Spectral Moment and Micro-Modulations Features for Robust ASR

Prof. Alexandros Potamianos

Technical University of Crete

Visit @ AT&T Labs

June 5, 2012

Acknowledgements

- Ph.D. work of Dr. Pirros Tsiakoulis (now @ Cambridge University) in collaboration with Dr. Dimitrios Dimitriadis @ NTUA/AT&T

- Publications

1. P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, " Spectral Moment Features Augmented by Low Order Cepstral Coefficients for Robust ASR ," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551-554, June 2010
2. P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, " Short-time instantaneous frequency and bandwidth features for speech recognition ," in *Proc. Automatic Speech Recogn. and Underst. Workshop* (ASRU-2009), Merano, Italy, Dec. 2009.
3. A. Potamianos and P. Tsiakoulis, “Robust Instantaneous Frequency and Bandwidth Estimation using Filterbank Arrays”, submitted to InterSpeech 2012.
4. D. Dimitriadis, P. Maragos, and A. Potamianos, " Robust AM-FM features for speech recognition ," *IEEE Signal Processing Letters*, vol. 12, pp. 621-624, Sept. 2005.

Outline

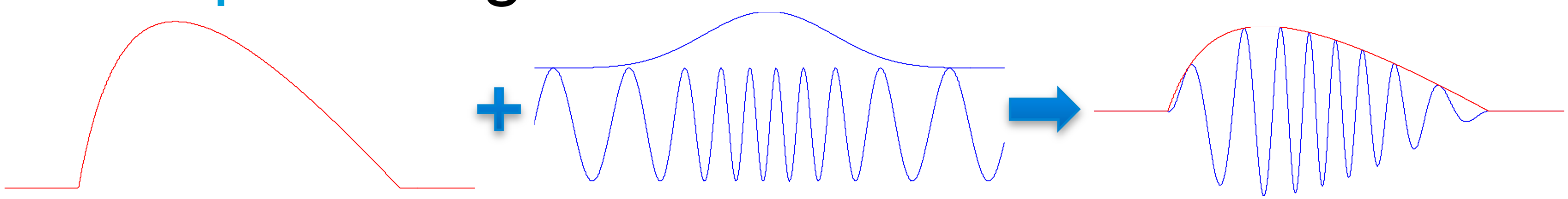
- Motivation
  - perceptual importance of frequency
- AM-FM and SMAC features
  - Instantaneous amplitude and frequency signals
  - Time vs frequency domain estimation
  - Spectral Moments features
- Recognition Experiments

Perceptual importance of frequency

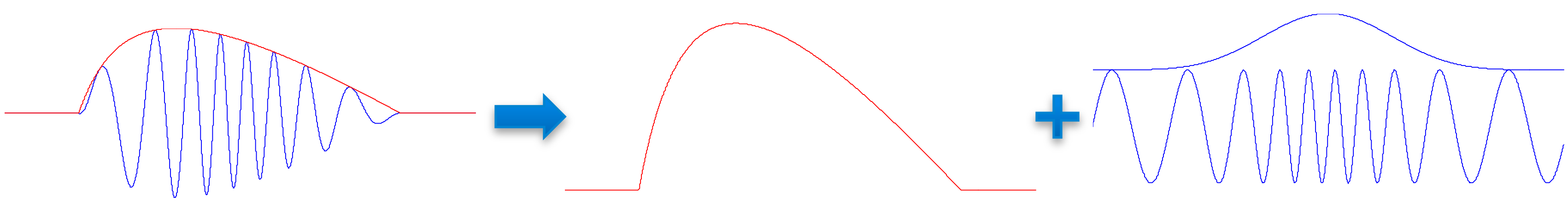
- Chimaeric sounds reveal dichotomies in auditory perception
  - [Smith Z. M., Delgutte B. and Oxenham A. J., Nature 2002]
  - [http://research.meei.harvard.edu/chimera/index.html]
- Speech recognition with amplitude and frequency modulations
  - [Zeng F.G. et al, PNAS 2005]
- Our work
  - [ICASSP 2009, ASRU 2009]
  - recent results

The AM-FM speech model

- The speech signal is modeled as a sum of resonant signals each one being an AM-FM composite signal

