Neuro-Inspired Audio Processing

David Anderson

Associate Professor of ECE
School of Electrical and Computer Engineering (ECE)
Georgia Institute of Technology

Acknowledgements

- Many have contributed:
  - Paul Hasler
  - Krishna Palem
  - Doug Chabries
  - Heejong Yoo
  - David Graham
  - Paul Smith
  - Richard Christiansen
  - Rich Ellis
  - Nikolaos Vasiloglou
  - …

Neuro-Inspired Signal Processing

Intelligent, robust, efficient systems

Signal Processing

Neuroscience

…

Neuromorphic Circuits

Machine Learning

What are the Problems/Opportunities?

- How can we learn from & apply knowledge from biological systems?
  - This is a major focus of our work here
  - Psychoacoustics is the basis for many products (e.g. mp3, aac)
  - Much left to do
- Non-linear processing
  - Analysis is very difficult – this makes it difficult to design non-linear systems and also to understand how existing systems (e.g. biological systems) work.
- Accuracy
  - How much is needed and how can robust systems be made from inaccurate subsystems?
- Parallel processing
  - Can we learn more about self-configurable / adaptive systems from biology?
- Timing
  - Most theory has been developed assuming continuous systems or regular samples.
  - We need to develop much more theory to describe processing using temporal encoding (e.g. spikes)

Analysis & Synthesis

Analysis only
- Signal understanding
- May be destructive

Analysis – Synthesis
- Signals modified for human consumption
- Analysis stage must be invertible
- Preserving perceptual integrity is important

Analysis vs. Analysis-Synthesis Problems

Analysis only
- Automatic speech recognition
- Audio scene understanding
- Signal localization
- Sound classification
- Stream analysis

Analysis-Synthesis
- Hearing compensation
- Signal enhancement
- Audio compression
- Beam forming
- Speech coding
- Signal separation

Physiologically Motivated Methods For Audio Pattern Classification

Sourabh Ravindran

Problem Statement

To build **audio classification** systems that are **low-power** and **robust** to changes in the environment.

Audio Classification

Audio classification deals with classifying a sound into one of the several pre-defined categories

Challenges

- Intra-class variability
  - Features should provide good inter-class discrimination but still maintain intra-class cohesion
- Features must be robust to noise
- Granularity Issue
  - Trade-off between complexity of system and granularity of classes
- Real-time response
  - Computationally efficient classification structures and feature extraction algorithms

Cooperative Analog-Digital Signal Processing Lab

Problems with conventional features

- Work well in noise free case but performance degrades in presence of noise
- Accuracy is reduced greatly when different classes are presented simultaneously

Why auditory modeling?

- Humans do an extremely good job of classifying sounds
- Physiologically inspired perceptual features are
  - Highly discriminative
  - Robust to noise

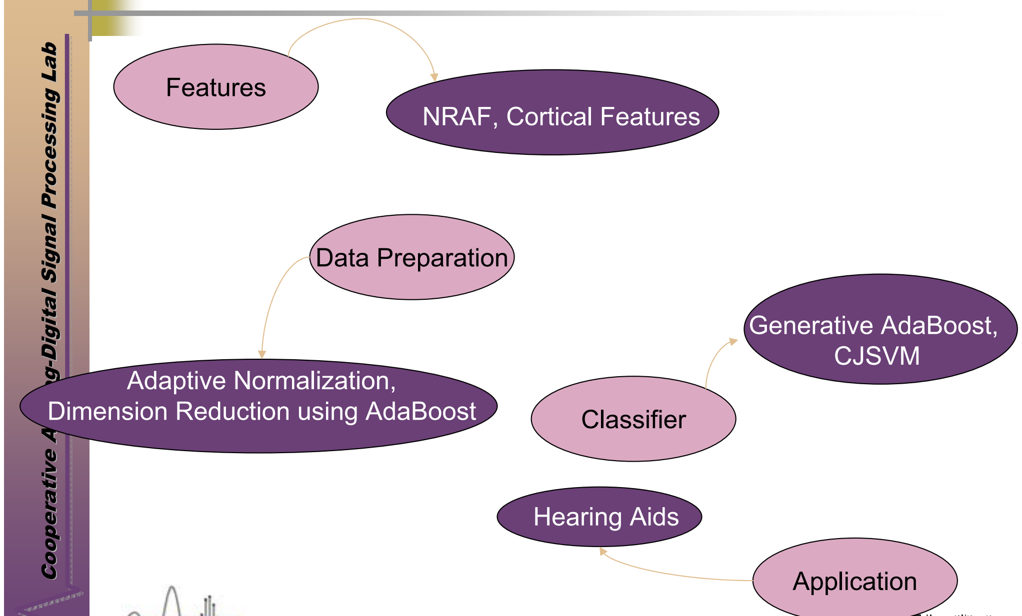
©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab

Part I – Perceptual Features

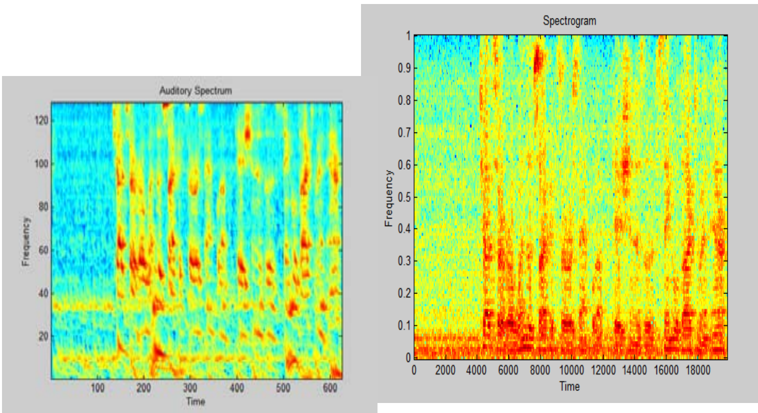
©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab



