Beamforming-Based Multichannel Acoustic Echo Cancellation

Yuval Konforti
M.Sc. Thesis Seminar

Supervisors: Prof. Israel Cohen and Dr. Baruch Berdugo
Outline

- Background and motivation
  - Acoustic Echo Cancellation
  - Region-of-Interest Beamforming
- Multichannel Acoustic Echo Cancellation with Beamforming in Dynamic Environments
- Array Geometry Optimization for Region-of-Interest Broadband Beamforming
- Conclusions
  - Research Contributions
  - Future research

Background and Motivation
Acoustic Echo Cancellation

- Echo can interfere and disrupt conversations.
- Stems from the acoustic coupling between a loudspeaker and a microphone.
- Echo is eliminated with an acoustic echo canceller (AEC).
  - Usually solved by a normalized least-mean-square filter (NLMS) approach.
  - Involves a double-talk detector.
Applications

Amazon

Apple Inc.

Webex by Cisco

Zoom
Multichannel Acoustic Echo Cancellation

- In real environments, two phenomena worsen performance:
  - Nonlinear loudspeaker distortion
  - Background noise
- Multichannel AEC can be used to mitigate these effects.
Spatial Filtering

- A signal is propagated in space.
- The signal arrives from a certain direction.
- A sensor array is utilized. The time difference of arrival (TDOA) between the sensors is in accordance with the direction.
Beamforming - a Sum of Digital Filters

- With one sensor:

\[ Z(f) = H(f)Y(f) \]

- With multiple sensors:

\[ Z(f) = \sum_{m=1}^{M} H_m^*(f)Y_m(f) = h^H(f)y(f) \]
Sensor Array Beamforming

- We define the steering vector

\[ \mathbf{d}(f, \cos \theta) = \begin{bmatrix} 1 & e^{-j2\pi f_0 \cos \theta} & \ldots & e^{-j(M-1)2\pi f_0 \cos \theta} \end{bmatrix}^T \]

- Each sensor utilizes a digital filter \( H_m(f) \).
- The response is measured by the beampattern:
  - Angle dependent.
  - Frequency dependent.

\[ \mathcal{B} [\mathbf{h}(f), \cos \theta] = \mathbf{d}^H (f, \cos \theta) \mathbf{h}(f) = \sum_{m=1}^{M} H_m(f) e^{j(m-1)2\pi f_0 \cos \theta} \]
Multiframe Filtering

- In the short-time fourier transform (STFT) domain, speech signals are correlative between frames.
- The inter-frame correlations must be updated, as they highly vary in time.
- Multiframe and Multichannel filters were used for speech separation and noise reduction(*), yet not for AEC.

AEC Performance Measures

- Echo-return loss enhancement (ERLE):
  - $y_1(t)$ - Echo component at reference microphone.
  - $y_{re}(t)$ - Residual echo component.

- Distortion Index (DI):
  - $u_1(t)$ - Desired component at reference microphone.
  - $u_f(t)$ - Filtered desired component.

$$\xi(t) = \frac{\text{LPF} \{ y_1^2(t) \}}{\text{LPF} \{ y_{re}^2(t) \}}$$

$$\nu(t) = \frac{\text{LPF} \{ [u_1(t) - u_f(t)]^2 \}}{\text{LPF} \{ u_1^2(t) \}}$$
Region-of-Interest Beamforming

- In many applications, the source location is unknown, but it can be assumed it is in a region-of-interest (ROI).
- Source localization / Direction-of-arrival (DOA) estimation may be employed to track a moving source.
- While the coefficients change, the array geometry is preserved.
Applications
Array Geometry Optimization

- Typically, the beamforming coefficients are found per given geometry. Usually symmetric geometries are considered:
  - Uniform Linear Arrays (ULAs).
  - Rectangular Arrays.
  - Uniform Circular Arrays (UCAs).
  - Uniform Concentric Circular Arrays (UCCAs).
- Differential Microphone Arrays (DMAs) produce an approximate frequency invariant (FI) response.
- The geometry of a microphone array has an important impact on beamforming performance.
Optimization Methods

- Two common methods for array geometry optimization:
  - Greedy-based approaches.
  - Genetic Algorithms.
- Greedy-based approaches find the best position for the placed microphone in each step, but the overall result may not be optimal.
- Genetic algorithms are unstable and may also converge to nonoptimal results.
- The study in (*) optimize the geometry for a ROI, but only for narrowband signals.

