# Defeating Ambient Noise: Practical Approaches for Noise Reduction and Suppression

Ivan Tashev
Microsoft Research
Redmond, USA



## Introduction

- Why signal enhancement is important:
  - Reducing the ambient noise from the captured audio signal is crucial for providing good sound in modern computing systems, critical for the needs of real time communication and speech recognition.
- Tutorial goal:
  - To present the key theoretical aspects and share our practical experience in the area of noise suppression and reduction for application in sound capture and processing systems.
- Target audience:
  - Engineers and researchers working in the area of audio signal processing planning or building audio systems for sound capturing.

## Introduction (2)

- Noise suppression as science and as art:
  - It is a science, because uses mathematical models and hypotheses, it is repeatable, i.e. we get the same results with the same input data
  - It is an art, because it is about human perception of the sound and requires evaluation from a human
- For speech signals the process is part of more general term speech enhancement



## Defeating ambient noise: tutorial agenda

- Basics
- Noise suppression
- Directional microphones
- Microphone arrays
- Advanced techniques
- Free joke and conclusions

# Basics

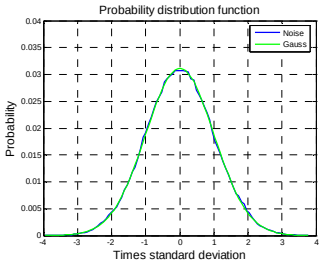
- Noise: definition and properties
- Signal: definition and properties
- Noise suppression and reduction, speech enhancement
- Audio processing in frequency domain: weighting, transformation, synthesis
- Bandpass filtering

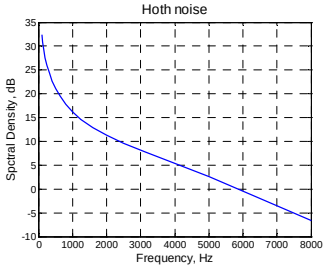


# Basics: noise properties

- Statistical model:
  - Zero mean Gaussian random process
  - Right: airplane noise PDF vs. Gaussian PDF
- In frequency domain:
  - White noise spectrum
  - Pink noise: 6 dB/oct decrease
  - Colored noise – with given spectrum
  - Hoth noise: typical room noise model
- Temporal characteristics:
  - Pseudo stationary compared to speech
  - Specific noises may be different: wind noise
- Spatial characteristics:
  - Ambient, isotropic: evenly distributed
  - Point noise sources - jammers

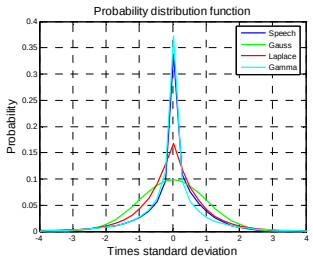
White Inside A320 NYC Café



## Basics: signal properties

- In most of the cases the signal is speech
- Statistical model (in long term):
  - Zero mean random Gaussian (Laplace, Gamma) process
- Frequency domain (in short term):
  - Voiced – e.g. vowels (harmonic structure)
  - Unvoiced – e.g. fricatives (noise type)
- Temporal:
  - Speech and nonspeech segments
- Spatial:
  - Point sound source (mouth or loudspeaker)




May 14th, 2006 ICASSP 2006, Toulouse, France 7

## Basics: classification

- Noise suppression: removing the noise based on statistical models of the noise and signal, spectral subtraction
- Noise reduction or cancellation: removing the noise based on knowledge or estimation of the corrupting signal
- Signal (speech) enhancement: more general term for any type of processing aiming improving some property of the signal
- Active noise cancellation: decreasing the noise level in certain area by sending opposite phase sound with loudspeakers – not discussed in this tutorial

May 14th, 2006 ICASSP 2006, Toulouse, France 8

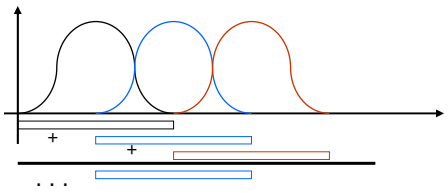
## Basics: processing flow

- Processing in frequency domain
- Audio frames:
  - 80-1024 samples, 5-25 ms
- Frequency domain transformations:
  - Fourier (FFT): symmetric spectra, zero $F_s/2$ bin, process the first half
  - MCLT (Malvar, 1992): shifts bins ½ frequency bin
  - Other: Hartley, wavelet, cepstra; no re-synthesis

May 14th, 2006 ICASSP 2006, Toulouse, France 9

## Basics: processing flow (2)

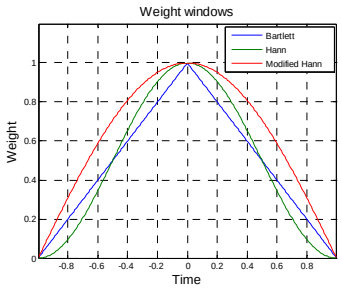
- Overall process (typical):
  - Extract the frame
    - Weighting
    - Transform
    - Process
    - Inverse transform
    - Synthesis (overlap-add) using ½ of the previous frame
  - Move one half frame forward, repeat




May 14th, 2006 ICASSP 2006, Toulouse, France 10

# Basics: processing flow (3)

- Weighting function:
  - Keeps the spectral peaks less smeared
  - Commonly used:
    - Bartlett (triangle)
    - Hann or Hanning (cos-shaped)
    - Modified Hann – sqrt(cos)-shaped, to be applied twice
  - If re-synthesis is not required
    - Natural, Bartlett, Parsen: sinc, sinc$^2$ and sinc$^4$ in frequency domain
    - Max-Fauque-Bertier (sinc): rectangular in frequency domain
    - Blackman and further generalization as Taylor sequence