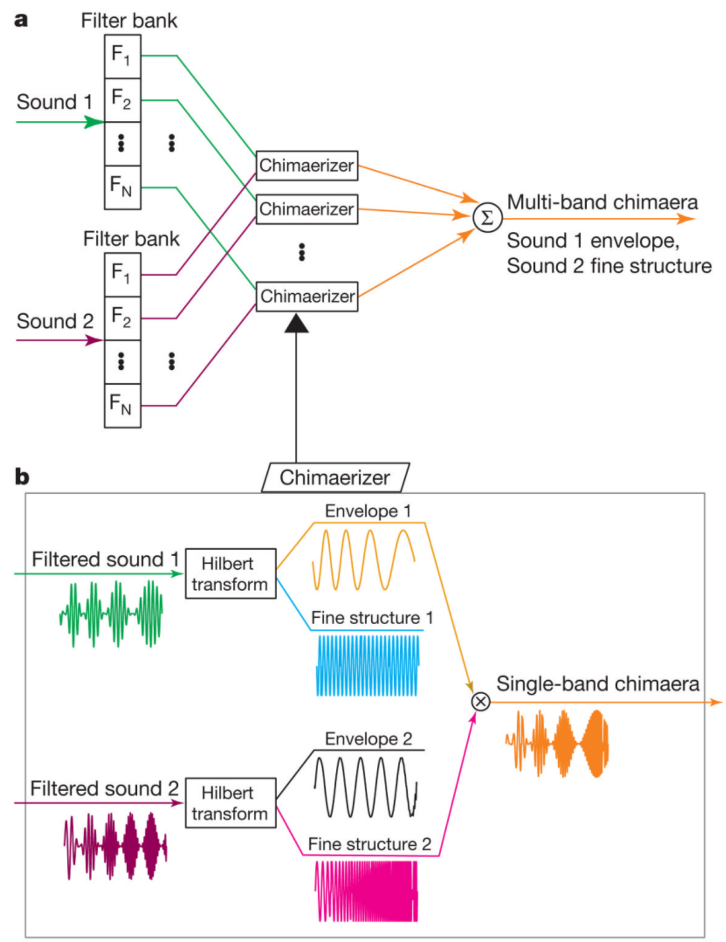


- The demodulation problem



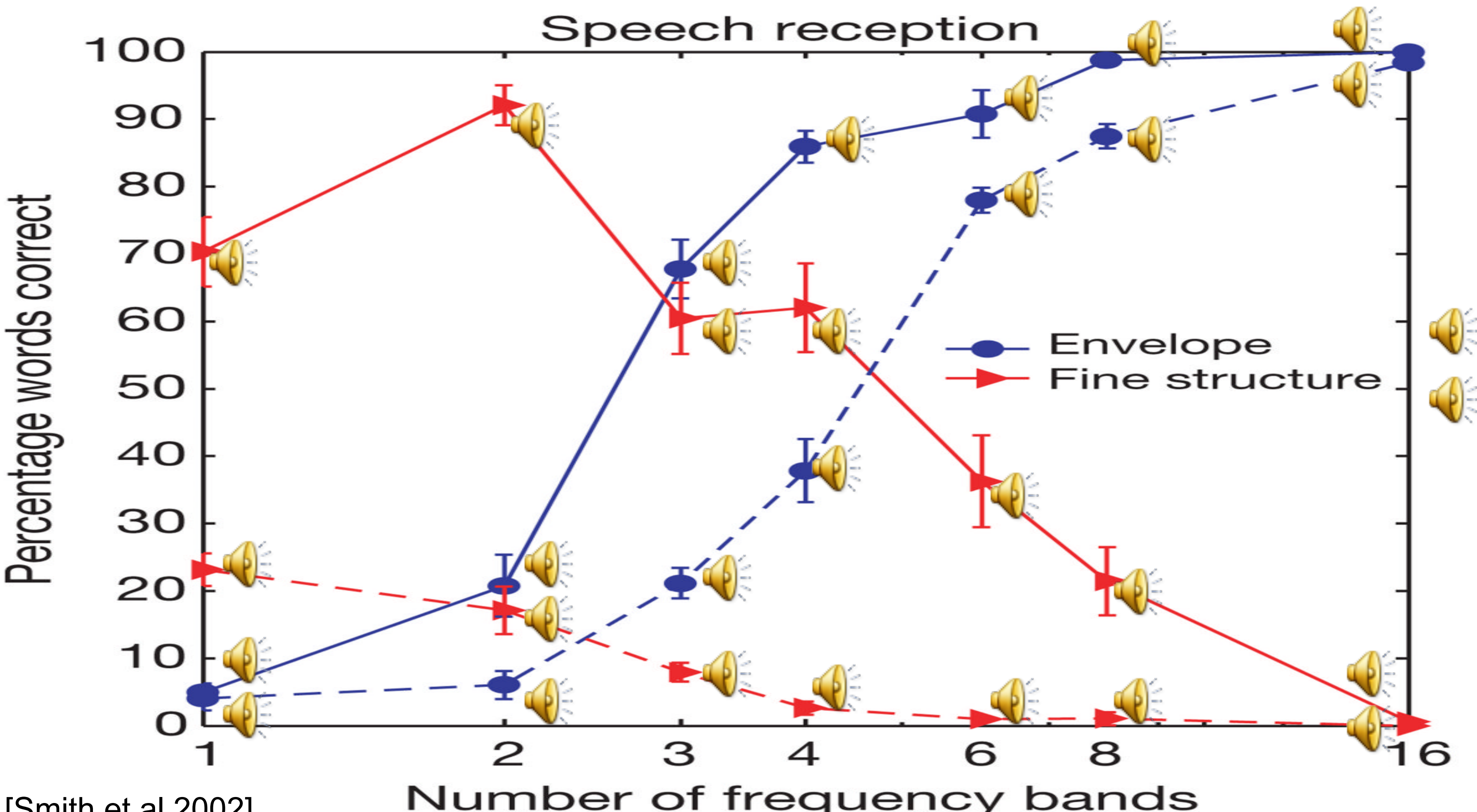
Chimaera synthesis

- Filterbank analyis
  - 80-8,820 Hz
  - number of filters: variable
- Hilbert Trasform – Analytic Signal
  - amplitude envelope
  - fine structure: $\cos(\varphi(t))$
- Two input signals
  - envelope from 1st
  - fine structure 2nd



[Smith et al 2002]

Chimaeras reception results: Speech-Noise, Speech-Speech



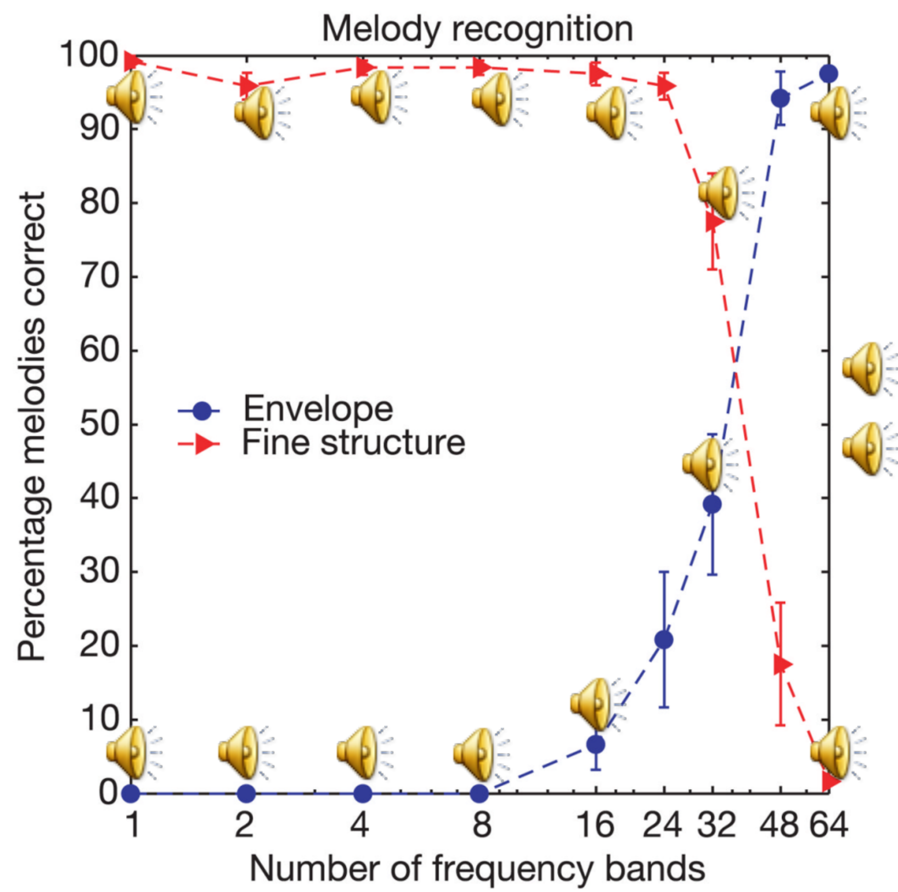
[Smith et al 2002]

Chimaeras reception results: Speech-Noise, Speech-Speech

- Reception highly depends on number of bands
- Speech envelope – Noise fine structure
  - reception improves as number of bands increases
  - good performance for very few bands 4
- Noise envelope – Speech fine structure
  - reverse behaviour
  - good reception with only 1-2 bands
- Speech – Speech
  - envelope dominates fine structure
- Amplitude conveys ‘what’ information

