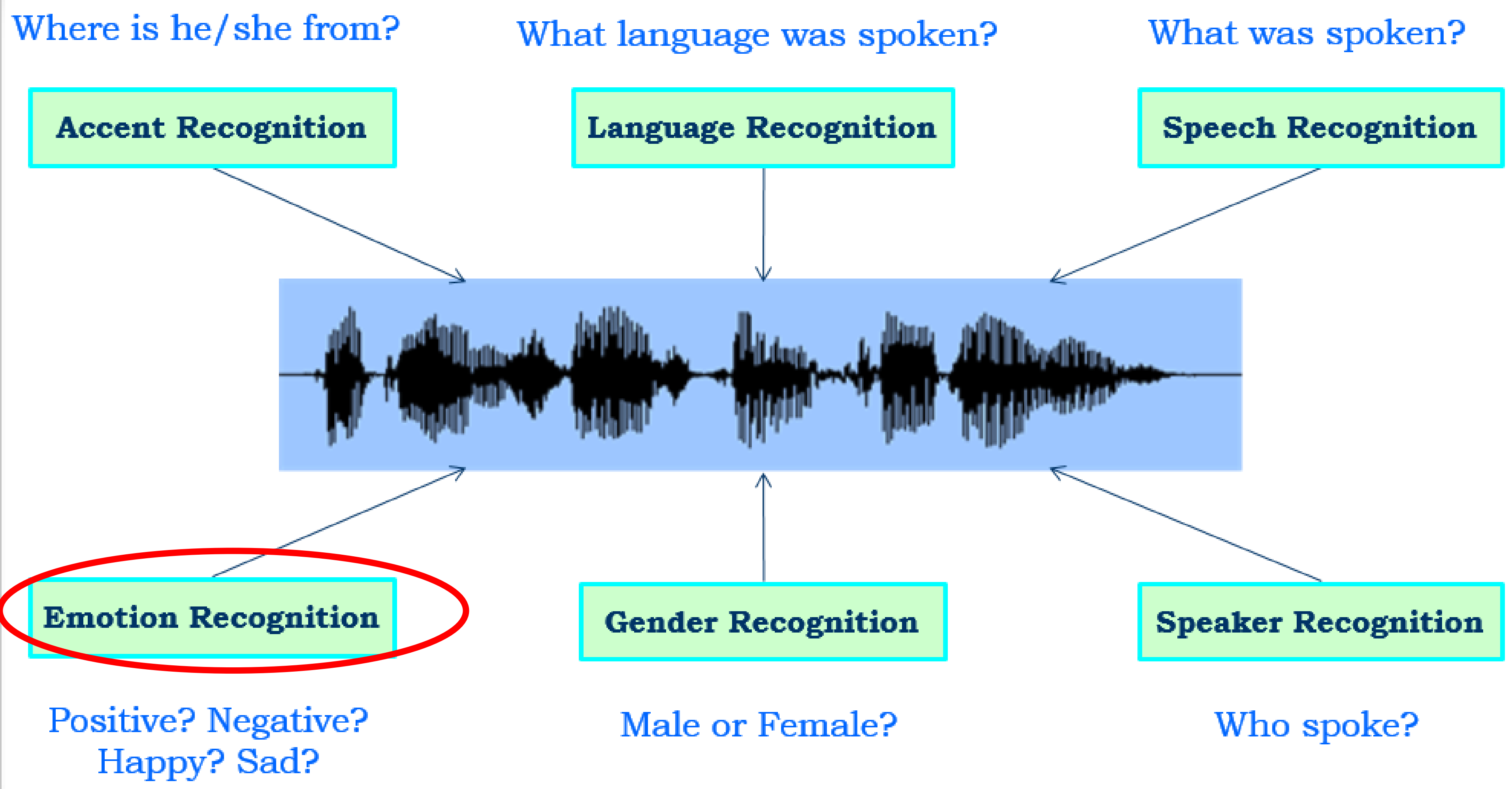
ICIC 2015

Paper ID: ES 794

# Multi-Taper Spectral Features for Emotion Recognition from Speech

**Santosh Chapaneri & Deepak Jayaswal**
St. Francis Institute of Technology
University of Mumbai

# Rich Information contained in Speech



Where is he/she from?

Accent Recognition

What language was spoken?

Language Recognition

What was spoken?

Speech Recognition

Emotion Recognition

Positive? Negative?
Happy? Sad?

Gender Recognition

Male or Female?

