# A Feature-Driven Hierarchical Classification Approach to Emotions in Speeches Using SVMs

S. Majuran and A. Ramanan
Department of Computer Science, University of Jaffna, Sri Lanka.
shayu.kiri@gmail.com, a.ramanan@jfn.ac.lk

SLSWCS
Student Workshop on Computer Science

## Objectives

- To test the relative performance of nine basic statistical measurements obtained from MFCC for emotion classification.
- To hierarchically order emotions in speeches using a decision tree with subset of statistical measurements of MFCC at each decision node.

1

## Introduction

This work presents an analysis of basic nine statistical measurements of MFCC [2] features and a feature-driven hierarchical classification scheme using SVMs to classify emotions in speeches. The proposed method for emotion classification is evaluated on two benchmark datasets: Berlin database [1] and Danish Emotional Speech (DES) database [5].

2

## Dataset

Berlin: Contains 535 utterances spoken by 10 different German actors in happy (71), sad (62), angry (127), bored (81), disgusted (46), fearful (69), and neutral (79) ways. Numbers in parantheses indicate the number of utterances per emotion. Each sentence consists an average of 10 words.

DES: Four professional speakers, two males and two females, were asked to speak predefined sentences and words in Danish for five emotions: neutral, angry, happy, sad, and surprised. A total of 260 sentences are available in the database, with 52 sentences per emotion class making up 28 minutes of speech material. Each speaker was asked to utter two words, nine short sentences and two passages in all five emotions.

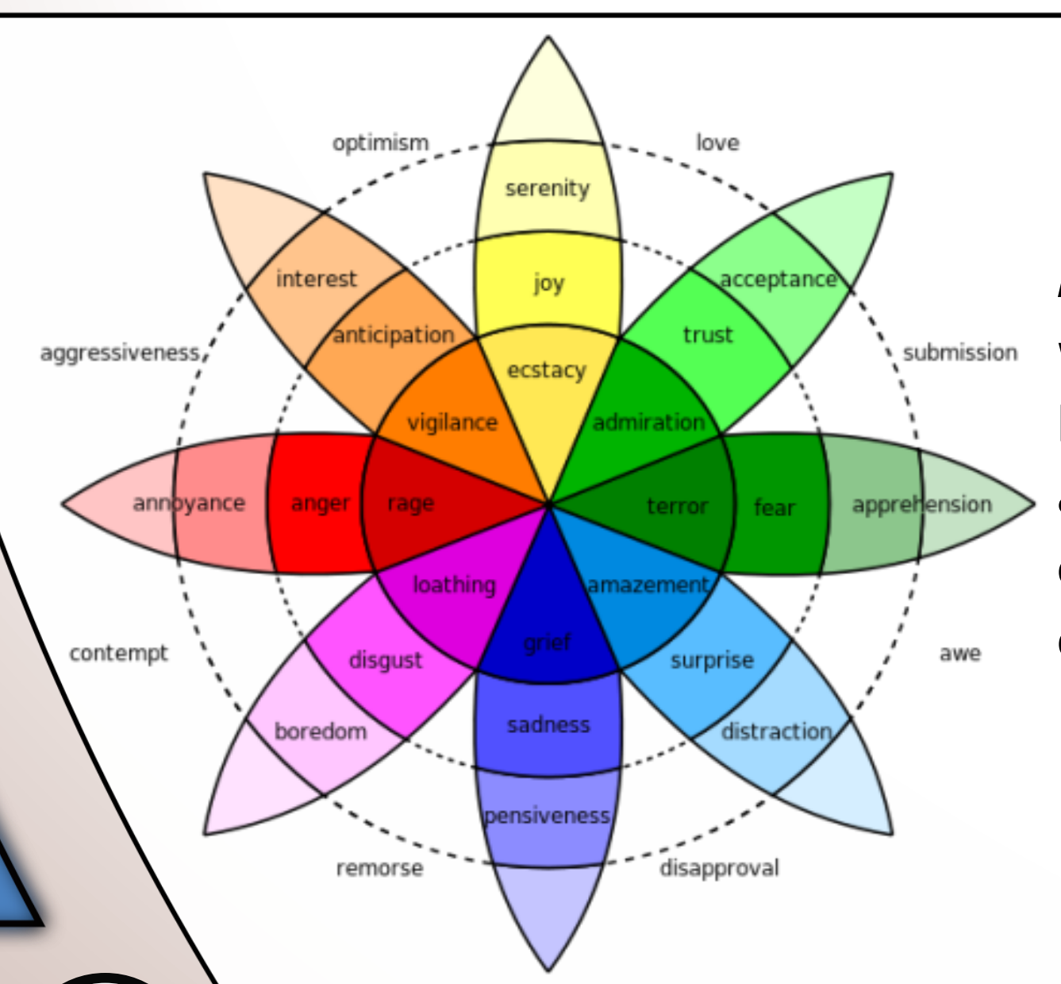
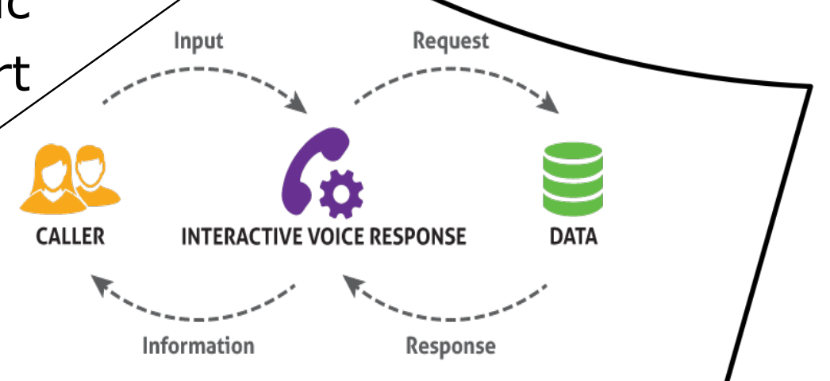


