Speaker Localization for Far-field and Near-field Wideband Sources Using Neural Networks

Güner Arslan[1]

F. Ayhan Sakarya[2,3]

Brian L. Evans[1]

[1]Embedded Signal Processing Laboratory, Dept. of Electrical and Computer Engineering,
The University of Texas at Austin

[2]Dept. of Electronics and Telecommunication Engineering, Yý ldý z Technical University

[3]Wireless Technology Laboratory, Lucent Technologies

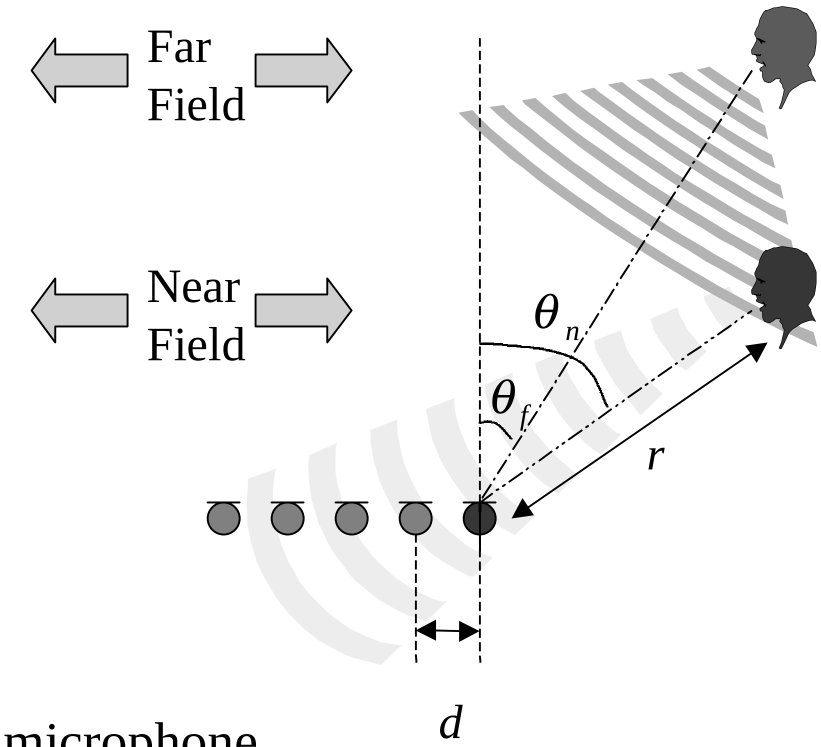
Introduction

- Speaker localization
  - Videoconferencing: steer camera to speaker
  - Acoustic echo cancellation: steer beam to speaker
- Spatial array source localization techniques
  - Computationally intensive
- Spatial array neural network techniques
  - Massive parallelism
  - Far-field assumption
  - Narrowband assumption

Speaker Localization

$$\tau_m^{far} = (m-1)\frac{d \sin\theta_f}{c}$$

$$\tau_m^{near} = \frac{r - \sqrt{r^2 - 2(m-1)rd\sin\theta_n + (m-1)^2 d^2}}{c}$$



c : Velocity of sound in air

d : Inter-sensor spacing

m : Microphone index number

r : Range of near-field speaker

τ_m : Time delay between reference and m^{th} microphone

θ_f : Far-field direction of arrival

θ_n : Near-field direction of arrival (DOA)

Feature Selection

- Desired properties of feature vectors
  - Mapping to desired DOAs
  - Independent of phase, frequency, bandwidth, and amplitude
  - Easy to compute

DOA ⟺ Time delay ⟺ Phase difference in frequency

- Cross-power spectrum between all pairs of neighboring sensors
  - Includes phase difference between sensors
  - Depends on frequency
  - computationally intensive

Feature Selection

- Normalized instantaneous cross-power spectrum

$$\phi_{m,m+1}(\Omega_i) = e^{-j\Omega_i(\tau_m - \tau_{m+1})}$$

DOA $\Longleftrightarrow$ Time delay $\Longleftrightarrow$ Phase difference $\Longleftrightarrow$ Cross-power spectrum

