Robust Speaker Recognition using Microphone Arrays

- Authors:
  - Iain McCowan*, Jason Pelecanos and Sridha Sridharan
- Affiliation:
  - The Speech Research Laboratory, RCSAVT
  - Queensland University of Technology
  - *Now with IDIAP, Switzerland

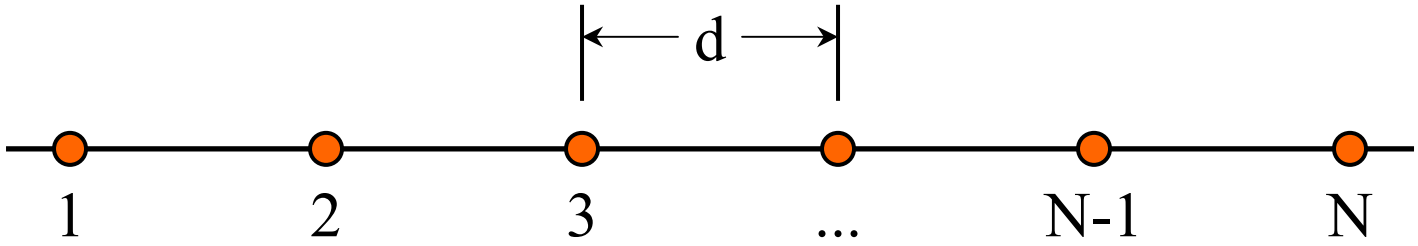
Overview

- Introduction
- Microphone Arrays and Beamforming
- Speaker Recognition
- Speaker Recognition and Microphone Arrays
- Experimental Evaluation
- Conclusions

Introduction

- Research is required into improving the robustness of speech and speaker recognition in noisy environments
- Hands-free operation is a desirable feature
- Microphone arrays have the potential to satisfy these criteria
- There is limited research examining the use of microphone array speech enhancement for improving speaker verification

Microphone Arrays and Beamforming



- A microphone array is effectively a sampled spatial aperture
- The response of a microphone array is directional in nature
- The *directivity pattern* is a plot of the array response as a function of direction

Microphone Arrays and Beamforming

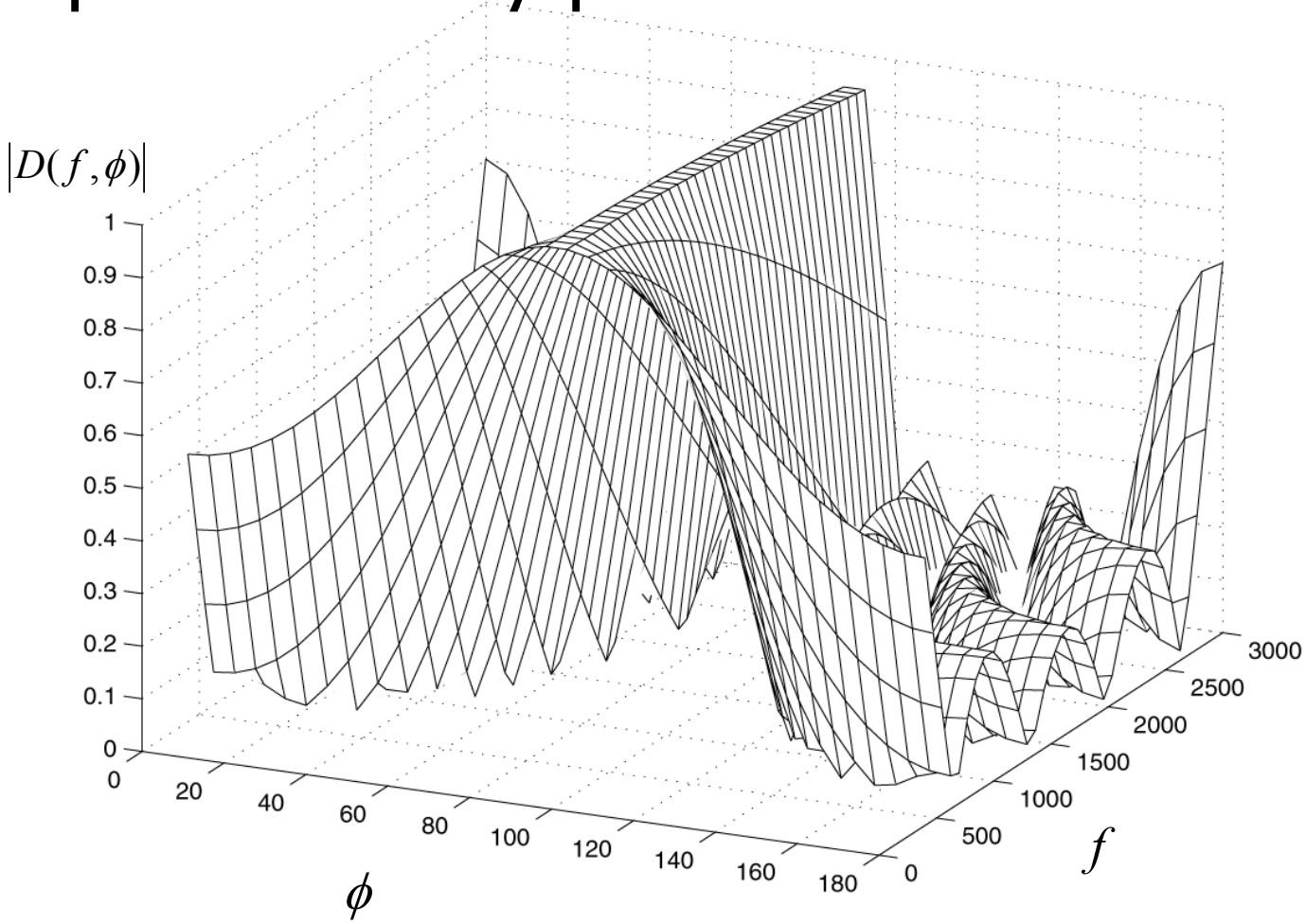
- The directivity of a linear, equi-spaced microphone array is