Chimaeras reception results: Melody-Melody

- Reversal of the relative importance between envelope and fine structure
- Melody reception from fine structure up to 32 bands
- Crossover point around 40 bands
  - bandwidths become narrower than the critical bandwidths



Summary of findings

- Speech envelope
  - conveys phonetic information ('what')
- Fine structure
  - less phonetic information
  - pitch perception / localization ('where')
  - rhyme, melody
- Listening tests [Zeng et al, 2005]
  - AM performs well in noise free situations
  - FM improves performance in noise

Acoustic representation for speech recognition

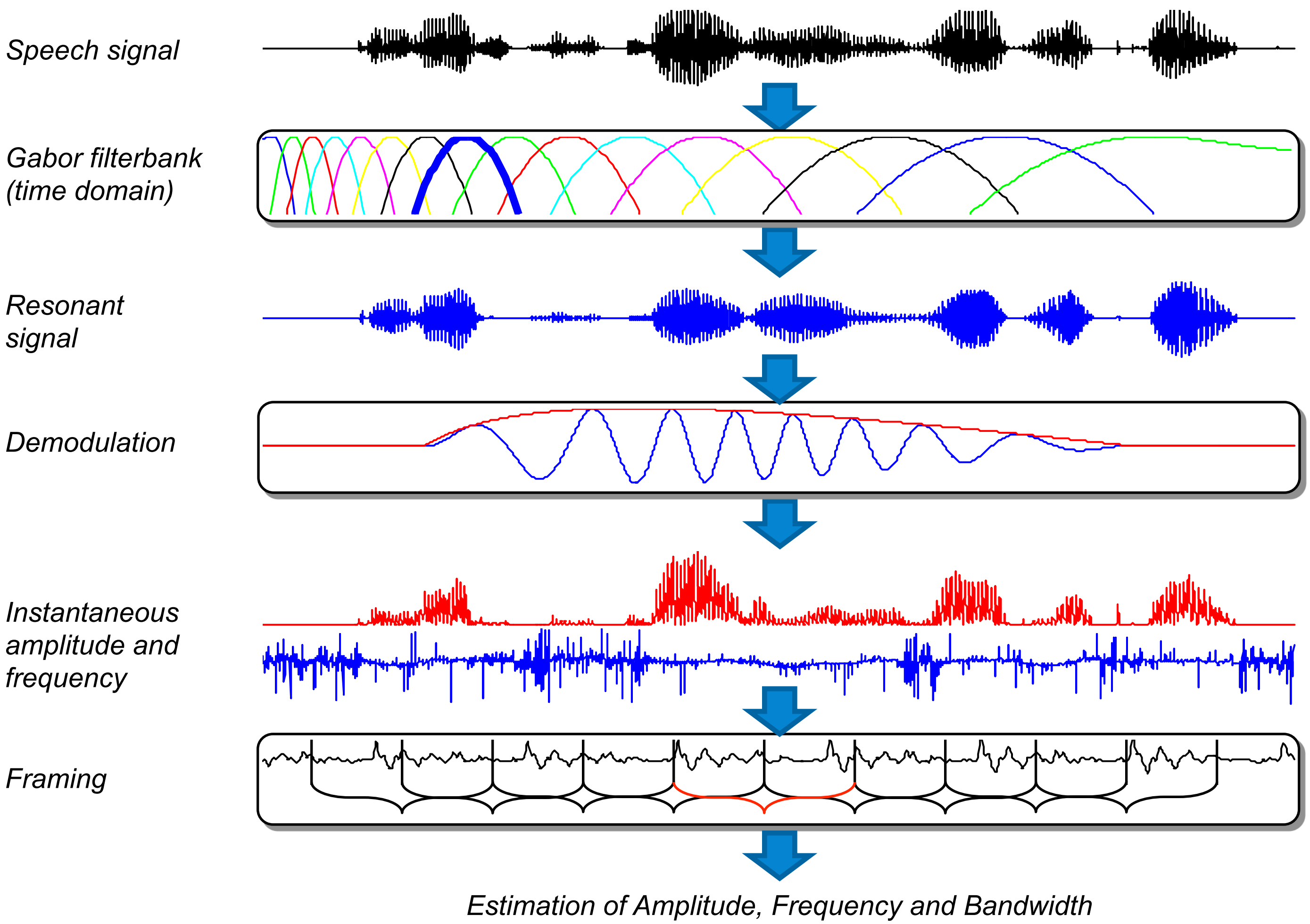
Related work

- MFCC – standard acoustic representation
  - [Davis & Mermelstein 1980]
  - energy measure with a triangular mel filterbank with 50% overlap
- AM-FM Features
  - [Dimitriadis et al 2005, 2006]
  - few bands – appended to MFCC vector
  - FMP – bandwidth over frequency ratio
- Frequency representation
  - [Paliwal et al 2003, Chen et al 2004]
  - triangular linear filterbank with 50% overlap (spectral centroids)

Acoustic representation

- Time domain
  - amplitude (energy)
  - frequency
  - bandwidth
- Frequency domain
  - Spectral moments
- Parameterization for ASR front-end
  - decorrelation (DCT)
  - filterbank

Time domain


Speech signal
Gabor filterbank
(time domain)
Resonant
signal
Demodulation
Instantaneous
amplitude and
frequency
Framing
Estimation of Amplitude, Frequency and Bandwidth

Estimation of Amplitude, Frequency and Bandwidth

- Mean squared amplitude (energy measure)

