Distant Speech Recognition Using Microphone Arrays

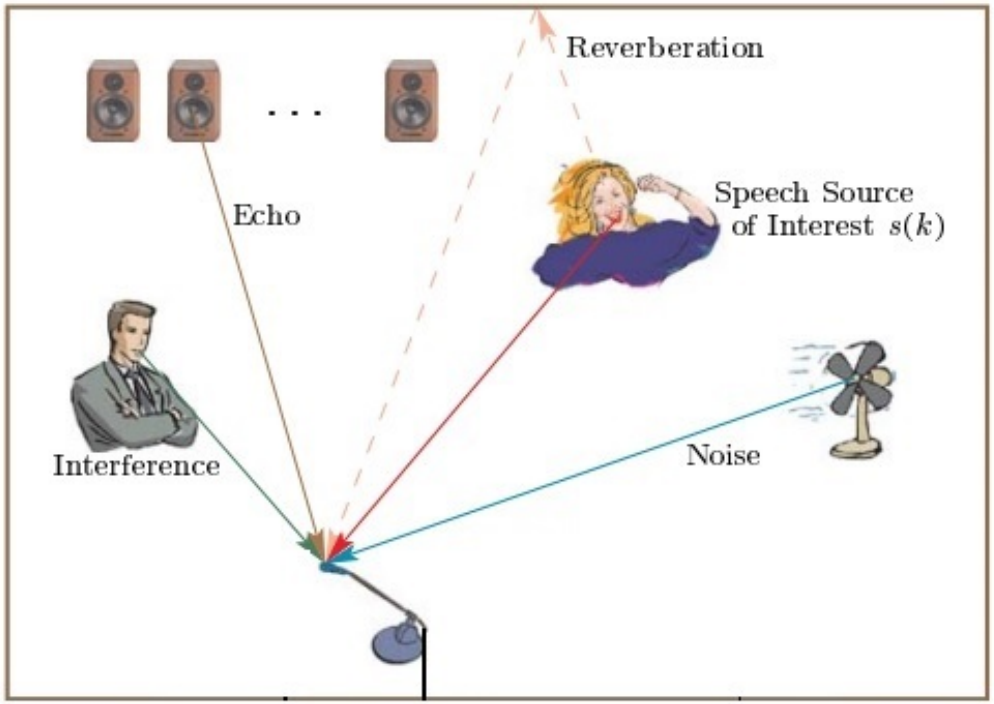
George Jose (153070011)
Guide : Prof. Preeti Rao

Indian Institute of Technology, Bombay

December 21, 2016

Outline

Far Field Speech Recognition : Challenges



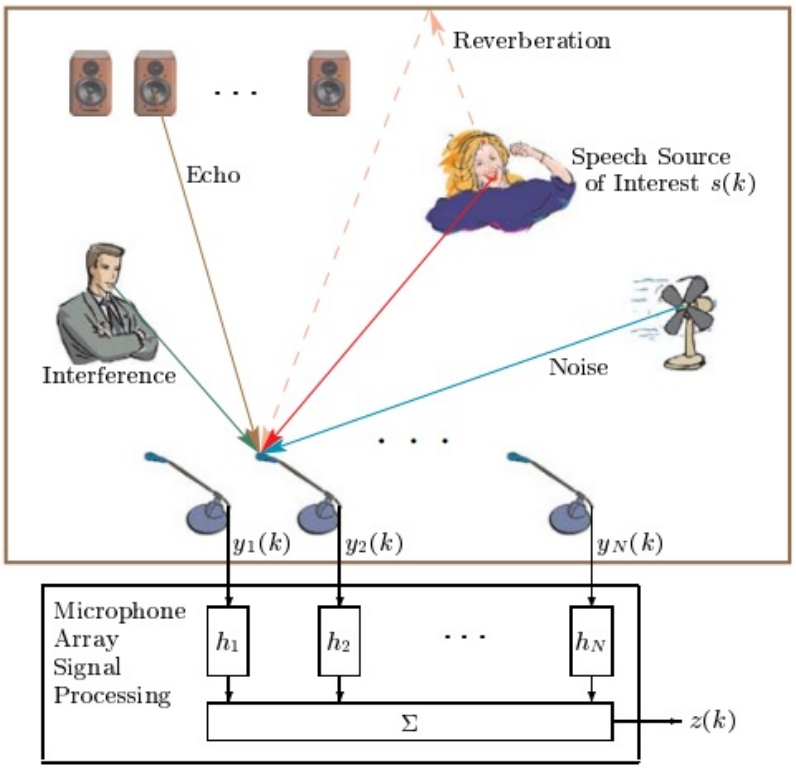
(Seltzer, 2003)

Major Challenges:

1. Noise
2. Reverberation
3. Echo
4. Interference Speaker

Solution

Exploit the separation in spatial domain



(Seltzer, 2003)