***Fig. 1.*** Wheel of emotions which consists of eight basic emotions and eight advanced emotions each composed of two basic ones created by Robert Plutchik.

### Applications for Speech recognition



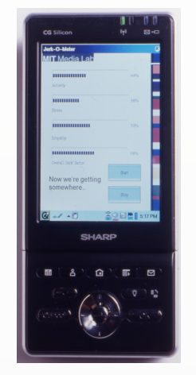
***EMOSPARK:*** AI console that uses face-tracking and language analysis to assess human emotion and deliver relevant content accordingly

***Court Reporter***

***FaceSay***

***Virtual Assistant***

ITSPOKE
Automated Tutor



***jerk-O-meter:*** Speech-Feature Analysis Provides Feedback on Your Phone Interactions

3

## Feature Extraction

- A high-pass filter is applied with a pre-emphasis coefficient of 0.97.
- Speech signal is converted into sequence of feature frames.
- The Hamming window is applied using 10ms of hop time to each frame which has the window length of 25ms.
- The first 13 MFCCs (first coefficient is replaced with the log energy) are extracted from frames.
- FFT converts a signal from original domain to a representation in the frequency domain.
- The bank of filters according to Mel scale are used to compute a weighed sum of filter spectral components and log of of these energies are computed.
- DCT converts the log Mel spectrum into time domain. ⟶ MFCC

