Introduction

- Headsets pose a significant impediment to user acceptance of speech interfaces.
- ASR performance with arrays of microphones several meters distant from a user’s mouth is nearly as good as that achievable with a CTM.
- Getting good performance requires:
  - *Voice activity detection* (VAD) to detect when a speaker is speaking;
  - *Voice prompt suppression* (VPS) to allow for “barge in”;
  - *Speaker tracking* (ST) to provide the speaker’s location for beamforming;
  - *Beamforming* to focus on the desired speech and suppress noise, competing speech, and reverberation.
- In this talk, we will focus beamforming.
Linear Aperture

\[ z = r \cos \theta \]

Plane Wave

\[ x = r \sin \theta \cos \phi \]

\[ y = r \sin \theta \sin \phi \]

Wavefront

Linear Aperture
Plane Wave

- Before taking up the case of conventional microphone arrays, let us consider the *linear aperture* of length $L$.
- The *wavenumber vector* \( \mathbf{k} \triangleq \frac{2\pi}{\lambda} \mathbf{a} \),

where $\lambda$ is the length of the propagating wave, indicates the direction of arrival and frequency of the wave.
- The magnitude $k \triangleq |\mathbf{k}| = \frac{2\pi}{\lambda} = \omega/c$, where $c$ is the speed of sound, indicates the frequency of the plane wave.
- The component of $\mathbf{k}$ along the $z$-axis is given by

\[
k_z \triangleq -|\mathbf{k}| \cos \theta = -\frac{2\pi}{\lambda} u,
\]

where $u \triangleq \cos \theta$ is the *direction cosine*. 
Making the simplest assumption, the same wave will arrive at all points on the aperture, but *not* simultaneously.

The time *delay* for the aperture point \((0, 0, z)\) with respect to the origin will be given by

\[
\tau(z) = k z = -\frac{\omega z \cos \theta}{c}.
\]  

The Fourier transform of the signal component arriving at point \(z\) can be expressed as

\[
F(\omega, k, z) = F(\omega) e^{-ikz},
\]

where \(i \triangleq \sqrt{-1}\).
If the signal components are weighted with a function $w^*_a(z)$ and then combined, then the result is the frequency wavenumber response function,

$$\gamma(\omega, k) \triangleq \int_{-\infty}^{\infty} w^*_a(z) e^{-ikz} \, dz. \quad (5)$$

Let us initially assume that

$$w_a(z) = \begin{cases} 1, & \forall \quad -L/2 \leq z \leq L/2, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

In this case we find

$$\gamma(\omega, k) = \int_{-L/2}^{L/2} e^{-ikz} \, dz = \text{sinc} \left( \frac{L}{2k} \right),$$

where $\text{sinc}(x) \triangleq \frac{\sin x}{x}$.
Beampatterns

At low frequencies, the beampattern has poor *directivity*.

At high frequencies, the beampattern has a pronounced sidelobe structure.

**Figure:** Beampatterns for the linear aperture.
As a uniformly sensitive array is difficult or impossible to build, let us consider sampling the aperture at $S$ points

$$z_s = \left( s - \frac{S - 1}{2} \right) d \quad \forall \quad s = 0, 1, \ldots, S - 1,$$

where $d \triangleq L/S$ is the *intersensor spacing*. 

**Uniform Linear Array**
Sampling Function

- This sampling is accomplished by defining the *sampled* sensitivity function

\[ w_s(z) \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \delta(z - z_s). \] (8)

- Substituting (8) into (5), provides

\[ \gamma_s(\omega, k) = \frac{1}{S} \exp \left\{ ikd \left( \frac{S - 1}{2} \right) \right\} \sum_{s=0}^{S-1} e^{-iks_d}. \]

- This can be readily simplified to

\[ \gamma_s(\omega, k) = \frac{1}{S} \cdot \frac{\sin \left( \frac{Sd}{2} k \right)}{\sin \left( \frac{d}{2} k \right)}. \]
Beampatterns for the Linear Array

Unlike those for the linear aperture, beampatters for the linear array are periodic.

Figure: Beampatterns for the linear aperture (dotted line) and linear array (solid line) with $S = 11$ and a) $d/\lambda = \frac{1}{2}$, b) $d/\lambda = 1$, and c) $d/\lambda = \frac{3}{2}$. 
Beampattern Steering

Clearly the look direction for the beampatterns in Figures 1 and 2 is given by $(\theta_L, \phi_L) = (\pi/2, 0)$, which is known as *broadside*.