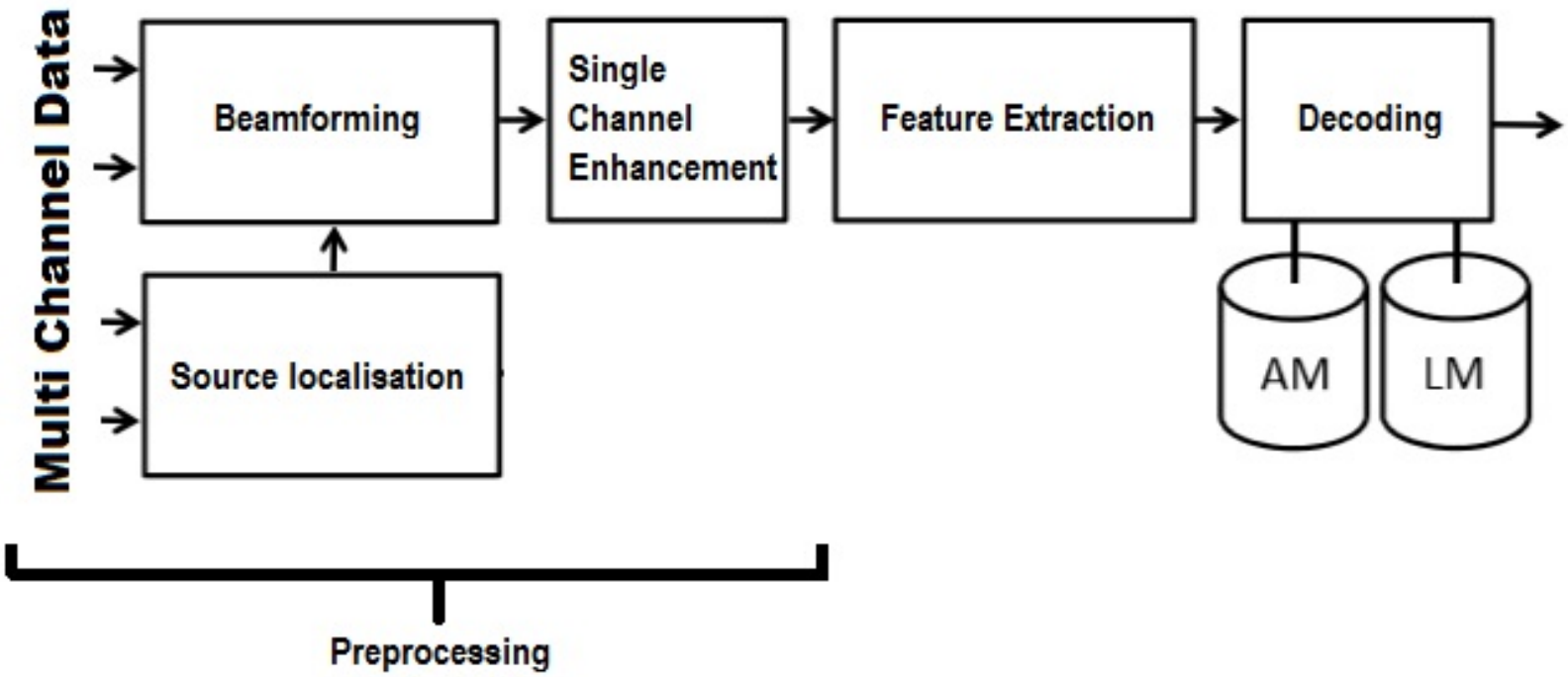
How ?

Use multiple microphones

Why ?

Signals from each source arrive with different delays at each microphone

System Overview



Two Stage Process

1. Source Localisation : Locate the source direction
2. Beamforming : Apply spatial filtering to attenuate the noise

Source Localisation

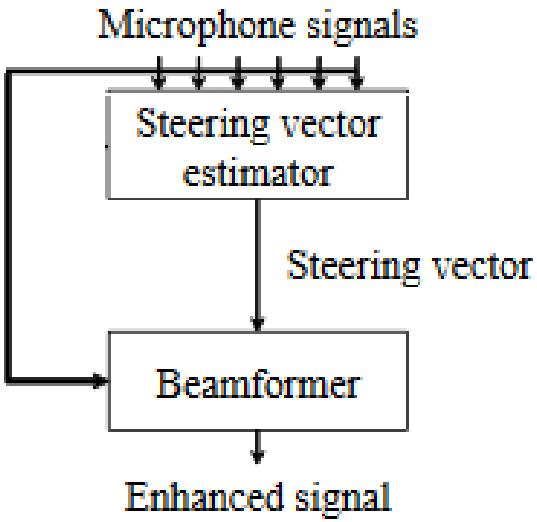


Figure: Front End Model (Higuchi et al., 2016)

- Steering Vector represented by :

$$\mathbf{d(f)} = [e^{-j\omega\tau_{1i}}\ e^{-j\omega\tau_{2i}}\ \ldots\ e^{-j\omega\tau_{Ni}}]^T$$

- τ_{ji} - Time Delay of Arrival (TDOA)
  $\tau_{ji} = \tau_j$ - τ_i
- Encodes source position information

TDOA Estimation

- Simplest Method : To find the cross correlation
  $R_{y_1y_2}(\tau) = \mathrm{E}[y_1(n)y_2(n-\tau)]$
- τ which maximizes $R_{y_1y_2}(\tau)$ provides an estimate of delay
- In practice, cross correlation computed by:
  $R_{y_1y_2}(\tau) = \mathrm{IDFT}\ \{G_{y_1y_2}(f)\}$
- In presence of reverberation it produces arbitrary peaks

Generalised Cross Correlation Phase Transform (GCC PHAT)

- Discards the amplitude and keeps only the phase
- Whitening of signal produces a sharp peak
  $R_{y_1y_2}(\tau) = \mathrm{IDFT}\ \{\frac{G_{y_1y_2}(f)}{|G_{y_1y_2}(f)|}\}$

Eigen Decomposition

- Estimate the covariance matrices of noisy speech and noise.
- Noisy speech covariance matrix estimate :

$$R_{sn}(k) = \frac{1}{L}\sum_{l=1}^{L}\mathbf{Y}(k,l)\mathbf{Y}^{H}(k,l)$$

- Estimate noise covariance matrix :

$$R_{n}(k) = \frac{1}{L}\sum_{l=1}^{L}\mathbf{V}(k,l)\mathbf{V}^{H}(k,l)$$

- Speech covariance matrix estimate : $R_s(k) = R_{sn}(k) - R_n(k)$
- Perform eigen decomposition of speech covariance matrix
- Find the eigen vector associated with maximum eigen value