©Georgia Institute of Technology
May 2005

Auditory Spectrum Vs Spectrogram

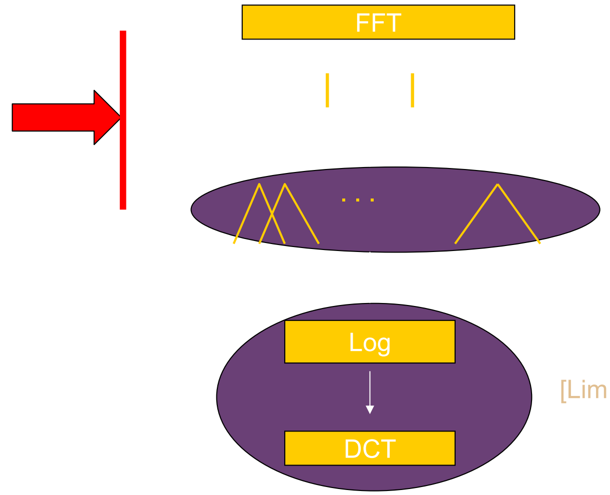


Auditory Spectrum

Spectrogram

Cooperative Analog-Digital Signal Processing Lab

Noise robustness of MFCCs



[Beet2002]

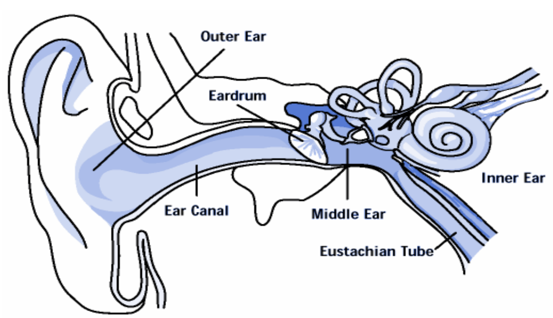
[Lim1979, Lockwood1993, Hansen2001]

©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab

Shihab's Early Auditory Model

[Shamma1996]



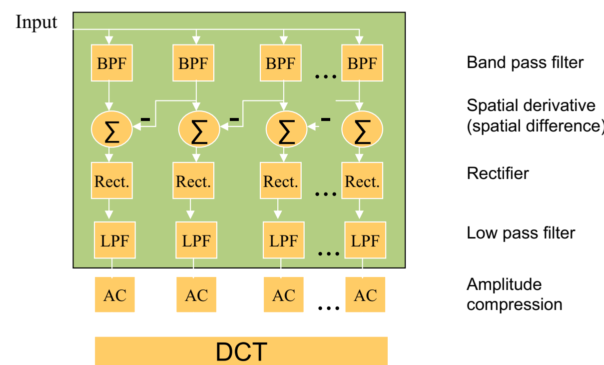
Input $h(t;s)$ | ∂t | g( ) | w(t) | ∂s | v(s) | HWR | $\int_T$

Cochlea | Hair cell stage | Cochlear nucleus

©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab

Noise-Robust Auditory Features (NRAF)



NRAF*

[*] Sourabh Ravindran, David V. Anderson and Malcolm Slaney, "Low Power Audio Classification for Ubiquitous Sensor Networks", ICASSP 2004, Montreal, Canada.

Paul Smith and Matt Kucic and Rich Ellis and Paul Hasler and David V. Anderson, "Cepstrum frequency encoding in analog floating-gate circuitry", ISCAS, 2002, Phoenix, AZ.

©Georgia Institute of Technology
May 2005

Motivation for Using BPFs

$$(\Delta a)(\Delta b) \geq \frac{1}{2}\left|[A,B]\right| \qquad A=t \qquad B=-j\frac{d}{dt}$$

(Uncertainty Principle)

$$(\Delta t)^2 = \int (t - E(t))^2 \, |s(t)|^2 \, dt$$

$$(\Delta \omega)^2 = \int (\omega - E(\omega))^2 \, |\hat{s}(w)|^2 \, d\omega$$

$$[A,B] = AB - BA = -j$$

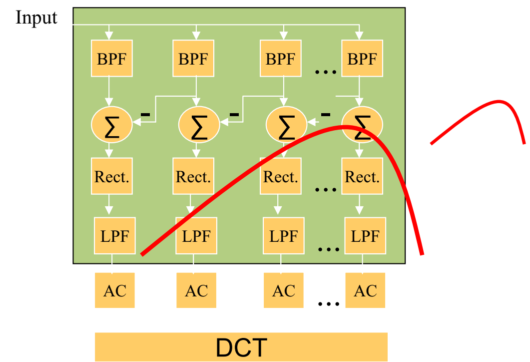
$$(\Delta t)(\Delta w) \geq \frac{1}{2}$$

(Time-frequency trade-off)

Leon Cohen, "Time-Frequency Distributions – A Review", Proceedings of IEEE, VOL. 77, NO. 7, JULY 1989

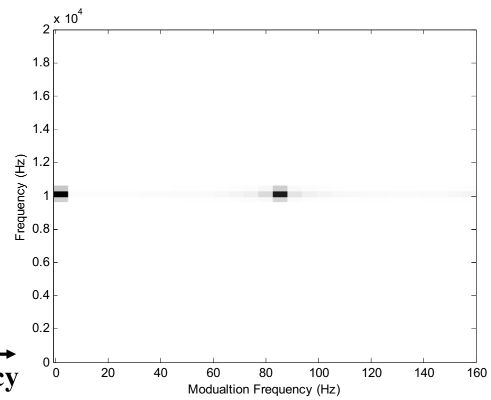
Richard Lyon, "A Computational Model of Filtering, Detection and Compression in the Cochlea", ICASSP, May 1982, Paris.