- Eliminates amplitude dependence
- Calculate at *K* different frequencies
- Skip averaging
  - rough estimate

Calculation of Feature Vectors

- Calculate *N*-point FFT at every sensor
- Find the indexes of the *K* largest FFT coefficients in absolute value at the reference sensor
- For every neighboring pair of sensors, complex conjugate multiply the FFT coefficients at these indexes
- Normalize all results with their absolute value
- Construct a vector containing the real and imaginary parts of the results and the index numbers

Array Speech Signal Model

- Signal received by sensor *m* at time index *n*

$$s_m[n] = \sum_{k=1}^{N_s} a_k \cos\left(2\pi \frac{f_k}{f_s} n - \phi_k - 2\pi \ f_k \tau_m \right) + v[n]$$

N_s : Number of frequencies in the wideband signal

f_k : The k^{th} frequency (uniform distribution on [200,2000])

a_k : Random amplitude of k^{th} frequency (uniform distribution on [0,1])

ϕ_k : Random phase of the k^{th} frequency (uniform distribution on $[0,2\pi]$)

f_s : Sampling frequency

τ_m : Time delay between the reference sensor and m^{th} sensor

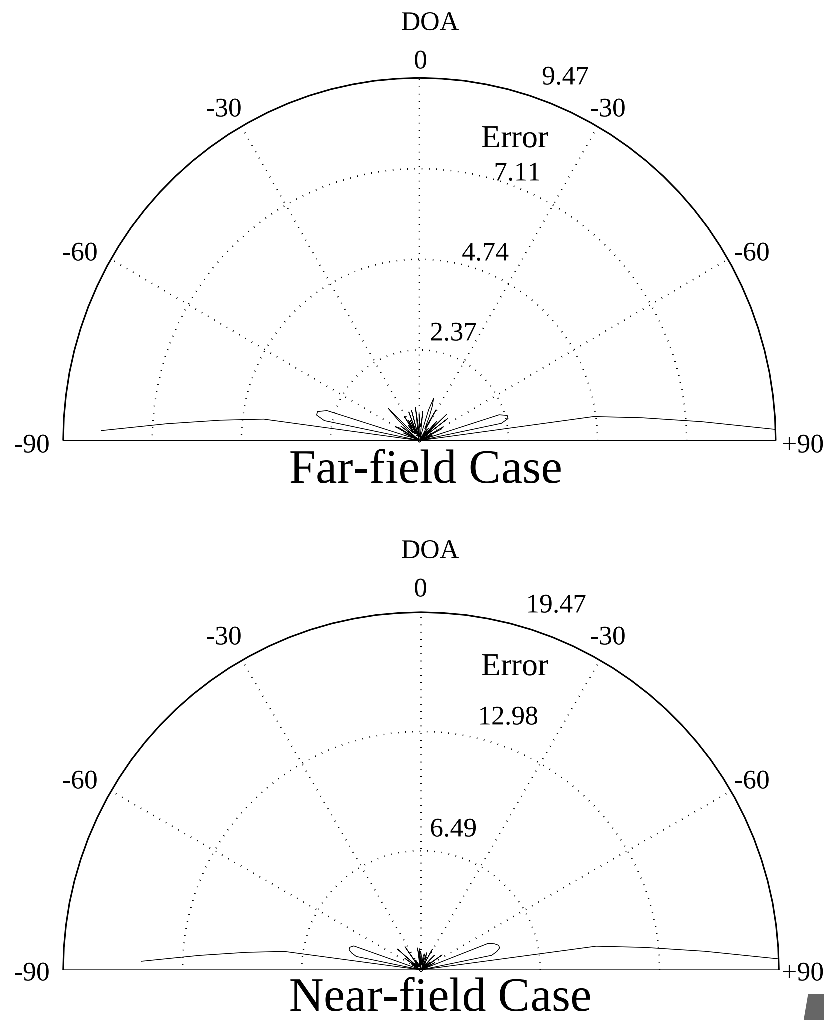
$v[n]$: White Gaussian noise

Training and Testing

- Training data
  - Vary DOA from -90^o to 90^o with 5^o increments
  - Include -90^o but exclude $+90^o$ because of ambiguity
  - Use 100 independent sets of 128 snapshots for each DO
- Training
  - Fast backpropagation for multilayer perceptron neural network
  - Repeat training 10 times with random initial weights
- Testing
  - 100 independent tests
  - Compute error as difference between estimated and real DOA
  - Average absolute error for DOA from -90^o to 90^o with 1^o steps