# Basics: bandpass filtering

- Bandpass filtering:
  - Do not process frequency bins below and above certain frequencies – zero them
  - Typical low limit: 100-300 Hz for speech
  - Typical high limit: 0.45$F_s$, reduces aliasing
  - Dynamic bandpass filtering
    - Measure SNR per bin
    - Adjust the low and high slopes
    - Apply the filter
  - No kidding!
    - Increases speech intelligibility
    - Saves artifacts and distortions
    - Saves efforts and some CPU time

## Basics: summary

- Noise and signal properties: statistical, frequency, temporal, and spatial
- Suppression vs. reduction vs. enhancement vs. cancellation
- Processing in frequency domain
  - Break in 50% overlapping frames – most common
  - Weighting function is important, sqrt(cos)-shaped most common
  - Overlap-add processing
- Bandpass filtering: increases intelligibility, reduces artifacts and saves efforts



## Noise suppression

- Gain based noise suppression
- *a priori* and *a posteriori* SNR
- Suppression rules
- ML and Decision Directed approach for *a priori* SNR estimation
- Uncertain presence of signal
- Voice activity detectors
- Accounting for the temporal characteristics
- Overall architecture
- Demos

## Noise suppression: gain based processing

- Given signal $x_n(t)$ and noise $d_n(t)$ mixed in $y_n(t)$
- Observed in frequency domain, $n$-th frame, $k$-th frequency bin: $Y_k = X_k + D_k$
- Noise suppression:
  - $\tilde{X}_k = \left(G_k |Y_k|\right)\frac{Y_k}{|Y_k|} = G_k . Y_k$
  - $G_k$ – time varying, non-negative, real value gain (or suppression rule)
  - The estimator keeps the same phase as $Y_k$: under Gaussian assumptions the best phase estimator is observed phase
- The goal of noise suppression is for each frame to estimate $G_k$ vector optimal in certain way



## Noise suppression: *a priori* and *a posteriori* SNR

- Signal and noise: statistically independent Gaussian processes
- Signals variances $\lambda_X(k), \lambda_D(k), \lambda_Y(k)$
- *a priori* and *a posteriori* SNRs

$$\xi(k) \triangleq \frac{\lambda_X(k)}{\lambda_D(k)} \qquad \gamma(k) \triangleq \frac{|Y(k)|^2}{\lambda_D(k)}$$

- The suppression rule is now function of two parameters: $G_k(\xi_k, \gamma_k)$

# Noise suppression: suppression rules

- Wiener (1945): $G(k) = \frac{|Y(k)|^2 - \lambda_D(k)}{|Y(k)|^2} = 1 - \gamma(k) = \frac{\xi(k)}{1+\xi(k)}$

  MMSE spectral amplitude estimator

- Derivation
  - Goal $\varepsilon\left\{\left[X_k - \hat{X}_k\right]^2\right\}$
  - Solution

$$G(k) = \frac{P_{XY}(k)}{P_{YY}(k)} = \frac{P_{YY}(k) - P_{DD}(k)}{P_{YY}(k)} = \frac{|Y(k)|^2 - \lambda_D(k)}{|Y(k)|^2} = 1 - \frac{\lambda_D(k)}{|Y(k)|^2} = 1 - \gamma(k)$$

- Problems:
  - Musical noises in the pauses
  - Distortion in the speech segments

Musical noises and distortions



# Noise suppression: suppression rules

- McAulay/Malpass (1980):

$$G(k) = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{|Y(k)|^2 - \lambda_D(k)}{|Y(k)|^2}}$$

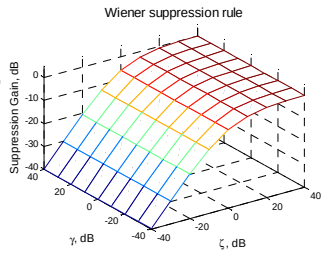
  ML spectral amplitude estimator

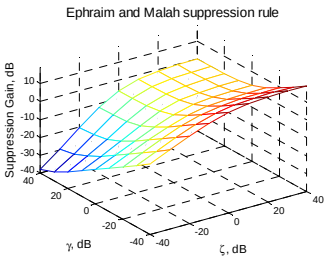
- Ephraim/Malah (1984):
  - Introduce *a priori* SNR

$$G_k = \frac{\sqrt{\pi \nu_k}}{2\gamma_k}\left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right]\exp\left(\frac{-\nu_k}{2}\right)$$

  MMSE short term spectral amplitude estimator

  Where: $\nu(k) \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k$



# Noise suppression: suppression rules (2)

- Ephraim/Malah (1985): $G_k = \frac{\xi_k}{1+\xi_k}.\exp\left(\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right)$

  MMSE short term log spectral amplitude estimator
- Computational complexity of Ephraim and Malah suppression rules
- Efficient alternatives, P. Wolfe/S. Godsill (2001):
  - Joint Maximum A Posteriori Spectral Amplitude Estimator
  - Maximum A Posteriori Spectral Amplitude Estimator
  - MMSE Spectral Power Estimator: $G_k = \sqrt{\frac{\xi_k}{1+\xi_k}\left(\frac{1+\nu_k}{\gamma_k}\right)}$
- Gaussian noise and Gamma speech distributions, Martin (2002)