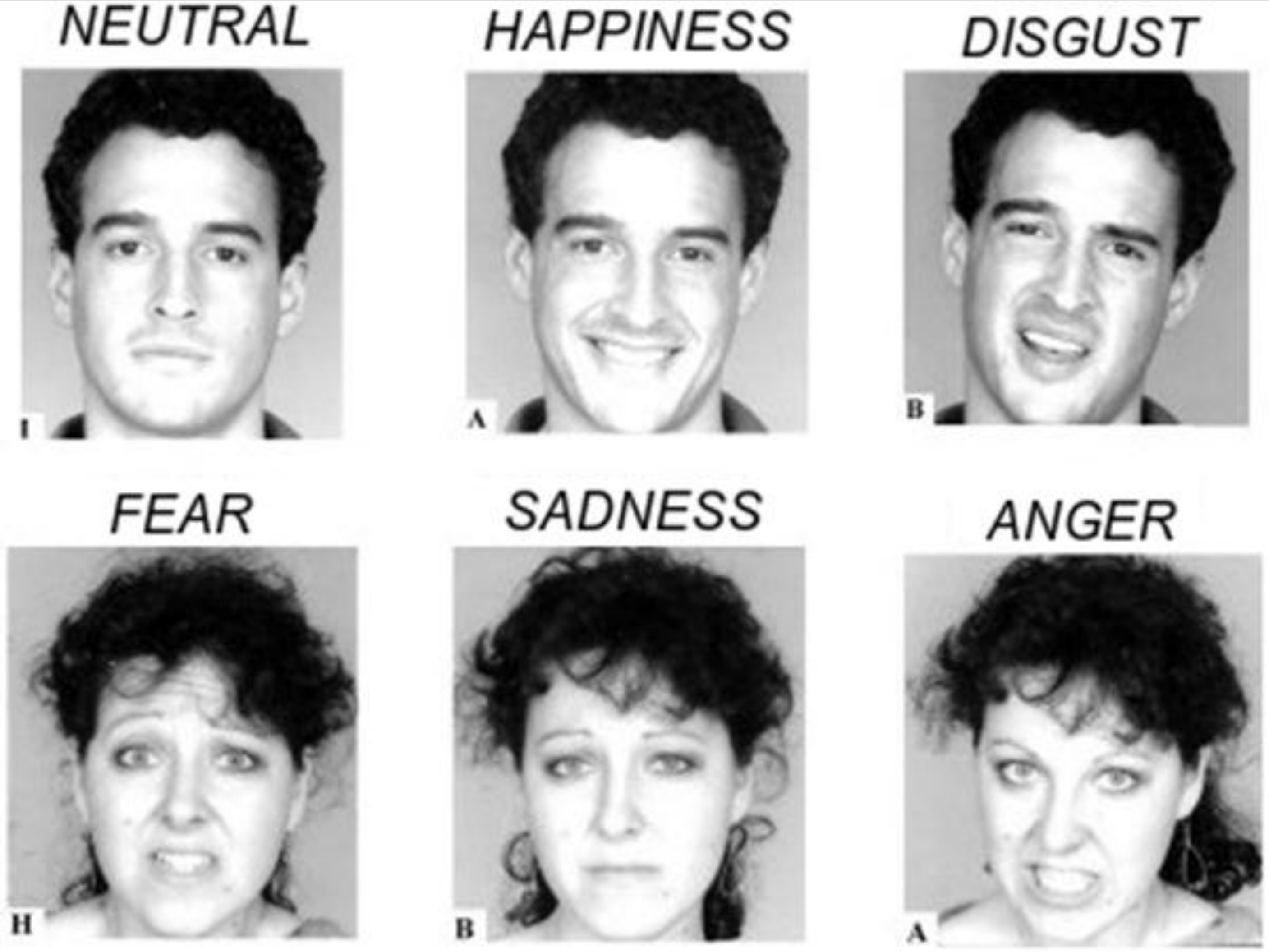
Speaker Recognition

Who spoke?

# Why Emotion Recognition?

- Detecting frustration of callers to automated help line
- Computer tutorials via virtual avatars
- Lie detection
- Humanoid Robots

# Basic Emotions



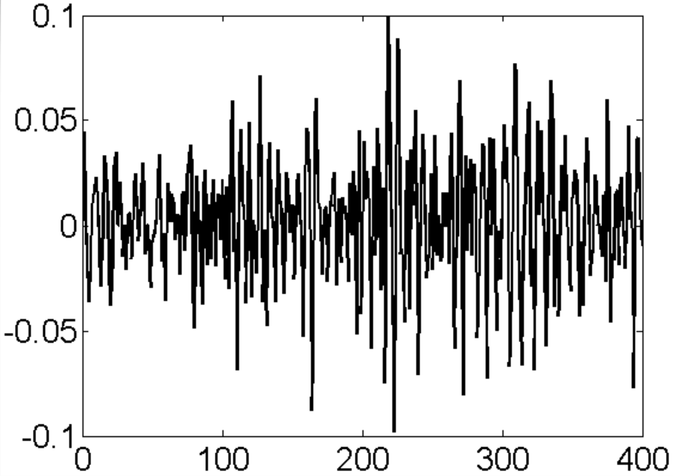
# Speech Database

- Berlin Emotional Database (EMO-DB) [1]
- Total 535 utterances:

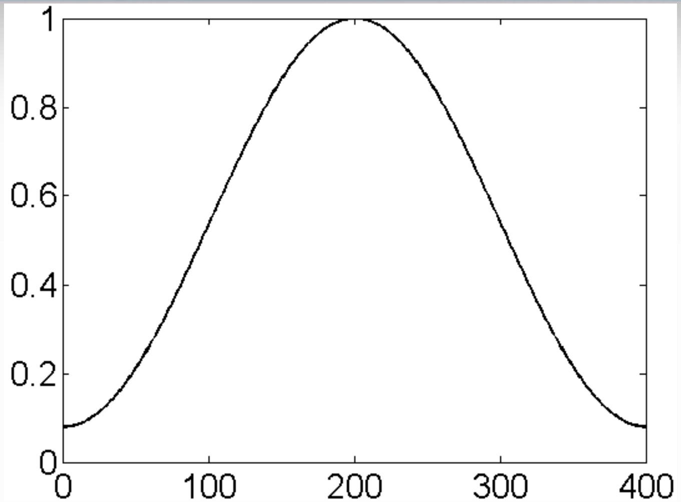
| Anger | Boredom | Disgust | Fear | Happiness | Sadness | Neutral |
|---|---|---|---|---|---|---|
| 127 | 81 | 46 | 69 | 71 | 62 | 79 |

- 70% used for training, 30% for testing
- Sampling frequency 16 kHz
- 16-bit resolution, mono channel samples

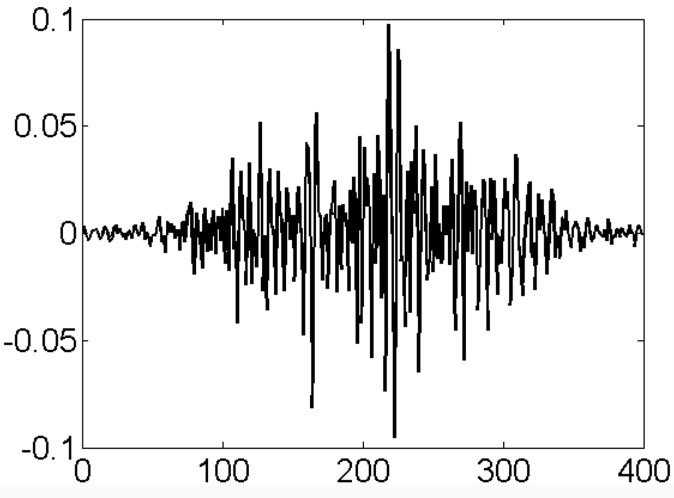
# Single Taper Spectrum



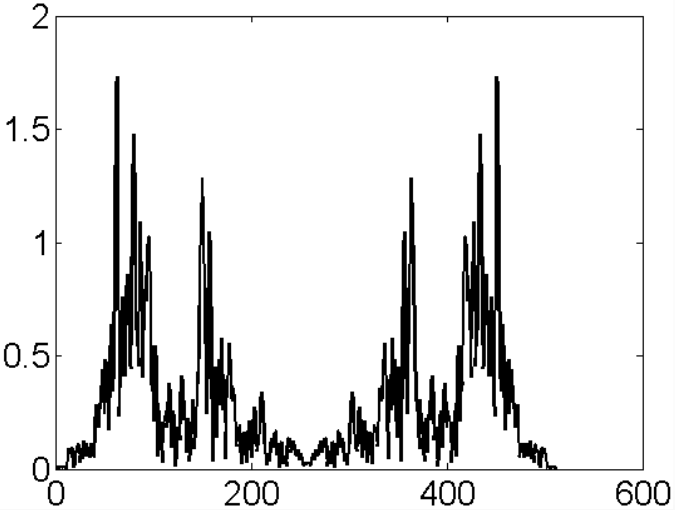
Speech frame (25 msec)



Hamming window



Windowed frame



Spectral estimate

# Single Taper Spectrum

$$\hat{S}(f) = \left| \sum_{n=0}^{N-1} w(n)s(n)e^{-j2\pi nf} \right|^2, \ |f| \le \frac{1}{2}$$