Setting the look direction to broadside is achieved with a uniform weighting of the linear aperture as in (6), or the uniform weighting of the sensor outputs in (8).

The look direction can readily be set to any desired direction $k = k_L$ by setting the sensor weights to

$$w_s(z; k_L) \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \delta(z - z_s).$$  (9)
Doing so yields the beampattern

\[ B(k; k_L) \triangleq v_k^H(k_L)v_k(k), \]  

where the \textit{array manifold vector} is defined as

\[ v_k(k) \triangleq \begin{bmatrix} e^{i\left(\frac{S-1}{2}\right)kd} & e^{i\left(\frac{S-1}{2} - 1\right)kd} & \cdots & e^{-i\left(\frac{S-1}{2}\right)kd} \end{bmatrix}^T. \]
Effect of Beampattern Steering

Figure: Effect of steering on the grating lobes for $S = 20$ plotted in Cartesian and polar coordinates.
Grating Lobes

- From the figure, it is apparent that for $d/\lambda \leq 1/2$, the behavior of the array is a very good approximation of that of the continuous aperture throughout the entire working range $-1 \leq u \leq 1$.

- On the other hand, while the behavior of the main lobe around $u = 0$ is good for $d/\lambda = 1, 3/2$, large spurious lobes with the same magnitude as the main lobe arise at points well-removed from the look direction.

- These lobes are known as *grating lobes* and arise from a phenomenon known as *spatial aliasing*.

- The *half wavelength rule* requires

$$\frac{d}{\lambda} \leq \frac{1}{2}.$$
Hence, we’ve learned that first element of beamforming is determining the speaker’s position. How can this be done?

The time delay of arrival (TDOA) between the microphones at positions $\mathbf{m}_1$ and $\mathbf{m}_2$ can be expressed as

$$T(\mathbf{m}_1, \mathbf{m}_2, \mathbf{x}) \triangleq \frac{\|\mathbf{x} - \mathbf{m}_1\| - \|\mathbf{x} - \mathbf{m}_2\|}{c}$$  \hspace{1cm} (12)

where $c$ is the speed of sound.

The definition (12) can be rewritten as

$$T_{mn}(\mathbf{x}) \triangleq T(\mathbf{m}_m, \mathbf{m}_n, \mathbf{x}) = \frac{D_m - D_n}{c},$$  \hspace{1cm} (13)

where $D_s \triangleq \|\mathbf{x} - \mathbf{m}_s\| \quad \forall \quad s = 0, \ldots, S - 1.$
Illustration of TDOAs

Simple geometric considerations enable TDOAs between pairs of microphones to be calculated once the speaker’s position $x$ is known.

$$m_0, m_m, m_{m+1}, m_{N-1}$$

$$D_0, D_m, D_{m+1}, D_{N-1}$$
Let $\hat{\tau}_{mn}$ denote the observed TDOA for the $m$th and $n$th microphones.

The TDOAs can be observed or estimated by applying the phase transform

$$\rho_{mn}(\tau) \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{Y_m(e^{j\omega\tau}) Y_n^*(e^{j\omega\tau})}{|Y_m(e^{j\omega\tau}) Y_n^*(e^{j\omega\tau})|} e^{j\omega\tau} d\omega, \quad (14)$$

where $Y_n(e^{j\omega\tau})$ denotes the short-time Fourier transform for the $n$th sensor.

Once $\rho_{mn}(\tau)$ has been calculated, the TDOA estimate is

$$\hat{\tau}_{mn} = \max_{\tau} \rho_{mn}(\tau). \quad (15)$$
Source Localization

- Instantaneous source localization can be performed by minimizing

\[ \epsilon(x) = \sum_{n=1}^{M} \frac{1}{\sigma_n^2} [\hat{\tau}_n - T_n(x)]^2, \quad (16) \]

where \( \sigma_n^2 \) denotes the error covariance associated with this observation, and \( \hat{\tau}_n \) is the observed TDOA as in (14) and (15).