# Noise suppression: *a priori* SNR estimation

- *a priori* SNR estimation:
  - ML approximation: $\hat{\xi}(k) = \frac{|Y(k)|^2 - \lambda_D(k)}{\lambda_D(k)}$
  - Decision-directed (Ephraim/Malah, 1984):

$$\hat{\xi}(k) = \alpha\frac{\left|\hat{X}^{(n-1)}(k)\right|^2}{\lambda_D{}^{(n-1)}(k)} + (1-\alpha)\max\left[0, \gamma^{(n)}(k)-1\right], \alpha \in [0,1)$$

- Noise variation estimation
  - Requires signal/noise classification of the audio frames/bins
  - In non-signal frames/bins update the noise model:

$$\lambda_D{}^{(n)}(k) = (1-\beta)\lambda_D{}^{(n-1)}(k) + \beta\left|Y^{(n)}(k)\right|^2$$

## Noise suppression: uncertain presence of signal

- McAulay/Malpass (1980)
- Observation $Y_k = X_k + D_k$ holds only if we have signal presented
- Real case:

$$Y_k = \begin{vmatrix} X_k + D_k & with\ signal, & state\ H_1 \\ D_k & just\ noise, & state\ H_0 \end{vmatrix}$$

- Modified MMSE suppression rule:

$$E\left\{\tilde{X}_k \mid Y_k\right\} = P(H_1 \mid Y_k).E\left\{\tilde{X}_k \mid Y_k, H_1\right\}$$



## Noise suppression: voice activity detectors

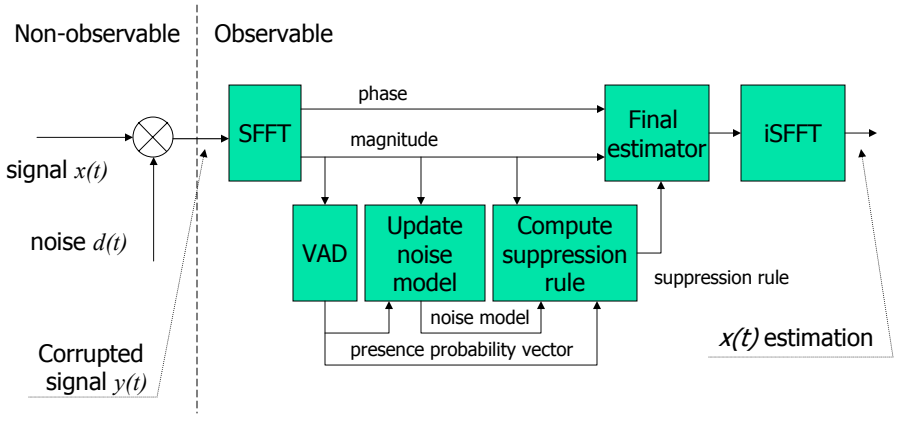
- Energy based, binary decision
  - Track minimal energy $E_{\min}^{(n)} = \begin{vmatrix} E_{\min}^{(n-1)} + \frac{\tau_{up}}{T}\left(E_{\min}^{(n-1)} - |Y|^2\right) & |Y|^2 > E_{\min}^{(n-1)} \\ E_{\min}^{(n-1)} + \frac{\tau_{down}}{T}\left(E_{\min}^{(n-1)} - |Y|^2\right) & E_{\min}^{(n-1)} > |Y|^2 \end{vmatrix}$
  - For classification apply threshold (2.5-7 $E_{min}$)
  - Can be done per frame or per bin
- Probabilistic based (Sohn et. all., 1999)
  - Compute likelihood ratio: $\Lambda_k = \frac{1}{1+\xi_k}\exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\}$
  - Apply hang-over scheme
  - Result: signal presence probability vector (per bin)
- See Martin (2001) as well

# Noise suppression: using temporal properties

- Suppression rule estimators use only the current frame: artifacts, distortions
- Temporal gain smoothing
  - Direct smoothing: $G_k^{(n)} = (1-\beta)G_k^{(n-1)} + \beta G_k$
  - HMM based: $G_k^{(n)} = \frac{a_{01} + a_{11}G_k^{(n-1)}}{a_{00} + a_{10}G_k^{(n-1)}} G_k$
  - Practical interpolation: $G_k^{(n)} = \sqrt{G_k^{(n-1)} G_k}$

May 14th, 2006 ICASSP 2006, Toulouse, France 23

# Noise suppression: overall architecture




May 14th, 2006 ICASSP 2006, Toulouse, France 24

# Noise suppression: practical tips and tricks

- Limit:
  - Suppression gains: keep above -60 dB
  - Probabilities: [1e-4,0.9999]
- Smooth (in time and/or frequency):
  - Noise models
  - Gains
- Simplify:
  - Do not use more complex models than necessary
  - Simpler model with more precise or faster parameters estimation usually works better



# Noise suppression: demonstrations

| Algorithm | Signal | Noise | SNR | Improvement |
|---|---|---|---|---|
| Not processed | -21.5 | -33.3 | 11.8 | |
| Wiener filtered | -22.1 | -52.3 | 30.2 | 18.2 |
| McAulay-Malpass | -21.6 | -36.0 | 14.4 | 2.6 |
| Ephraim-Malah | -22.0 | -47.0 | 25.0 | 13.2 |
| MMSE SPE | -22.2 | -44.8 | 22.5 | 10.7 |