Beamforming Performance Measures

- **White Noise Gain:**
  \[
  \mathcal{W} [h(x, \omega, \theta)] \triangleq \frac{\left| d^H(x, \omega, \theta) h(x, \omega, \theta) \right|^2}{h^H(x, \omega, \theta) h(x, \omega, \theta)}
  \]

- **Directivity Factor (DF):**
  \[
  \mathcal{D} [h(x, \omega, \theta)] \triangleq \frac{\left| d^H(x, \omega, \theta) h(x, \omega, \theta) \right|^2}{h^H(x, \omega, \theta) \Gamma(x, \omega) h(x, \omega, \theta)}
  \]

  where
  \[
  \Gamma_{i,j}(x, \omega) = \frac{\sin(\omega (x_i - x_j) / c)}{\omega (x_i - x_j) / c}, \quad 1 \leq i, j \leq M.
  \]

- **Directivity Index:**
  \[
  \mathcal{DI}_{[\omega_L, \omega_H]} [h(x, \omega, \theta)] \triangleq \frac{\int_{\omega_L}^{\omega_H} \left| d^H(x, \omega, \theta) h(x, \omega, \theta) \right|^2 d\omega}{\int_{\omega_L}^{\omega_H} h^H(x, \omega, \theta) \Gamma(x, \omega) h(x, \omega, \theta) d\omega}
  \]
Multichannel Acoustic Echo Cancellation with Beamforming in Dynamic Environments
A Beamforming-Based Approach

- $y(k, n)$ - Echo component received by the array.
- $u(k, n)$ - Desired component received signal by the array.
- $v(k, n)$ - Background noise received signal by the array.
- $d(k, n)$ - total received signal by the array.

\[
\hat{U}(k, n) = h^H(k, n) d(k, n)
\]

\[
d(k, n) = y(k, n) + u(k, n) + v(k, n)
\]
Linear-Constraint-Minimum-Variance Beamforming

- Theoretically enforces:
  - Complete echo cancellation.
  - No distortion.
  - Minimum residual noise.
- Ideal ERLE and DI.
- Requires accurate estimates of $g(k)$ and $q(k)$.

$$
\mathbf{h}^{\ast} (k, n) = \arg \min_{\mathbf{h}(k, n)} ||\mathbf{h} (k, n)||^2
$$

s.t.

$$
C_1 [\mathbf{h} (k, n)] : \mathbf{h}^H (k, n) \mathbf{g}(k) = 0
$$
$$
C_2 [\mathbf{h} (k, n)] : \mathbf{h}^H (k, n) \mathbf{q}(k) = 1
$$

$$
\mathbf{h}^{\ast} (k, n) = \mathbf{C} (k) [\mathbf{C}^H (k) \mathbf{C} (k)]^{-1} \mathbf{i}_c
$$

where

$$
\mathbf{C} (k, n) = [\mathbf{q} (k), \mathbf{g} (k)]
$$

$$
\mathbf{i}_c = [1, 0]^T
$$
AEC Scheme

Procedure:
1. Steering vectors $g(k)$ and $q(k)$ are estimated.
2. Beamformer $h(k, n)$ is designed.
3. Estimate $\hat{U}(k, n)$ is produced.

No double talk detection used.
Utilizing Multiple Sensors - A Frame Invariant Expression

- Neglecting nonlinear loudspeaker distortion.
- Neglecting background noise.
- Assuming a static environment in the last $L$ frames.
- Utilizing the Multiplicative Transfer Function (MTF) approximation (*)

$$D_m(k, n - l + 1) = G_m(k) X(k, n - l + 1) + Q_m(k) S(k, n - l + 1)$$

For any two microphones $m_1$ and $m_2$:

$$\frac{Q_{m_1}(k)}{Q_{m_2}(k)} = \frac{D_{m_1}(k, n - l + 1) - G_{m_1}(k) X(k, n - l + 1)}{D_{m_2}(k, n - l + 1) - G_{m_2}(k) X(k, n - l + 1)}$$

Utilizing Multiple Frames

- For any two recent frames $l_1$ and $l_2$:

$$
\begin{align*}
\frac{D_{m_1}(k, n - l_1 + 1) - G_{m_1}(k)X(k, n - l_1 + 1)}{D_{m_2}(k, n - l_1 + 1) - G_{m_2}(k)X(k, n - l_1 + 1)} &= \\
\frac{D_{m_1}(k, n - l_2 + 1) - G_{m_1}(k)X(k, n - l_2 + 1)}{D_{m_2}(k, n - l_2 + 1) - G_{m_2}(k)X(k, n - l_2 + 1)}
\end{align*}
$$