Asymmetrical Shape



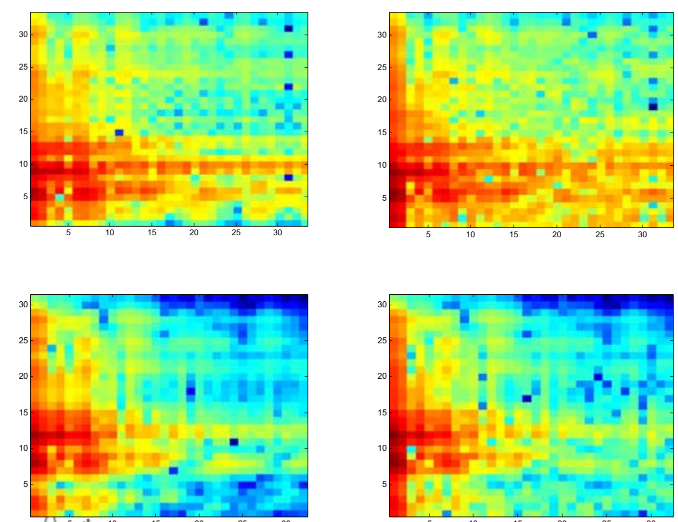
NRAF*

Modulation Spectra

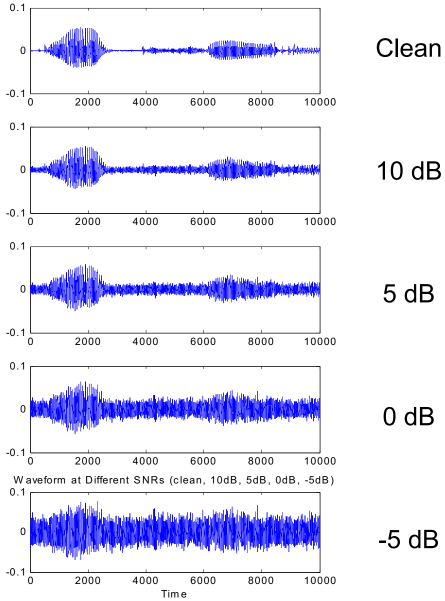


Modulation transform of a 10kHz tone modulated by a sinusoid at 80Hz

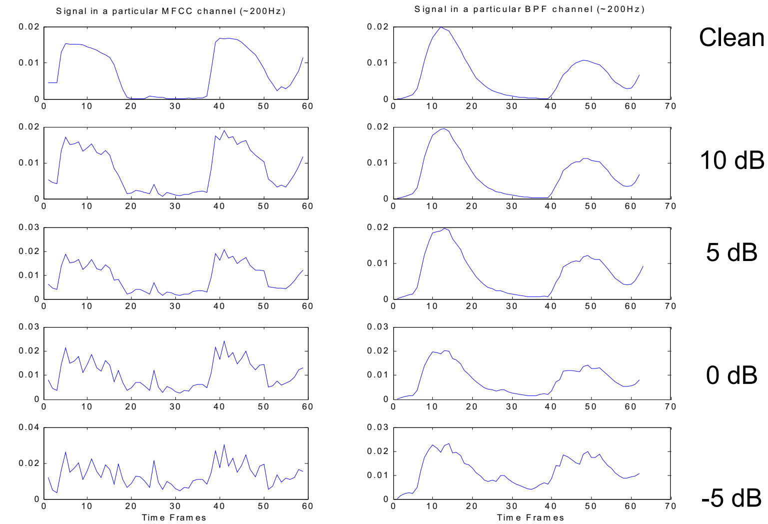
Modulation Spectra Comparison (babble noise)



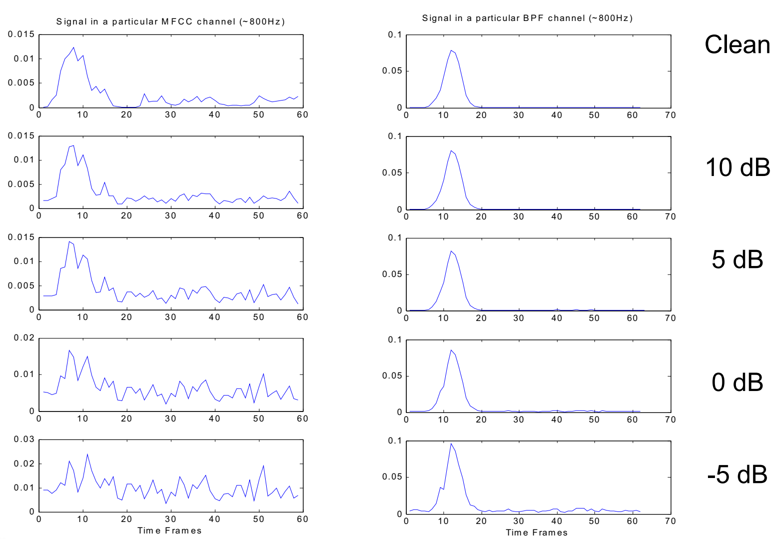
Speech signal at various SNRs



Signal in a particular Channel (~200 Hz)



Signal in a particular Channel (~800 Hz)