Note: All measurement units are dB


Input file


Wiener


McAulay/Malpass


Ephraim/Malah


MMSE SPE

## Noise suppression: summary

- Noise suppression as time varying, real value, non-negative gain (or suppression rule) based operation
- *a priori* and *a posteriori* SNRs estimation is essential – the decision-directed approach
- Signal may or may not be present – voice activity detectors are critical
- Estimation of precise noise model is with high importance
- Smoothing in time improves listening results



## Directional microphones

- Microphone types
- Pressure gradient microphone
- Parameters for directional microphones
- First order directional microphones
- Classification and parameters
- Bottom line

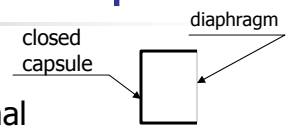
## Directional microphones: microphone types

- Microphone is a device that converts the air pressure to a electric signal
- Microphone types:
  - Carbon – in first phones
  - Crystal – piezoelectric effect based
  - Dynamic – inverted loudspeaker
  - Condenser – measurement grade mics
  - Electret – the most common today

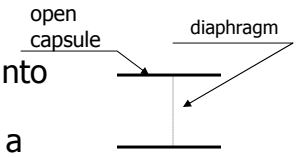


## Directional microphones: pressure gradient microphone

- Pressure microphone
  - Converts pressure to electric signal
  - Can be designed as diaphragm in closed capsule
  - Acoustical monopole
- Pressure gradient microphone
  - Converts the pressure difference into electric signal
  - Can be designed as diaphragm in a open capsule
  - Acoustical dipole



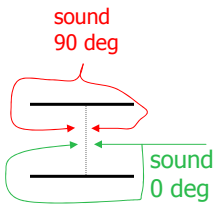

acoustical monopole, omnidirectional




acoustical dipole, directional

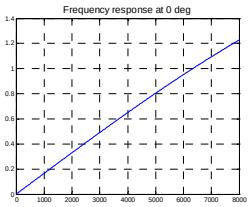
# Directional microphones: pressure gradient microphone (2)

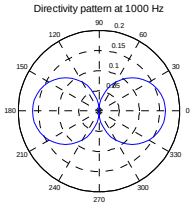
- Directivity pattern of pressure gradient microphone
  - Has figure-8 directivity pattern
  - Frequency response: 6 dB/oct slope towards low frequencies




$$U(f,\theta) = 1 - \exp(-j2\pi f \frac{d\cos(\theta)}{v})$$

$d = 9$ mm
$v = 342$ m/s
$f = 0\text{-}8000$ Hz
$\theta = 0\text{-}360^{O}$






# Directional microphones: first order microphones

- First order microphone as combination of delayed $\tau$ and subtracted two signals from two microphones at distance $d$
- Directivity pattern

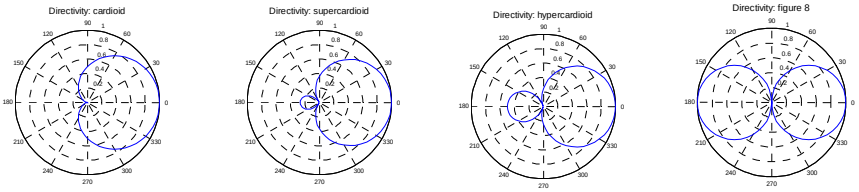
$$U(f,\theta) = 1 - \exp\left(-j2\pi f\left(\tau + \frac{d\cos(\theta)}{v}\right)\right)$$

$$U(f,\theta)_{Norm} \approx \alpha + (1-\alpha)\cos(\theta)$$

Omnidirectional and directional microphones

# Directional microphones: classification

| Type | α | DI | Note |
|---|---|---|---|
| omnidirectional | 1.00 | 0.0 | No directivity |
| cardioid | 0.50 | 4.8 | Zero at 180 deg |
| supercardioid | ~0.35 | 5.7 | Highest front-to-back ratio, zeros at ±125 deg |
| hypercardioid | 0.25 | 6.0 | Highest DI, zeros at ± 109 deg |
| figure 8 | 0.00 | 4.8 | Zero at 90 deg, acoustic dipole |



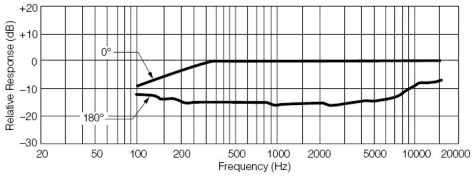

cardioid supercardioid hypercardioid figure 8 (dipole)

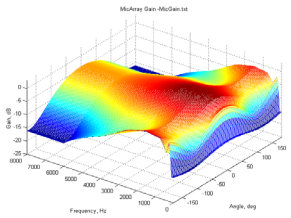
May 14th, 2006 ICASSP 2006, Toulouse, France 33

# Directional microphones: parameters

- Directivity pattern $U(f,c)$
- Directivity index $DI(f) = 10.\log_{10}\left(\frac{P(f,\varphi_T,\theta_T)}{\frac{1}{4\pi}\int_0^{\pi} d\theta \int_0^{2\pi} d\varphi \cdot P(f,\varphi,\theta)}\right)$
- Sensitivity, -45 dBV/Pa typical
- SNR, 60 dB typical
- Frequency response: front/back

$$P(f,\varphi,\theta) = |U(f,c)|^2, \quad \rho = \rho_0 = \text{constant}$$



May 14th, 2006 ICASSP 2006, Toulouse, France 34

## Directional microphones: summary

- In the Noise suppression section we learned that 6 dB noise suppression is a good achievement
- An cardioid microphone gives 4.8 dB noise reduction without distortions and artifacts
- In real systems design using directional microphones is important
- The microphone directivity pattern is further denoted as $U(f,c)$, $f$ – frequency, $c$ – look-up direction $c = \{\theta, \varphi, \rho\}$