$$D(f,\phi) = \sum_{n=1}^{N} w_n(f)\, e^{j\frac{2\pi f}{c}(n-1)d\cos\phi}$$

- Depends on
  - N : number of sensors
  - d : inter-element spacing
  - f : frequency of interest

Microphone Arrays and Beamforming

- Sample directivity pattern



Microphone Arrays and Beamforming

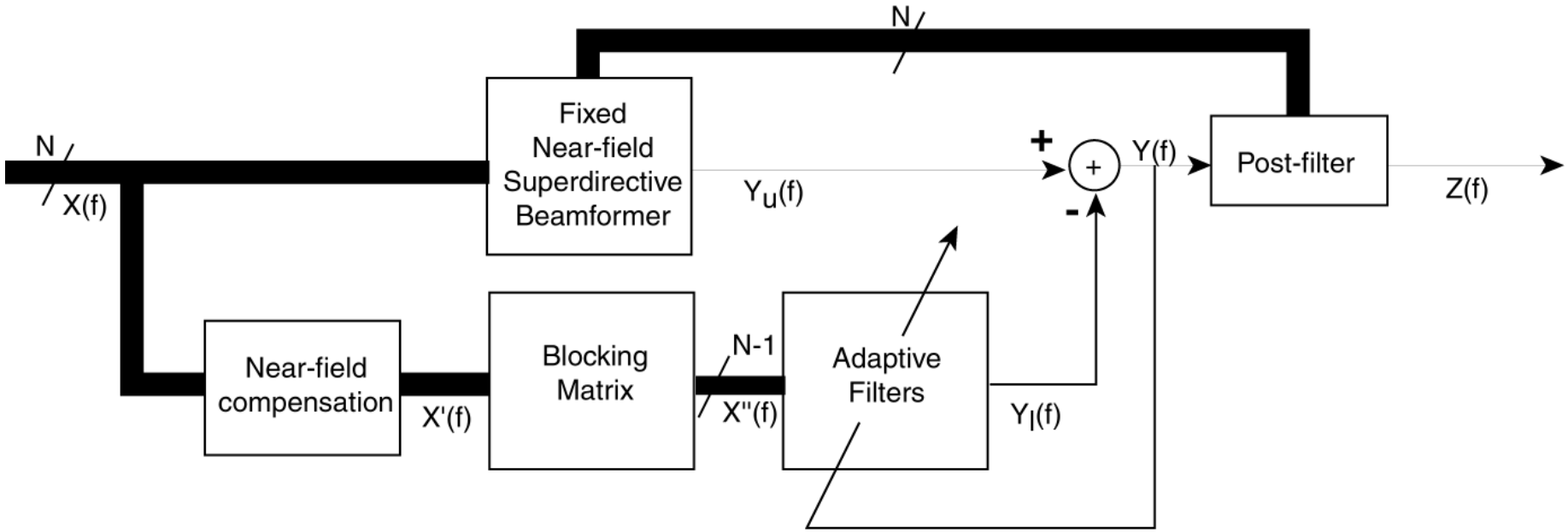
- Beamforming algorithms determine the complex sensor weights $w_n(f)$ to implement a desired *steering* and *shaping* of the directivity pattern.
  - Thus we can enhance signals based purely on directional information
- Beamforming techniques can be fixed or adaptive

