comparison of intensity values for extreme emotions (Anger and Sadness) vs Subband index..

- (a) Burmese female
- (b) Mandarin female
- (c) Burmese male
- (d) Mandarin male

It can be observed that Anger (high arousal emotion) has higher intensity values in higher frequency bands while Sadness (low arousal emotion) has higher intensity values in lower frequency bands.

K - Means Clustering

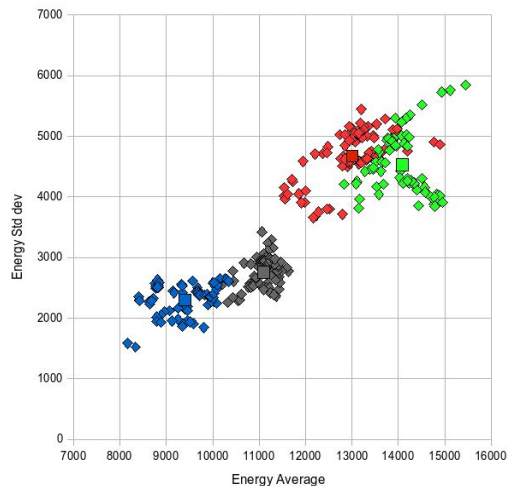
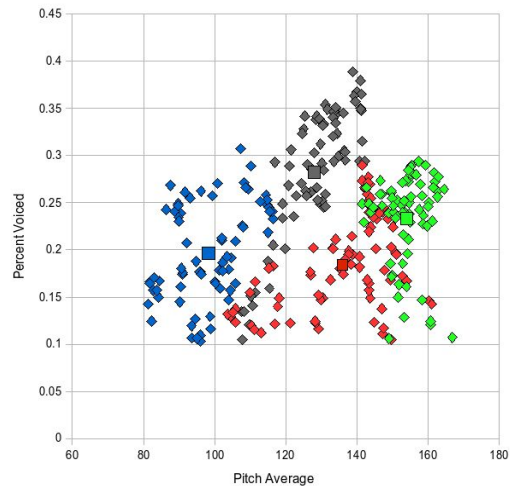
- ***k*-means clustering** is a method of **vector quantization**. *k*-means clustering aims to **partition n observations into k clusters** in which each observation belongs to the cluster with the nearest **mean**, serving as a **prototype** of the cluster.
- This method when applied to extracted MFCCs can yield a useful technique to label emotion in a speech signal.



K - Means Performance

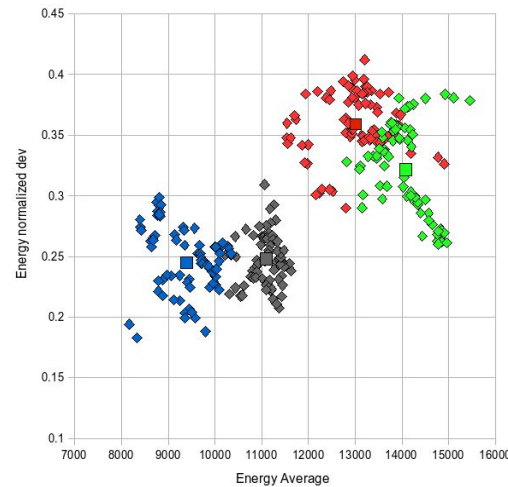
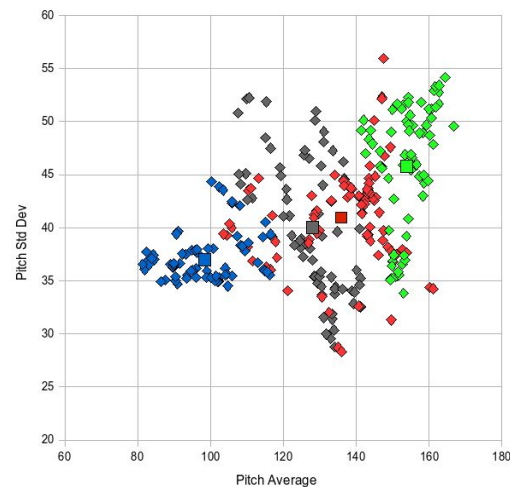
Six different “opposing” emotion pairs were chosen: despair and elation, happy and sadness, interest and boredom, shame and pride, hot anger and elation, and cold anger and sadness.