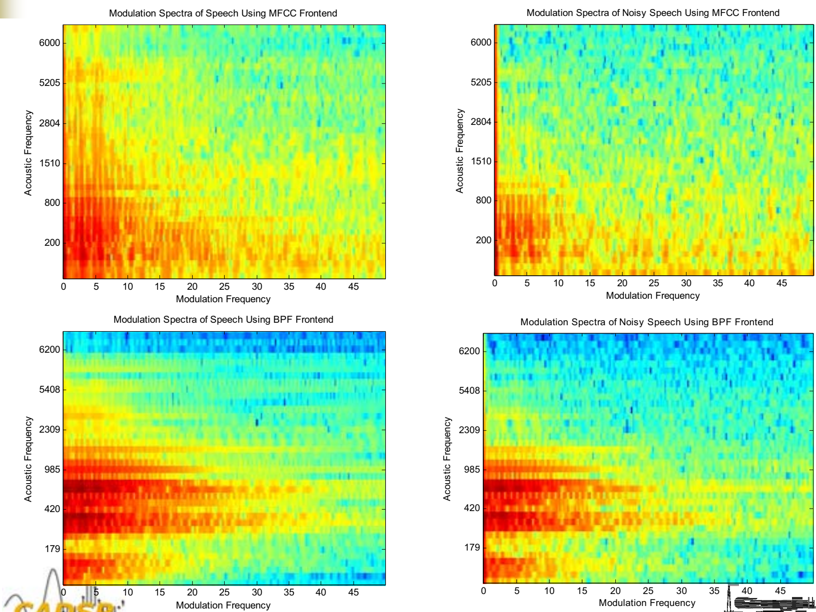


Proposed Solution – BPF-MFCC

- Replace FFT by filter-bank
- Do peak detection in each channel
- Root compression
- This is similar to the analog implementation of MFCC proposed in [Smith2002].

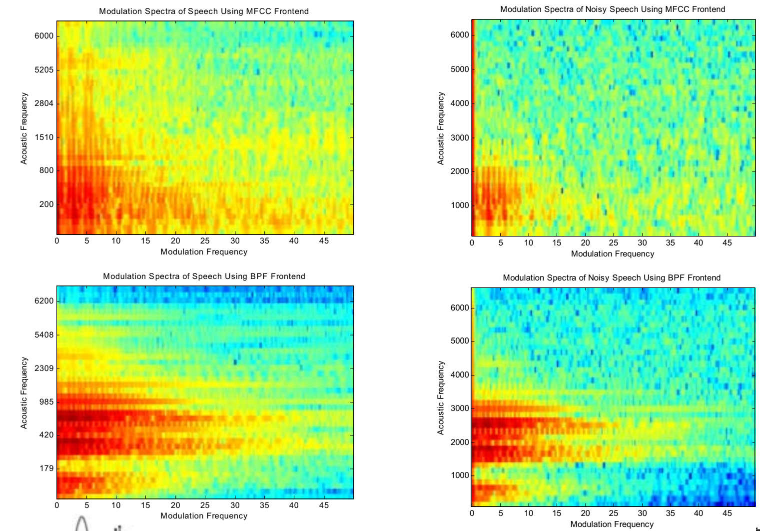
Cooperative Analog-Digital Signal Processing Lab

Modulation Spectra Comparison (pink noise)



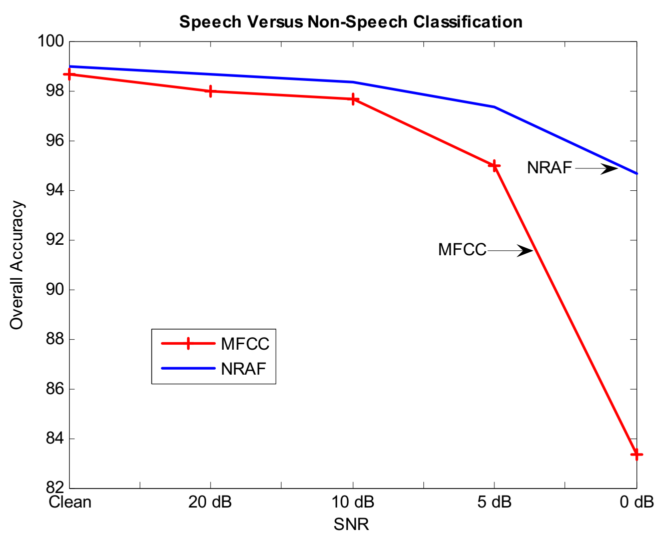
Cooperative Analog-Digital Signal Processing Lab

Modulation Spectra Comparison (white noise)



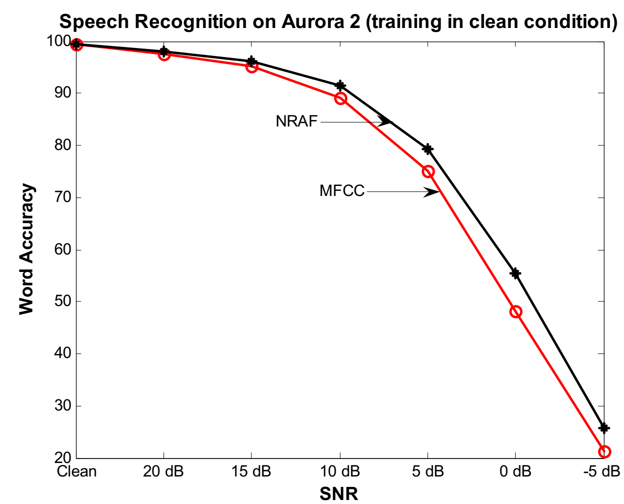
Cooperative Analog-Digital Signal Processing Lab

Speech versus non-speech classification



Cooperative Analog-Digital Signal Processing Lab

Connected Digits Recognition



Information theoretic measure of clustering [Dom2001]

Conditional Entropy:

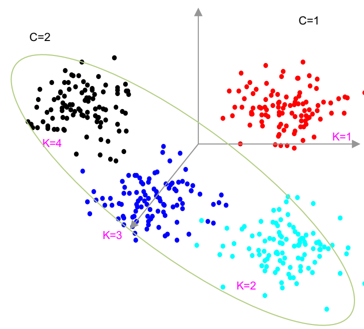
$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} p(c,k) log p(c|k)$$

$$H^e(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} \frac{h(c,k)}{n} log \frac{h(c,k)}{h(k)}$$

$$H^e(C|K) = H^e(C,K) - H^e(K)$$

Mutual Information:

$$I^e(C;K) = H^e(C) - H^e(C|K)$$

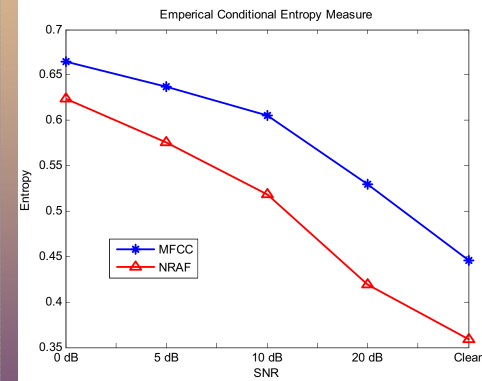


$$H^e(C) = -\sum_{c=1}^{|C|} \frac{h(c)}{n} log \frac{h(c)}{n}$$

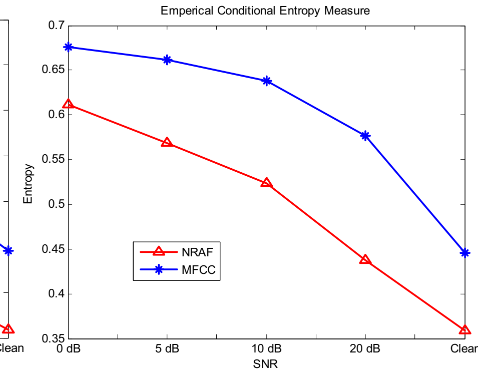
Cooperative Analog-Digital Signal Processing Lab

Empirical Conditional Entropy Measure

Pink Noise



White Noise

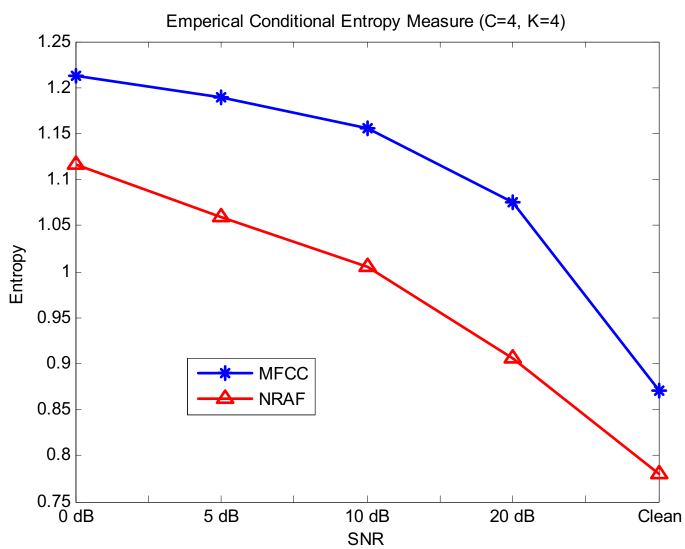


C = 2 , K = 4

©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab

What about class discrimination (for C>2)?



©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab

Noise Modulation Filtering

Let, $x(t) = s(t) + n(t)$

Assuming, $s(t) = \sum_i e_{s_i}(t) v_i(t)$

Output at the spatial derivative stage is,

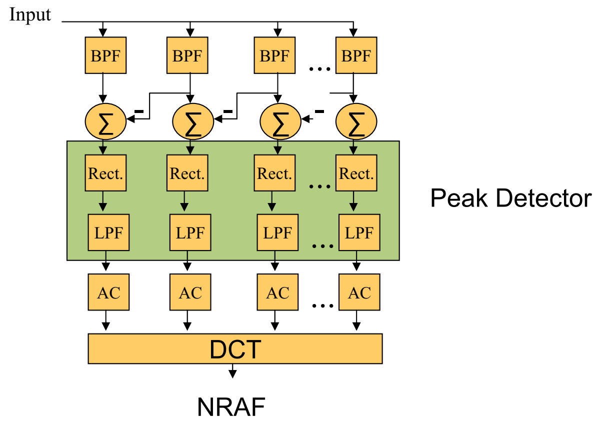
$$\left(s_i(t) + n_i(t)\right) - \left(s_{i+1}(t) + n_{i+1}(t)\right)$$

Peak detector output is given by,

$$\left(e_{s_i}(t) - e_{s_{i+1}}(t)\right) + \left(e_{n_i}(t) - e_{n_{i+1}}(t)\right)$$

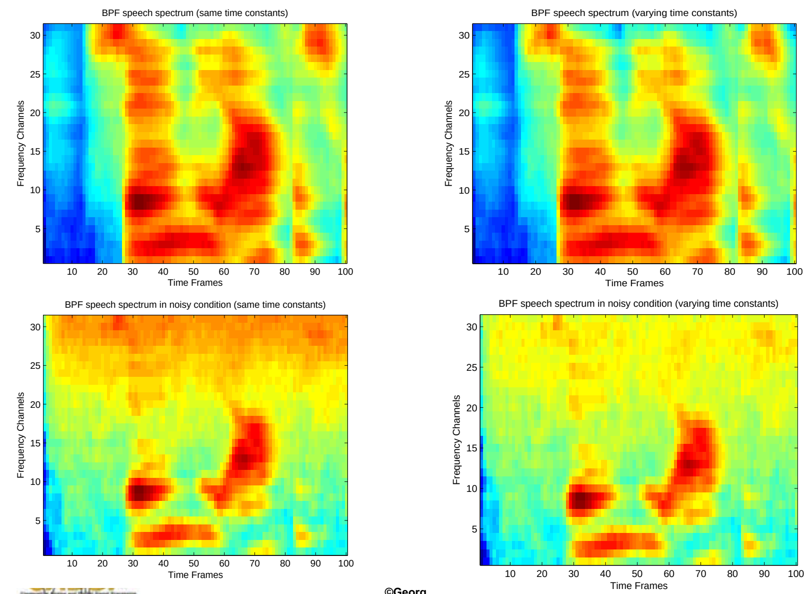
©Georgia Institute of Technology
May 2005

