Microphone Array For Headset With Spatial Noise Suppressor

Ivan Tashev, Michael L. Seltzer, and Alex Acero

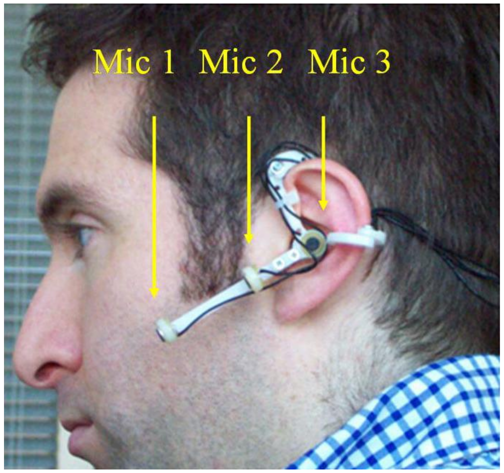
Microsoft Research

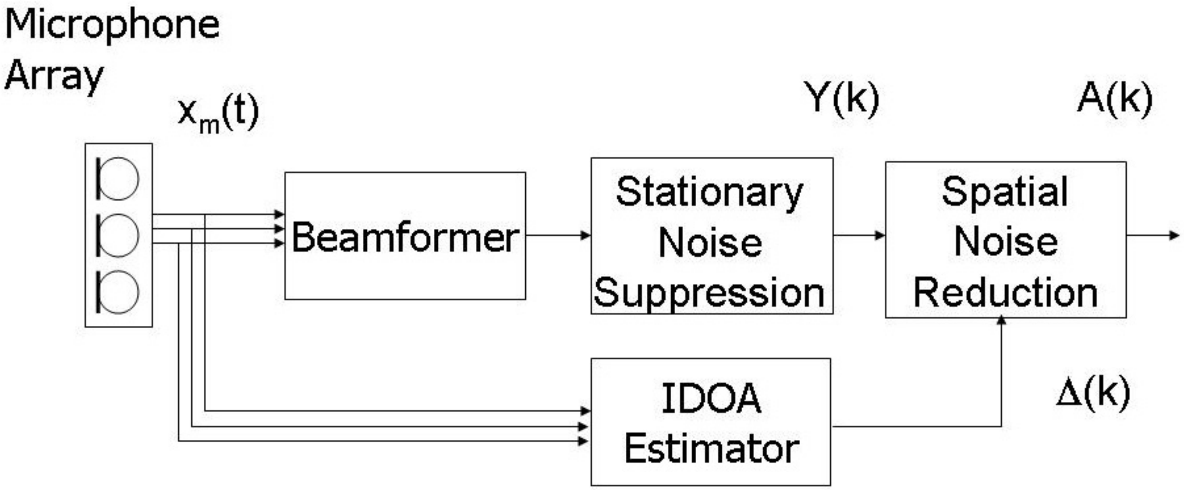
Problem and solutions

- Problem: Providing good quality sound capture with a small headset
  - A short boom loses 6 dB in SNR
- Solution: Using multiple microphones for beamforming and spatial filtering
- Constrains: Low CPU usage, memory footprint and price, long battery life

Solution architecture

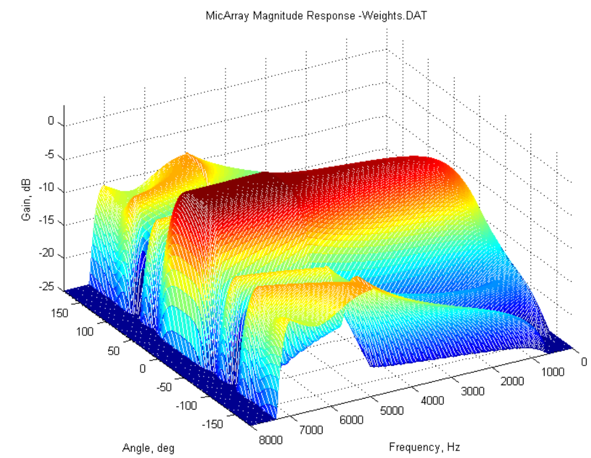
- 3 element microphone array
- Time invariant beamformer
- Stationary noise suppressor
- Spatial noise suppressor

