Calibration of Microphone Arrays for Improved Speech Recognition

Michael L.Seltzer[1] and Bhiksha Raj[2]

1. Department of Electrical and Computer Engineering and School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15217 USA

2. Mitsubishi Electric Research Lab

Cambridge, MA 02139 USA

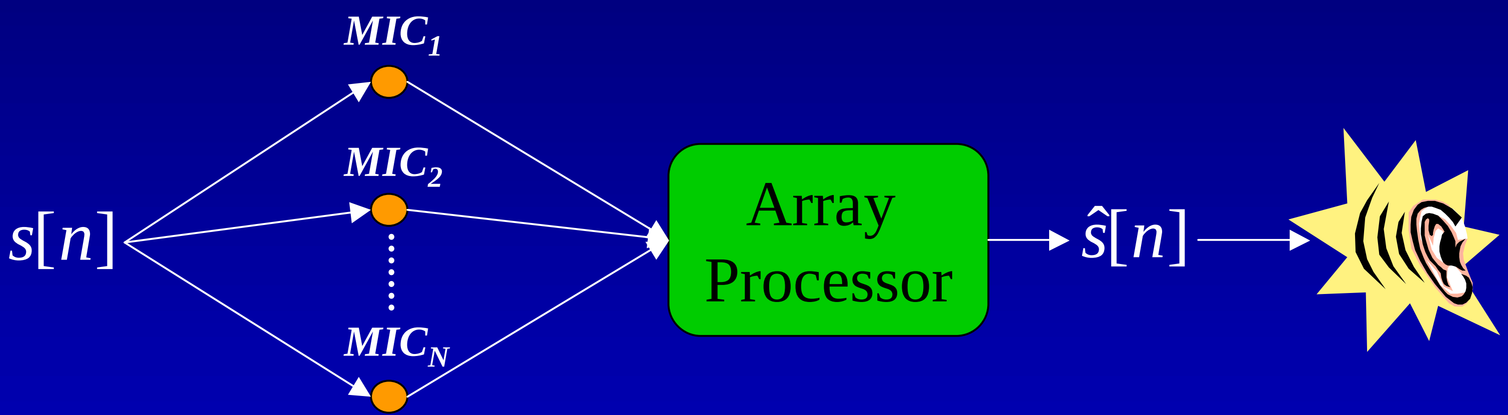
4 September 2001

Introduction

- Current speech recognition technology is capable of good performance in quiet conditions with close-talking microphones.
- In many applications, the environment is noisy and the use of a close-talking microphone is impossible or inconvenient.
- As the distance between the user and the microphone grows, the signal is increasingly susceptible to distortions from the environment.
- Using an array of microphones, rather than a single microphone, has been proposed as a solution to this problem.

Microphone Array Processing

- Combine multiple signals captured by the array to obtain a higher quality output signal, as judged (typically) by a *human listener.*

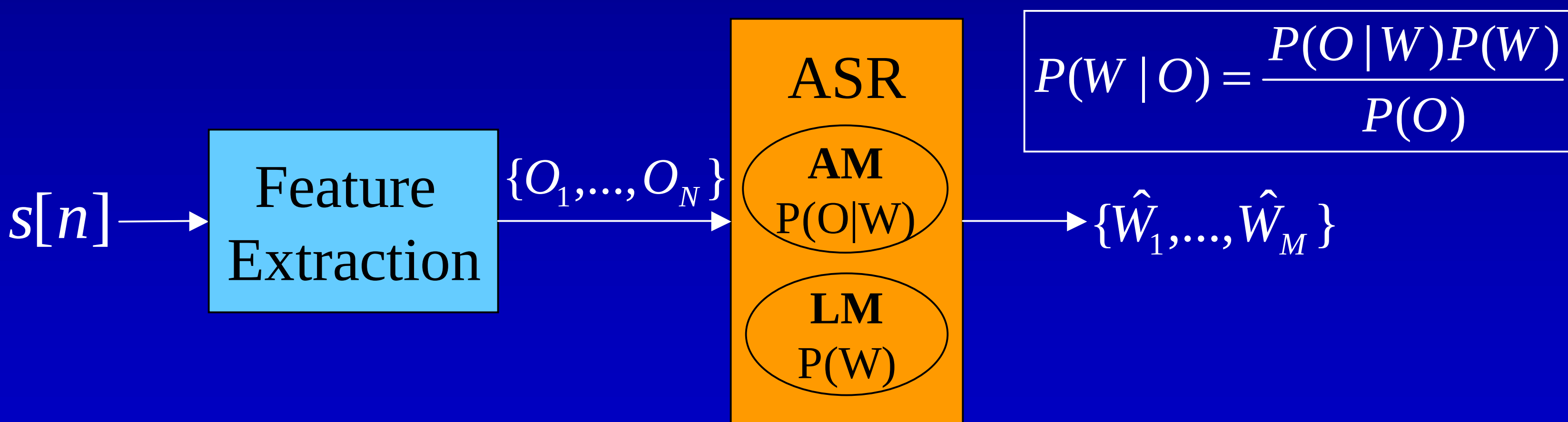


- Many array processing methods exist:
  - Fixed/adaptive schemes, de-reverberation techniques, blind source separation.
- The objective of these methods is *speech enhancement*, a *signal processing* problem.