Varying Time-constants



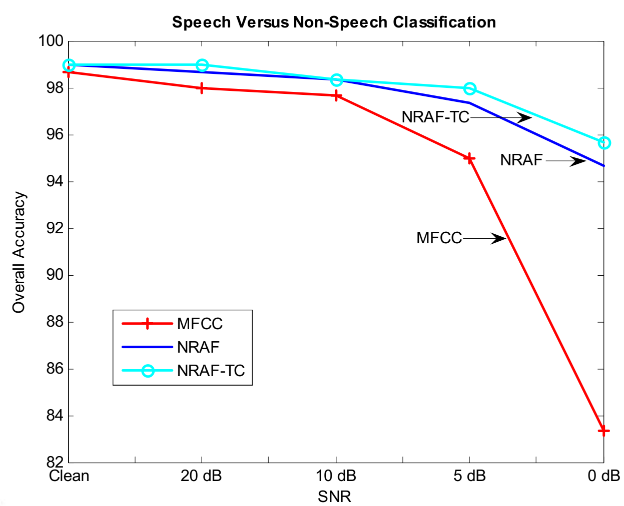
Cooperative Analog-Digital Signal Processing Lab

Example showing usefulness of varying TC

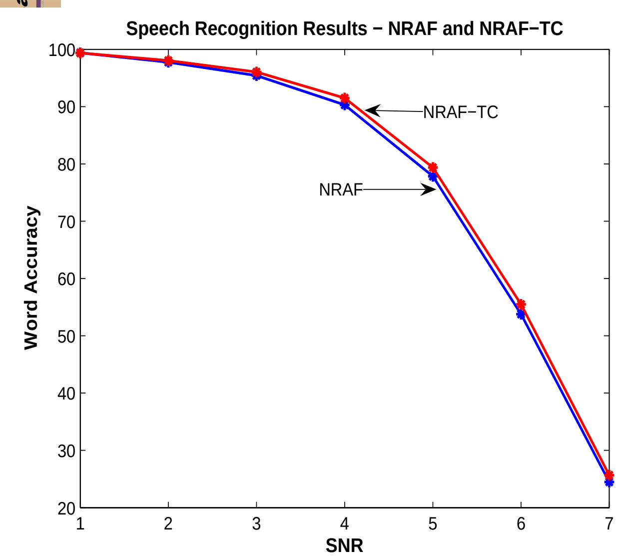


Cooperative Analog-Digital Signal Processing Lab

Results - Speech versus non-speech classification

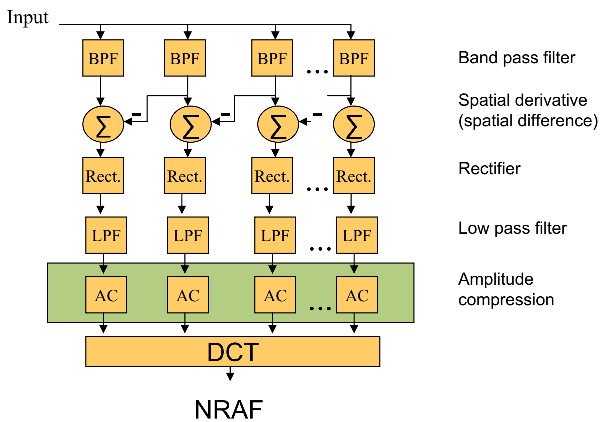


Results – Speech recognition



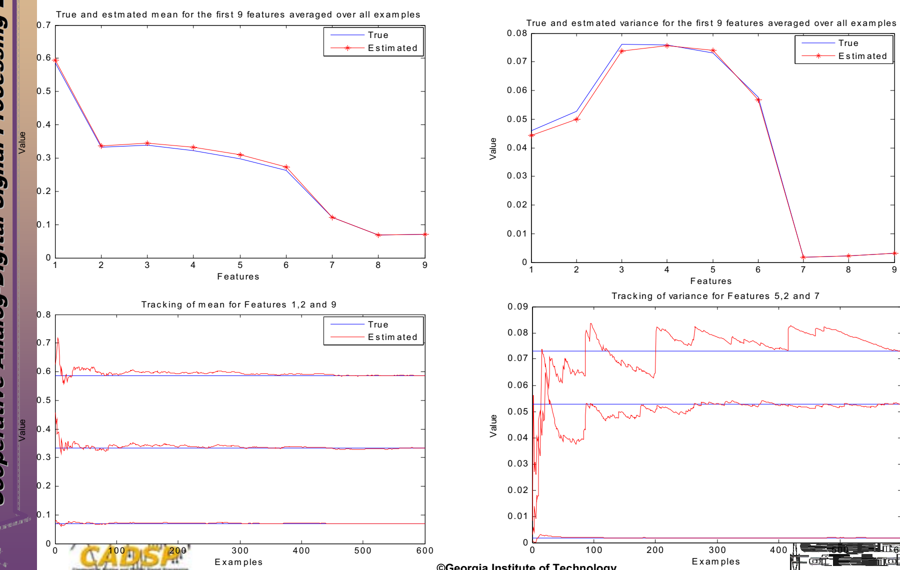
Condition	Significance level	Differences
Clean	Not significant	-
20 dB	0.4	2.94
15 dB	0.2	13.72
10 dB	0.05	38.18
5 dB	0.1	42.40
0 dB	0.1	51.49
-5 dB	0.2	29.44