- Quadratic elements of $G_{m_1}(k)G_{m_2}(k)$ are reduced.
- A linear equation with respect to $G_{m_1}(k)$ and $G_{m_2}(k)$ is obtained:

$$
\begin{align*}
G_{m_1}(k) \left[ X(k, n - l_1 + 1)D_{m_2}(k, n - l_2 + 1) - X(k, n - l_2 + 1)D_{m_2}(k, n - l_1 + 1) \right] + \\
G_{m_2}(k) \left[ X(k, n - l_2 + 1)D_{m_1}(k, n - l_1 + 1) - X(k, n - l_1 + 1)D_{m_1}(k, n - l_2 + 1) \right] = \\
D_{m_1}(k, n - l_1 + 1)D_{m_2}(k, n - l_2 + 1) - D_{m_3}(k, n - l_2 + 1)D_{m_2}(k, n - l_1 + 1)
\end{align*}
$$
A Linear Set of Equations

- Overall, for any pick of $1 \leq m_1, m_2 \leq M$ and $1 \leq l_1, l_2 \leq L$, we get such an equation.

- Uninformative equations:
  - $m_1 = m_2$
  - $l_1 = l_2$
  - Swapping $m_1 \leftrightarrow m_2$
  - Swapping $l_1 \leftrightarrow l_2$

- Must have more equations than variables: $M \leq \binom{M}{2} \binom{L}{2} \geq L(L - 1)(M - 1) \geq 4$

- Must have multiple frames, multiple sensors, and $M \geq 3$ or $L \geq 3$. 
A Least-Mean-Squares Approach

- Equations may contradict due to assumptions.
- A least-mean-squares (LMS) solution provides a good estimate.
- As $L$ grows:
  - More equations are added.
  - The environment is assumed to be static for longer periods.
- As $M$ grows:
  - More equations are added.
  - More variables are added.
Matrix Formulation

Inputs: $M, L,$

$$X (k, n - l + 1), D_m (k, n - l + 1) \quad 1 \leq l \leq L \quad 1 \leq m \leq M$$

Outputs: $\tilde{G}_m (k, n)$

$1 \leq m \leq M$

Create list $M_{\text{list}}$ of $\binom{M}{2}$ non-repetitive pairs $(m_1, m_2)$

Create list $L_{\text{list}}$ of $\binom{L}{2}$ non-repetitive pairs $(l_1, l_2)$

$$A (k, n) \leftarrow 0_{\binom{M}{2} \times M}$$

for $i = 1, 2, \ldots, \binom{M}{2}$ do

$$(m_1, m_2) \leftarrow M_{\text{list}} (i)$$

for $j = 1, 2, \ldots, \binom{L}{2}$ do

$$(l_1, l_2) \leftarrow L_{\text{list}} (j)$$

$$A_{\left[\binom{i}{2} + j, m_1\right]} (k, n) \leftarrow X (k, n - l_1 + 1) D_{m_2} (k, n - l_2 + 1) - X (k, n - l_2 + 1) D_{m_2} (k, n - l_1 + 1)$$

$$A_{\left[\binom{i}{2} + j, m_2\right]} (k, n) \leftarrow X (k, n - l_2 + 1) D_{m_1} (k, n - l_1 + 1) - X (k, n - l_1 + 1) D_{m_1} (k, n - l_2 + 1)$$

$$b_{\left[\binom{i}{2} + j\right]} (k, n) \leftarrow D_{m_1} (k, n - l_1 + 1) D_{m_2} (k, n - l_2 + 1) - D_{m_1} (k, n - l_2 + 1) D_{m_2} (k, n - l_1 + 1)$$

end for

end for

$$\left[\tilde{G}_1 (k, n), \tilde{G}_2 (k, n), \ldots, \tilde{G}_M (k, n)\right]^T \leftarrow A^\dagger (k, n) b (k, n)$$
Simulation Configuration

- Dynamic environment.
- Double-talk.
- Low Signal-to-Echo Ratio in 4 scenarios:
  1. Speakerphone at $A$, talker at $C$.
     -17.89 dB.
  2. Speakerphone at $A$, talker at $D$.
     -18.7 dB.
     -15.27 dB.
     -16.89 dB.
Results - Echo Cancellation

a. The total received signal in the reference microphone $d_1(t)$
b. The echo component signal in the reference microphone $y_1(t)$
c. The desired component signal in the reference microphone $u_1(t)$
d. The beamformer output signal $\hat{u}(t)$.

- Good echo cancellation despite significant acoustic coupling.
Results - Performance as Function of Microphones

a. ERLE.
b. DI.