$$A[i] = \log \sum_{n=0}^{N} (a_i[n])^2$$

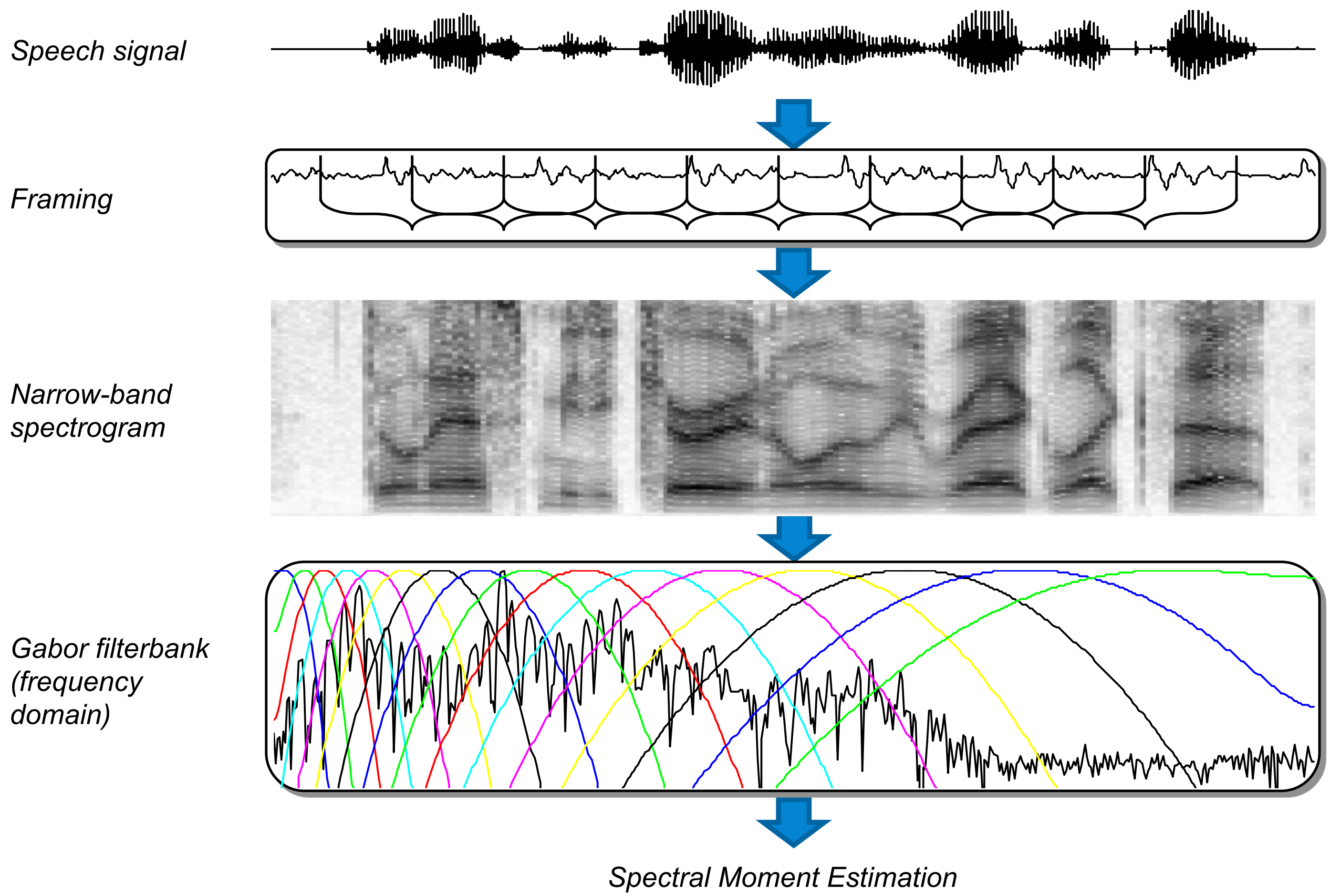
- Mean weighted frequency (biased)

$$F_w[i] = \frac{\sum_{n=0}^{N} (f_i[n] - F_i)(a_i[n])^2}{\sum_{n=0}^{N} (a_i[n])^2} = \frac{\sum_{n=0}^{N} f_i[n] (a_i[n])^2}{\sum_{n=0}^{N} (a_i[n])^2} - F_i$$

- Bandwidth

$$B_w^f[i] = \left(\frac{\sum_{n=0}^{N} \left[(f[n] - F_i)^2 (a[n])^2 \right]}{\sum_{n=0}^{N} (a[n])^2} \right)^{\frac{1}{2}} \quad B_w^a[i] = \left(\frac{\sum_{n=0}^{N} (\dot{a}[n]/2\pi)^2}{\sum_{n=0}^{N} (a[n])^2} \right)^{\frac{1}{2}}$$

Frequency domain



Spectral Moment Estimation

- Band passed signal of k-th filter

$$x_k(n) = x(n) * h_k(n) \leftrightarrow X_k(\omega) = X(\omega)H_k(\omega)$$

- Spectral moment of order m

$$S^m(k) = \int_0^\pi |X_k(\omega)|^\gamma \omega^m d\omega$$

- Central spectral moment

$$S_c^m(k) = \int_0^\pi |X_k(\omega)|^\gamma (\omega - \omega_k)^m d\omega$$

- Normalized spectral moments

$$N^m(k) = \frac{S^m(k)}{S^0(k)} \qquad N_c^m(k) = \frac{S_c^m(k)}{S_c^0(k)}$$

Time and Frequency domain duality
[see work of Cohen, Boashash]

- Amplitude – Energy (zero order moment)

$$\sum (a_k[n])^2 \leftrightarrow \sum |X_k[\Omega_n]|^2$$

$$A \leftrightarrow S^0 \equiv N^0$$

- Frequency – 1st spectral moment

$$\frac{\sum f_k[n]\,(a_k[n])^2}{\sum (a_k[n])^2} \leftrightarrow \frac{\sum \Omega_n |X_k[\Omega_n]|^2}{\sum |X_k[\Omega_n]|^2}$$

$$F_w \leftrightarrow N^1 \equiv \omega_k + N_c^1$$

- Bandwidth – 2nd spectral moment …

1^{st} vs 0^{th} spectral moment

$$\frac{dS^0(k)}{d\omega_k} = \frac{d}{d\omega_k}\int_0^\pi |X_k(\omega)|^\gamma d\omega$$
$$= \int_0^\pi \frac{d|X_k(\omega)|^\gamma}{d\omega_k} d\omega$$
$$\simeq \int_0^\pi |X(\omega)|^\gamma \frac{d|H_k^+(\omega)|^\gamma}{d\omega_k} d\omega$$

$$\frac{d|H_k^+(\omega)|^\gamma}{d\omega_k} = (\sqrt{\pi}/2\alpha)^\gamma \frac{de^{-\gamma(\omega-\omega_k)^2/4\alpha^2}}{d\omega_k}$$
$$= (\sqrt{\pi}/2\alpha)^\gamma 2(\gamma/4\alpha^2)(\omega-\omega_k)e^{-\gamma(\omega-\omega_k)^2/4\alpha^2}$$
$$= (\gamma/2\alpha^2)(\omega-\omega_k)|H_k^+(\omega)|^\gamma$$

$$\frac{dS^0(k)}{d\omega_k} \simeq \frac{\gamma}{2\alpha^2}\int_0^\pi |X(\omega)|^\gamma |H_k^+(\omega)|^\gamma (\omega-\omega_k) d\omega$$
$$\simeq \frac{\gamma}{2\alpha^2}\int_0^\pi |X_k(\omega)|^\gamma (\omega-\omega_k) d\omega = \frac{\gamma}{2\alpha^2} S_c^1(k) \longrightarrow S_c^1(k) \simeq \frac{2\alpha^2}{\gamma}\frac{dS^0(k)}{d\omega_k}$$

$$\downarrow$$

$$\boxed{N_c^1(k) \simeq \frac{2\alpha^2}{\gamma S^0(k)}\frac{dS^0(k)}{d\omega_k} = \frac{2\alpha^2}{\gamma}\frac{d\log(S^0(k))}{d\omega_k}}$$