All Speakers						
Experiment	Features	Distance Measure	Centroid	Iterations	Recognition Accuracy	Variance
despair-elation	MFCC	L1 norm	UDC	100	75.76%	1.74%
happy-sadness	MFCC	L1 norm	UDC	1	77.91%	14.34%
interest-boredom	Pitch	L1 norm	UDC	100	71.21%	2.48%
shame-pride	MFCC	L1 norm	UDC	1	73.15%	3.23%
hot anger-elation	MFCC	L1 norm	UDC	1	69.70%	10.75%
cold anger-sadness	MFCC	L1 norm	UDC	1	75.66%	3.35%



◆ Sad
 ◆ Happy
 ◆ Angry
 ◆ Neutral
 ■ Sad
 ■ Happy
 ■ Angry
 ■ Neutral

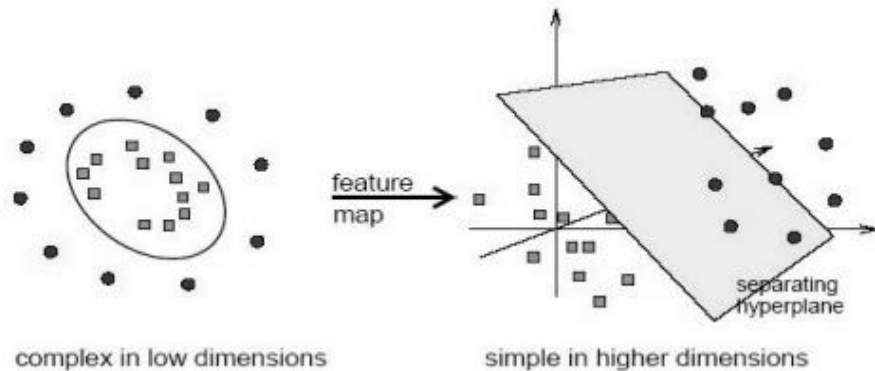
K - Means Clustering for MFCC vectors



Reference : <http://crteknologies.fr/projets/emospeech/>

Support Vector Machines (SVMs)

- Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

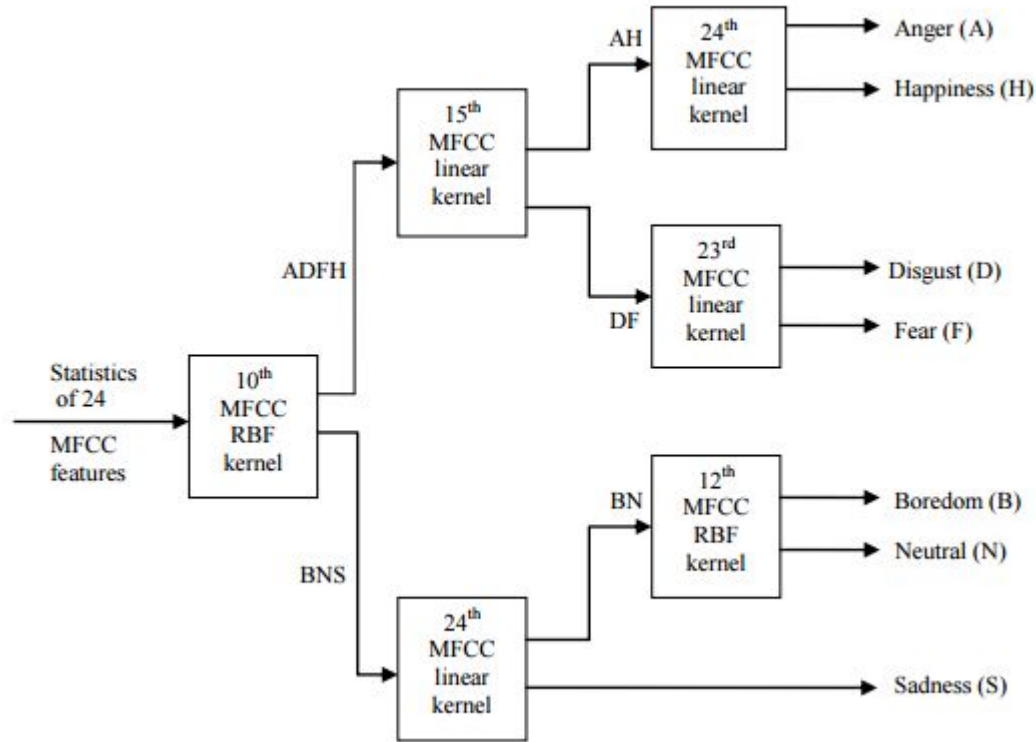


SVM (Results) - I

- A **Confusion Matrix**, also known as a error matrix , is a specific table layout that allows visualization of the performance of an algorithm.
- The results were obtained using a SVM classifier on **Berlin Emotion Database** containing 406 files for 5 emotion classes
- Though the SVM is **binary classifier** it can be also used for classifying multiple classes. Each feature is associated with its class label e.g. angry, happy, sad, neutral, fear.