Automatic Speech Recognition (ASR)

- Parameterize speech signal and compare parameter sequence to statistical models of speech sound units to hypothesize what a user said.

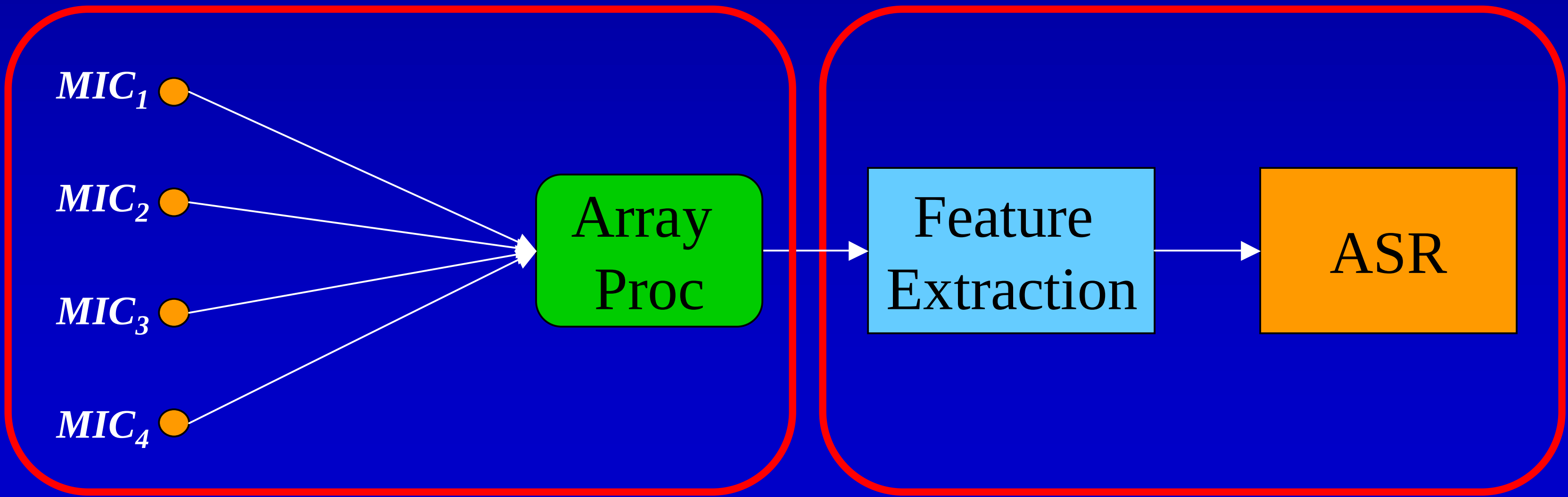
- The speech signal is interpreted by a *machine*.



- The objective is *accurate recognition*, a *statistical pattern classification* problem.

ASR with Microphone Arrays

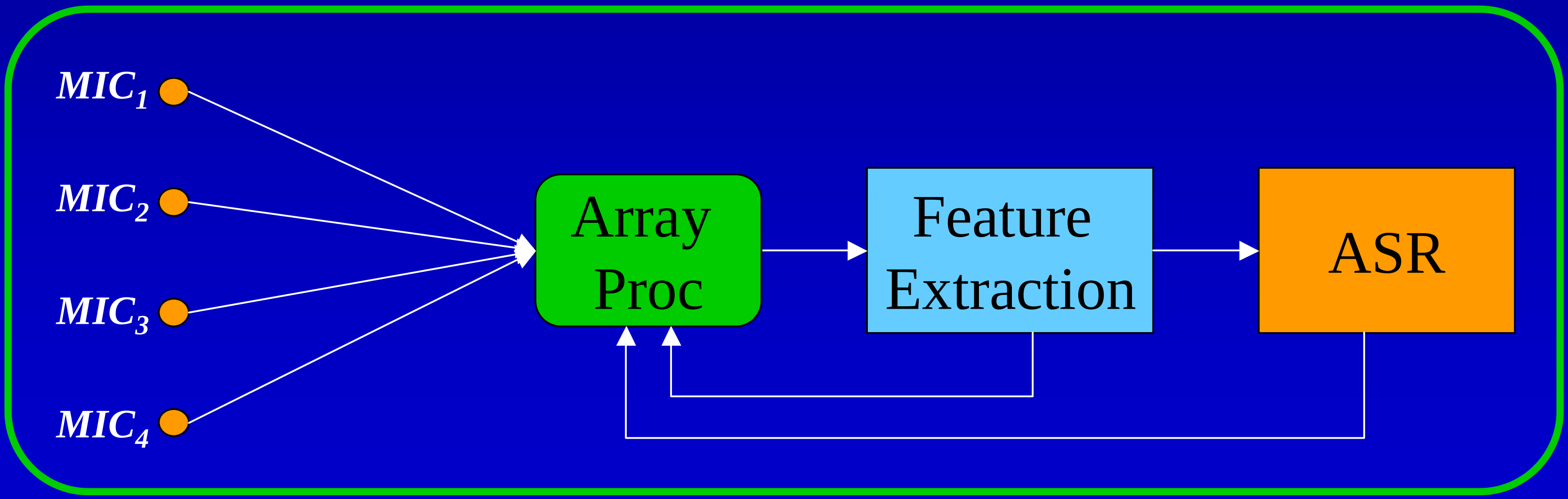
- Recognition with microphone arrays has been performed by “gluing” the two systems together.
- We believe this is not the ideal approach.
  - Systems have different objectives.
  - Each system does not exploit information present in the other.