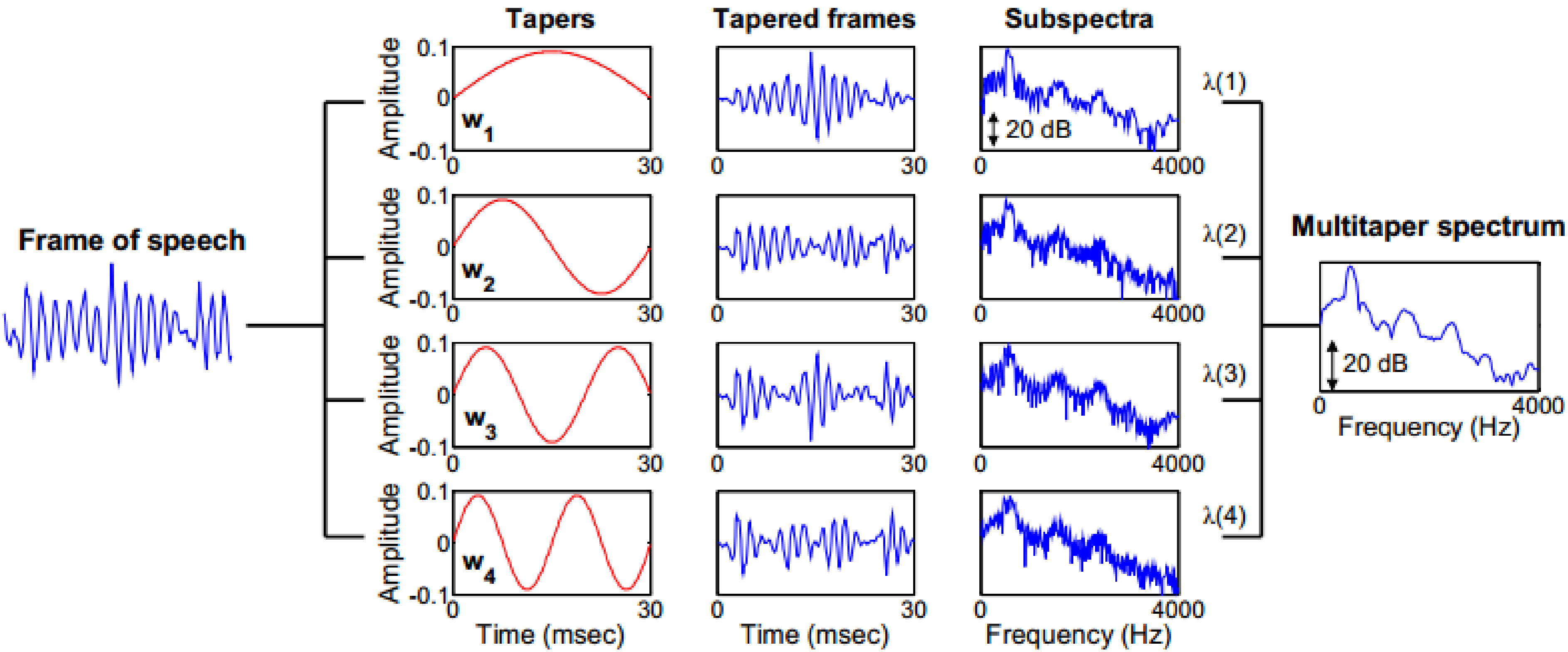
- Problem: Spectral estimate has large variance relative to true spectrum [2]
- Trivial solution: Use Welch’s periodogram to reduce variance => but increases bias [3]

- Use concept of **multi-taper spectral estimates**; an idea proposed in 1982 by Thomson and later by several others [4 – 6]

# Multi-Taper Concept

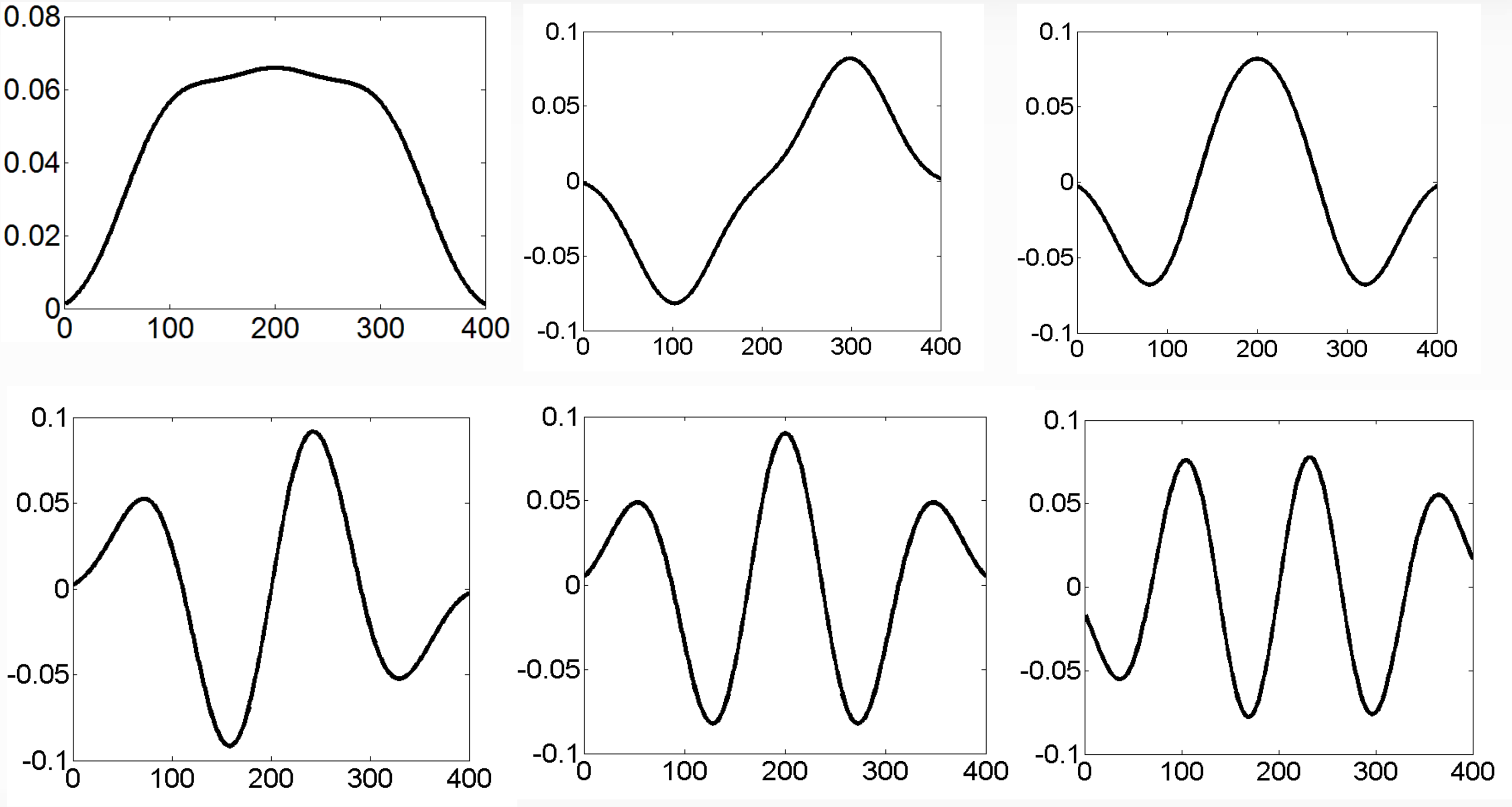
- From statistics: If a random variable *X* has variance of $\sigma^2$, then the statistical average of *n* independent samples of *X* will have variance of $\sigma^2/n$
- Use multiple orthonormal tapers => resulting in eigen-spectra
- Take weighted average of these to obtain a spectral estimate with reduced variance
- Orthonormal => Uncorrelated => Less Variance [7]
- Types of Multi-tapers:
  - Thomson
  - Multi-Peak
  - Sine Weighted Cepstrum Estimator (SWCE)