SVM (Results) - II



24 MFCC features are extracted from each utterance of the database.

Statistical measurements like mean, median, maximum value, minimum value, range, inter-quartile range, standard deviation, kurtosis and skewness are performed over these features.

Reference : "SVM Scheme for Speech Emotion Recognition using MFCC Feature" - IJCA, , A. Milton, Sharmy Roy

SVM (Performance)

Emotion	Emotion Recognition (%)						
	A	D	F	H	B	N	S
A	83.5	0	0	3.15	13.4	0	0
D	0	82.7	0	6.17	0	6.17	4.93
F	2.2	2.17	47.8	30.43	17.4	0	0
H	8.7	1.45	2.9	73.9	10.1	2.9	0
B	16.9	1.4	1.4	22.5	57.7	0	0
N	0	22.6	0	3.22	0	74.2	0
S	0	56.9	0	2.53	0	1.26	39.2

Table. Overall Confusion Matrix for different Emotion Classes

Overall accuracy of **68%** was obtained using RBF (Radial Basis Function) SVM.

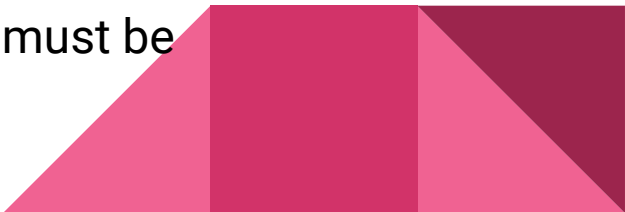
Reference : "SVM Scheme for Speech Emotion Recognition using MFCC Feature" - IJCA, , A. Milton, Sharmy Roy

Convolutional Neural Networks (CNN)

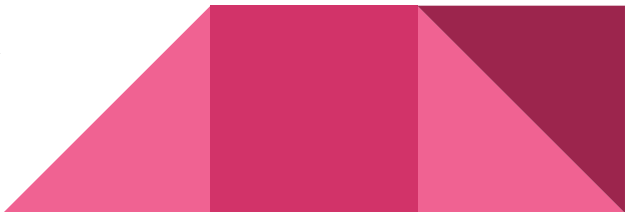
- Convolutional Neural Networks (CNNs) are popular for classification of data with high temporal and spatial correlation and complex feature-to-class properties.
- **State-of-the-art** techniques for speech recognition these days use GPU accelerated CNNs along with pre-trained models trained on huge datasets.
- Baidu and Google [Ng et. al. [link](#)] both use variants of convnets for speech recognition.
- **Best in-class Accuracy** : 78 +- 3.2% per class

Intricacies and Possible Issues

Emotion recognition is a challenging task, the following are some known issues and problems which developers of such systems face :

- Differentiating between various emotions and deciding which particular speech features are more useful is not clear. (i.e separation of linguistic content from **background noise** and further extraction of **phonemes representing emotion**).
 - Various factors like **gender, accent** (thus language), loudness, etc need to be considered for successfully deciding upon the emotion the speech presents. However, an ideal speech emotion recognizer system must be able reduce the dependance of such factors.
- 

Some Applications

- Recognize voices, detect anger in speech and prioritize angry calls!
In real time applications such as call analysis in the emergency services like ambulance and fire brigade, verification of emotions to analyze genuineness of requests is important.
 - **Medicine:**
Can be used in rehabilitation, help monitoring and **counselling** patients/clients.
Can be used in therapy of **Autism**, for people who struggle to express/interpret emotions
 - **Law:**
Deeper discovery of depositions and can be used as a non-invasive lie detector test.
- 

Future Work

- Higher accuracy can be obtained using the combination of more features and using deeper neural architectures. To sum up, future work is to extract the delta features from each utterance and then use the SVM hierarchical structure for classification.
 - Based on the results obtained in this report it seems like sparse coding could yield better performance. But this can only be concluded after verifying on some different database.
 - Expression of emotions is an universal phenomenon, which may be independent of speaker, gender and language. Cross lingual emotion recognition study may be another interesting work for further research.
- 