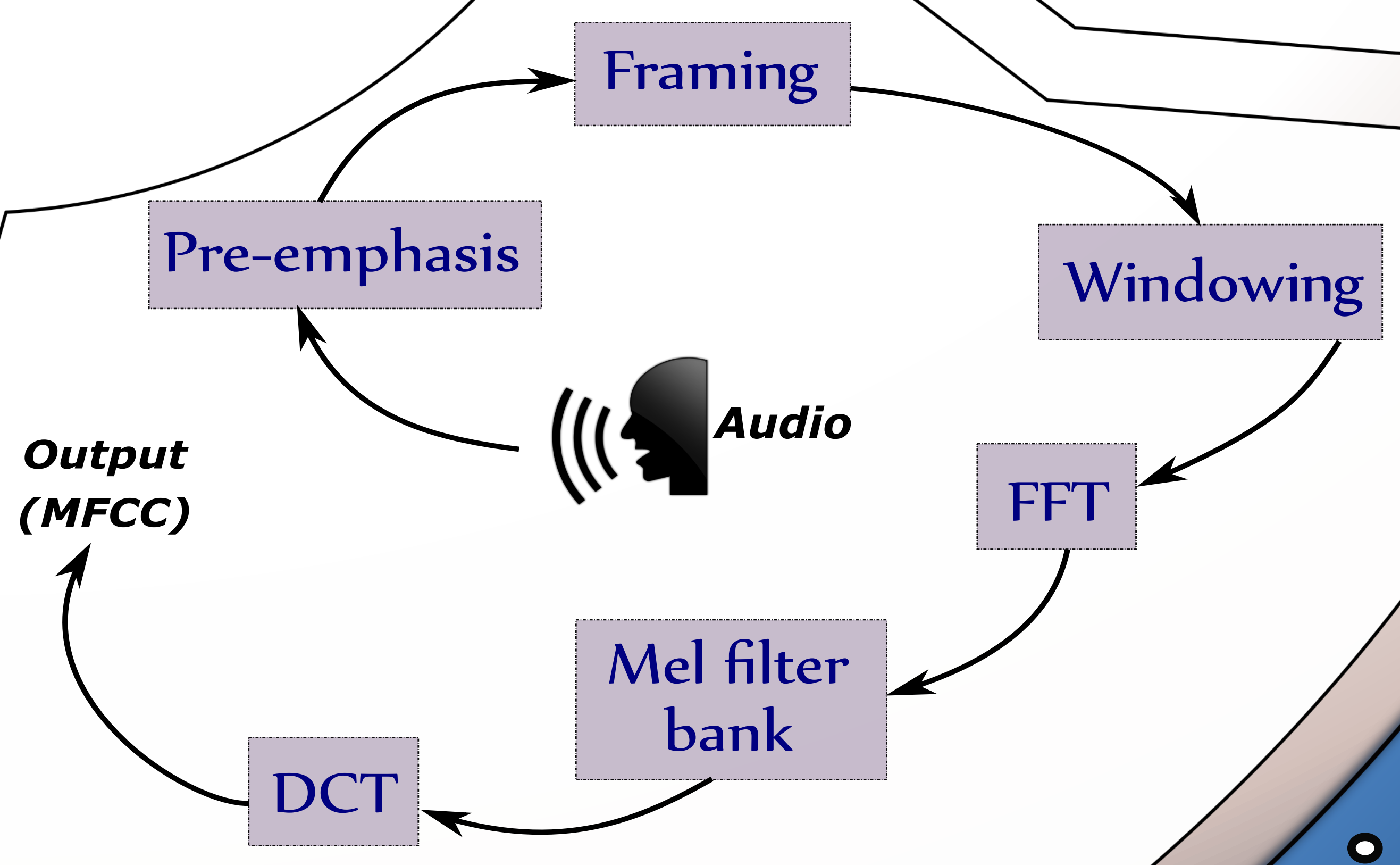


***Fig. 2.*** Block diagram of feature extraction.

Mel frequency for any given frequency $f$ in Hz:

$$Mel(f) = 2595 \times \log(1 + f/700)$$

4

## Structure of Decision Tree

A decision tree is a model of decisions and their consequences that creates a right-balancing tree structure of OVA-based classifier scheme at each node to make sure the higher performance [4].

- At the root node, one selected emotion is compared with all other emotions.
- After classifying the emotion it proceeds to the next level and the emotion at root is eliminated from the training set and the appropriate emotion is used as the decision node for the next stage.
- The leaf node of the decision tree considers two emotions in the classification using the optimal model.

The hierarchical tree needs (N-1) classifiers in training and at most (N-1) nodes for decision making (N - # of classes).