# Multi-Taper Concept

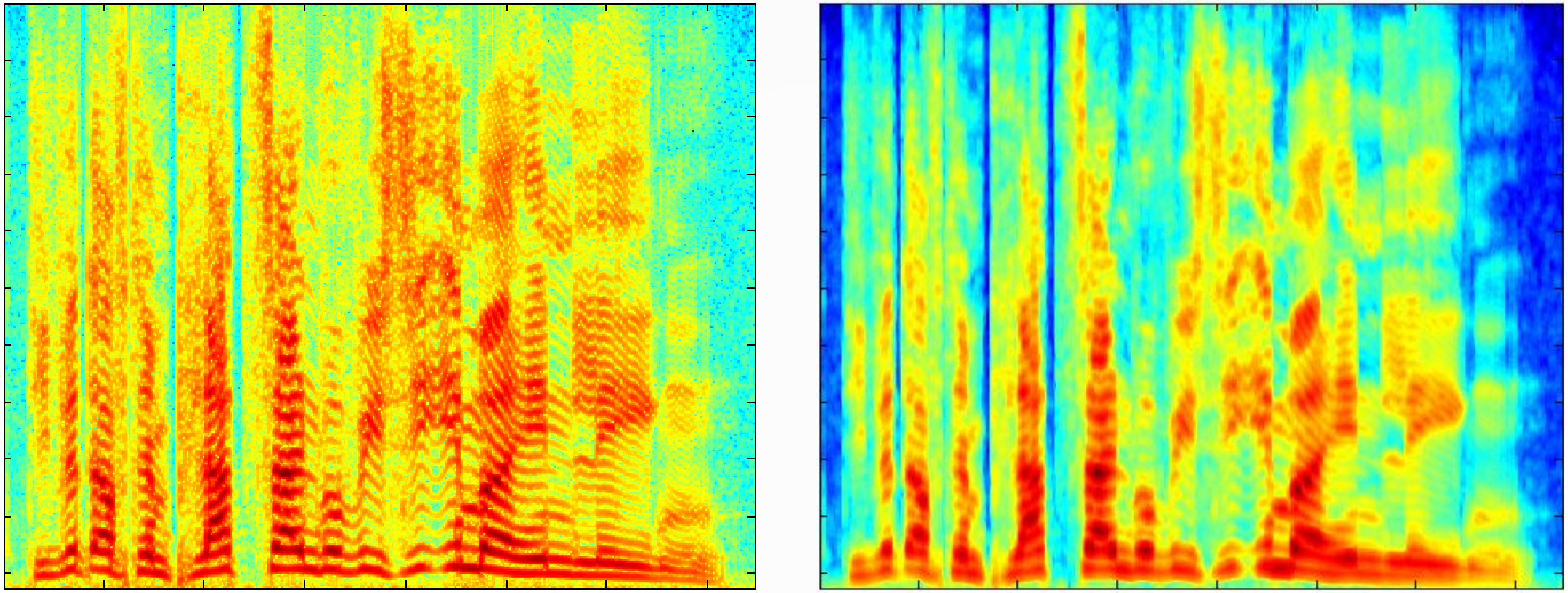


$$\hat{S}_{MT}(f) = \frac{1}{\sum_{k=1}^{K} \lambda_k} \left( \sum_{k=1}^{K} \lambda_k \left\{ \left| \sum_{n=0}^{N-1} w_k(n) s(n) e^{-j2\pi nf} \right|^2 \right\} \right), \ |f| \le \frac{1}{2}$$