A new approach

- Consider array processor and speech recognizer to be components of a single interconnected system which allows information to pass in both directions.

- Develop an array processing scheme specifically targeted at improved speech recognition performance without regard to conventional array processing objective criteria.



ASR-based Array Processing

- The simplest beamforming technique (delay and sum) simply averages the signals together:

$$y[n] = \frac{1}{N}\sum_{i=1}^{N} x_i[n - T_i]$$

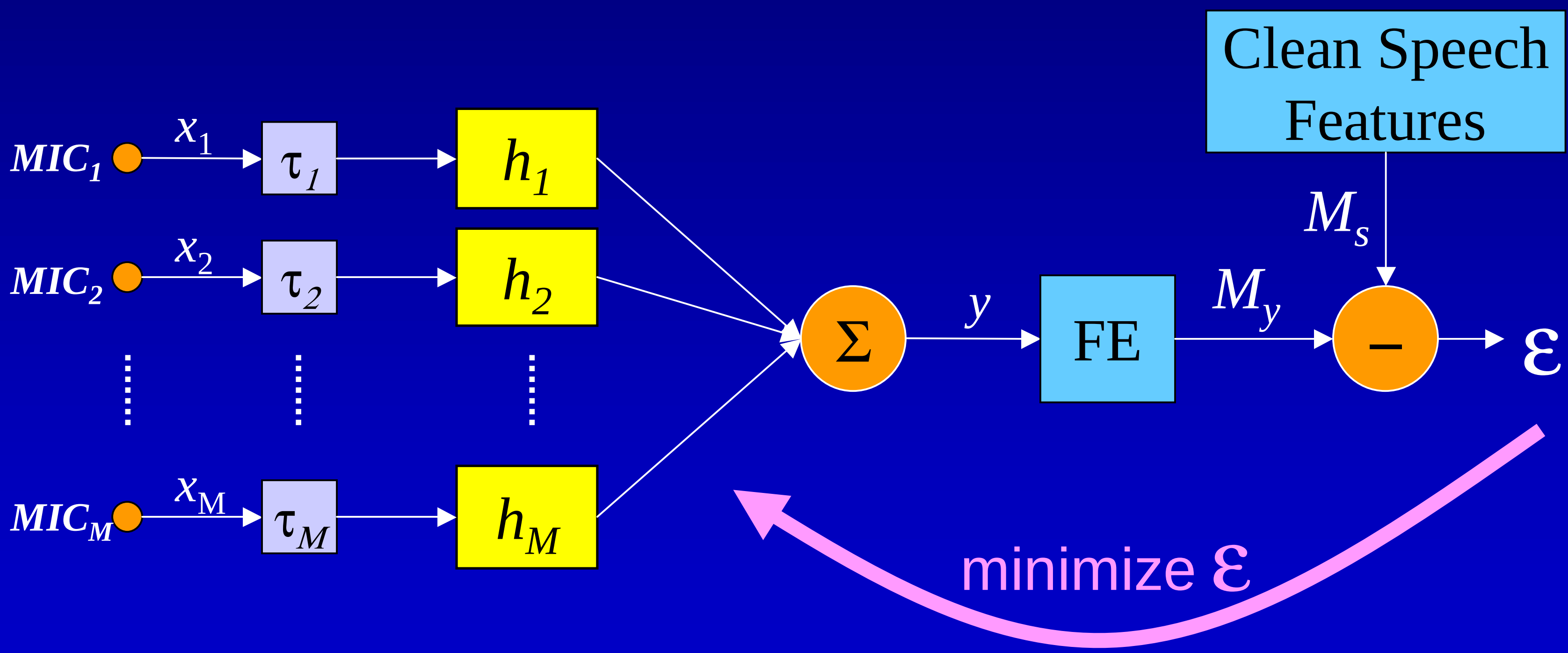
- Others weight or filter the signals before combining:

$$y[n] = \sum_i \alpha_i x_i[n - T_i] \qquad y[n] = \sum_i h_i[n] \otimes x_i[n - T_i]$$

- How do we choose the weights or filter coefficients to improve speech recognition performance?

What criterion do we want?

- Want an objective function that uses parameters *directly related to recognition*



An Objective Function for ASR

- Define Q as the SSE of the log Mel spectra of clean speech s and noisy speech y

$$Q = \sum_f \sum_l \left(M_y[f,l] - M_s[f,l]\right)^2$$

where y is the output of a filter-and-sum microphone array and $M[f, l]$ is the l^{th} log Mel spectral value in frame f.