Beamforming

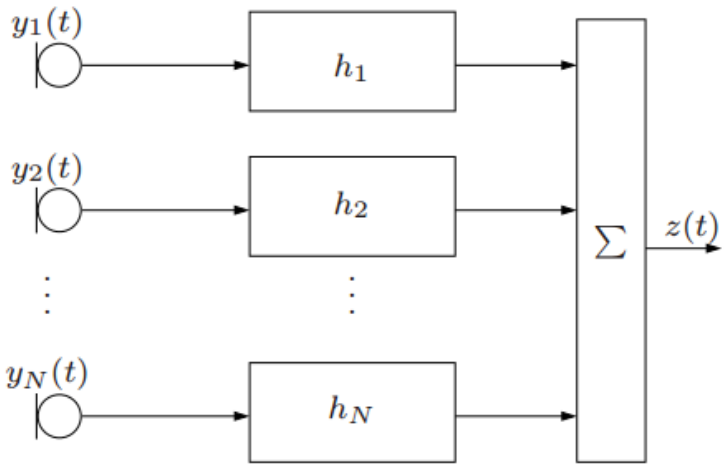


Figure: Beamformer Model (Cohen et al., 2009)

Signal at jth microphone :

$$y_j(n) = s_i(n - \tau_{ji}) + v_j(n)$$

In STFT domain,

$$Y_j(k,l) = S_i(k,l)e^{\frac{-j2\pi k\tau_{ji}}{N}} + V_j(k,l)$$

In vector notation,

$$\mathbf{Y}(k,l) = \mathbf{d(k)}S_i(k,l) + \mathbf{V}(k,l)$$

Objective

Perform spatial filtering by steering the response of the microphone array towards the speaker direction

Acoustic Beamforming

Beamformer Output :

$$\begin{aligned} Z(k,l) &= \mathbf{h}^H(k)\mathbf{Y}(k,l) \\ &= \mathbf{h}^H(k)(\mathbf{d}(k)S_i(k,l)) + \mathbf{V}(k,l)) \\ &= \mathbf{h}^H(k)(\mathbf{X}(k,l) + \mathbf{V}(k,l)) \end{aligned}$$

Beamformer should not distort the speech signal

$$\mathbf{h}^H(k)\mathbf{d}(k) = 1$$

Output power :

$$\begin{aligned} P = E[Z(k,l)Z^H(k,l)] &= E[\mathbf{h}^H(k)\mathbf{Y}(k,l)\mathbf{h}^H(k)] \\ &= \mathbf{h}^H(k)R_x(k)\mathbf{h}(k) + \mathbf{h}^H(k)R_v(k)\mathbf{h}(k) \end{aligned}$$

Noise power at the output should be minimum

MVDR Beamformer

Targets:

- Minimize the noise power at the output of the beamformer
- Constraint : Signal should not be distorted

Optimization Problem

$$\mathbf{h}(f) = \underset{\mathbf{h}(k)}{arg\,min}\, \mathbf{h}^H(k)\boldsymbol{R}_v(k)\mathbf{h}(k) \quad \text{subject to } \mathbf{h}^H(k)\mathbf{d}(k) = 1$$

Solving the optimization problem gives :

Minimum Variance Distortionless Response (MVDR) Beamformer

$$\mathbf{h}_{MVDR}(f) = \frac{\boldsymbol{R}_v^{-1}(k)\mathbf{d}(k)}{\mathbf{d}^H(k)\boldsymbol{R}_v^{-1}(k)\mathbf{d}(k)}$$

Without Source Localisation

1) Alternate MVDR formulation (Souden, Benesty, & Affes, 2010)

$$\mathbf{h}_{MVDR}(f) = \frac{R_v^{-1}(k)R_s(k)}{trace\{R_v^{-1}(k)R_s(k)\}}\mathbf{e}_{ref}$$

- Minimizes the noise at the output

2) Generalized Eigen Value beamforming (Jahn Heymann & Haeb-Umbach, 2016b)

$$\mathbf{h}_{GEV}(k) = \arg\max_{\mathbf{h(k)}} \frac{\mathbf{h}^H(k)R_s(k)\mathbf{h}(k)}{\mathbf{h}^H(k)R_v(k)\mathbf{h}(k)}$$

- Maximizes the SNR at the output

ASR Experiments

- ASR Model: GMM-HMM triphone model
- Model Parameters : 500 senones x 8 Gaussians
- Train Data: TIDigits adults training set (clean speech)
- Test Data : TIDigits adults test set simulated for :
  - SNR - 20dB
  - T_{60} - 750ms
  - Source Distance : 2m
  - Source Angle : 45^o