Gain Adaptation

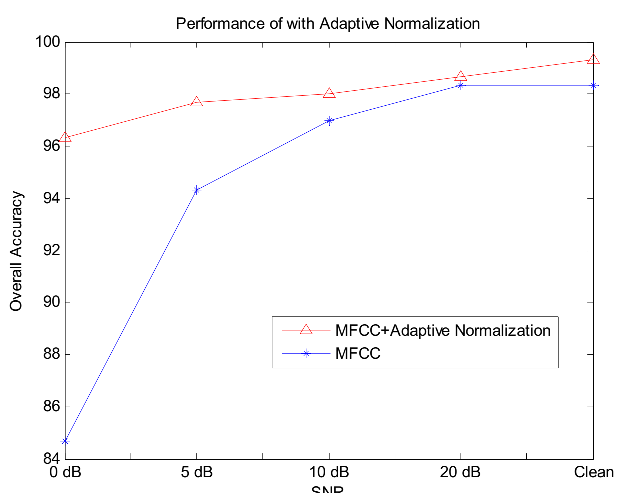


Adaptive Normalization

- Use Kalman filter to track the mean and variance of the test data.



Adaptive Normalization Results (Speech vs non-speech classification)



$$s(t) = \sum e_k(t) v_k(t)$$

$$\log \hat{e}_k(t) = \alpha \log e_k(t) + \log \beta \qquad (1)$$

$$\boxed{\begin{aligned} \hat{e}_{k_{\max}} &= e_{k_{\max}} \\ \hat{e}_{k_{\min}} &= K e_{k_{\min}} \end{aligned}}$$

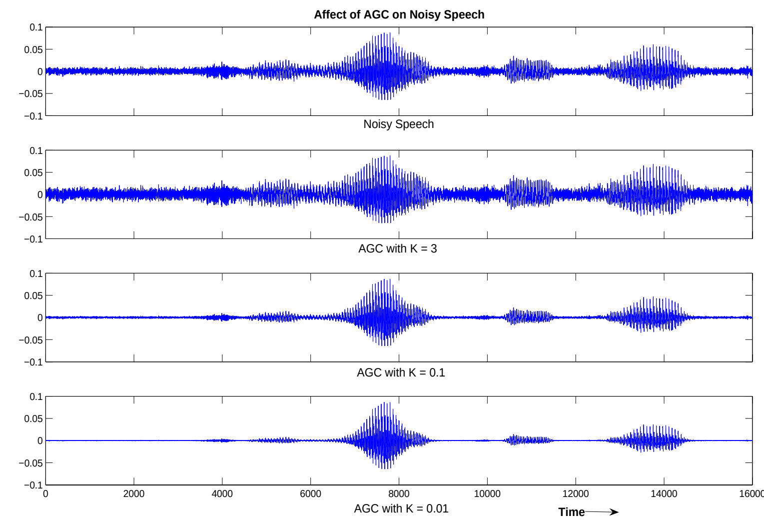
$$\beta = e_{k_{\max}}^{1-\alpha}$$

$$\alpha = 1 - \frac{\log(K)}{\log(M)} \qquad M = \frac{e_{k_{\max}}}{e_k}$$

$$G = \left(\frac{e_{k_{\max}}}{e_k} \right)^P \quad \text{(2)}$$

$$P = \frac{\log(K)}{\log(M)}$$

$$W(f) \approx \frac{(SNR(f))^2}{(SNR(f))^2 + 1} \quad \text{(3)}$$



Results – Speech Recognition

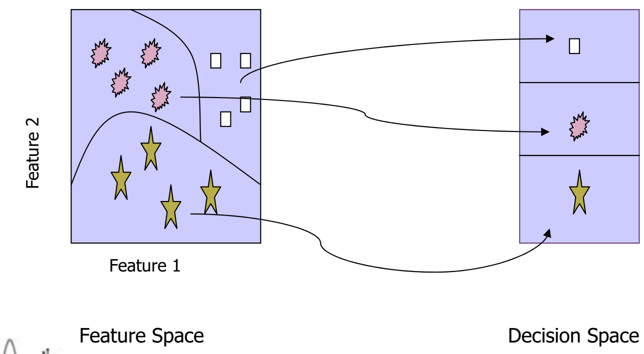
	NRAF	NRAF-AGC		
		K=0.05	K=0.01	K=0.005
Clean	99.51	99.48	99.42	99.23
20 dB	97.73	98.13	98.10	98.04
15 dB	95.73	96.50	96.56	96.90
10 dB	90.76	92.39	92.54	93.03
5 dB	79.71	83.02	83.79	84.92
0 dB	59.69	64.54	65.67	69.08
-5 dB	37.80	41.51	42.19	44.24

Noise Suppression

Part II – Classification Structure

Pattern Classification

- Pattern Classification can be viewed as the mapping of the feature space into the decision space.



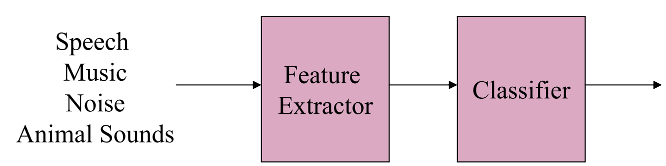
Classification Methods

- Gaussian Mixture Models
  - Models each class with a N-dimensional Gaussian
- Artificial Neural Network Classifier
  - Auditory features tend to work better with neural
  - nets based classifier/ recognizer
- AdaBoost based classifier
- Support Vector Machines
- …

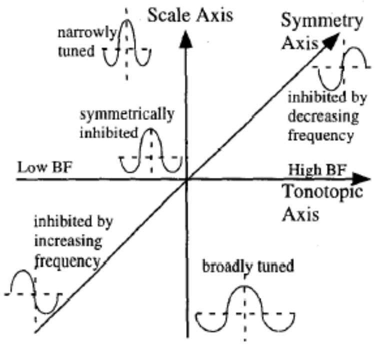
Description of problem

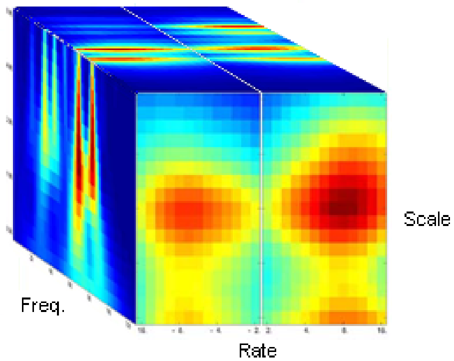
Humans are much more effective at audio understanding than machines. We can distinguish subtle changes in speech or a variety of other sounds that are difficult to quantify for a computer.