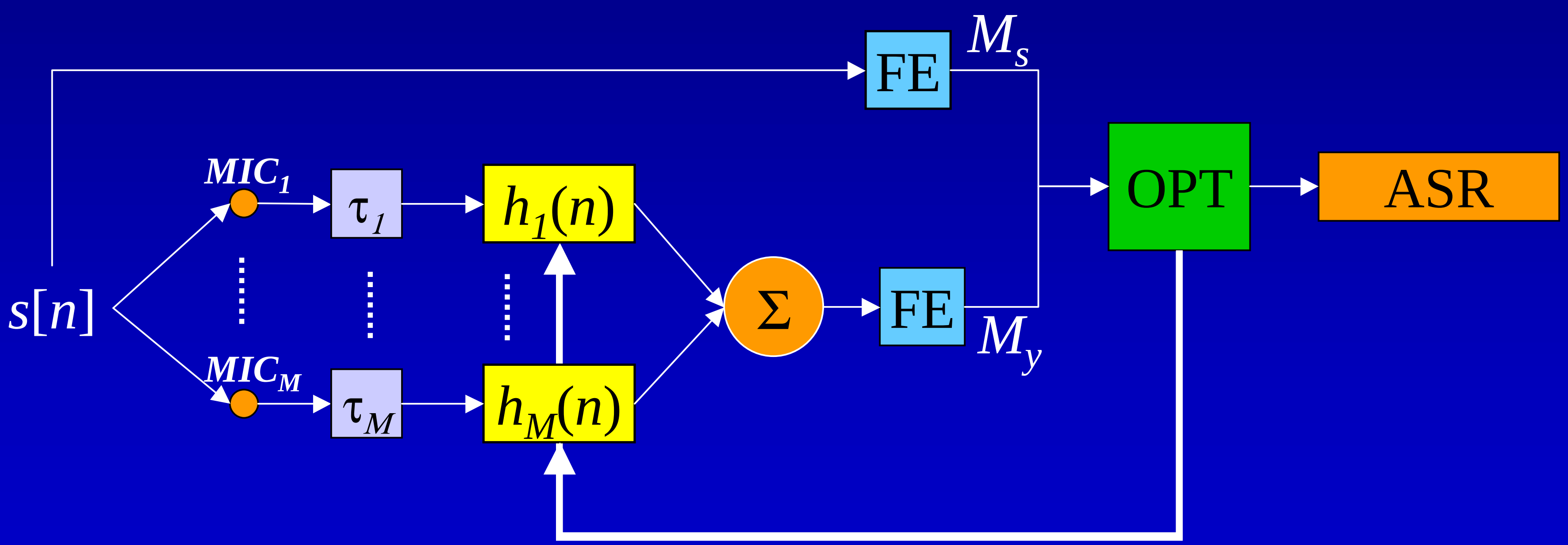
- $M_y[f, l]$ is a function of the signals captured by the array and the filter parameters associated with each microphone.

Calibration of Microphone Arrays for ASR

- Calibration of Filter-and-Sum Microphone Array:
  - Have a user speak an utterance with known transcription.
    - With or without close-talking microphone
  - Derive optimal set of filters.
    - Minimize the objective function with respect to the filter coefficients.
    - Since objective function is non-linear, use iterative gradient-based methods.
  - Apply to all future speech.

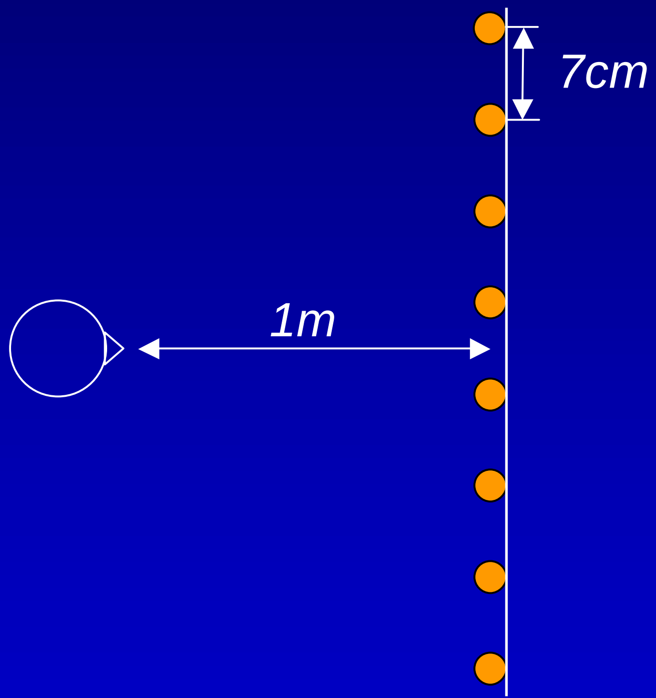
Calibration Using Close-talking Recording

- Given the close-talking mic recording for the calibration utterance, derive an “optimal” filter for each channel to improve recognition