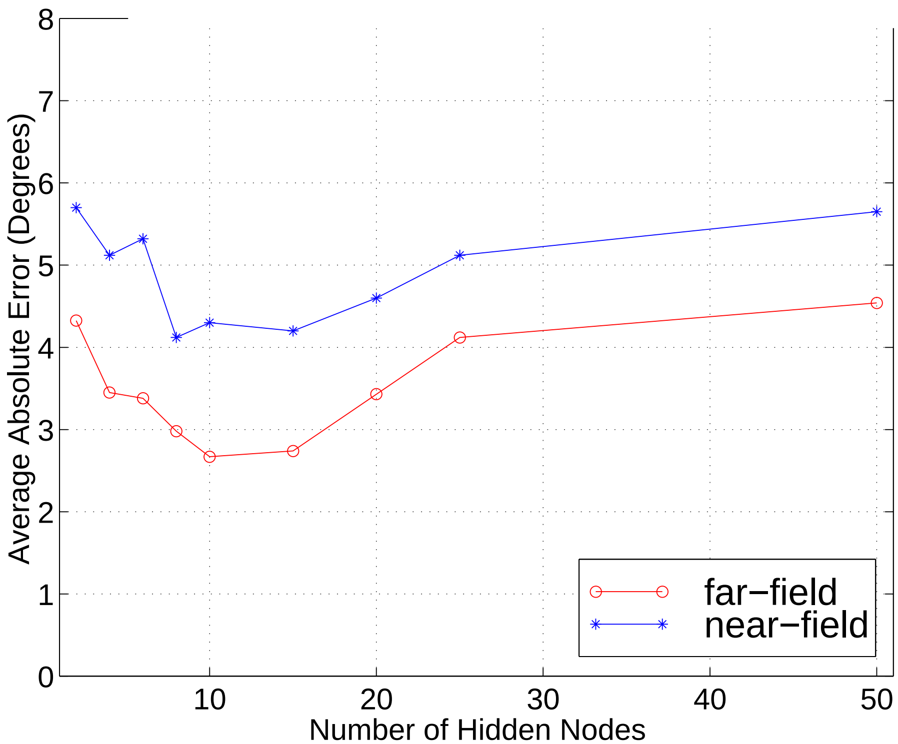
Error Distribution

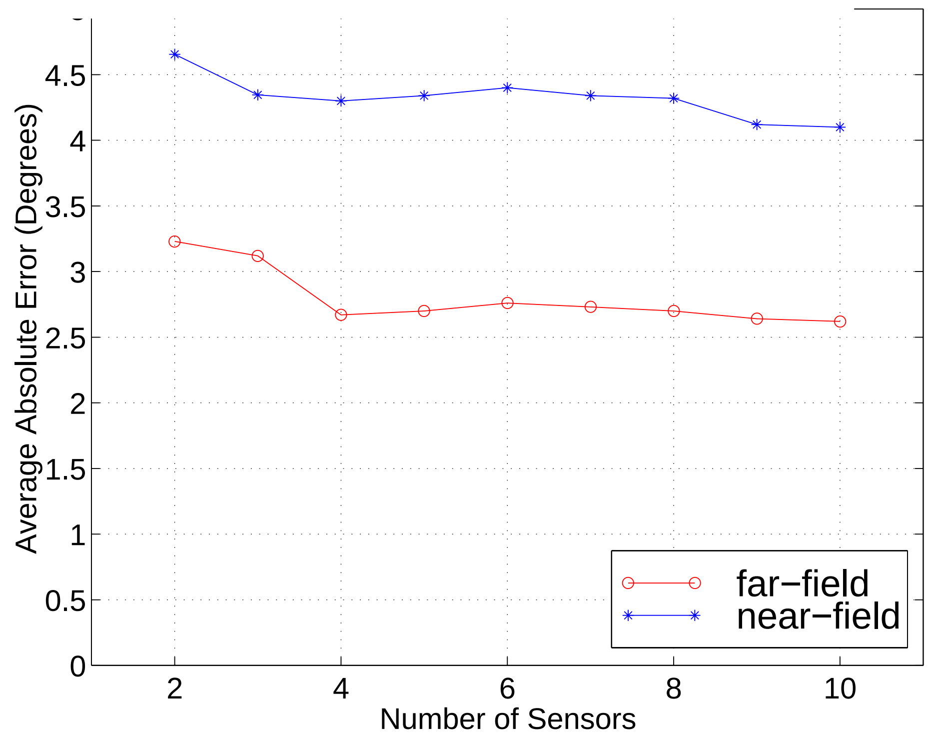
- Error measure
  - Average absolute error over 100 independent tests
  - Most error occurs near -90 and +90 degrees due to ambiguity at -90º and 90º