This research is focused on developing front-end *feature extraction* and *classification systems* for audio signals inspired by the human auditory system.



Cortical Model [Shamma1997]





Cooperative Analog-Digital Signal Processing Lab

AdaBoost Classifier [Viola2000]

- Given examples $(x_1, y_1),\ldots,(x_n, y_n)$ where $y_i = 0,1$ for negative and positive examples respectively.
- Initialize weights $w_{1,i} = 1/(2m), 1/(2n)$ for $y_i = 0,1$ respectively, where m and n are the number of negatives and positives respectively.
- For t = 1 to T

1. Normalize weights,

$$w_{t,i} = w_{t,i} / (\sum_j w_{t,j})$$

2. Train h_j ; error, $\varepsilon_{t,j} = \sum_i w_{t,i} \mid h_j(x_i) - y_i \mid$
3. Choose classifier h_t, with the least ε_t
4. Update weights: $w_{t+1,i} = w_{t,i}(\beta_t)^{(1-e_i)}$

$$\beta_t = \varepsilon_t / (1-\varepsilon_t)$$

$e_i = 0$ if x_i if classified correctly,
1 otherwise

©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab

The final strong classifier is:

$$h(x) = 1 \quad \text{if} \quad \sum_{t=1}^{T} \alpha_t h_t(x) \geq (1/2) \sum_{t=1}^{T} \alpha_t$$

where, $\alpha_t = \log(1/\beta_t)$

$= 0$ else

Convert to multi-class problem by using several 1-versus-1 classifiers.
Deadlocks resolved by normalized confidence measure.

©Georgia Institute of Technology
May 2005

Cooperative Analog-Digital Signal Processing Lab

Main Results I

AWARE

Using boosting for classification and features derived from an advanced auditory model we achieved 97.7 % classification. Confusion matrix is as shown below,

True Class →

Classified As →

	Noise	Animal	Music	Speech
Noise	344	20	0	0
Animal	0	157	2	0
Music	0	3	352	0
Speech	0	0	0	246

We see that most of the errors are when animal sounds are wrongly classified as noise. The misclassified sounds were even hard for human listeners to categorize.

©Georgia Institute of Technology
May 2005

Main Results II

Phonak Database

	Phonak (30 sec data)	Version 1 (1 sec data)	Version 2 (1 sec data)	Version 3 (1 sec data)	Version 4 (1 sec data)	Version 5 (30 sec data)
Music	**80 %**	**87.9 %**	**92.1 %**	**93.3 %**	**84.8 %**	**100 %**
Speech	**90 %**	**82.9 %**	**84.5 %**	**85.4 %**	**88.1 %**	**91.6 %**
Noise	**80 %**	**79 %**	**84.05 %**	**84.05 %**	**91.8 %**	**91.6 %**
Noisy Speech	**65 %**	**84.1 %**	**80.6 %**	**82.5 %**	**86.5 %**	**100 %**
Overall	**78.8 %**	**83 %**	**85.3 %**	**86.3 %**	**87.8 %**	**95.8 %**

Using the Phonak database, we outperformed their classification using only 1 second segments. (They require 30 seconds of data to make the classification.)

Results

Phonak Database

	Phonak (30 sec data)	Version 1 (1 sec data)	Version 2 (1 sec data)	Version 3 (1 sec data)	Version 4 (1 sec data)	Version 2 (30 sec data)
Overall	78.85 %	83 %	85.3 %	86.3 %	87.7 %	95.8 %

Tel-03 Database

	GMM	AdaBoost 1	AdaBoost 2	AdaBoost 3	Cascade
Overall	92.7 %	93.3 %	93.6 %	95.5 %	97.8 %

Complete Table (Hit Rate)

Phonak Database

	Phonak (30 sec data)	Version 1 (1 sec data)	Version 2 (1 sec data)	Version 3 (1 sec data)	Version 4 (1 sec data)	Version 5 (30 sec data)
Music	80 %	87.9 %	92.1 %	93.3 %	84.8 %	100 %
Speech	90 %	82.9 %	84.5 %	85.4 %	88.1 %	91.6 %
Noise	80 %	79 %	84.05 %	84.05 %	91.8 %	91.6 %
Noisy Speech	65 %	84.1 %	80.6 %	82.5 %	86.5 %	100 %
Overall	78.8 %	83 %	85.3 %	86.3 %	87.8 %	95.8 %

Complete Table (False Rate)

Phonak Database

	Phonak (30 sec data)	Version 1 (1 sec data)	Version 2 (1 sec data)	Version 3 (1 sec data)	Version 4 (1 sec data)	Version 2 (30 sec data)
Music	10 %	2.7 %	3.4 %	3.3 %	2.8 %	0 %
Speech	7.8 %	1.6 %	2.0 %	1.9 %	3.4 %	0 %
Noise	10 %	6.2 %	5.7 %	5.1 %	4.4 %	0 %
Noisy Speech	7.8 %	11.2 %	8.3 %	7.8 %	5.6 %	4.1 %

GMM and AdaBoost

Sound Classification (4 classes)	
Classifier	% Correct
GMM	92.25
AdaBoost	93.06

NRAF

AdaBoost	97.68

NRAF + Cortical Features

Feature Fusion