5

## Experimental Setup

- We calculated the nine basic statistical measures: mean, standard deviation, minimum, maximum, median, inter-quartile range, range, and kurtosis for each Mel-frequency vector of coefficients.
  - All the training data were scaled to [-1,1] and the test data were adjusted using the same linear transformation.
    - Experiments were carried out using linear OVA-based SVM classifier [3] with the optimal value C=1.0 for the linear kernel using 10-fold cross-validation.

| Classes | Emo |
|---|---|
| Fear | 1 |
| Disgust | 2 |
| Happy | 3 |
| Bored | 4 |
| Neutral | 5 |
| Sad | 6 |
| Anger | 7 |

***Table I.*** Considered indices for emotions (Emo) to berlin dataset

***Table II.*** Classification rate on Berlin using subset of statistical measurements that are used in the decision tree

| Emo | Feature set | Rate |
|---|---|---|
| 6 | m1, m3, m4, m8, m9 | 0.98±0.02 |
| 2 | m1, m3, m8 | 0.95±0.03 |
| 7 | m1, m2 | 0.94±0.04 |
| 3 | m6, m5, m3, m8, m2, m9, m1, m4 | 0.93±0.05 |
| 1 | m1, m6, m9, m4 | 0.93±0.03 |
| 4 | m2, m4, m9 | 0.70±0.10 |
| 5 | m2, m4, m8 | 0.68±0.08 |