- It could be beneficial to also use past TDOAs to estimate the current position of the source.
A Kalman filter is governed by the state and observation equation,

\[ x_k = F_{k|k-1} x_{k-1} + u_{k-1}, \quad \text{and} \]
\[ y_k = H_{k|k-1}(x_k) + v_k, \]

respectively, where

- \( F_{k|k-1} \) denotes the transition matrix,
- \( u_{k-1} \) denotes the process noise,
- \( H_{k|k-1}(x) \) denotes the vector observation functional, and
- \( v_k \) denotes the observation noise.
Prediction and Correction

- By assumption $F_{k|k-1}$ is known, so that the *predicted state estimate* is obtained from

$$\hat{x}_{k|k-1} \equiv F_{k|k-1} \hat{x}_{k-1|k-1},$$

(19)

where $\hat{x}_{k-1|k-1}$ is the *filtered state estimate* from the prior time step.

- The new filtered state estimate is calculated from

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + G_k \left[ y_k - H(\hat{x}_{k|k-1}) \right],$$

(20)

where $G_k$ denotes the *Kalman gain*.

- The Kalman gain can be calculated through a well-known recursion.
A block diagram illustrating the prediction and correction steps in the state estimate update of a conventional Kalman filter are shown in the figure.

**Figure**: Predictor-corrector structure of the Kalman filter.
A schematic of the generalized canceller is shown in the figure.

- The *blocking matrix* is chosen to achieve the orthogonality condition $\mathbf{B}^H \mathbf{w}_q = 0$. 
Conventional beamformers minimize the variance of their output subject to a distortionless constraint.

But there other possible optimization criteria:

- The kurtosis of a random variable $X$ is

$$\text{kurt}(X) \triangleq \frac{\mathbb{E}\{|X|^4\}}{\left(\mathbb{E}\{|X|^2\}\right)^2} - 3.$$

- The negentropy of $X$ is

$$\text{negent}(X) \triangleq H(X_{\text{Gauss}}) - H(X).$$

where $H(X)$ is the entropy of $X$ and $X_{\text{Gauss}}$ is a Gaussian random variable with the same variance as $X$.

Kurtosis and negentropy measure non-Gaussianity.
Characteristics of Speech

Figure: Histograms of real parts of subband frequency components of clean speech and a) pdfs, b) noise-corrupted speech and c) reverberated speech.
Distant Speech Recognition Task

- The task is a simple listen and repeat exercise.
- An experimenter repeats a simple phrase and the child subject must repeat it.
- Far-field data is captured with a 64-channel linear microphone with a 2 cm intersensor spacing.
- The vocabulary size is approximately 150 words.
- The four passes of speech recognition are always the same; on the techniques for beamforming vary.
## Distant Speech Recognition Results

<table>
<thead>
<tr>
<th>Beamforming Algorithm</th>
<th>%WER Adult</th>
<th>%WER Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Distant Mic.</td>
<td>3.4</td>
<td>14.2</td>
</tr>
<tr>
<td>Delay-and-Sum</td>
<td>2.2</td>
<td>7.6</td>
</tr>
<tr>
<td>Superdirective</td>
<td>2.1</td>
<td>6.5</td>
</tr>
<tr>
<td>Maximum Kurtosis</td>
<td>0.6</td>
<td>5.3</td>
</tr>
<tr>
<td>Close-talking Mic.</td>
<td>1.9</td>
<td>4.2</td>
</tr>
</tbody>
</table>

**Table:** Distant speech recognition results.
A plane wave impinging with a polar angle of $\theta$ on an array of microphones can be expressed as

$$G_{pw}(kr, \theta, t) = e^{i(\omega t + kr \cos \theta)}$$

$$= \sum_{n=0}^{\infty} i^n (2n + 1) j_n(kr) P_n(\cos \theta) e^{i\omega t}, \quad (21)$$

where $j_n$ and $P_n$ are respectively the spherical Bessel function of the first kind and the Legendre polynomial, both of order $n$, and $k \triangleq 2\pi/\lambda$ is the wavenumber.
Scattered Wave

- Assume the plane wave encounters a rigid sphere with a radius of $a$.
- The scattered wave will have the pressure profile

$$G_s(kr, ka, \theta, t) = - \sum_{n=0}^{\infty} i^n (2n+1) \frac{j_n'(ka)}{h_n'(ka)} h_n(kr) P_n(\cos \theta) e^{i\omega t},$$