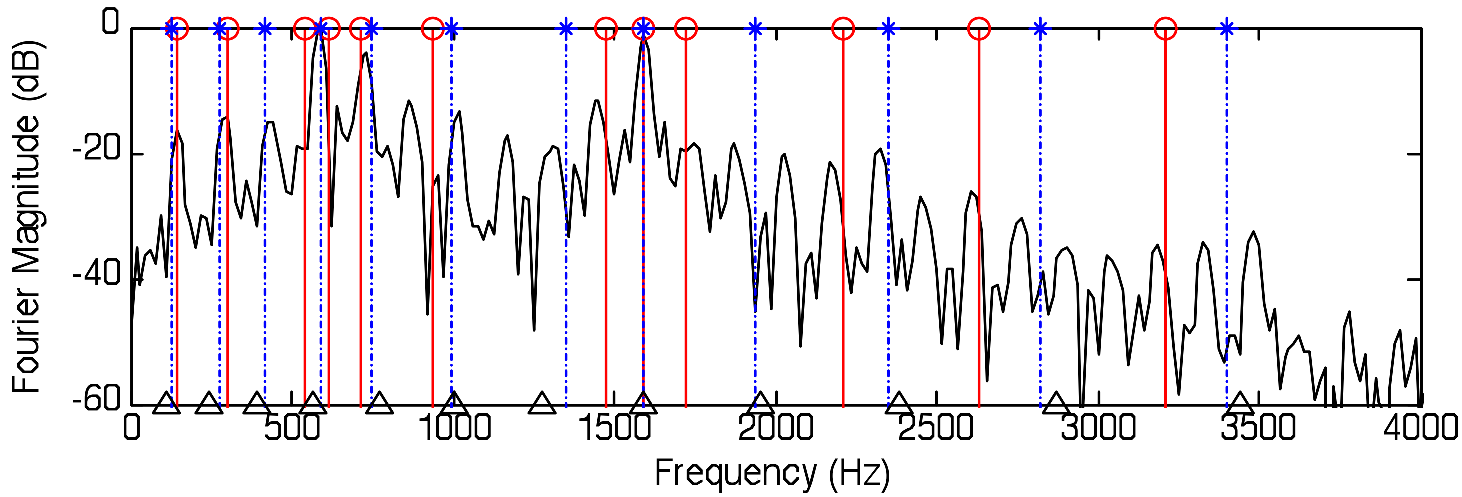
1st vs 0th spectral moment

- Proportional to the log power spectrum

$$N_c^1(k) \simeq \frac{2\alpha^2}{\gamma} \frac{d\log(S^0(k))}{d\omega_k}$$

- Depends on
  - the γ constant (usually is 2)
  - the bandwidth of the filter
- The energy information is lost
  - spectral tilt information not directly observable