- **Blue** - $M = 2$, **Red** - $M = 3$, **Yellow** - $M = 4$, **Purple** - $M = 5$, $L = 4$.
- Significant change only between $M = 2$ and $M = 3$.
- Performance limit. Increasing $M$ also increases number of variables in estimation process.
Results - Performance as Function of Frames

a. ERLE.

b. DI.


- $L = 2$ are insufficient for the proposed approach.

- Performance limit. Increasing $L$ also relies more strongly on a longer period where the environment is static.
Results - Method Comparison (*)

a. ERLE.

b. DI.

- The NLMS filter in (*) was trained in a period like the first segment, but with no near-end speech.

- Improvement in both ERLE and DI.
- Significant improvement after the speakerphone moves due to changing reflections. The NLMS filter is irrelevant once the reflections have changed.

Array Geometry Optimization for Region-of-Interest Broadband Beamforming
A Variable Geometry

- $M$ omnidirectional microphones are placed nonuniformly across a linear aperture $A$ in placements $x_m$, $m = 1, 2, \ldots, M$.

- The steering vector is

\[
d(x, \omega, \theta) = \begin{bmatrix} e^{-j \frac{\omega}{c} x_1 \cos \theta}, e^{-j \frac{\omega}{c} x_2 \cos \theta}, \ldots, e^{-j \frac{\omega}{c} x_M \cos \theta} \end{bmatrix}^T
\]
Coefficients Dependent on Geometry

For different geometries, the coefficients are designed differently. For a geometry $x$ and look direction $\theta$, we define the coefficients vector as

$$\mathbf{h}(x, \omega, \theta) \triangleq \begin{bmatrix} H_1(x, \omega, \theta), H_2(x, \omega, \theta), \ldots, H_M(x, \omega, \theta) \end{bmatrix}^T$$

With geometry $x$, the beampattern directed toward $\tilde{\theta}$ has a response at angle $\theta$ of

$$B\left[\mathbf{h}(x, \omega, \tilde{\theta}), \theta\right] = \mathbf{d}_H^H(x, \omega, \theta) \mathbf{h}(x, \omega, \tilde{\theta})$$
Problem Formulation

- Our objective is to find the optimal array geometry $x$, that maximizes the worst-case directivity index, in an ROI $\Theta$. Each beamformer, directed toward $\theta$, must admit to the distortionless constraint, have sufficient WNG, and maintain minimal distances.

$$\mathbf{x}^* = \arg \max_{\mathbf{x}} \min_{\theta \in \Theta} \mathcal{D}[\omega_L, \omega_H] \left[ \mathbf{h}(\mathbf{x}, \omega, \theta) \right]$$

s.t. $\mathcal{B} [\mathbf{h}(\mathbf{x}, \omega, \theta), \theta] = 1 \quad \forall \theta \in \Theta, \forall \omega \in \Omega$

$\mathcal{W} [\mathbf{h}(\mathbf{x}, \omega, \theta)] \geq \delta \quad \forall \theta \in \Theta, \forall \omega \in \Omega$

$|x_i - x_j| \geq d_c \quad \forall i, j \in [1, M], i \neq j$

$0 \leq x_m \leq A \quad \forall m \in [1, M]$

- Not convex. Cannot be solved by convex optimization algorithms.
Formalizing a Solvable Problem - Constraints

- Consider $N$ candidate microphone locations, $Q$ frequencies, and $P$ look directions.

- To guarantee number of microphones:
  \[ C_1 [s] : s^H i_N = M \]

- To guarantee minimal spacing:
  \[ C_2 [s] : s^H U \leq i_C^T \]

- To guarantee the distortionless constraint:
  \[
  C_3 [h_{\text{tot}} (\omega, \theta)] : d_{\text{tot}}^H (\omega_q, \theta_p) h_{\text{tot}} (\omega_q, \theta_p) = 1 \quad \forall p \in [1, P], \quad \forall q \in [1, Q]
  \]

- To guarantee the desired WNG:
  \[
  C_4 [h_{\text{tot}} (\omega, \theta)] : h_{\text{tot}}^H (\omega_q, \theta_p) h_{\text{tot}} (\omega_q, \theta_p) \leq \frac{1}{\delta} \quad \forall p \in [1, P], \quad \forall q \in [1, Q]
  \]

- To guarantee beamformer use of selected placements:
  \[
  C_5 [s, h_{\text{tot}} (\omega, \theta)] : |H_{\text{tot}, i} (\omega_q, \theta_p)|^2 \leq \frac{S_i}{\delta} \quad \forall i \in [1, N], \quad \forall p \in [1, P], \quad \forall q \in [1, Q]
  \]