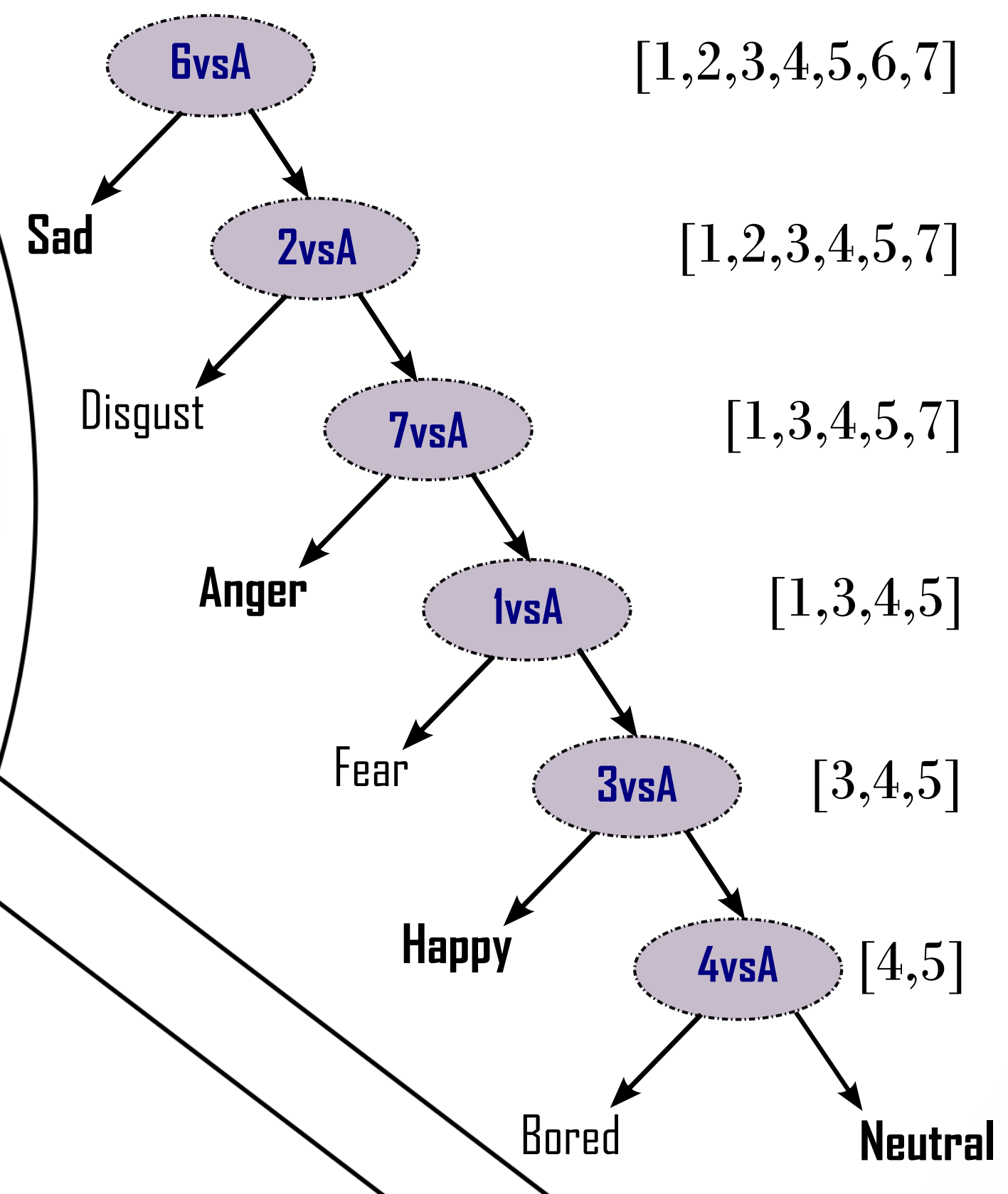


***Fig. 3.*** Architecture of the classification to find the best emotion at each node out of seven classes of emotions from Berlin.

6

All the seven emotions are classified using binary classification and the average recognition rate for emotions is ***90.69%***.

***Table III.*** Classification rate of emotions on Berlin dataset

| Emotions | Rate |
|---|---|
| Fear | 0.92±0.03 |
| Anger | 0.93±0.04 |
| Happiness | 0.86±0.03 |
| Sadness | 0.95±0.02 |
| Neutral | 0.86±0.04 |
| Boredom | 0.88±0.04 |
| Disgust | 0.95±0.03 |

The influence of each statistical measurements of MFCC in emotion classification is also experimented by using 117-D features and OVA-based SVM classifier.

***Table IV.*** Classification rate of statistical measures of MFCC on Berlin

| No. | Statistical Measures | Rate |
|---|---|---|
| m1 | Mean | 0.89±0.04 |
| m2 | Standard Deviation | 0.87±0.03 |
| m3 | Minimum | 0.88±0.03 |
| m4 | Maximum | 0.86±0.05 |
| m5 | Median | 0.89±0.04 |
| m6 | Inter-quartile range | 0.88±0.03 |
| m7 | Kurtosis | 0.86±0.03 |
| m8 | Skewness | 0.87±0.04 |
| m9 | Range | 0.88±0.03 |

7

OVA-SVMs: Entire MFCC features of 117-dimension were classified using binary OVA-based SVMs.

Hierarchical Scheme (without FS): 117-dimensional features were classified using the hierarchical decision tree of emotions.

Hierarchical Scheme (with FS): Evaluated using optimal feature set of statistical measures using the decision tree.

***Table V.*** Analysis of overall performance

| Dataset | Classification Rate | | |
|---|---|---|---|
| | OVA-SVMs | Hierarchical Scheme (without FS) | Hierarchical Scheme (with FS) |
| Berlin | 70.88% | 84.17% | 87.20% |

## Discussion and conclusion

- Classification rates of basic nine statistical measures for emotions varies 86-89% and ***mean*** shows better performance while taken into account the estimates of uncertainty for very few trails.
- Testing results indicate that emotions can be hierarchically organised in the order: sad, disgust, anger, fear, happiness, bored, and neutral, respectively.
- Also performed the all above experiments on Danish Emotional Speech (DES) dataset and it shows 81.08% as the average recognition rate of emotions. DES shows 70.88%, 84.17% and 87.20% for the classification of 117-D MFCC features using OVA-based SVMs, 117-D features using the hierarchical decision tree and using the optimal feature set of statistical measures through

## References

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, "A database of German emotional speech", in European conference on speech and language processing, pp. 1517-1520, 2005.

[2] S. Davis and P. Mermelstein, "Comparision of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE transactions on Acoustics, Speech and Signal Processing, vol. 28, pp. 357-366, 1980.

[3] C. Hsu, C. Chang and C. Lin, "A Comparison of Methods for Multi-class Support Vector Machines", IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 1-4, 2002.

[4] A. Hassan and R. I. Damper, "Multi-class and hierarchical SVMs for emotion recognition", in Proceedings of Interspeech, pp. 2354-2357, 2010.

[5] I. Engberg, A. Hansen, O. Andersen and P. Dalsgaard, "Danish Emotional Speech Database DES," in European Conference on Speech and Language Processing, pp. 1695-1698, 1997.