ASR Results

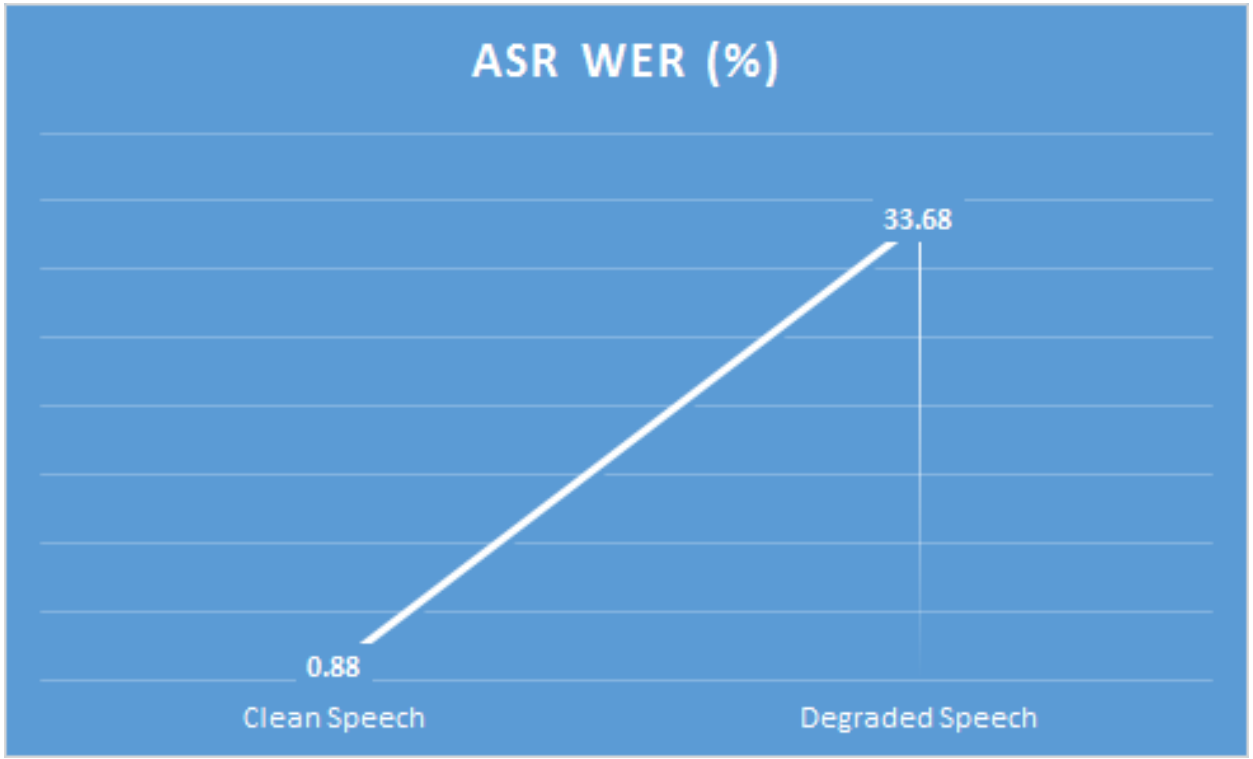


Figure: WER (%) of ASR model tested on noisy and reverberant speech with SNR of 20dB and $T_{60} = 750$ms for a source at an angle 45^o and 2m away from 8-channel circular array of radius 10cm

ASR Results

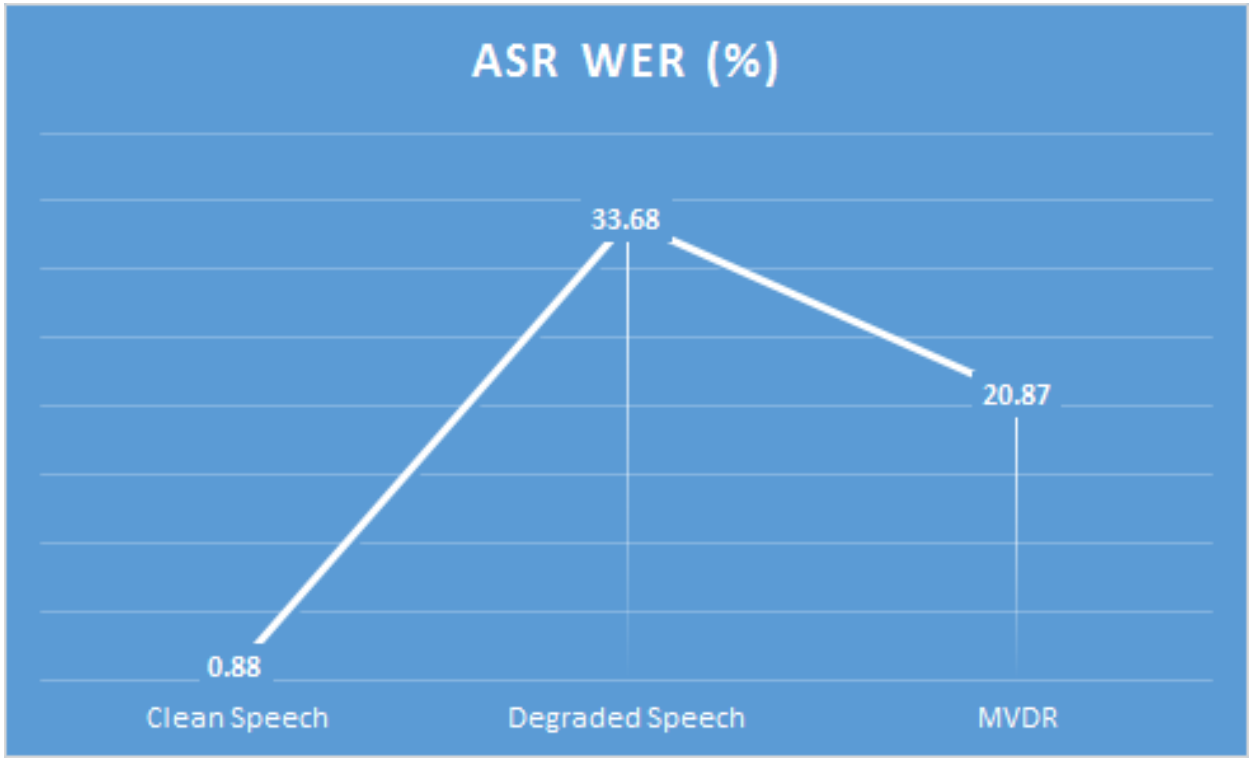


Figure: WER (%) of ASR model tested on noisy and reverberant speech with SNR of 20dB and $T_{60} = 750$ms for a source at an angle 45^o and 2m away from 8-channel circular array of radius 10cm