- $s = [1, 0, 1, 0, 1]^T$
When the distortionless constraint is met, the directivity index is determined only by the denominator. Therefore, to maximize the worst-case directivity index we should minimize:

$$R[h_{tot}(\omega, \theta)]$$

$$= \max_{p \in [1, P]} \sum_{q=1}^{Q} h_{tot}^{H}(\omega_q, \theta_p) \Gamma_{tot}(\omega_q) h_{tot}(\omega_q, \theta_p)$$
Optimal Array Design

- The optimal design is found by solving a mixed-integer convex optimization problem.
- The non-zero elements of the optimal binary vector $s^*$ yield the optimal microphone locations $x^*$. The non-zero elements of the optimal coefficients $h^*_\text{tot}(\omega, \theta)$ yield the optimal coefficients $h^*(x^*, \omega, \theta)$.

\[
\min_{s, h_{\text{tot}}(\omega, \theta)} R [h_{\text{tot}}(\omega, \theta)] \\
\text{s.t. } C_1 [s], C_2 [s], C_3 [h_{\text{tot}}(\omega, \theta)], C_4 [h_{\text{tot}}(\omega, \theta)], C_5 [s, h_{\text{tot}}(\omega, \theta)]
\]
Coefficient Post-Processing

- Since the worst-case look direction is considered, beamformers directed toward other directions may not yield the best possible directivity. To circumvent this, we introduce a post-processing scheme.

- To maximize directivity while maintaining sufficient WNG, the coefficients are found by the robust superdirective beamformer:

\[
\mathbf{h}_\epsilon (x^*, \omega, \theta) = \frac{\mathbf{\Gamma}_\epsilon^{-1} (x^*, \omega) \mathbf{d} (x^*, \omega, \theta)}{\mathbf{d}^H (x^*, \omega, \theta) \mathbf{\Gamma}_\epsilon^{-1} (x^*, \omega) \mathbf{d} (x^*, \omega, \theta)}
\]

- For every frequency and look direction, \( \epsilon \) is found by a bisection search:

\[
0 \leq \epsilon \leq \frac{\lambda_1 - \sqrt{M/\delta} \lambda_M}{\sqrt{M/\delta} - 1}
\]

where \( \mathbf{\Gamma} (x^*, \omega) = \mathbf{Q} (x^*, \omega) \mathbf{\Lambda} (x^*, \omega) \mathbf{Q}^T (x^*, \omega) \) is the eigenvalue decomposition such that \( \mathbf{\Lambda} = \text{diag} [\lambda_1, \lambda_2, ..., \lambda_M] \), \( \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_M \).
Results - Geometry

- $M = 6$ microphones, aperture length $A = 17.5\, [cm]$, minimal distance $d_c = 0.5\, [cm]$. Frequencies from $f_L = 2\, [kHz]$ to $f_H = 6\, [kHz]$, look directions up to $\theta_H = 30^\circ$. Minimum WNG is $\delta = -10\, [dB]$. $N = 40$, $Q = 15$, $P = 15$.

- Some sensors close together yield high directivity, like DMAs.
- Some sensors further apart, due to WNG constraint.
Results - Directivity Index

- Better directivity in the worst-case direction.
- Better directivity in all ROI directions.
Results - WNG and Directivity Factor

- (a) ULA – Excellent WNG, but spatial aliasing damages DF at higher frequencies.
- (b) Dense geometry – Barely sufficient WNG, cannot produce high directivity due to WNG constraint.
- (c) Proposed – Good WNG, good broadband directivity.
Conclusions
Research Contributions

- A novel beamforming-based multichannel AEC method was proposed.
  - Capable of operating in dynamic environments.
  - Does not require double-talk detection.
  - Robust to extremely low SER.
  - Achieves higher ERLE and lower DI compared to an existing method.

- The array geometry for broadband ROI was optimized.
  - A convex framework was used, enabling the convergence to the global optimum.
  - A broadband frequency range was considered.
  - Sensors were placed close / further apart, to achieve compromise between directivity and noise robustness.
Future Research

- Modelling nonlinear loudspeaker distortion and background noise. The LMS approach works well only if those components are relatively small.

- Developing more accurate inter-frame and inter-sensor relations. For example, instead of the MTF approximation (*) using the Cross-Multiplicative Transfer Function (CMTF) approximation (**).

- Exploring other array types. The search is limited to non-uniform linear arrays. This will enable more complex ROIs too.

- Geometry optimization by other approaches, such as learning-based methods or deep neural networks (DNNs). Our framework runtime grows considerably with more optimization variables.


Questions?

Thank you!

Special thanks to my supervisors Prof. Israel Cohen and Dr. Baruch Berdugo