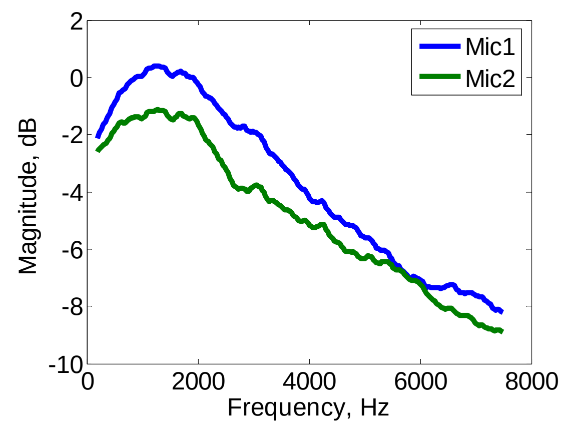




Time-Invariant Beamformer

- Processing in frequency domain $$Y(f) = \sum_{m=1}^{M} W_m(f) X_m(f)$$
- Weights computed using deterministic algorithm
- Trade-off: better directivity for more instrumental noise
- Compensation for the diffraction around the head





Stationary Noise Suppressor

- High time constant for building the noise model
- Probabilistic Gaussian classifier for VAD
  - Variants: using a bone microphone or an accelerometer for robust detection
- Suppression rule based on MMSE Spectral Power Estimator (P. Wolfe and S. Godsil, 2003) – efficient version of Ephraim and Malah suppression rule (1984)

Noise suppression

- Signal $x_n(t)$ and noise $d_n(t)$ mixed in $y_n(t)$
- Observed: $Y_k = X_k + D_k$
- Noise suppression: $\hat{X}_k = H_k \cdot Y_k$
- H_k – suppression rule, real vector.
  Keep the same phase as Y_k
- Signal variances $\lambda_X(k), \lambda_D(k), \lambda_Y(k)$
- *a priory* and *a posteriori* SNRs

$$\xi(k) \triangleq \frac{\lambda_X(k)}{\lambda_D(k)}, \gamma(k) \triangleq \frac{|Y(k)|^2}{\lambda_D(k)}, \nu(k) \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k$$

Common Suppression Rules

- Weiner suppression rule (1945): $H(k)=\dfrac{\lambda_X(k)}{\lambda_X(k)+\lambda_D(k)}$
- Ephraim and Malah rule (1984):

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

- Efficient alternatives, P. Wolfe & S. Godsil (2003):
  - Joint Maximum A Posteriori Spectral Amplitude Estimator
  - Maximum A Posteriori Spectral Amplitude Estimator
  - Minimum Mean Square Error Spectral Power Estimator

$$H_k=\sqrt{\frac{\xi_k}{1+\xi_k}\left(\frac{1+\nu_k}{\gamma_k}\right)}$$

Variation and SNR estimations

- Noise variation estimation
  - Use VAD to classify the audio frames
  - For non-voiced frames update the noise model: $\lambda_D(n,k) = (1-\beta)\lambda_D(n-1,k) + \beta\left|Y(n,k)\right|^2$
- *a priory* SNR estimation:
  - Approximate: $\hat{\xi}(k) = \dfrac{\left|Y(k)\right|^2 - \lambda_D(k)}{\lambda_D(k)}$
  - Decision-directed (Ephraim and Malah):

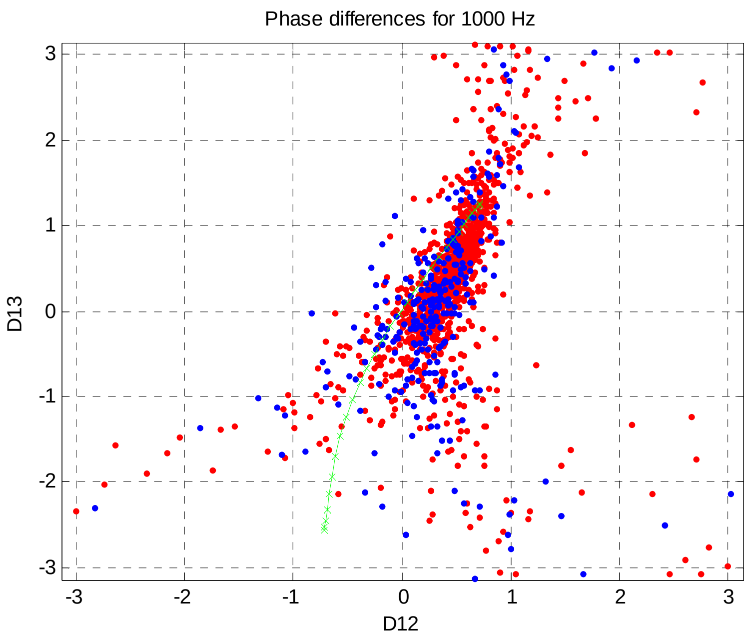
$$\hat{\xi}(k) = \alpha \frac{\left|\hat{X}(n-1,k)\right|^2}{\lambda_D(n-1,k)} + (1-\alpha)\max\left[0, \gamma(n,k)-1\right], \alpha \in [0,1)$$