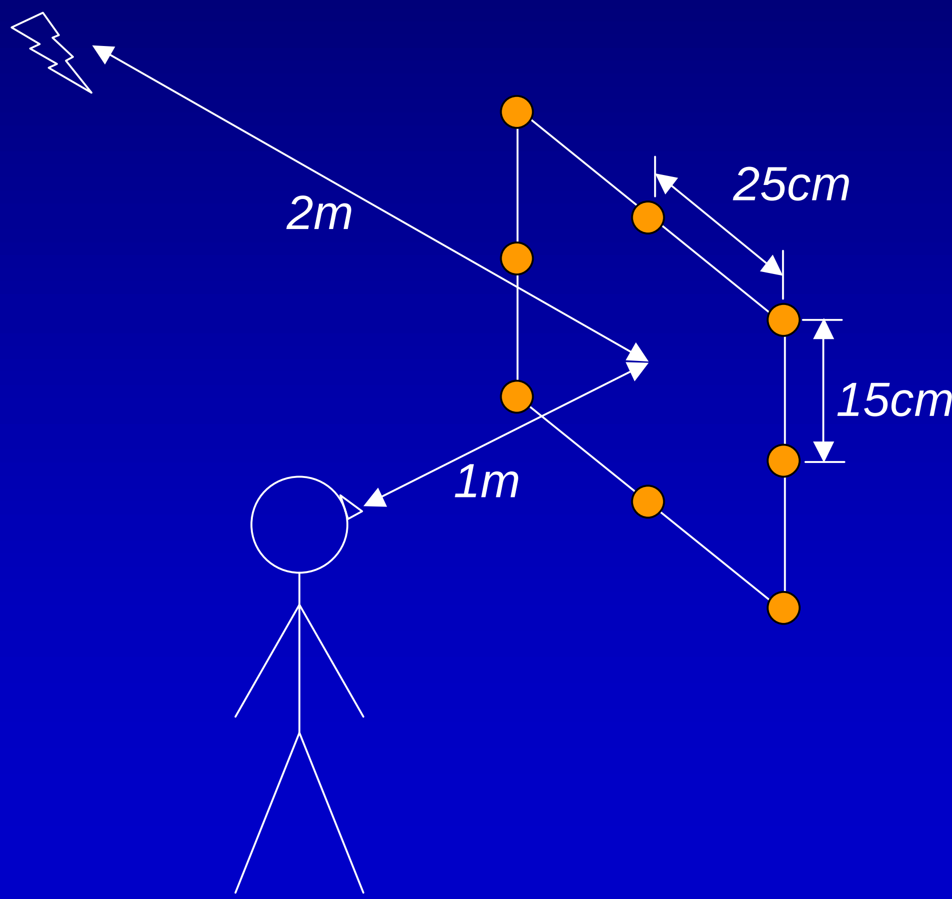
Multi-microphone data sets

- TMS
  - Recorded in the CMU Speech Lab
    - Approx. 5m x 5m x 3m
    - Noise from computer fans, blowers ,etc.
  - Isolated letters and digits, keywords
  - 10 speakers * 14 utterances = 140 utterances
  - Each utterance has close-talking mic control waveform



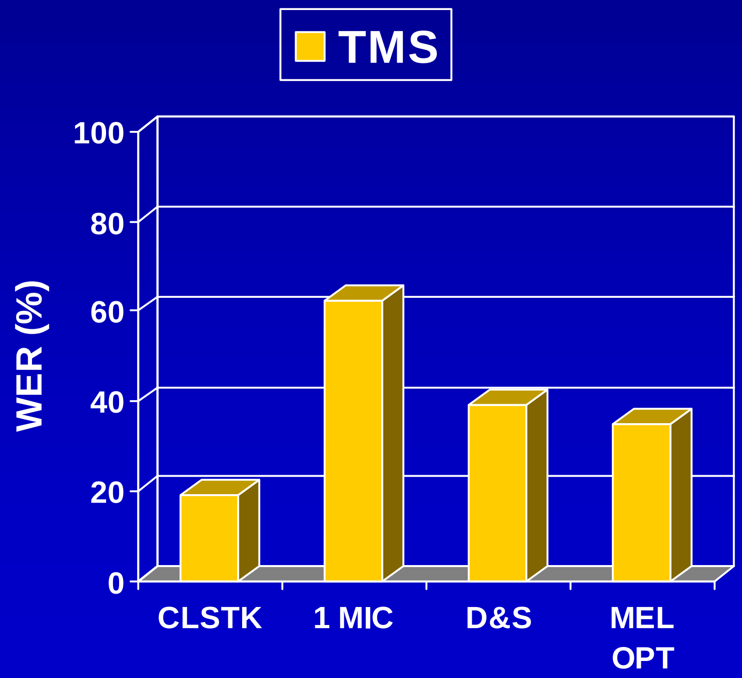
Multi-microphone data sets (2)

- WSJ + off-axis noise source
  - Room simulation created using the image method
    - 5m x 4m x 3m
    - 200ms reverberation time
    - WGN source @ 5dB SNR
  - WSJ test set
    - 5K word vocabulary
    - 10 speakers * 65 utterances = 650 utterances
  - Original recordings used as close-talking control waveforms



Results

- TMS data set, WSJ0 + WGN point source simulation
  - Constructed 50 point filters from a single calibration utterance
  - Applied filters to all test utterances