Training

- 128 snapshots, 6 dominant frequencies, 0.05 m inter-sensor spacing

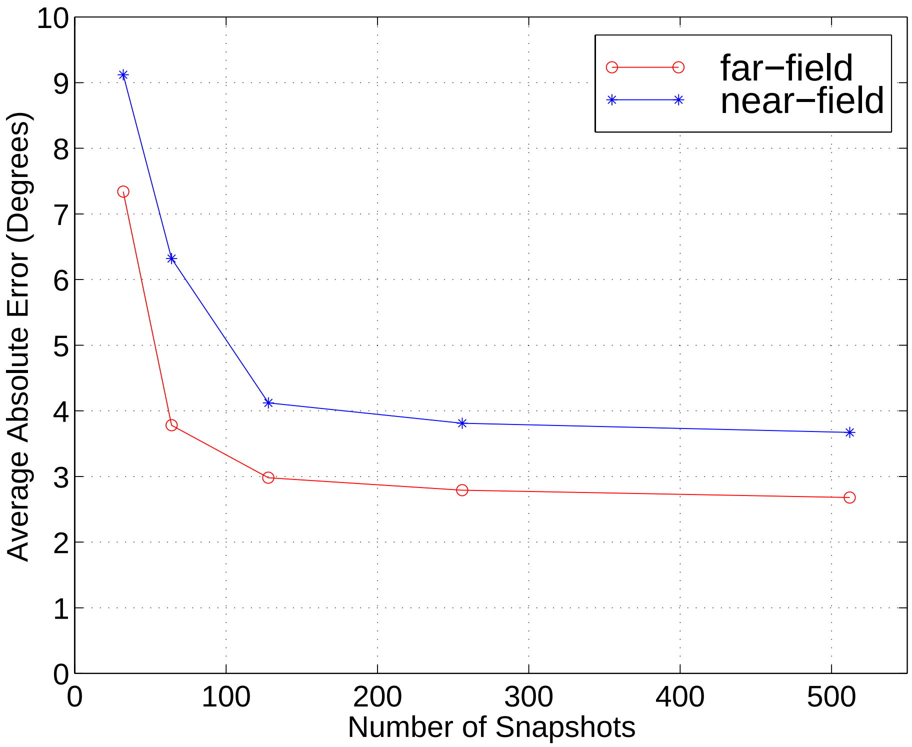


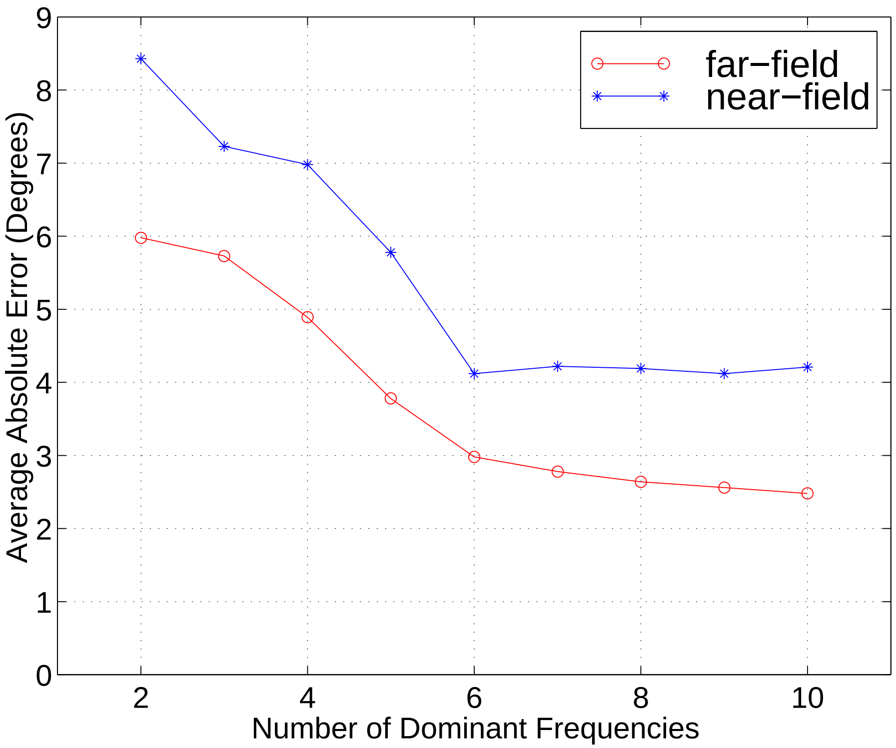


- Best performance with 10 hidden nodes and 4 sensors

Training

- 10 hidden units, 4 sensors, 0.05m inter-sensor spacing

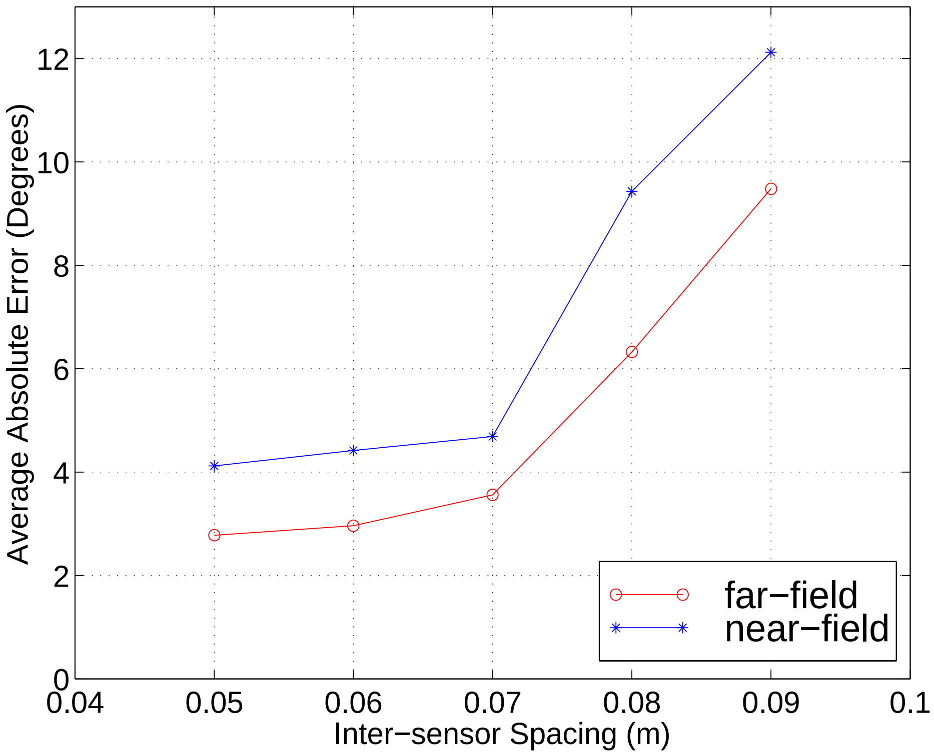


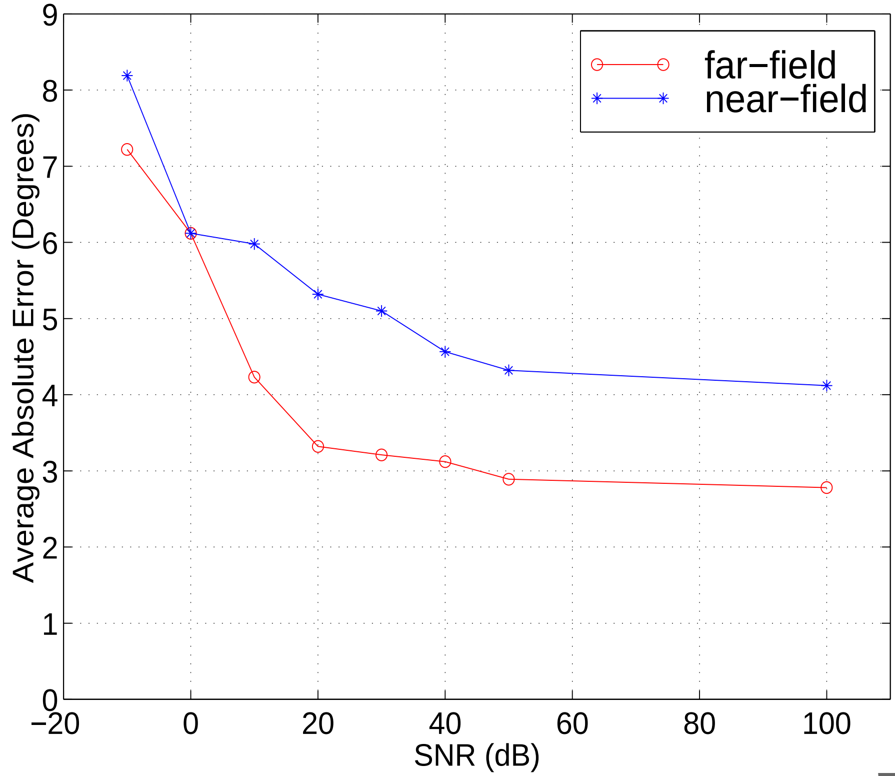


- Diminishing returns after 128 snapshots and 6 dominant frequencies

Training

- 4 sensors, 10 hidden units, 4 sensors, 6 dominant frequencies





- Best inter-sensor spacing in 0.05 m

Computational Requirements

Multiplication	$2MN\log_2 N$	$+$	$4KM$	$+$	$((2M-1)K+1)L$
Addition	$2MN\log_2 N$	$+$	$4KM$	$+$	$((2M-1)K+1)L$
Division	0	$+$	$K(M-1)$	$+$	0
Lookup table	No		No		Yes

N-point Fast Fourier Transform at *M* sensors

Instantaneous cross-power spectrum estimation

Multilayer perceptron forward propagation

- For M=4, N=128, K=6, L=10, f_s=8000 Hz : 1 MFLOPS/s

Conclusion

- Localization for near-field and far-field speakers
- Choose instantaneous cross-power spectrum samples as feature vectors
- Design a neural network to map feature vectors to DOAs
- Estimate DOAs with average error of less than 6^o
- May be implemented in real-time in software of hardware
- Approach could be generalized for direction finding of non-speech signals