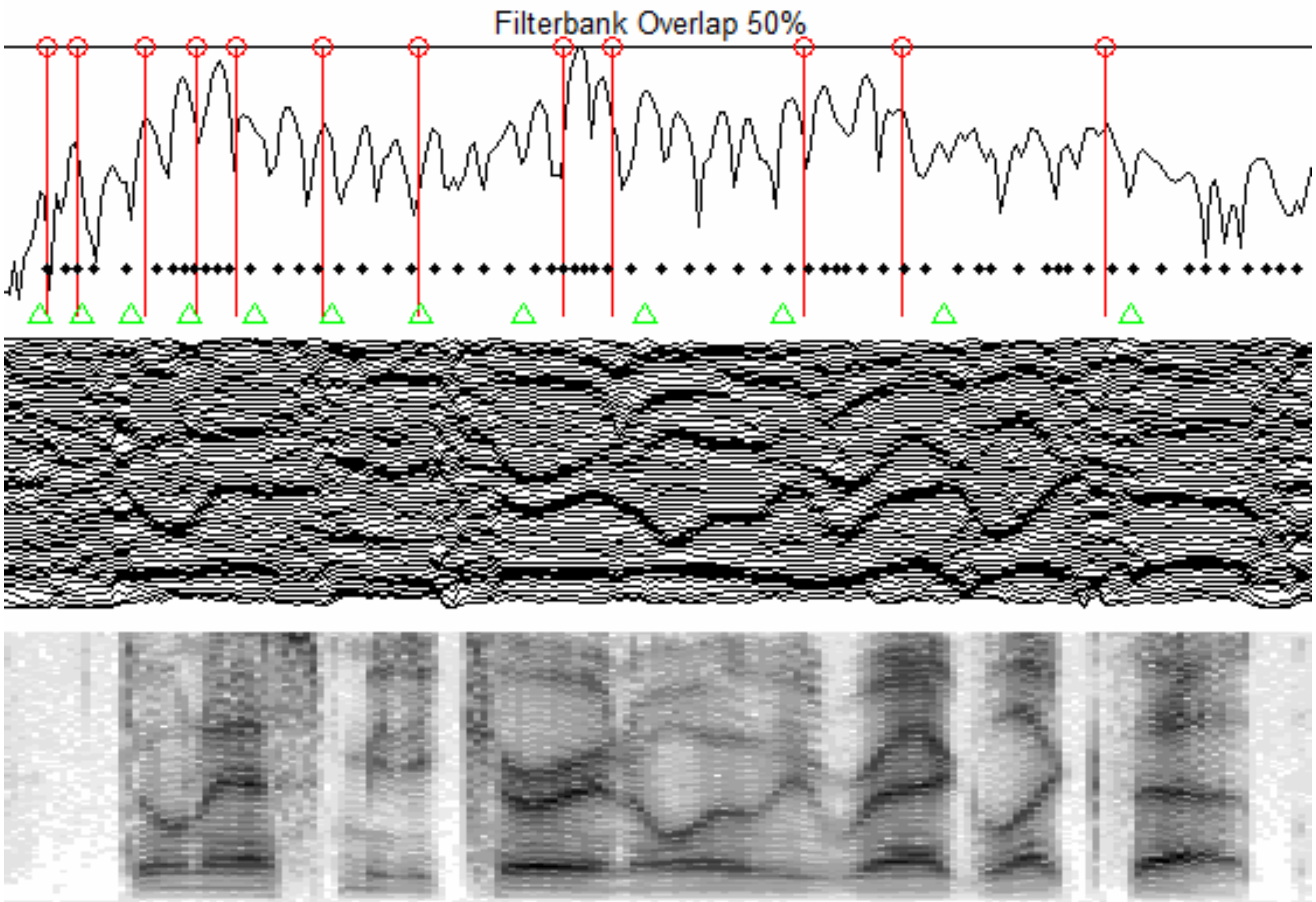
The role of the filter's bandwidth



- Filter's bandwidth
  - wider → formants
  - narrower → pitch harmonics

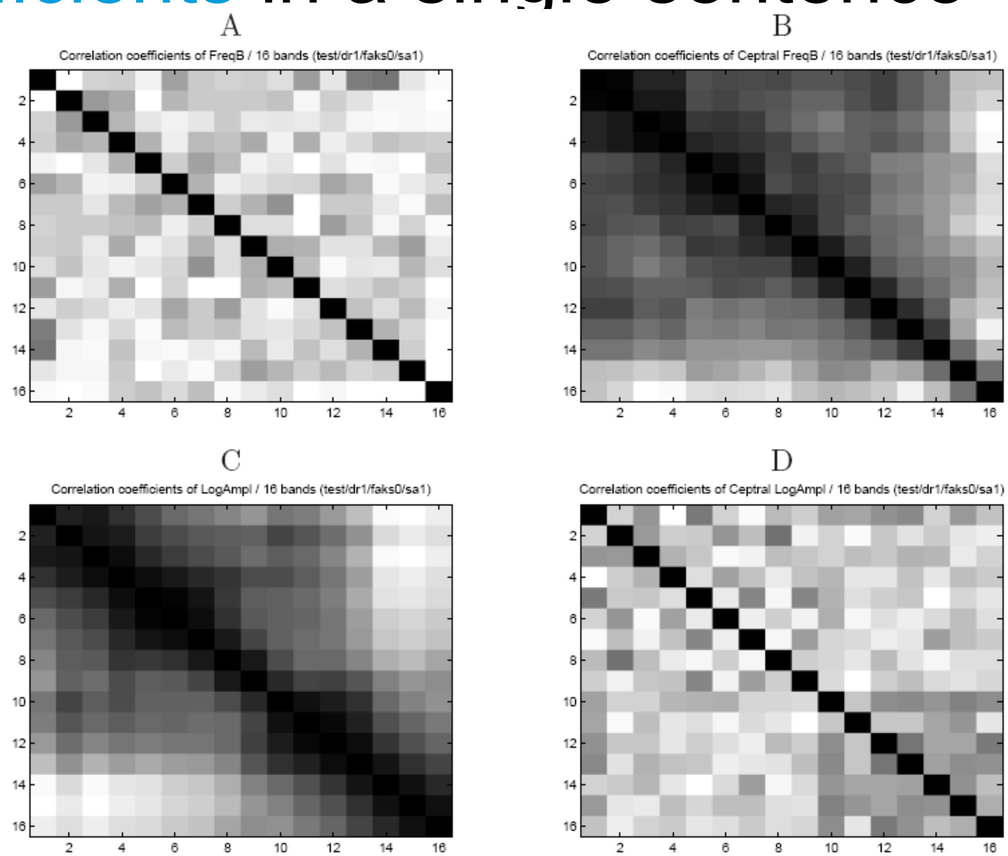
$$\alpha \to 0 \Rightarrow N_c^1(k) \to 0 \Rightarrow N^1(k) \to \omega_k$$

Speech Pyknogram: 2nd spectral Moment



The decorrelation problem

- Correlation coefficients in a single sentence
  - A: frequency
  - B: DCT of frequency
  - C: amplitude
  - D: DCT of amplitude
  - Amplitude components are highly correlated
  - Frequency components do not require correlation



Recognition experiments

Optimizing the filter's bandwidth

- TIMIT (61 phonemes)
  - 3 state HMMs / 16 Gaussians
- Bandwidth → frequency overlap
  - frequency requires higher overlap ~70%
  - amplitude is not seriously affected

Filterbank Overlap	50%	60%	70%	80%
A_{DCT}	**60.09**	**60.38**	59.95	58.86
F_w	49.57	59.40	**61.21**	60.86
B_w	37.37	46.51	51.14	**53.03**

Number of filters

	16	20	26
MFCC	60.20	60.58	**60.66**
A_{DCT}	60.09	60.68	**61.16**
F_w	61.21	**61.34**	59.88
N_c^1 (SM)	60.54	**61.02**	60.38
B_w	51.14	**51.22**	49.05
B_w^f	**48.17**	47.67	44.14
B_w^a	48.06	**49.37**	48.15
B_w^{a+}	50.49	**51.31**	50.95

- Amplitude in dB and transformed with DCT (equivalent to MFCC)
- Frequency
  - 70% overlap
  - no DCT
  - outperforms amplitude
- Bandwidth features have a notworthy performance (70% overlap, and no DCT)
- We focus on frequency based utilizing the first spectral moment

Energy and spectral envelope

	16	20	26
MFCC+E	64.06	**64.28**	64.10
MFCC+C0	64.16	**64.29**	64.24
F_w+E	63.78	**63.99**	62.55
F_w+C0	**64.28**	64.11	62.73
SM+C0	64.17	64.41	63.60
SMAC (SM+C0-C1)	**64.82**	64.80	64.58
SM+C0-C2	64.74	**65.19**	64.84
SM+C0-C3	64.64	65.06	**65.00**

SMAC

- Spectral Moment features Augmented by low order Cepstral coefficients
  - first order normalized central spectral moment
  - plus few cepstral coefficients

- Key advantages
  - retain the feature vector in the frequency domain
  - zero mean (due to the central moment)
  - robustness

AURORA 2

- Connected word recognition task
  - word HMMs / 16 states
  - various types and levels of noise
- SMAC: 12 filters up to 4 kHz + C0 + C1
  - significant gain for all noise levels

	20 dB	15 dB	10 dB	5 dB
MFCC (39)	94.07	85.04	65.51	38.45
PLP (39)	+0.09	+0.26	+1.63	+2.73
RASTA-PLP (39)	+2.59	+7.00	+11.62	+6.73
SMAC (42)	**+2.98** (3.17%)	**+8.01** (9.42%)	**+13.27** (20.26%)	**+8.03** (20.88%)

AURORA 3

- Car noise (Spanish and Italian tasks)
  - WM (well-matched), MM (medium-mismatched), HM (high-mismatched) conditions
  - same configuration as in the AURORA 2 task
- Performance improvement from WM to HM

	Spanish Task			Italian Task		
	WM	MM	HM	WM	MM	HM
MFCC (39)	86.88	73.72	42.23	**93.64**	82.02	39.84
PLP (39)	+5.16	+10.12	+10.49	-5.40	-9.51	-0.86
RASTA-PLP (39)	+7.06	+14.53	+30.70	-9.88	-6.75	**+23.49**
SMAC (42)	**+7.37** (8.48%)	**+15.49** (21.01%)	**+35.45** (83.95%)	-5.50 (-5.87%)	**+0.28** (0.34%)	+11.79 (29.56%)

Wiener Filtering

- Noise suppression using WF
- SMAC still outperforms MFCC

AURORA 2	20 dB	15 dB	10 dB	5 dB
WF+MFCC (39)	97.70	95.31	89.13	74.37
WF+SMAC (42)	-0.18 (-0.18%)	**0.38** (+0.40%)	**1.62** (+1.82%)	**3.09** (+4.15)