## Microphone arrays

- Definition and types
- Delay-and-sum beamformer
- Terminology
- Time-invariant beamformers, demo
- Sound source localization
- Adaptive beamformers
- Spatial filtering, demo

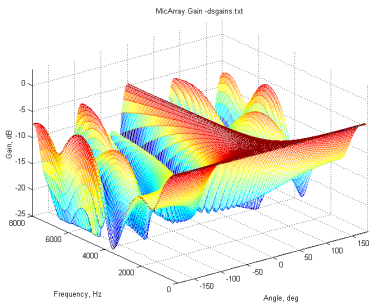
# Microphone arrays: definition and types

- Set of synchronously sampled microphones
- Types:
  - linear, planar, 3D
  - compact and large
  - uniform, nonuniform and random spacing
  - near field and far field
- Advantage: allow spatial filtering, reducing the noises and reverberation
- Disadvantage: require more microphones and more processing time

May 14th, 2006 ICASSP 2006, Toulouse, France 37

# Microphone arrays: delay-and-sum beamformer

- The most intuitive approach
- Shift the signals to align them and sum
- Advantages:
  - Simple and efficient
- Problems:
  - Variable directivity
  - Big sidelobes
  - Low efficiency




May 14th, 2006 ICASSP 2006, Toulouse, France 38

# Microphone arrays: terminology

- Beamforming: making the microphone array to listen to given look-up direction
- Beamsteering: electronically change the look-up direction the microphone array listens to
- Nullsteering: suppressing the sounds coming from given direction
- Sound source localization: techniques to detect, localize and track one or multiple sound sources using microphone array



# Microphone arrays: general parameters

- Generalized form:

$$Y(f) = \sum_{i=0}^{M-1} W(f,i) X_i(f)$$

$M$ – number of microphones
$X_i(f)$ – spectrum of i-th channel
$W(f,i)$ – weight coefficients matrix
$Y(f)$ – output signal

- Parameters:
  - Directivity pattern $B$:

$$B(\theta, f) = W^H(f) \cdot D(f, \theta),$$

$$D(f,\theta) = \frac{e^{-j2\pi f \frac{\|c - p_m\|}{v}}}{\|c - p_m\|} U(f,c)$$

  - Main Response Axis – direction $\theta_{\max}$ towards max sensitivity, look-up direction
  - Beamwidth: area -3 dB around MRA

# Microphone arrays: general parameters (2)

- Ambient noise gain: isotropic noise reduction

$$H_C(f) = \frac{1}{4\pi} \int_0^{2\pi} \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} B(\varphi, \theta, f) d\theta d\varphi$$

- Non-correlated (sensor) noise gain

$$H_N(f) = \sqrt{\sum_{i=0}^{M-1} W(f,i)^2}$$

- Total noise gain: combination of the two above

$$H(f) = \sqrt{\frac{(H_C(f).N_C(f))^2 + (H_N(f).N_N(f))^2}{N_C(f)^2}}$$

- The beamformer design is to find weight matrix to satisfy certain criteria & constrains



# Microphone arrays: time invariant beamformer

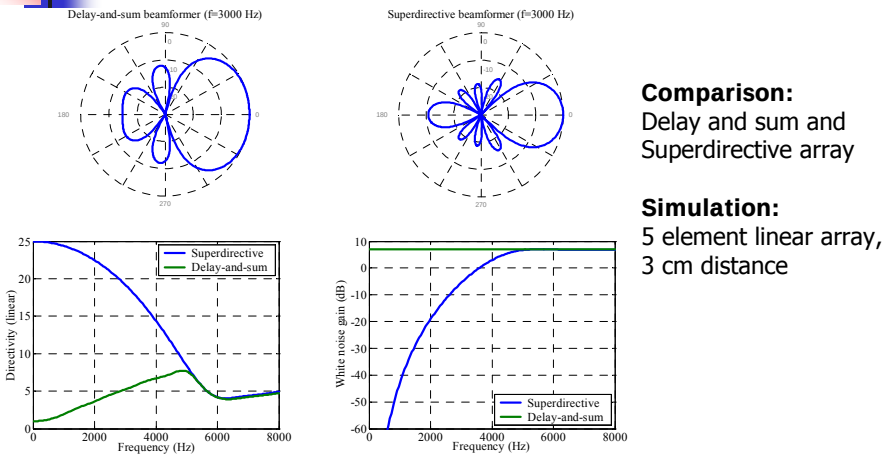
- Design criteria:
  - Max noise suppression: highly non-linear
  - Replaced with directivity pattern matching – reducing the optimization dimensions
  - Isotropic noise assumption
- Constrains:
  - Unit gain and zero phase shift towards MRA
  - Frequently: in the beamwidth area
- Two controversial trends: decreasing the ambient noise gain increases the non-correlated noise gain.
- Optimum? – Minimize the total gain

# Microphone arrays: time invariant beamformer (2)

- Superidirective beamformer (Cox, 1986)
  - $\min_W (W^H \Phi_{XX} W) \quad subject \quad to \quad W^H D = 1$
  - $\Phi_{XX}$ is the power spectral density matrix of the input signals assuming isotropic noise
  - Constrained LMS algorithm, antenna array
  - Achieves maximum directivity
  - Chu, 1997; Elko, 2000