# Multi-Peak Multi-Tapers



# Multi-Taper Spectral Estimate



Spectrograms of speech signal with
Hamming Single-taper (left), Multi-peak Multi-taper (right)

# Features for Emotion Recognition

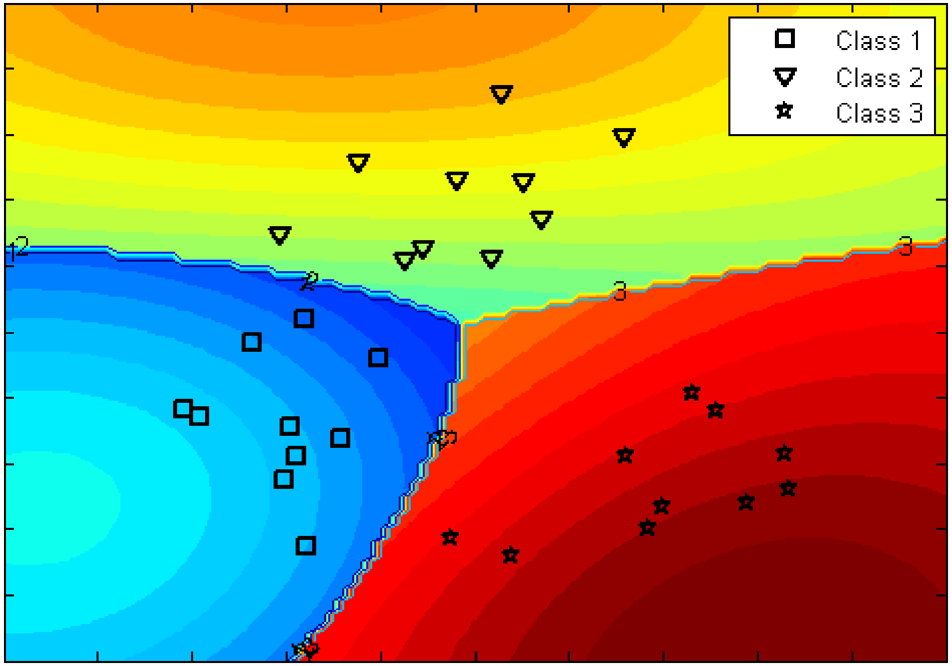
- Spectral features computed from single as well as multi-taper spectral estimates: Energy, Centroid, Spread, Skewness, Kurtosis, Rolloff, Decrease, Slope, Variation, Flatness, Crest, Entropy, MFCC (Mel Frequency Cepstral Coefficients), OBSC (Octave Based Spectral Contrast)
- 12 (statistical features) + 13 (MFCC) + 12 (contrast features) per frame
- To reduce dimensionality, take statistics of each spectral feature contour using mean, standard deviation, median and inter-quartile range
- This results in (12 + 13 + 12) x 4 = **148 spectral features per emotion speech signal**