ASR Results

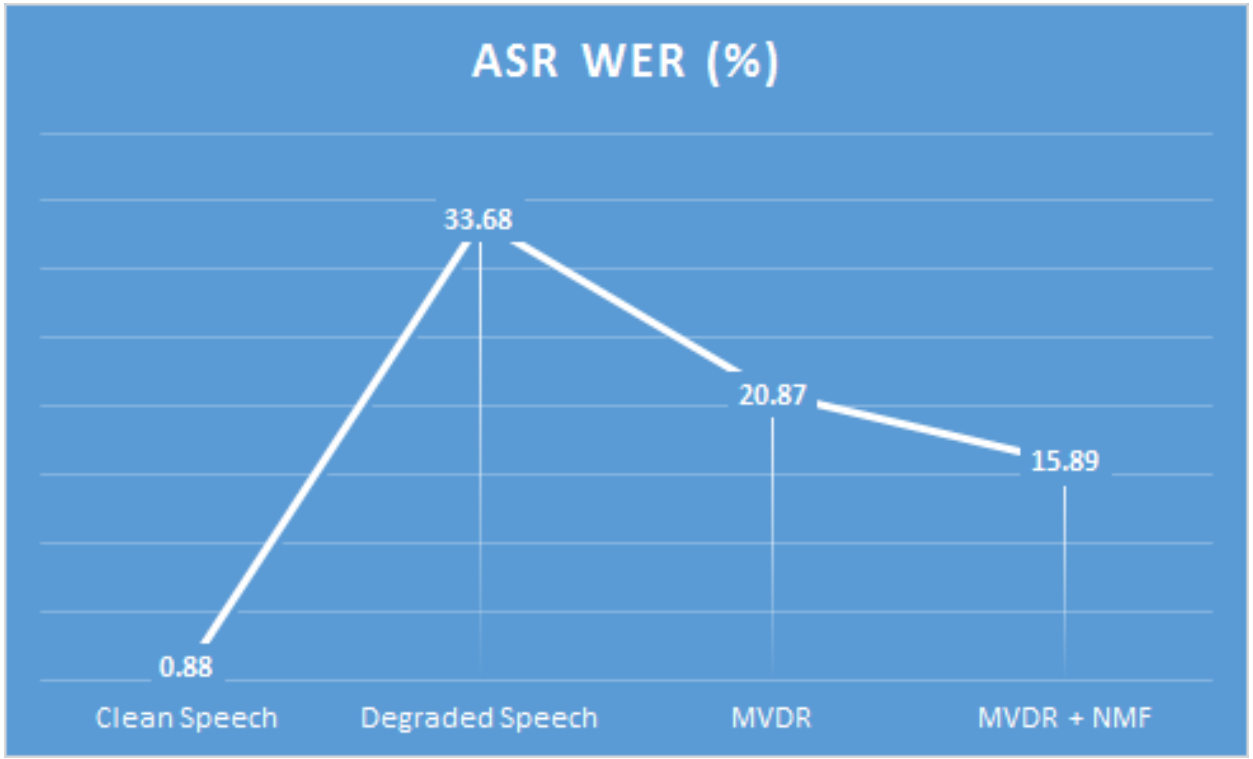


Figure: WER (%) of ASR model tested on noisy and reverberant speech with SNR of 20dB and $T_{60} = 750$ms for a source at an angle 45^o and 2m away from 8-channel circular array of radius 10cm