# Microphone arrays: time invariant beamformer (3)

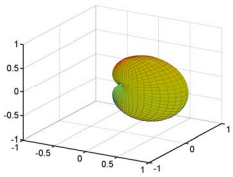


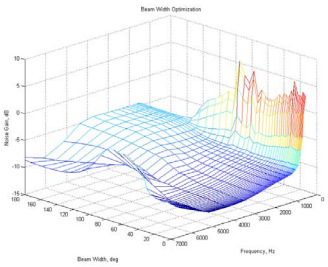

**Comparison:**
Delay and sum and Superdirective array

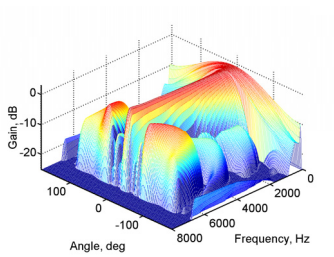
**Simulation:**
5 element linear array, 3 cm distance

# Microphone arrays: time invariant beamformer (4)

- Design example (Tashev/Malvar, 2005)
  - Four element linear array
  - Beamwidth vs. Frequency vs. Total Noise Gain
  - Directivity pattern vs. Frequency
  - Directivity pattern in 3D for 1000 Hz








Demonstrations:
a) Parallel recording
b) Real-time SSL



# Microphone arrays: time invariant beamformer (5)

- Advantages:
  - No VAD required
  - Stable, reliable, predictable, measurable
  - Guaranteed parameters
  - Fast switching to different speaker
  - Low CPU requirement
- Real-world problems:
  - Requires Sound Source Localizer to find and track the desired sound source
  - Sensor's & equipment's noises limit the performance
  - Microphones manufacturing tolerances:
    - Calibration during manufacturing
    - Auto calibration during use (Tashev, 2004)

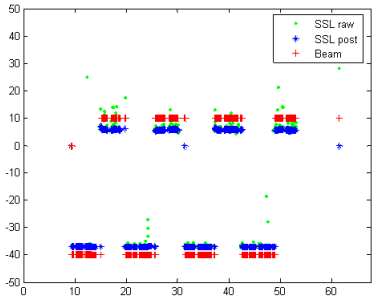
# Microphone arrays: source localization

- Time delay estimates based
  - Cross-correlation function
  - Weighting: ML, PHAT (Knap/Carter, 1976)
  - Combining the pairs
    - Brandstein et. all., 1996
    - Burchfield et. all., 2001 – uses optimization, works in 2D
    - Rui/Florencio, 2003 – sum or cross-correlation functions towards hypothesis
- Beamsteering based
  - Compute the output energy of set of beams
  - Find the maximum
  - Do interpolation for increased precision
  - Variant: two dimensional search



# Microphone arrays: source localization (2)

- Problems: noise and reverberation
- Post-processing the raw SSL results
  - Particle filtering
  - Kalman filtering
  - Real-time clustering
- Camera-assisted approach
  - Face detection software
  - Fusion SSL and video data




•Real SSL results: raw, post-processed, snapped to 10 degrees beams.
•Two persons talking at 6 and -38 degrees, distance 12 feet, conference room.
•Four element linear array.

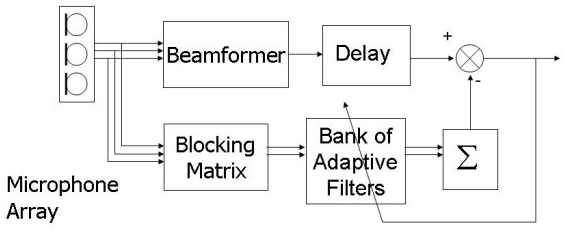
# Microphone arrays: adaptive algorithms

- Frost algorithm (Frost, 1972)
  - $\min_{W}(W^H \Phi_{XX} W) \quad subject \quad to \quad W^H D = 1$
  - $\Phi_{XX}$ is the power spectral density matrix of the input signals
  - Gradient descent optimization, i.e. constrained LMS algorithm
  - Designed for antenna array



# Microphone arrays: adaptive algorithms (2)

- Generalized Side Lobe Canceller (Griffiths/Jim, 1982)
  - Time-invariant beamformer
  - Nulls are sharper than beams
  - Blocking matrix – place null towards the sound source
  - Adaptive filters to minimize residual in the beamformer output



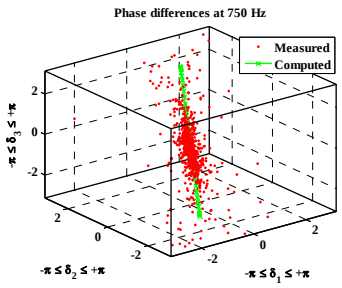
# Microphone arrays: adaptive algorithms (3)

- Advantages
  - Use fully the geometry under the specific noise
  - Very good with point noise sources
  - No calibration required
- Real-world problems
  - Higher requirement for CPU, memory
  - More complex for implementation
  - Slower adaptation and switching to next sound source
  - Non-predictable and non-guaranteed parameters
  - Similar to fixed beamformers performance with ambient type of noise