(22)

where $h_n = h_n^{(1)}$ denotes the Hankel function of the first kind while the prime indicates the derivative of a function with respect to its argument.
Combining (21) and (22) yields the total sound pressure field

\[ G(kr, ka, \theta) = \sum_{n=0}^{\infty} i^n (2n + 1) b_n(ka, kr) P_n(\cos \theta), \quad (23) \]

where the \( n \)th modal coefficient is defined as

\[ b_n(ka, kr) \triangleq j_n(kr) - \frac{j'_n(ka)}{h'_n(ka)} h_n(kr). \quad (24) \]
Modal Coefficients
Let us now define the spherical harmonic of order $n$ and degree $m$ as

$$Y_n^m(\theta, \phi) \triangleq \sqrt{\frac{(2n+1)(n-m)!}{4\pi(n+m)!}} P_n^m(\cos \theta) e^{im\phi},$$

where $P_n^m$ is the associated Legendre function.

The addition theorem for spherical harmonics states

$$P_n(\cos \gamma) = \frac{4\pi}{2n+1} \sum_{m=-n}^{n} Y_n^m(\theta_s, \phi_s) \bar{Y}_n^m(\theta, \phi),$$

where $\bar{Y}$ denotes the complex conjugate of $Y$. 
Upon substituting (26) into (23), we find

$$G(kr_s, \theta_s, \phi_s, ka, \theta, \phi) =$$

$$4\pi \sum_{n=0}^{\infty} i^n b_n(ka, kr_s) \sum_{m=-n}^{n} Y_n^m(\theta_s, \phi_s) \bar{Y}_n^m(\theta, \phi).$$

Hence, $b_n(ka, kr_s)$ serves as a “weighting function” for all spherical harmonics of order $n$.

The latter fact implies the directivity of a spherical microphone array is also poor at low frequencies.

The spherical harmonics $Y_0 \triangleq Y_0^0$, $Y_1 \triangleq Y_1^0$, $Y_2 \triangleq Y_2^0$ and $Y_3 \triangleq Y_3^0$ are shown in Figure 6.
Orthonormality

The spherical harmonics possess the all important property of orthonormality, which implies

\[ \delta_{n,n'} \delta_{m,m'} = \int_{\Omega} Y_n^m(\theta, \phi) \bar{Y}_{n'}^{m'}(\theta, \phi) \, d\Omega \]  

(28)

where \( \Omega \) denotes the surface of a sphere.
Modal Decomposition

Any sound field $V(kr_s, \theta_s, \phi_s)$, which is square-integrable over a sphere with radius $r_s$, admits the modal decomposition

$$V(kr_s, \theta_s, \phi_s) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} V_n^m(kr_s) Y_n^m(\theta_s, \phi_s)$$

when observed at $(r_s, \theta_s, \phi_s)$, where

$$V_n^m(kr_s) \equiv \int_{\Omega_s} V(kr_s, \theta_s, \phi_s) \bar{Y}_n^m(\theta_s, \phi_s) \, d\Omega_s$$

is the $(n, m)$th coefficient of the decomposition.
A plane wave admits the decomposition

\[ V(kr_s, \theta_s, \phi_s) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} G_n^m(\theta, \phi, ka, kr_s) Y_n^m(\theta_s, \phi_s) \]  

where

\[ G_n^m(\theta, \phi, ka, kr_s) = 4\pi i^n b_n(ka, kr_s) \bar{Y}_n^m(\theta, \phi). \]
For the case of discrete microphones on the surface of a sphere we have

\[ V_n^m(kr_s) \triangleq \frac{4\pi}{S} \sum_{s=1}^{S} V(kr_s, \theta_s, \phi_s) \bar{Y}_n^m(\theta_s, \phi_s), \quad (33) \]

Similarly, the orthonormality condition becomes

\[ \frac{4\pi}{S} \sum_{s=1}^{S} Y_n^m(\theta_s, \phi_s) \bar{Y}_{n'}^{m'}(\theta_s, \phi_s) = \delta_{n,n'} \delta_{m,m'}, \quad (34) \]
Practical Spherical Arrays
Consider the squared-error metric

\[
\epsilon(\theta, \phi, k) \triangleq \sum_{l=0}^{L-1} \left\| v_{k,l} - g_{k,l}(\theta, \phi) B_{k,l}(\theta, \phi) e^{i\omega_l Dk} \right\|^2,
\]

where \( v_{k,l} \) denotes the modal coefficients (33) obtained from a spherical array.