ASR Results

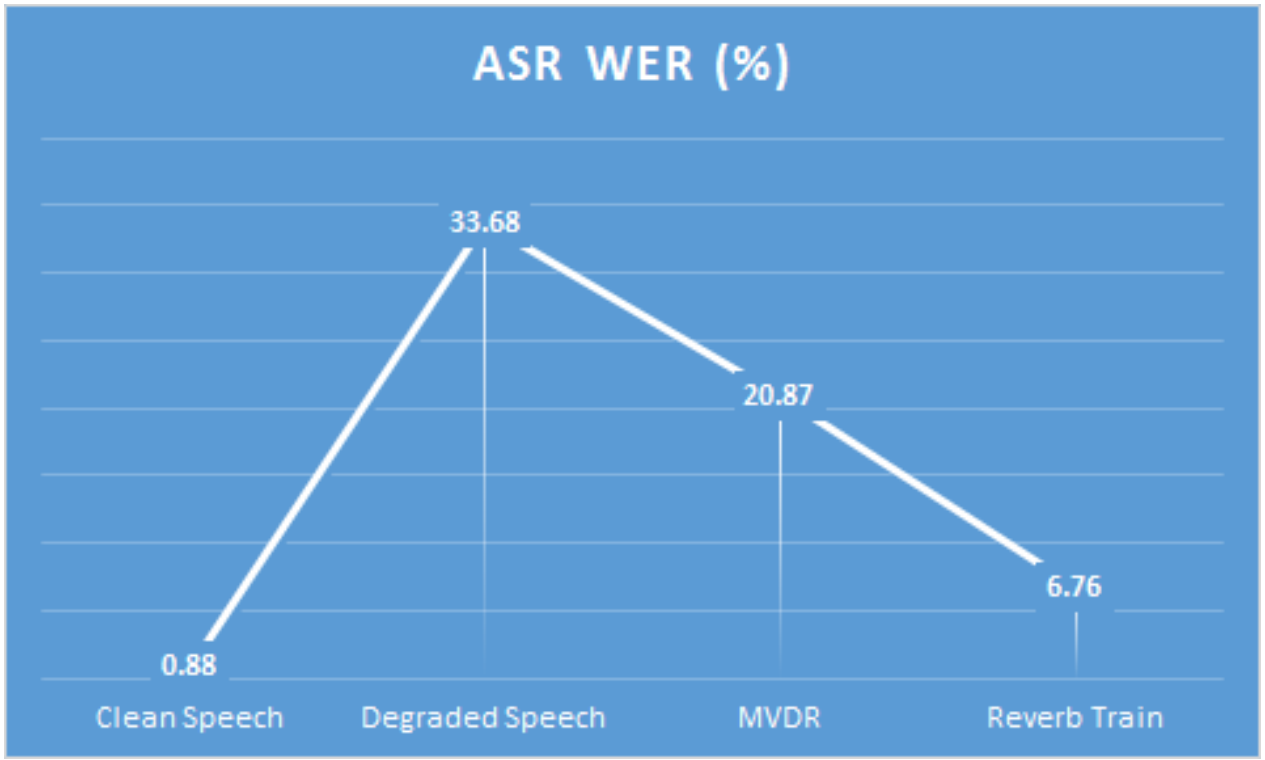


Figure: WER (%) of ASR model tested on noisy and reverberant speech with SNR of 20dB and $T_{60} = 750$ms for a source at an angle 45^o and 2m away from 8-channel circular array of radius 10cm

ASR Results

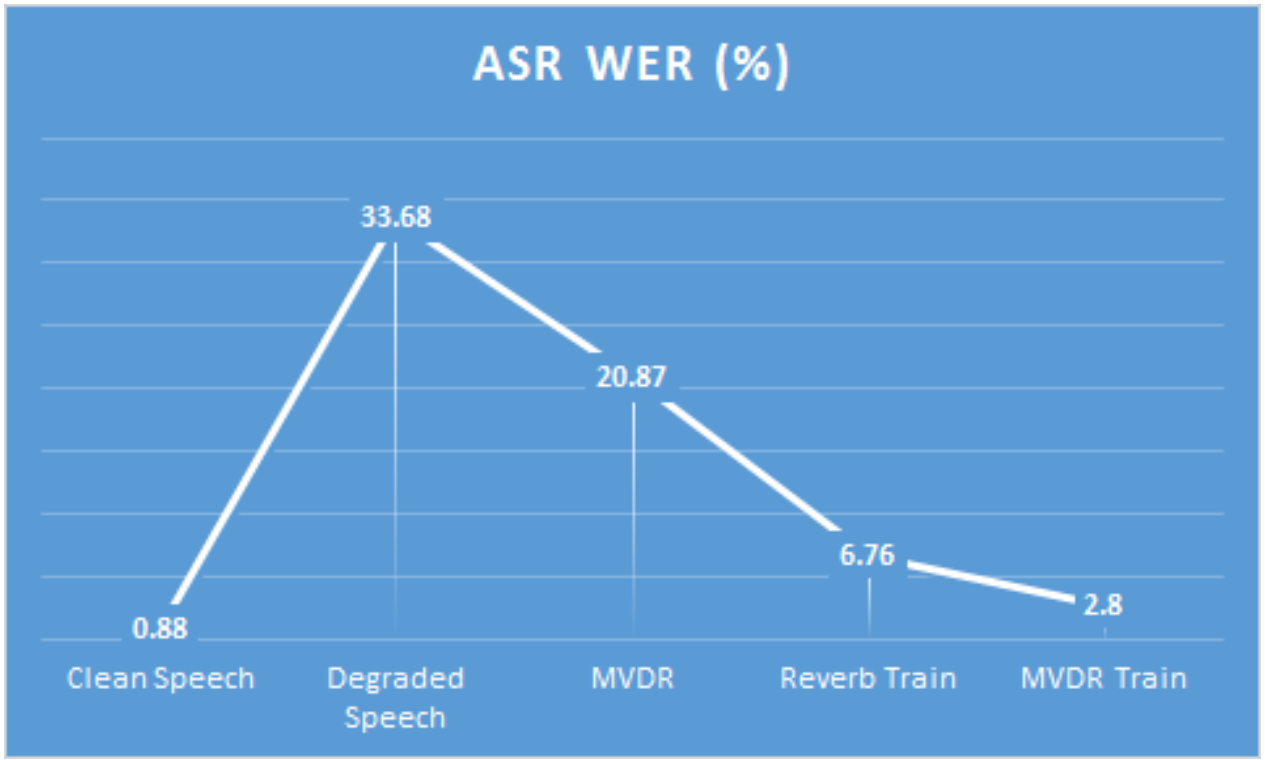


Figure: WER (%) of ASR model tested on noisy and reverberant speech with SNR of 20dB and $T_{60} = 750$ms for a source at an angle 45^o and 2m away from 8-channel circular array of radius 10cm

Observations

Two factors which affect the performance are :

1. An accurate estimate of steering vector

Method	**WER(%)**
GCC PHAT	20.87
Oracle	17.15

Table: Comparison of WER (%) for MVDR beamformer with steering vector estimated using GCC PHAT and the actual delays