# Microphone arrays: non-linear spatial filtering

- Implemented as non-linear post-processor
- Based on Instantaneous Direction Of Arrival (IDOA) estimation per bin
  - $\Delta(f) \triangleq [\delta_1(f), \delta_2(f), \ldots, \delta_{M-1}(f)]$
  - where $\delta_{j-1}(f) = \arg(X_1(f)) - \arg(X_j(f))$
- Compute the probability and apply in the same way as in noise suppression under uncertain presence of signal



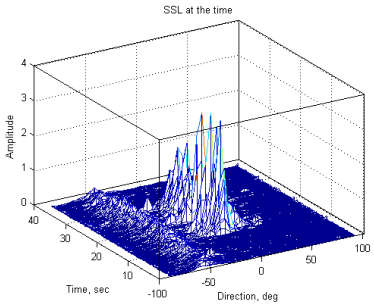
# Microphone arrays: non-linear spatial filtering (2)

- Generalized suppression with spatial information and known look-up direction
- Demo:
  - Recording conditions:
    - Human speaker at 0 degrees, 1.5 m
    - Radio at -45 degrees, 2 m
    - Office: normal noise and reverberation
    - Four element linear microphone array
  - Same audio recording, two sequences:
    - video: direction-frequency-power; audio: one microphone
    - video: direction-power for SSL; audio: array output




# Microphone arrays: non-linear spatial filtering (3)

- Advantages
  - Better directivity than time-invariant beamformer
  - Good source separation
  - Low CPU overhead
- Real-world problems
  - Requires channel matching, i.e. calibration
  - Non-linear processing (artifacts, musical noises)




direction-time-power

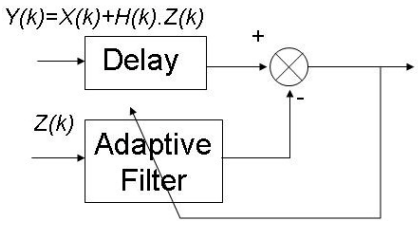
## Advanced techniques

- Adaptive noise reduction
- Psychoacoustic based noise suppressor
- Noise suppressor optimized for speech recognition
- Noise suppression with speech model
- Spatial noise suppression



## Advanced techniques: adaptive noise reduction

- Add a microphone to capture the noise signal (HDD in a laptop, engine in a car)
- Two inputs system:
  - voice + noise: $y(t)=x(t)+h(t)*z(t)$
  - noise only: $z(t)$
- Use LMS, RLS or NLMS adaptive filter
- Double talk detector necessary if leakage of $x(t)$ in $z(t)$



## Advanced techniques: adaptive noise reduction (2)

- Advantages:
  - Linear! No musical noises or distortions
  - Works with non-stationary noises
  - Low CPU requirement
- Real-world issues
  - Needs a second microphone
  - Limited applicability: when we can capture the noise only signal
  - Has some audible residuals and artifacts



## Advanced techniques: psychoacoustic noise suppressor

- Concept:
  - More energy removed -> more musical noises and distortions
  - Masking effects in frequency and time domains in human perception of sound
  - Why remove noises we can't hear?
- Real-life issues
  - Needs MOS tests for evaluation
  - Duplicates codec functionality – the new audio codecs use the same effect

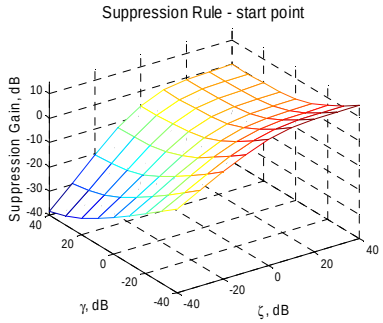
# Advanced techniques: noise suppressor for ASR

- General idea: optimize parameterized suppression rule for best recognition rate (Tashev/Droppo/Acero, 2006)
  - More training data improves average recognition, harms clean speech recognition
  - Rprop optimization algorithm: enhanced version of gradient descent
  - Objective function: Maximum Mutual Information (MMI) from ASR, closely related to the recognition accuracy
  - Optimization parameters: the suppression rule
  - Starting point: MMSE Spectral Power Estimator rule
- Baseline: 99.5% clean, 52.5% average
- Starting point: 96.9% clean, 74.9% average
- Achieved optimal point: 99.0% clean, 77.7% average

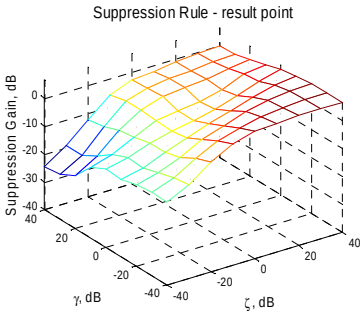
May 14th, 2006 ICASSP 2006, Toulouse, France 59

# Advanced techniques: noise suppressor for ASR (4)

- MMSE SPE




- After 20 Iterations




May 14th, 2006 ICASSP 2006, Toulouse, France 60

## Advanced techniques: using speech model

- General idea:
  - Detect and parse the speech signal: fricatives, vowels, glides, nasals, stops
  - Measure the parameters
  - Synthesize clean speech signal
- Real-world issues:
  - If we can do reliably the parsing – we solved the noise robust ASR problems ☺
  - Even text-to-speech systems do not have very good pronunciation, doing this without language model is more difficult