AURORA 3	Spanish Task			Italian Task		
	WM	MM	HM	WM	MM	HM
WF+MFCC (39)	94.84	88.31	78.32	**95.89**	**89.81**	**73.52**
WF+SMAC (42)	**+0.03** (0.03%)	**+2.78** (3.15%)	**+3.33** (4.25%)	-4.46 (-4.65%)	-3.39 (-3.77%)	-11.29 (-15.35)

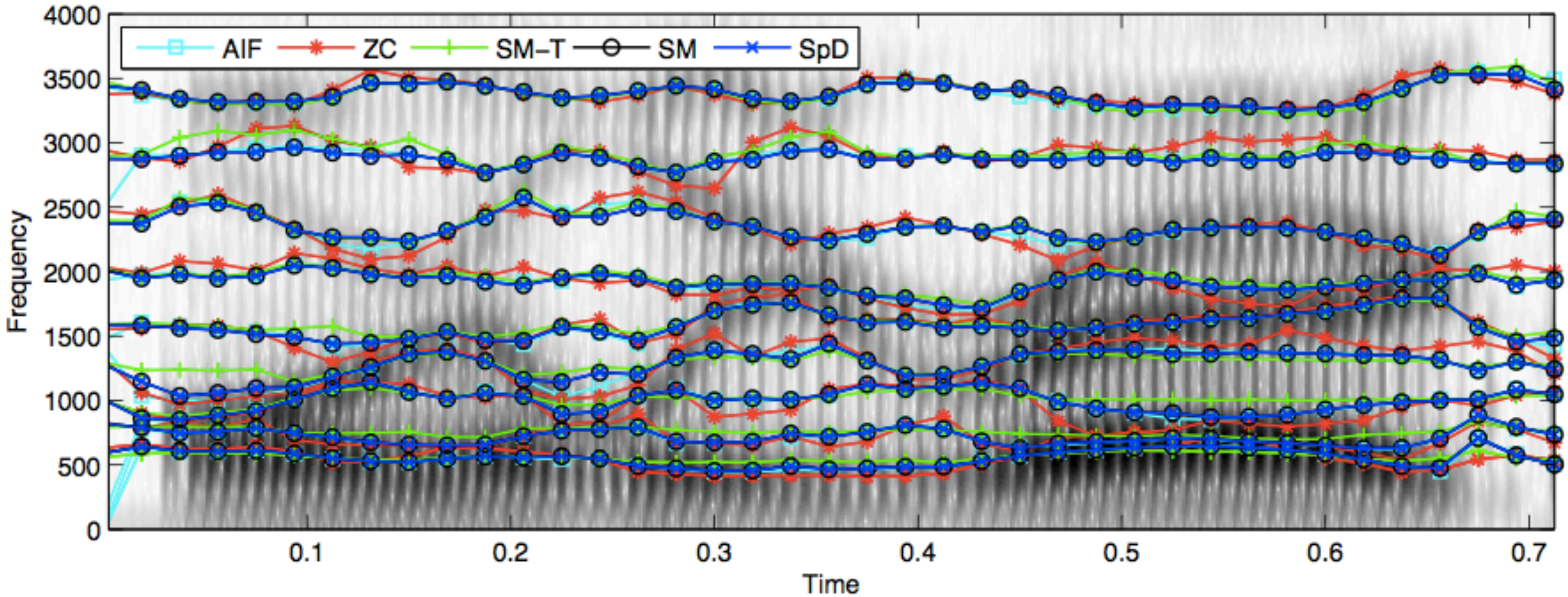
[Dimitriadis et al 2007]

Improved instantaneous frequency estimation

Feature Estimation Methods

- Time-domain: average weighted instantaneous frequency (AIF)
- Frequency domain:
  - spectral moment (SM)
  - spectral derivative (SpD)
- Zero-crossings (ZC)

Feature Trajectories & Performance



- Performance on TIMIT (+noise), Aurora 2,3 tasks:
  - SM/SpD is the top performer, closely followed by AIF, ZC is significantly worse

Relation with auditory front-ends

- Zero-crossings

  [Ghitza 1986, Kim et al 1999]

- Cochlear model, Auditory Spectrogram

  [Yang et al 1992, Wang & Shamma 1994, Ru 2001]

1. Auditory filtering: $y_1(t,x) = s(t) *_t h(t;x)$
2. Time-differentiation & averaging $y_2(t,x) = g(\partial_t y_1(t,x)) *_t w(t)$
3. Frequency differentiation & averaging $y_3(t,x) = \partial_x y_2(t,x) *_x \nu(x)$

Filterbank Arrays

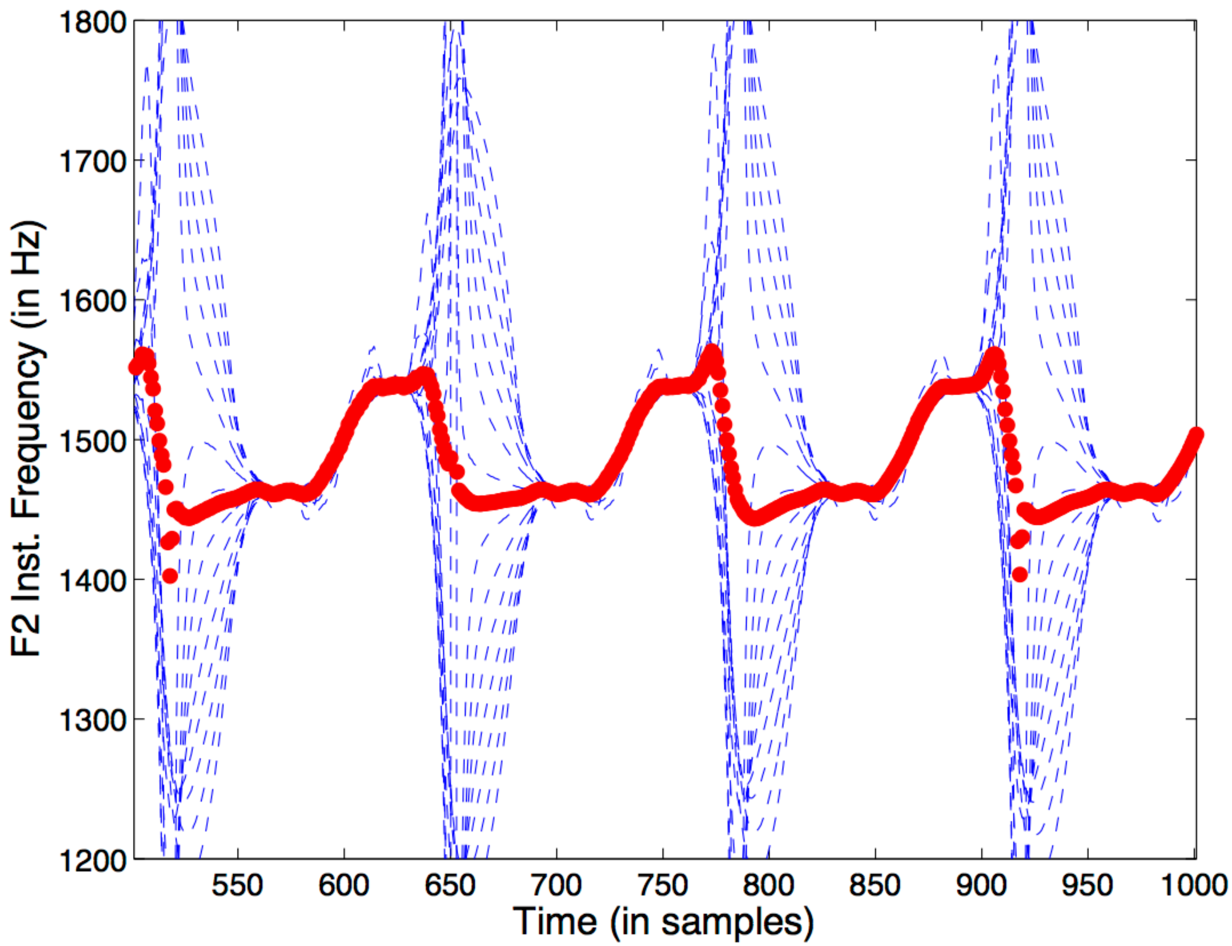
- Average (in frequency) inst. frequency and amplitude estimates over neighboring filters