# Feature Classification

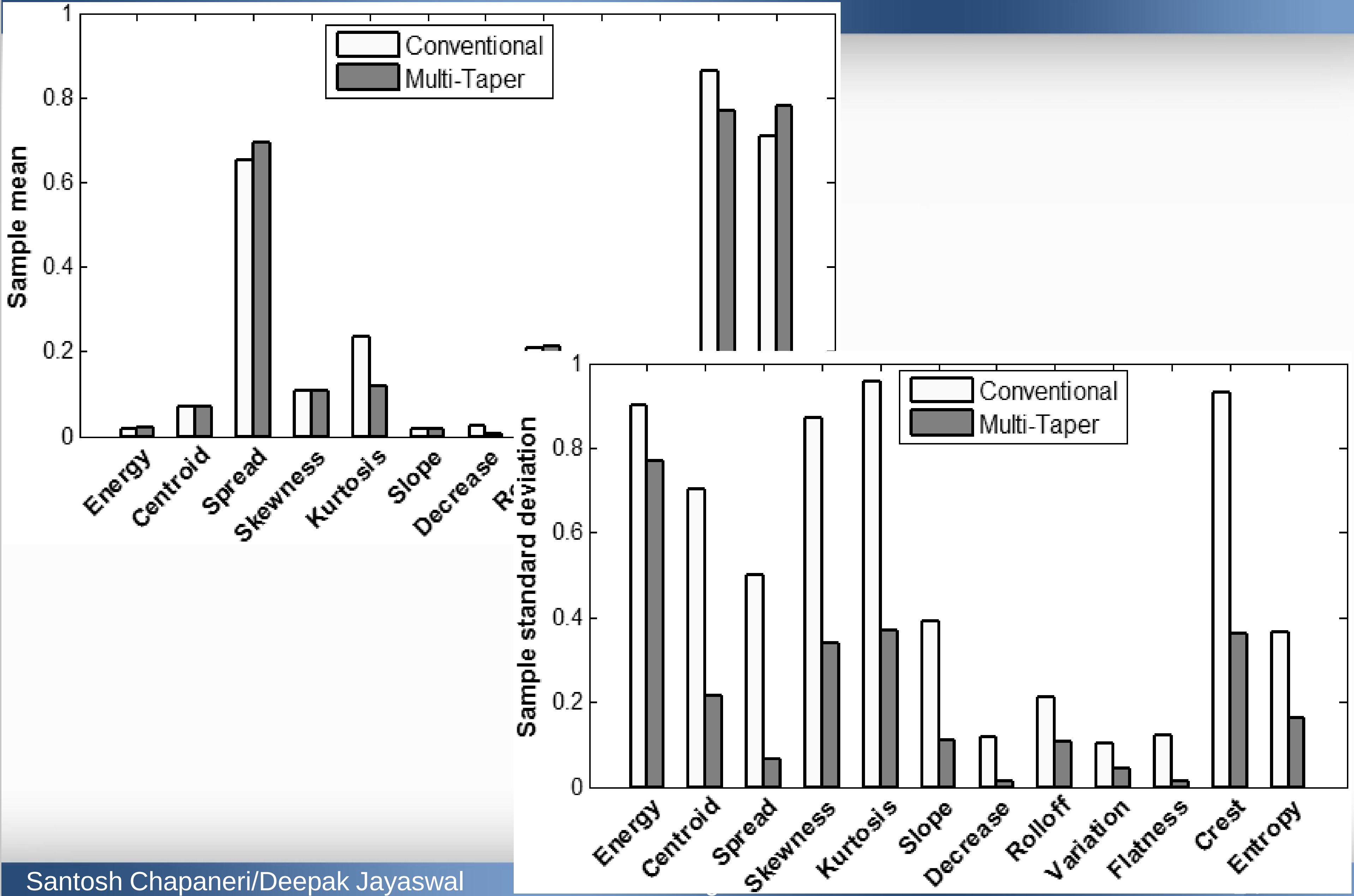
- Support Vector Machine => determines the optimal separating hyperplane
- RBF kernel, 10-fold cross validation
- One-against-one for multiclass classification

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i$$

$$\text{s.t. } y_i\left(\mathbf{w}^T\varphi(\mathbf{x}_i)+b\right)\ge 1-\xi_i,\ \xi_i\ge 0$$



# Results: Variance Reduction



# Results: Classification Accuracy

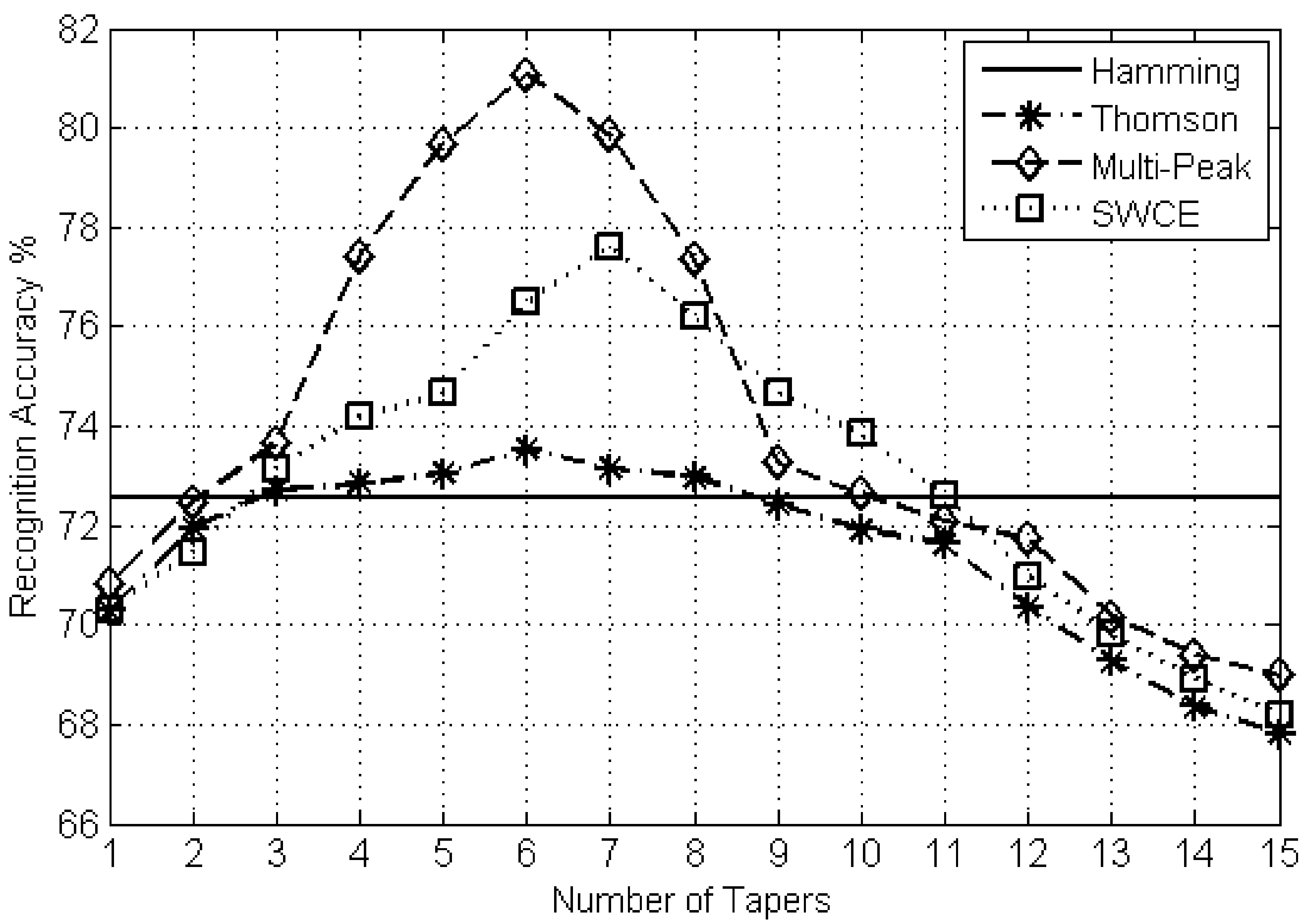
| **Emotions** | **Hamming (1 taper)** | **Thomson (6 tapers)** | **Multi-peak (6 tapers)** | **SWCE (6 tapers)** |
|---|---|---|---|---|
| **Two** | 100% | 100% | **100%** | 100% |
| **Three** | 90.32% | 91.43% | **93.45%** | 92.33% |
| **Four** | 82.5% | 83.55% | **85.32%** | 82.80% |
| **Five** | 82.17% | 84% | **88.38%** | 85.53% |
| **Six** | 75.65% | 76.74% | **81.43%** | 79.83% |
| **Seven** | 72.57% | 73.53% | **81.08%** | 77.49% |
| **+ / - Activation** | 93.28% | 95.34% | **96.94%** | 94.67% |
| **+ / - Valence** | 92.11% | 94.50% | **95.52%** | 95.92% |
| **+ / - Emotion** | 95.71% | 96.35% | **98.21%** | 97.25% |