2. An accurate estimate of noise covariance matrix

Experiment Details

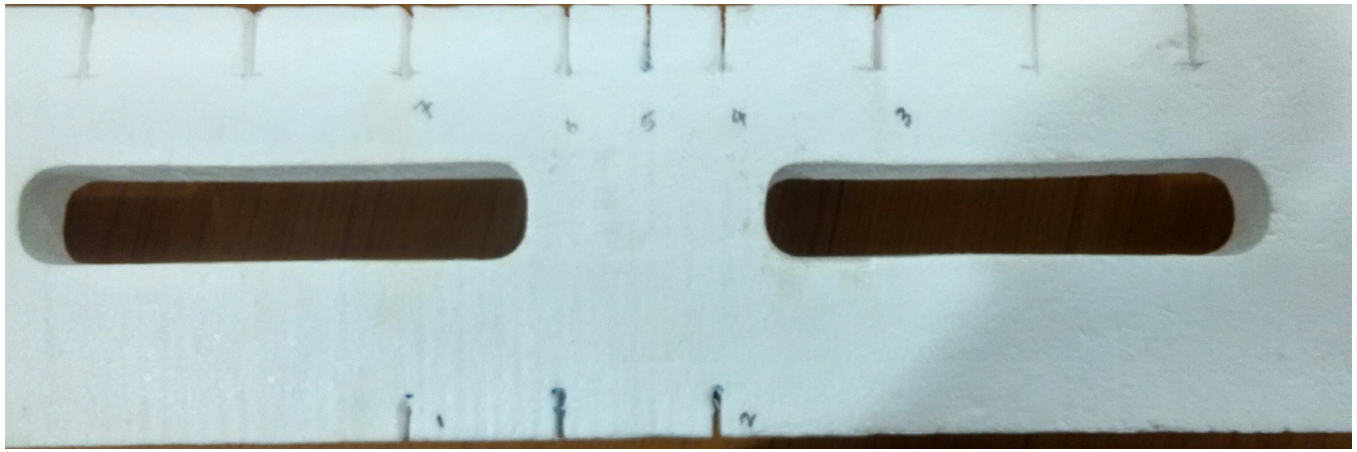


Figure: Array Microphone Setup

- One close talking microphone
- Seven distant microphones
- Non overlapping data from 15 different speakers
- 20 sentences from TIMIT database
- Two overlapping scenarios with 2 speakers

Spectral Mask Based Beamformers

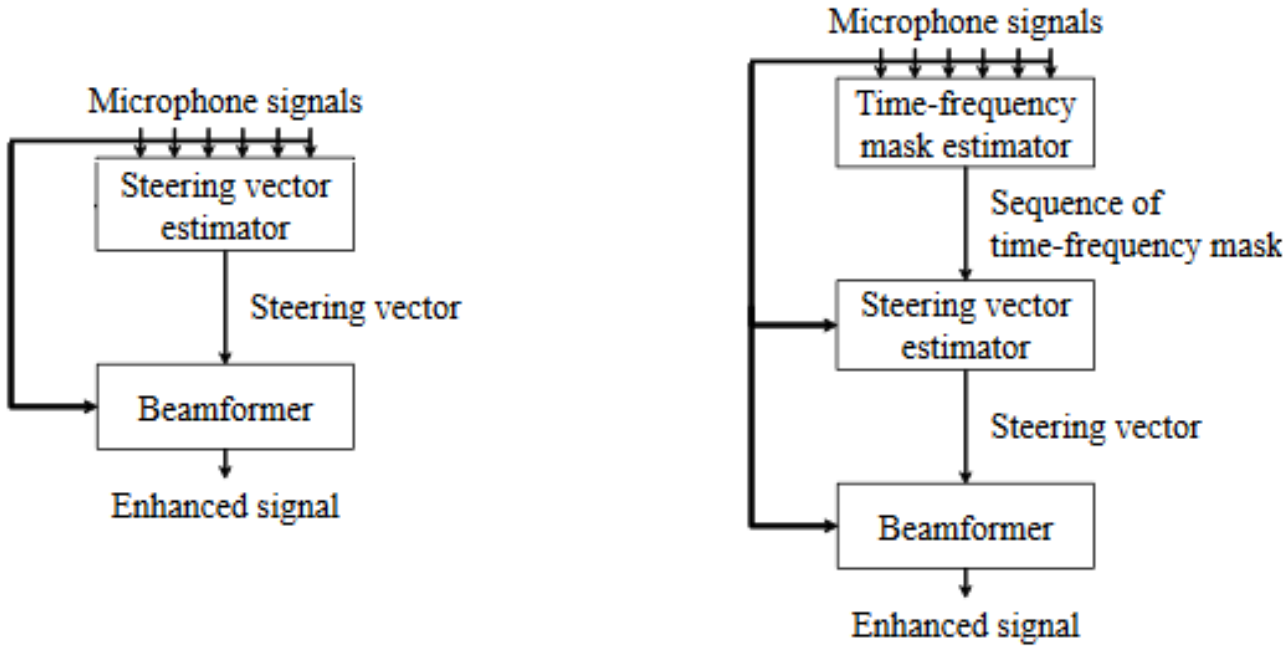


Figure: Traditional Beamformer (left) vs Mask based Beamformer (right) (Higuchi et al., 2016)

- Mask based beamformers estimate a spectral mask for each time-frequency bin prior to source localization.

Spectral Mask

- Each time-frequency bin can be classified into two categories:
  1. Noise
  2. Speech + Noise

Spectral Mask

- Spectral mask gives the probability that a particular time-frequency bin contains speech (speech mask) or noise (noise mask)

- Uses knowledge of sparsity of speech in time-frequency domain
- Improves source localization and spatial filtering performance

Advantages of Spectral Mask

- GCC PHAT gives equal weightage to all frequency bins.
- Speech will not be present in all frequency bins.
- This leads to erroneous source localization estimates.