Microphone Arrays and Beamforming

- Near-field adaptive beamforming (NFAB)
  - Adaptive beamforming technique
  - Optimised directivity for near-field source
  - Good low frequency performance
    - Important for speech
  - Previous research has shown good performance for speech enhancement compared to other beamforming techniques (eg delay-sum)

Microphone Arrays and Beamforming

- Block diagram



Microphone Arrays and Beamforming

- Sample sound files
  - 'speech-like' noise from NOISEX
    - Noisy input 
    - Beamformer (NFAB) output 
  - 'Cocktail party' recording
    - Noisy input 
    - Beamformer (NFAB) output 

Speaker Recognition

- The speaker verification system uses models established from the adapted Gaussian Mixture Model (GMM) approach as proposed by Reynolds.
- A target model is formed by adapting the parameters of a universal speaker model to the target speaker's speech
- The GMM is the fundamental building block of the system

Speaker Recognition

- A GMM approximates the Probability Density Function (PDF) of a multi-variate feature by the addition of K Gaussian component distributions
- Thus, the probability density of an observation is

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{K} w_i g(\vec{x}, \vec{\mu}_i, \Sigma_i)$$

with

$$g(\vec{x}, \vec{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu}_i)' (\Sigma_i)^{-1} (\vec{x} - \vec{\mu}_i) \right)$$

Speaker Recognition

- The speech is parameterised into frames of 24 coefficients
- There are 12 Mel-Frequency Cepstral Coefficients derived from 20 filterbanks (spaced over the 0-4000Hz band) with their corresponding 12 delta-coefficients
- The GMMs use 512 components to create an estimate of the PDF

Speaker Recognition

- The speaker hypothesis testing is made according to the likelihood ratio statistic

$$\frac{p(X \mid \lambda)}{p(X \mid \overline{\lambda})} \; <> \; Operating\ Threshold$$

- Speaker scoring is performed by using the expected frame-based log-likelihood ratio score

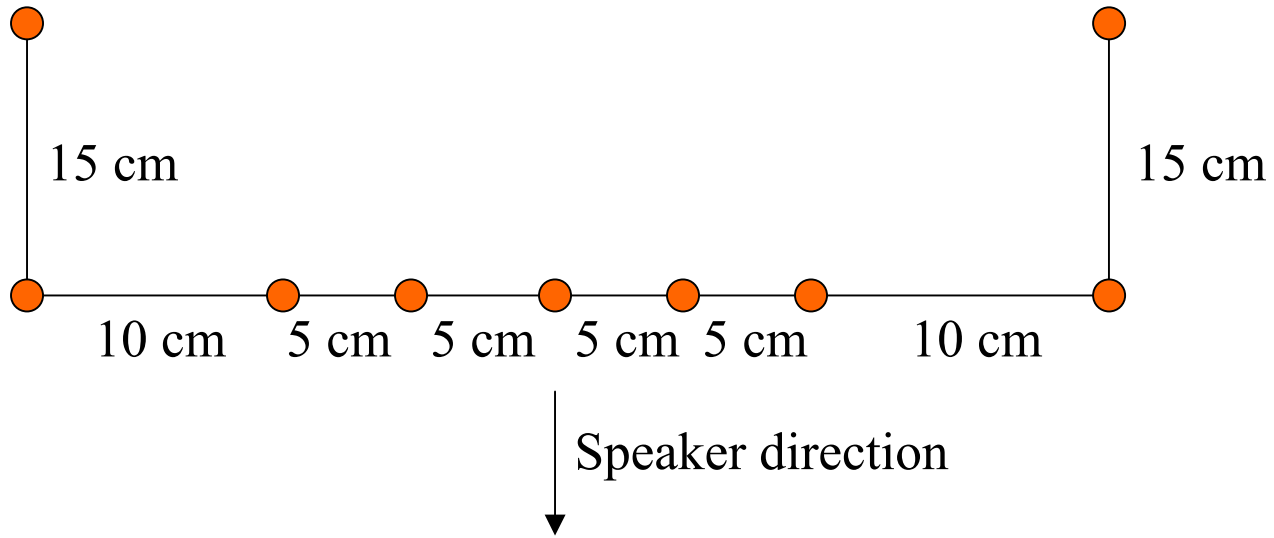
$$\Lambda = \frac{1}{T}\sum_{t=1}^{T}\left(\log p(\vec{x}_t \mid \lambda_{tar}) - \log p(\vec{x}_t \mid \lambda_{ubm})\right)$$