# Results: Confusion Matrix

A:Anger, B:Boredom, D:Disgust, F:Fear, H:Happy, N:Neutral, S:Sad

| | **A** | **B** | **D** | **F** | **H** | **N** | **S** | **Recall** |
|---|---|---|---|---|---|---|---|---|
| **A** | **16** | | 1 | | 2 | | | 84.21 |
| **B** | | **9** | | | | | 2 | 81.82 |
| **D** | | 1 | **14** | 1 | 1 | | 1 | 77.78 |
| **F** | | | 1 | **10** | 2 | 1 | 1 | 66.67 |
| **H** | 1 | | | 1 | **17** | | | 89.47 |
| **N** | | 3 | | | | **11** | | 78.57 |
| **S** | | 1 | | 1 | | 0 | **13** | 86.67 |
| **Prec.** | 94.12 | 64.29 | 87.50 | 76.92 | 77.27 | 91.67 | 76.47 | |

# Results: Impact of Number of Tapers



# Conclusion

- Multi-taper spectral estimates result in better performance relative to single-taper
- Due to reduced variance, spectral estimate and thus features are more discriminatory per emotion
- Multi-peak multi-tapers outperform other techniques
- **Future scope**:
  - Indian native speech language
  - App for Aakash tablets

# References

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech”, *Proc. Interspeech-2005*, Lisbon, Portugal, pp. 1-4, Jan 2005

[2] B. Babadi, and E. Brown, “A review of multi-taper spectral analysis”, *IEEE Tran. Biomedical Engg.*, vol. 61, no. 5, pp. 1555-1564, May 2014

[3] P. Abry, P. Gonçalves, and P. Flandrin, “Wavelets, spectrum analysis and 1/*f* processes”, *Lecture Notes in Statistics: Wavelets and Statistics*, vol. 103, pp. 15-29, Aug 1995

[4] D. Thomson, “Spectrum estimation and harmonic analysis”, *Proc. IEEE*, vol. 70, no. 9, pp. 1055–1096, Sep 1982

[5] M. Hansson, and G. Salomonsson, “A multiple window method for estimation of peaked spectra”, *IEEE Trans. Signal Proc.*, vol. 45, no. 3, pp. 778-781, Mar 1997

[6] K. Riedel, and A. Sidorenko, “Minimum bias multiple taper spectral estimation”, *IEEE Trans. Signal Proc.*, vol. 43, no. 1, pp. 188-195, Jan 1995

[7] T. Kinnunen, R. Saeidi, F. Sedlak, K. Lee, J. Sandberg, M. Sandsten, and H. Li, “Low-variance multi-taper MFCC features: a case study in robust speaker verification”, *IEEE Tran. Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1990-2001, Sep 2012

Thank You