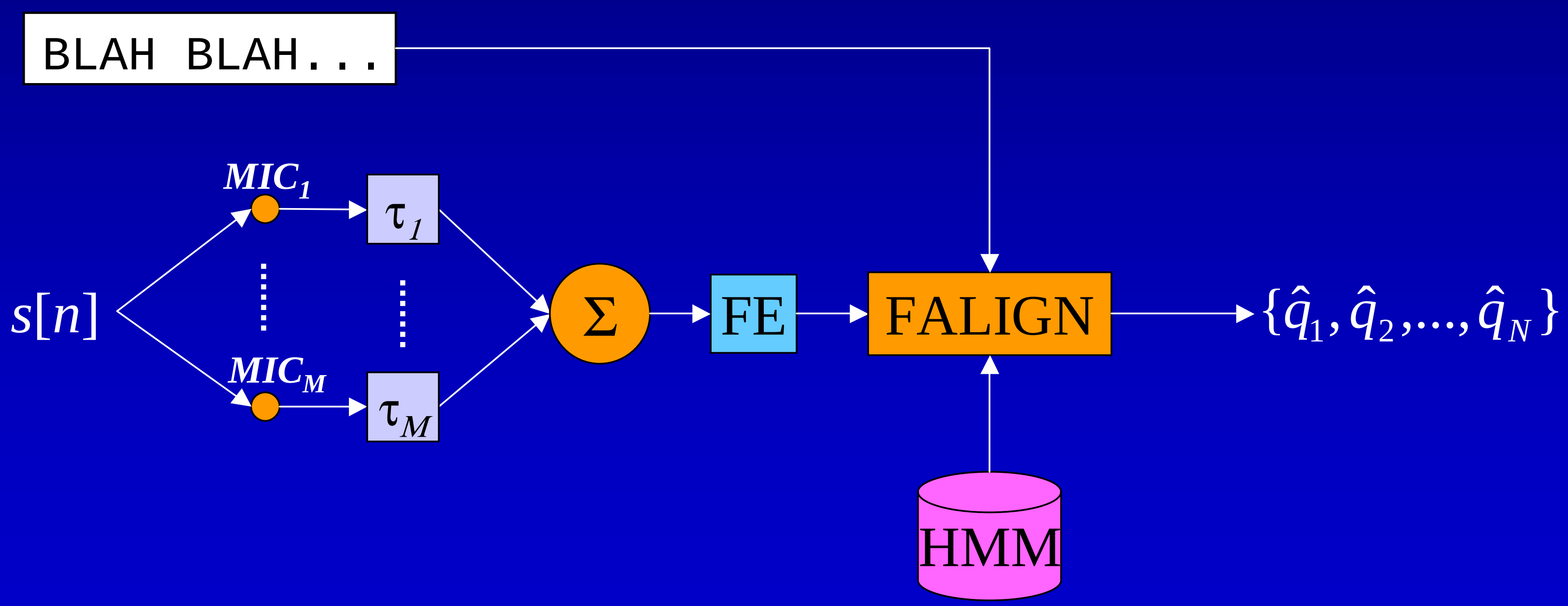


Calibration without Close-talking Microphone

- Obtain initial waveform estimate using conventional array processing technique (*e.g.* delay and sum).
- Use transcription and the recognizer to estimate the sequence of target clean log Mel spectra.
- Optimize filter parameters as before.

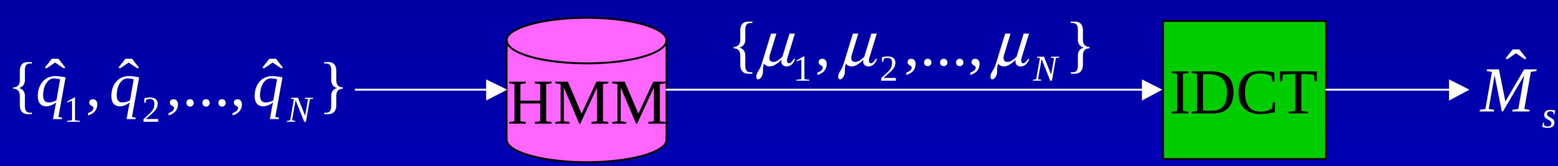
Calibration w/o Close-talking Microphone (2)

- Force align the delay-and-sum waveform to the known transcription to generate an estimated HMM state sequence.



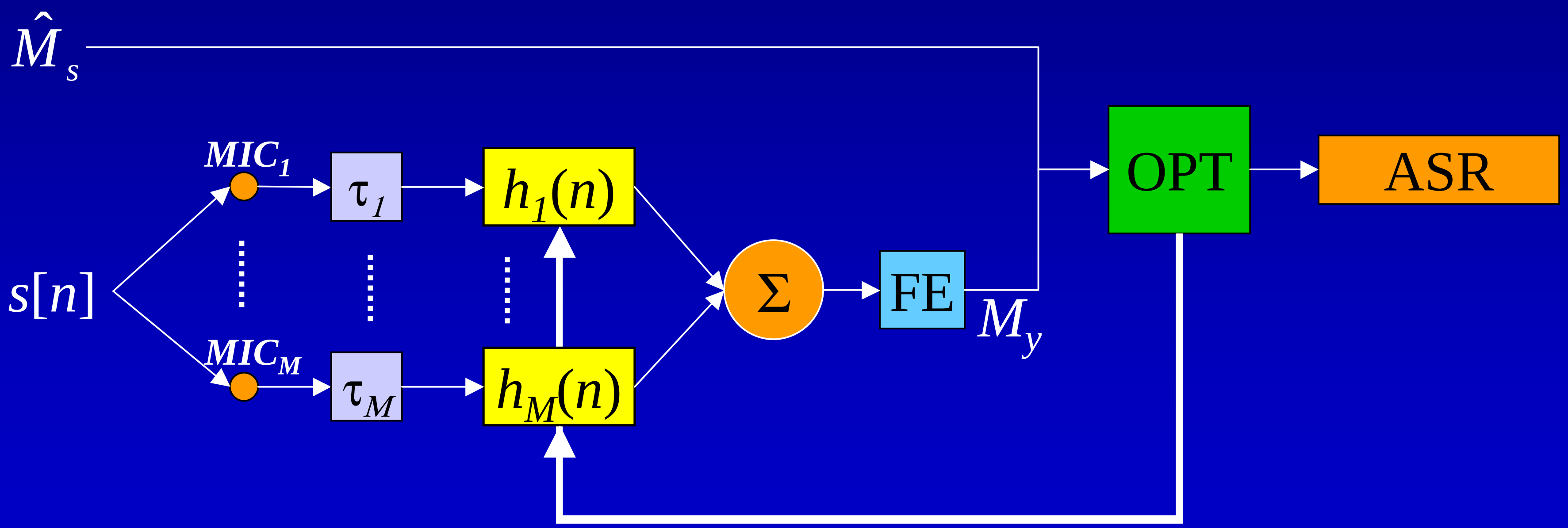
Calibration w/o Close-talking Microphone (3)