Speaker Recognition and Microphone Arrays

- Potential to improve...
  - robustness to noisy environments
  - ease of use (hands-free acquisition)
- Research to date has been limited
  - Several issues require investigation
    - Use of sophisticated beamforming techniques
    - Realistic data for experiments
    - Use of state of the art speaker recognition systems
    - To date, no known experiments for speaker verification

Experimental Evaluation

- NFAB beamforming technique
- 9 element microphone array
- Speaker is 70 cm from centre microphone



Experimental Evaluation

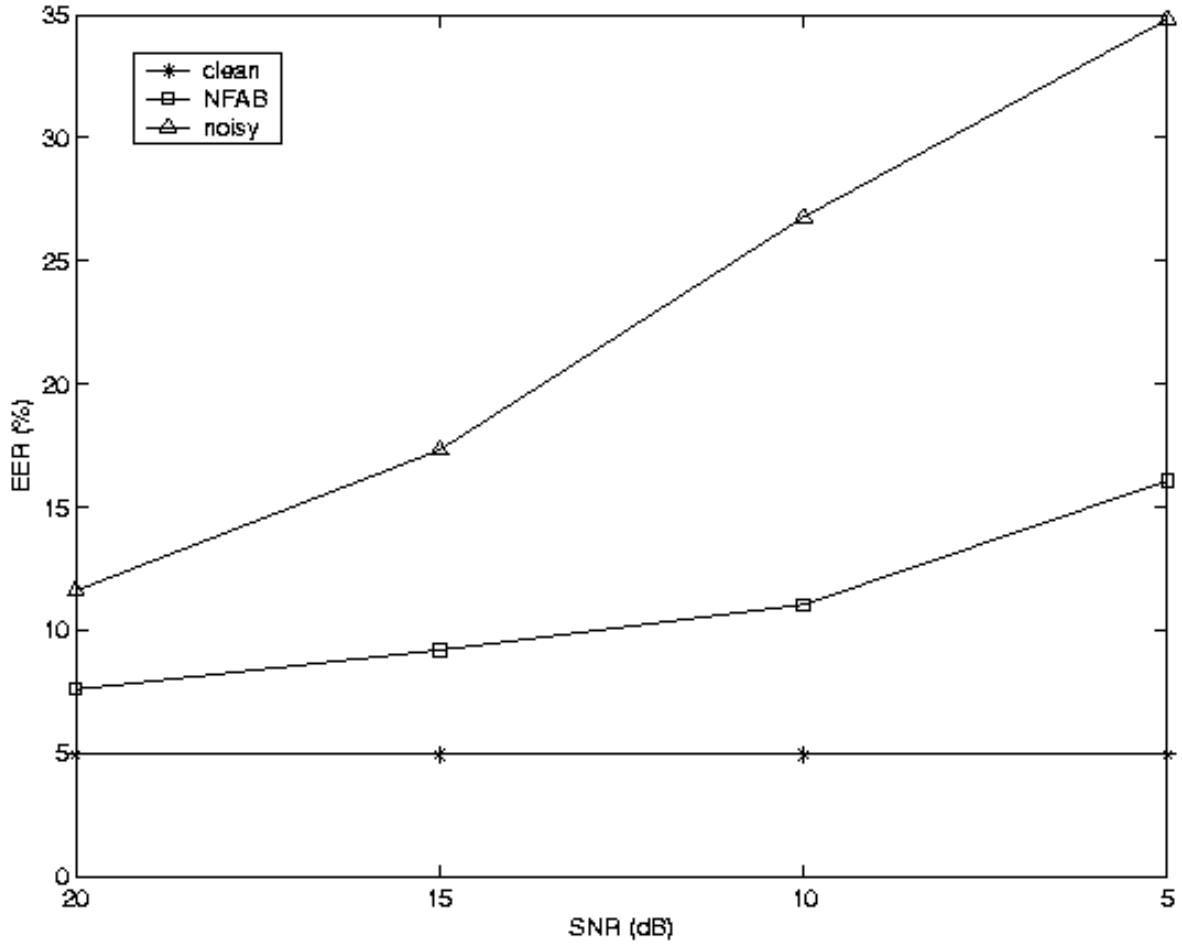
- Noise configuration
  - Ambient noise
    - Multi-channel recording in office
    - Contains computer noise, and some background speech and noise from the air-conditioning unit
  - Localised noise
    - 'Speechlike' noise (NOISEX database) located at 2.7m from the centre microphone at an angle of 56^{o}