The maximum likelihood estimate of \( B_{k,l}(\theta, \phi) \) is defined as

\[
\hat{B}_{k,l}(\theta, \phi) \triangleq \frac{g_{k,l}^H(\theta, \phi) v_{k,l}}{\left\| g_{k,l}(\theta, \phi) \right\|^2} \cdot e^{-i\omega_l Dk}.
\]
Given the simplicity of (36), we might plausibly modify the standard extended Kalman filter as such:

1. Estimate the scale factors $B_{k,l}$ as in (36).
2. Use this estimate to update the state estimates $(\hat{\theta}_k, \hat{\phi}_k)$ of the Kalman filter.
3. Perform an iterative update for each time step as in the *iterated extended Kalman filter* (IEKF) by repeating Steps 1 and 2.
IEKF Update

The state update of the IEKF involves the steps

\[
S_k(\eta_i) = \tilde{H}_k(\eta_i) K_{k|k-1} \tilde{H}_k^H(\eta_i) + V_k, \tag{37}
\]

\[
G_k(\eta_i) = K_{k|k-1} \tilde{H}_k(\eta_i) S_k^{-1}(\eta_i), \tag{38}
\]

\[
s_k(\eta_i) = y_k - H_k(\eta_i), \tag{39}
\]

\[
\zeta_k(\eta_i) \triangleq s_k(\eta_i) - \tilde{H}_k(\eta_i) (\hat{x}_{k|k-1} - \eta_i), \tag{40}
\]

\[
\eta_{i+1} \triangleq \hat{x}_{k|k-1} + G_k(\eta_i) \zeta_k(\eta_i), \tag{41}
\]

where \( \tilde{H}_k(\eta_i) \) is the linearization of \( H_k(\eta_i) \) about \( \eta_i \).

The local iteration is initialized at \( i = 1 \) by setting

\[
\eta_1 = \hat{x}_{k|k-1}.
\]
Matrix Factorization Lemma

Given any two $N \times M$ matrices $A$ and $B$ with dimensions $N \leq M$, 

$$AA^H = BB^H$$

iff there exists a unitary matrix $\theta$ such that 

$$A\theta = B.$$
Square-Root Implementation of the IEKF

Let \( K_{k|k-1} = K_{k|k-1}^{1/2} K_{k|k-1}^{H/2} \) where \( K_{k|k-1}^{1/2} \) denotes the Cholesky factor of \( K_{k|k-1} \).

Then

\[
A \theta = \begin{bmatrix}
V^{1/2} & \bar{H}_k(\eta_i) K_{k|k-1}^{1/2} & 0 \\
0 & F K_{k|k-1}^{1/2} & U_k^{1/2}
\end{bmatrix} \theta
\]

\[
= \begin{bmatrix}
S_k^{1/2}(\eta_i) & 0 & 0 \\
FG_k(\eta_i) S_k^{1/2}(\eta_i) & K_{k+1|k}^{1/2} & 0
\end{bmatrix} = B. \quad (42)
\]

Only the final estimate of \( K_{k+1|k}^{1/2} \) is saved for use in the succeeding time step.
State Update

The final position update is accomplished as follows:
Through forward substitution we can find that $\zeta'(\eta_i)$ achieving

$$\zeta_k(\eta_i) = S_k^{1/2}(\eta_i)\zeta_k(\eta_i), \quad (43)$$

where $\zeta_k(\eta_i)$ is defined in (40).

Hence, we can find that $\zeta''(\eta_i)$ achieving

$$F\zeta''(\eta_i) = B_{21}\zeta'(\eta_i)$$

through back substitution on $F$.

Finally, we update $\eta_i$ according to

$$\eta_{i+1} = \hat{x}_{k|k-1} + \zeta''(\eta_i), \quad (44)$$

where $\eta_1 = \hat{x}_{k|k-1}$. 
Consider once more the squared-error metric

$$
\epsilon(\theta, \phi, k) \triangleq \sum_{l=0}^{L-1} \left\| v_{k,l} - g_{k,l}(\theta, \phi) B_{k,l} e^{i\omega_l D_k} \right\|^2,
$$

Clearly, the proposed speaker tracking algorithm is an example of *analysis by synthesis*.

The synthesized sound field can readily be extended to include other effects:
- Diffuse noise;
- Sources of coherent interference;
- Other speakers.
Spherical Array Beampatterns

- Delay-and-Sum
- Hypercardioid
- Symmetric MVDR
- Asymmetric MVDR