- Extract the means from the single Gaussian HMMs of the estimated state sequence.
  - Since the models have been trained from clean speech, use these means as the target clean speech feature vectors.



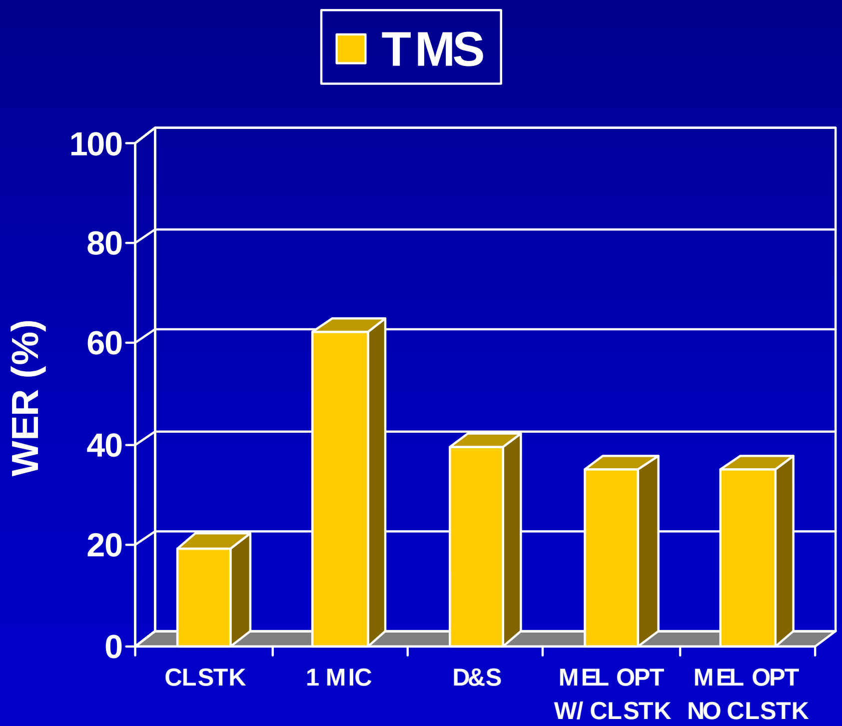
Calibration w/o Close-talking Microphone (4)

- Use estimated clean speech feature vectors to optimize filters as before.