Spatial Noise Reduction

- With microphone array the signals have position, i.e. one more dimension
- Instant Direction Of Arrival (IDOA) space:

$$\Delta(f) \triangleq [\delta_1(f), \delta_2(f), \ldots, \delta_{M-1}(f)]$$

where $\delta_{j-1}(f) = \arg(X_1(f)) - \arg(X_j(f))$



Estimation and suppression

- Signal and noise variances

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

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

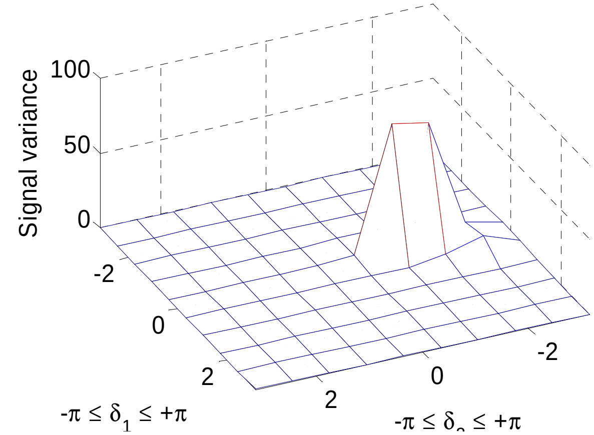
- *a priory* and *a posteriori* SNR

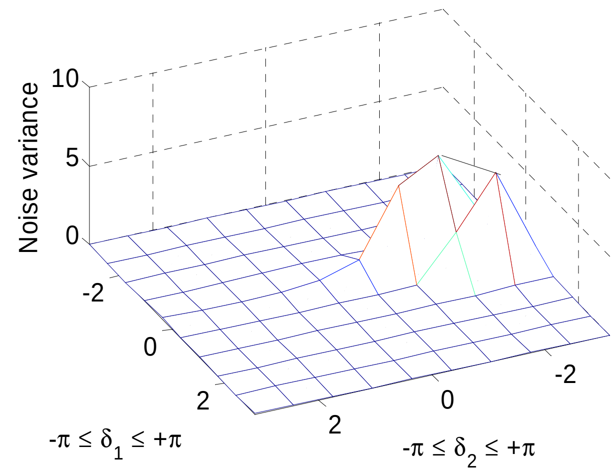
$$\xi(f \mid \Delta) \triangleq \beta \frac{\lambda_Y(f \mid \Delta) - \lambda_D(f \mid \Delta)}{\lambda_D(f \mid \Delta)} + (1-\beta)\max[0, \gamma(f \mid \Delta)], \beta \in [0,1)$$

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

- Suppression rule

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

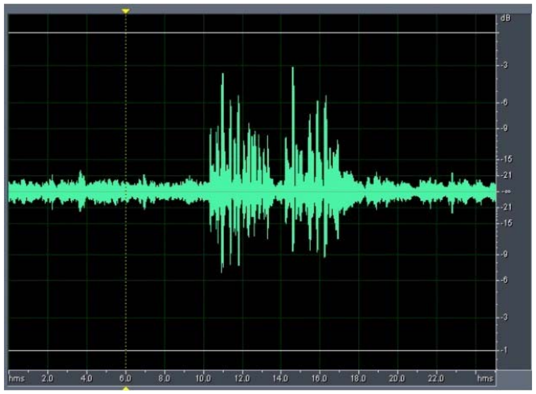




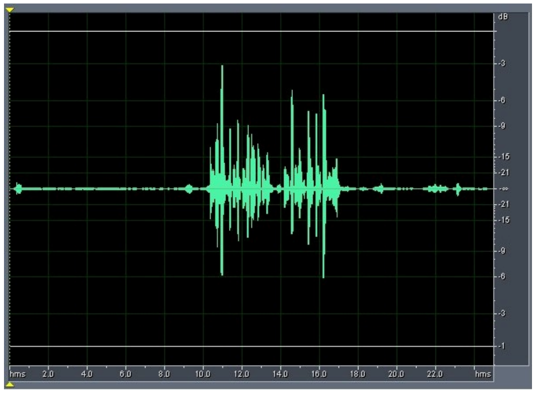
Results

- SNR improvement

	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



How it sounds?

- 75 dB cocktail party noise

Input Output

- 90 dB in-car noise

Input Output

Conclusions

- Presented processing algorithm provides noise reduction up to 14 dB
- It uses a priory known properties of the estimated signal and suppressed noise:
  - Stationary noise
  - Short-term-stationary signal
  - Spatially-stationary noise and signal
- Well balanced suppression from each stage provides low level distortions and artifacts
- Low scalability due to dimensions increasing

Finally

The art of noise suppression and reduction is knowing when to stop.

Thank you for your attention!

Questions?