Experimental Evaluation

- Male speakers from TIMIT database
- 3 sets of experiments
  - Varying levels of ambient noise
  - Varying levels of localised noise
  - Both ambient and localised noise
- Results compare EER for:
  - Clean input to centre microphone, noisy input at centre microphone and beamformer output

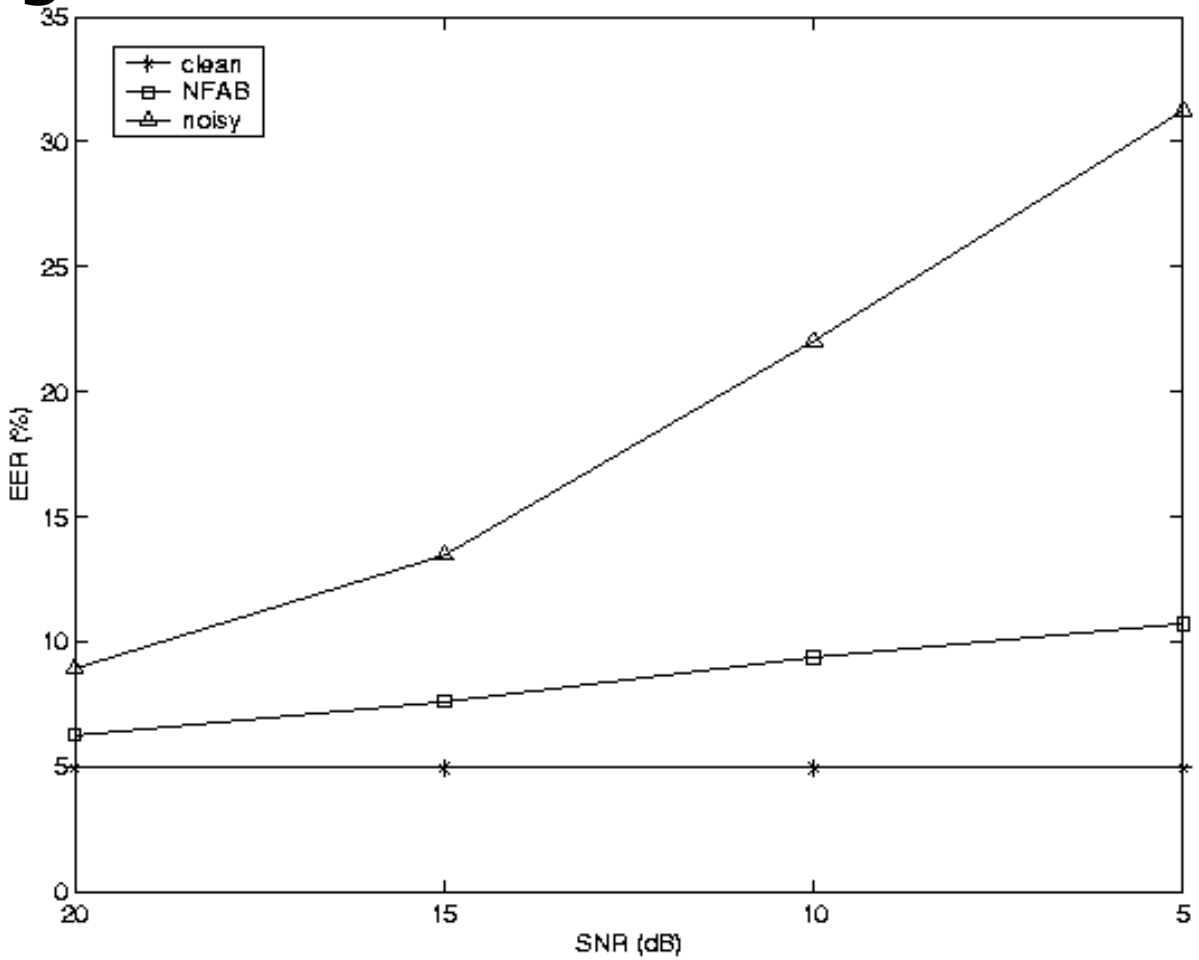
Experimental Evaluation

- Varying ambient noise



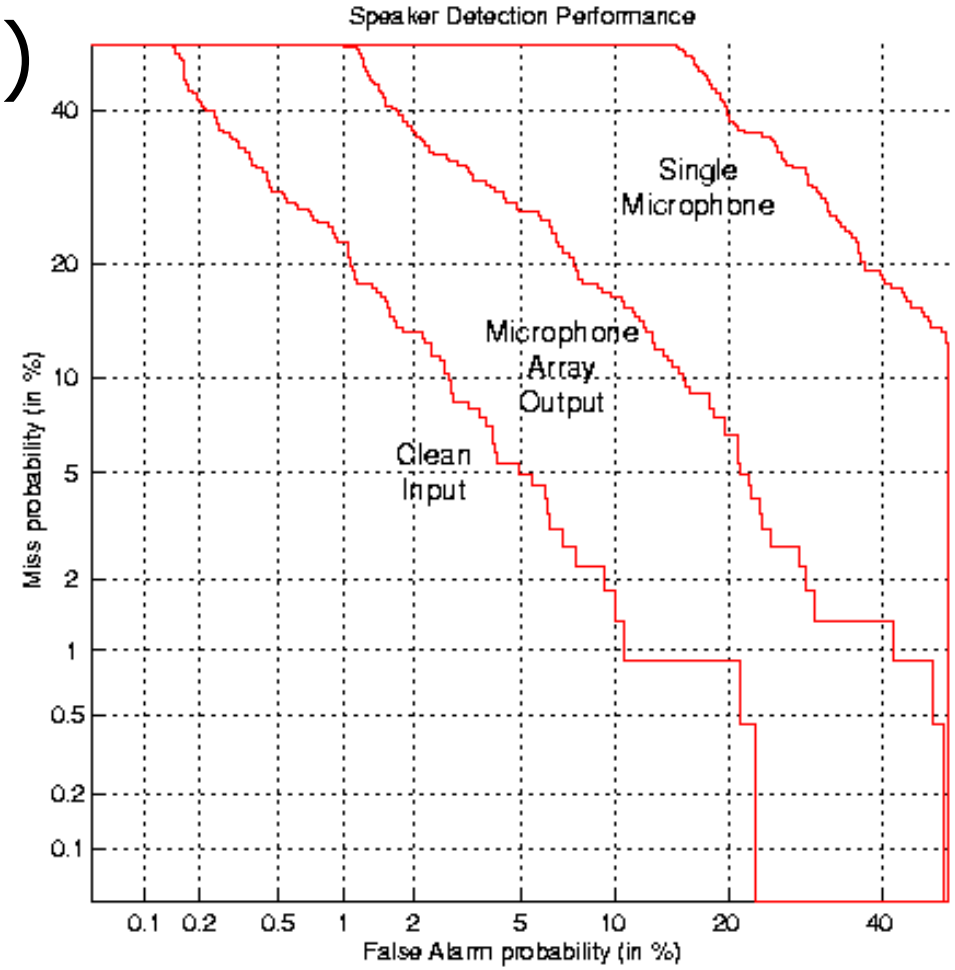
Experimental Evaluation

- Varying localised noise



Experimental Evaluation

- DET curve - ambient and localised noise (SNR=7dB)



Conclusions

- The research has...
  - reviewed the current state of research into microphone array speaker recognition
  - addressed some important issues by evaluating a sophisticated beamforming technique with a state-of-the-art speaker verification system
  - demonstrated the benefits of the array in
    - improving performance in high noise conditions
    - providing hands-free signal acquisition
- In summary
  - Arrays can improve the usability and robustness of practical speaker recognition systems