$$F_A = \frac{\sum_k \left(\int_{t_0}^{t_0+T} f(t,k)[a(t,k)]^2 dt \right)}{\sum_k \left(\int_{t_0}^{t_0+T} [a(t,k)]^2 dt \right)}$$

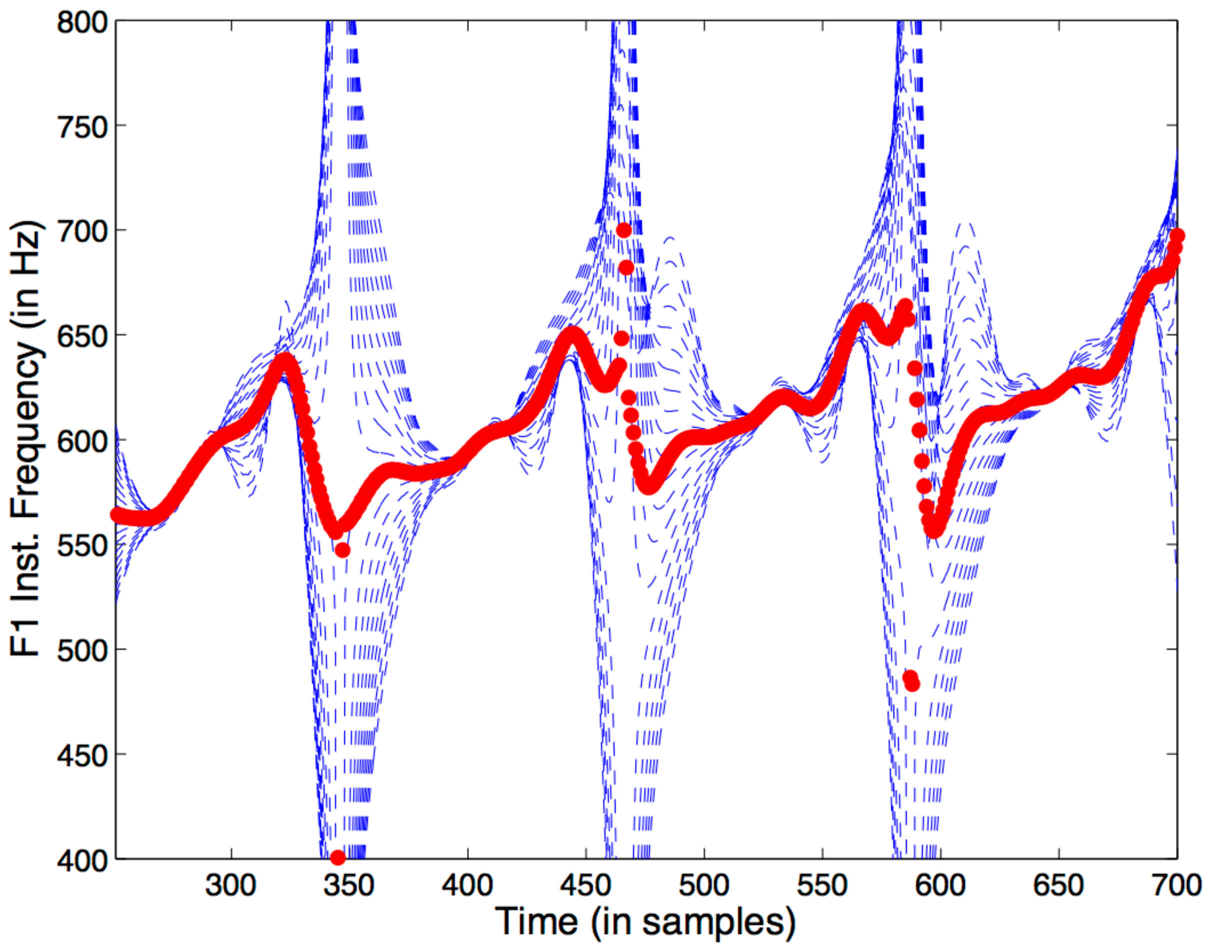
- Inverse variance weighting (variance estimated over neighboring filters)

$$F_{n,m} = \frac{\int_{t_0}^{t_0+T} f(t)[a(t)]^n [v_f(t)]^{-m} dt}{\int_{t_0}^{t_0+T} [a(t)]^n [v_f(t)]^{-m} dt}$$

IF estimation of synthetic resonance



IF estimation of real speech signal



Results

- Estimation error variance reduction using filterbank arrays
  - x 4-7 times for frequency and bandwidth estimates, e.g., AIF, using averaging of neighboring filters
  - x 1.5-2 times using inverse variance weighting
- Speech recognition
  - FMP feature set: second spectral moment over first spectral moment [Dimitriadis et al. 2005]
  - When used as stand-alone feature using filterbank arrays improves performance significantly: 40% => 60% (AURORA 3 Spanish Task)

Summary

- The SMAC frequency-domain front-end
  - equivalent performance in clean recording conditions
  - more robust in noisy situations
- Parameterization
  - larger frequency overlap (wider filters)
  - the SM vector remains in the frequency domain
  - addition of few cepstral coefficients

Discussion

- Equivalence between frequency and energy so what is different?
  - more robust in a variety of noise types
  - VTLN, spectral masking, frequency warping, etc
- What else is to be investigated?
  - theoretic noise analysis
  - alternative fusion of frequency and energy
  - higher order moments
  - other speech applications