## Advanced techniques: using speech model (2)

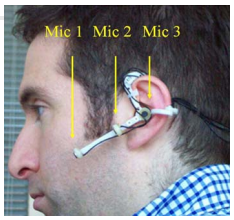
- Drucker (1968):
  - Detect and parse the speech signal: fricatives, vowels, glides, nasals, stops
  - Use separate enhancing filters for each category
  - Hard decision for presence and class
- McAulay/Malpass (1980) introduced soft decision rules and using several filters in parallel
- Some techniques:
  - Use the harmonic structure of vowels, time warping to make them flat, clean, un-warp
  - Use vocal tract model for generating fricatives and other consonants
  - Using language model (too specific)

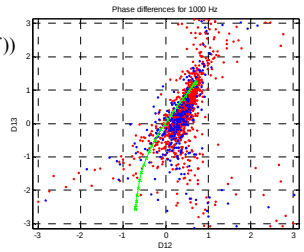
# Advanced techniques: spatial noise suppression

- Microphone array for headset (Tashev/Seltzer/Acero, 2005)
  - 3-element microphone array
  - Bone sensor for reliable VAD
  - Working in IDOA space

$$\Delta(f) \triangleq [\delta_1(f), \delta_2(f), \ldots, \delta_{M-1}(f)] \quad \delta_{j-1}(f) = \arg(X_1(f)) - \arg(X_j(f))$$

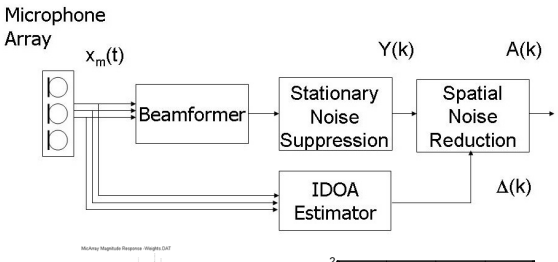
  - Multidimensional generalization of classic noise suppression
    - Building position-dependent noise models
    - Apply suppression rule

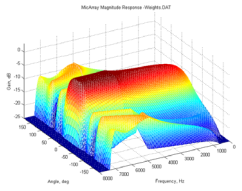


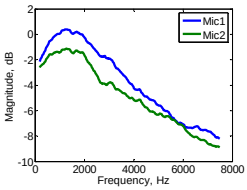




May 14th, 2006 ICASSP 2006, Toulouse, France 63

# Advanced techniques: spatial noise suppression (2)






- General architecture
- Beamformer directivity
- Diffraction around the head correction

May 14th, 2006 ICASSP 2006, Toulouse, France 64

# Advanced techniques: spatial noise suppression (3)

- Signal and noise variances

$$\lambda_Y(f \mid \Delta) \triangleq E\left[|Y(f \mid \Delta)|^2\right]$$

$$\lambda_D(f \mid \Delta) \triangleq E\left[|D(f \mid \Delta)|^2\right]$$

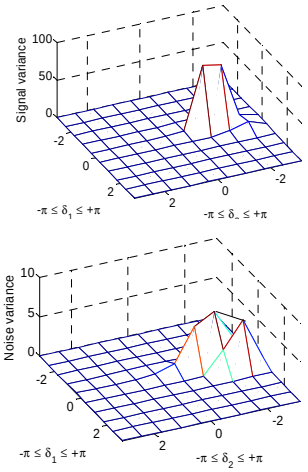
- *a priory* and *a posteriori* SNR

$$\xi(f \mid \Delta) \triangleq \beta \frac{\lambda_Y(f \mid \Delta) - \lambda_D(f \mid \Delta)}{\lambda_D(f \mid \Delta)} + (1-\beta)\max[0, \gamma(f \mid \Delta)], \beta \in [0,1)$$

$$\gamma(f \mid \Delta) \triangleq \frac{|Y(f \mid \Delta)|^2}{\lambda_D(f \mid \Delta)}$$

- Suppression rule

$$H(f \mid \Delta) = \sqrt{\frac{\xi(f \mid \Delta)}{1+\xi(f \mid \Delta)}\left(\frac{1+\vartheta(f \mid \Delta)}{\gamma(f \mid \Delta)}\right)}$$




May 14th, 2006 ICASSP 2006, Toulouse, France 65

# Advanced techniques: spatial noise suppression (4)

- SNR improvement, all units in dB

| | *BM* | *BF* | *NS* | *SR* |
|---|---|---|---|---|
| **Office, 55 dB** | 25.2 | 22.5 | 29.4 | 34.7 |
| **Café, 75 dB** | 7.2 | 12.3 | 17.5 | 22.8 |
| **Car, 90 dB** | 3.2 | 6.4 | 11.1 | 16.4 |

BM – best microphone,
BF – beamformer
NS – noise suppressor
SR – spatial noise suppressor

- Demo: parallel recording with BT headset

May 14th, 2006 ICASSP 2006, Toulouse, France 66

## Advanced techniques: summary

- Improving further the noise suppression and reduction increases complexity, requires more information.
- The algorithms become more specialized: for car, for speech, for ASR, for specific noises.
- Use good judgment when use or design them:
  - Do I need this?
  - How specific is the application?
- Remember: more complex model with more parameters means slower computation and adaptation. Use with caution.
- Still very exciting new algorithms, solving problems unsolved so far.



## Defeating ambient noise: final remarks

- The art of noise suppression is to know when to stop.
- None of the methods is universal, use cascading and make sure not to destroy important properties.
- Build processing blocks, think the whole system: well balanced suppression across the processing chain.
- Noise suppression is about human perception: use your ears and MOS tests.

## Finally

Thank you for choosing this tutorial!

Thank you for the attention!

## Questions?

Contact info: ivantash@microsoft.com
http://research.microsoft.com/users/ivantash