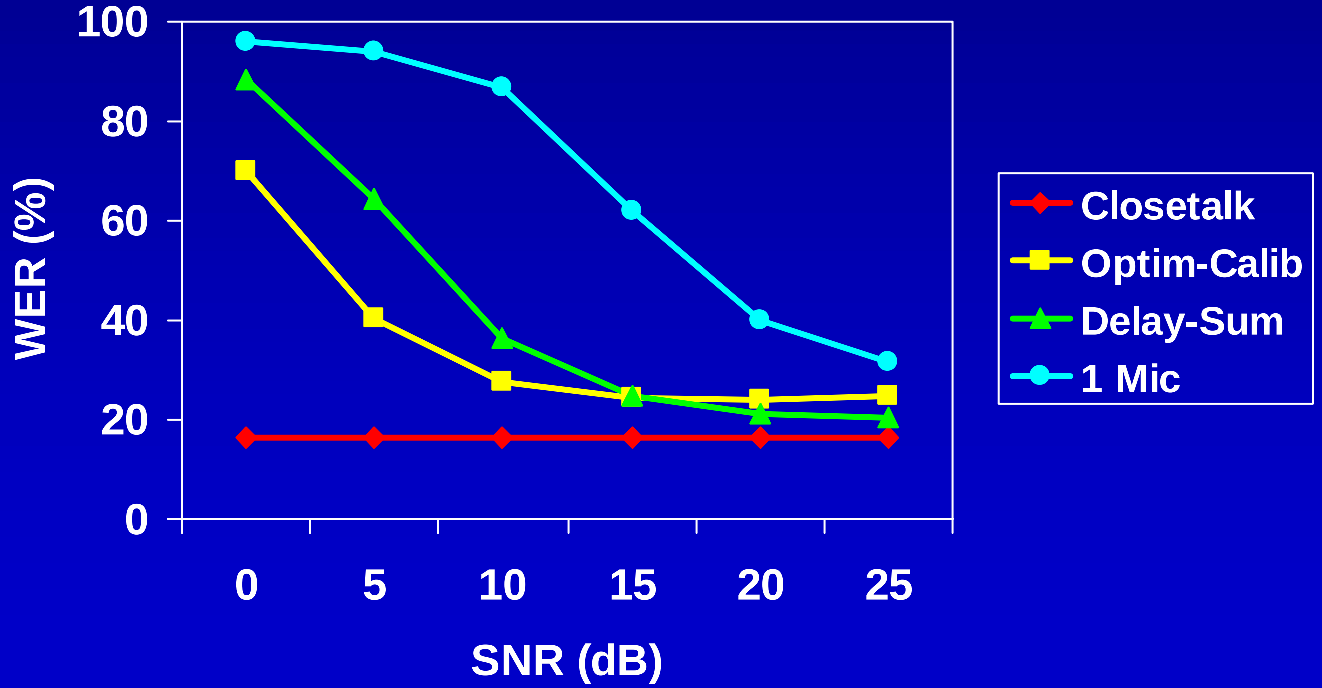
Results

- TMS data set, WSJ0 + WGN point source simulation
  - Constructed 50 point filters from calibration utterance
  - Applied filters to all utterances



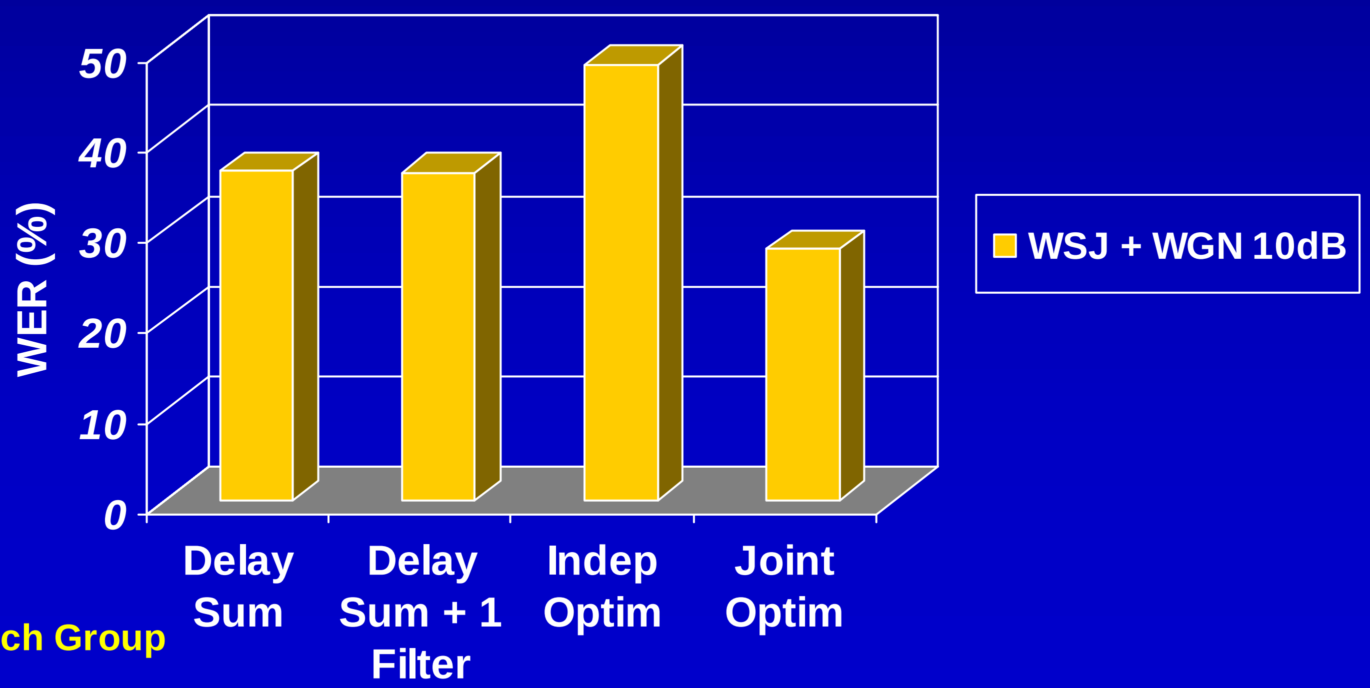
Results (2)

- WER vs. SNR for WSJ + WGN
  - Constructed 50 point filters from calibration utterance using transcription only
  - Applied filters to all utterances



Is Joint Filter Estimation Necessary?

- We compared 4 cases:
  - Delay and Sum
  - Optimize 1 filter for Delay and Sum Output
  - Optimize Microphone Array Filters Independently
  - Optimize Microphone Array Filters Jointly



Summary and Future Work

- We have presented a new microphone array calibration scheme specifically designed for speech recognition.
- We have achieved improvements in WER of up to 37% over conventional Delay and Sum processing using this method.
- Successfully fedback information from the recognizer all the way back to the waveform level.
- We plan to investigate the following extensions to the algorithm: reverberation compensation, unsupervised optimization, filter adaptation.