Better Steering Vector Estimate

Use spectral masks to give more weightage to bins where speech probability is high

- Noise covariance estimate happens when speech is absent
- Even in speech frames, some bins may not contain speech

Better Noise Covariance Matrix Estimate

Use spectral masks to give more weightage to bins where noise probability is high

Ideal Binary Masks

- Ideal binary mask for noise is defined as :

$$IBM_N(k,l) = \begin{cases} 1, & \text{if } \frac{||S||}{||N||} < 10^{Th_N(f)} \\ 0, & \text{else} \end{cases}$$

- Ideal binary mask for speech is defined as :

$$IBM_S(k,l) = \begin{cases} 1, & \text{if } \frac{||S||}{||N||} > 10^{Th_S(f)} \\ 0, & \text{else} \end{cases}$$

- $\mathrm{Th}_N(\mathrm{f})$ & $\mathrm{Th}_S(\mathrm{f})$ are chosen to give low false positive rate

Spectral Mask Estimation

1) Neural Network Based Mask Estimation (Jahn Heymann & Haeb-Umbach, 2016a)

Objective

Train a neural network for single-channel mask prediction to predict both speech and noise masks for that channel

- Train neural networks using ideal binary masks as targets.
- Network trained with the binary cross-entropy loss function.
- During recognition, predict masks for each channel.
- Perform median pooling to obtain a single mask.
- LSTMs and BLSTMs can be used to predict the mask.

Spectral Mask Estimation

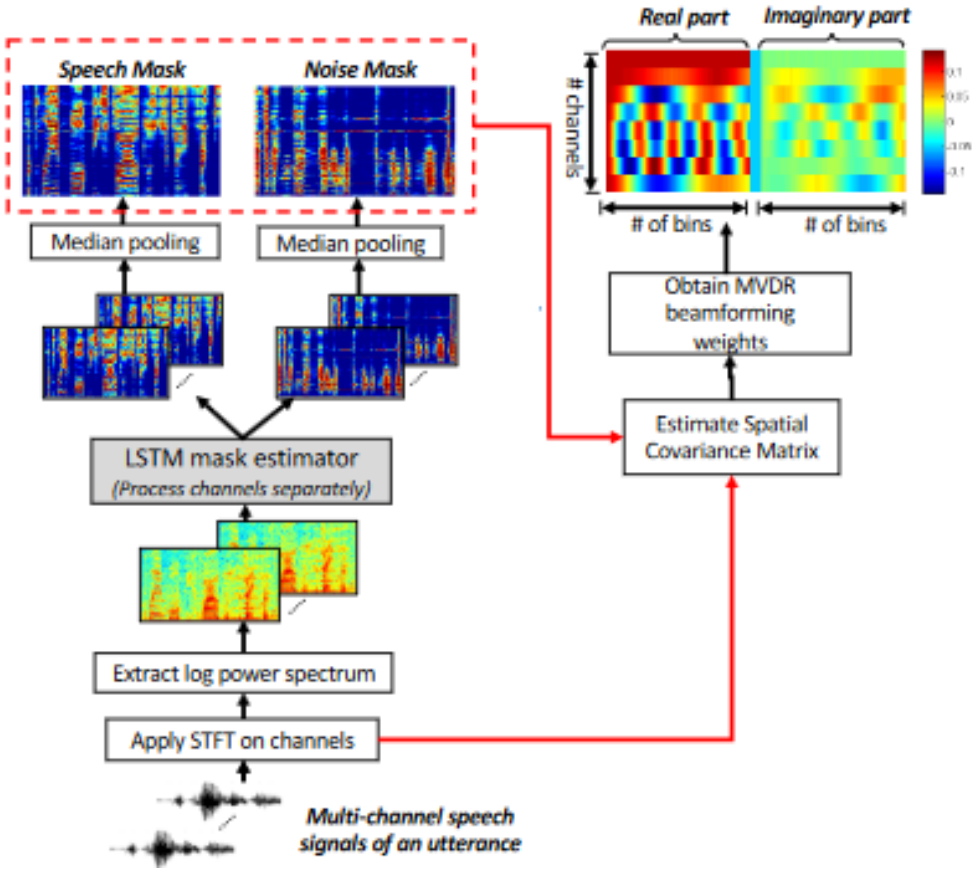


Figure: LSTM based mask estimator (Xiong Xiao, 2016)

Spectral Mask Estimation

2) Complex Gaussian Mixture Model (CGMM) (Higuchi et al., 2016)

- Generative model for soft mask estimation : $\mathbf{y}(k,l) = \mathbf{d}^v(k)s^v(k,l)$
- Clustered into two classes: speech+noise class and noise class.
- Assumes s(k,l) follows a complex Gaussian Distribution : $s^v(k,l) \sim \mathcal{N}_c(0, \phi^v_{k,l})$
- Multichannel observed signal $\mathbf{y}(k,l)$ follows : $P(\mathbf{y}(k,l)|i_{k,l} = v) = \mathcal{N}_c(0, R^v_k \phi^v_{k,l})$
- Joint distribution is a CGMM : $P(\mathbf{y}(k,l)) = \pi^{(n)}\mathcal{N}_c(0, R^{(n)}_k \phi^{(n)}_{k,l}) + \pi^{(s+n)}\mathcal{N}_c(0, R^{(s+n)}_k \phi^{(s+n)}_{k,l})$
- Parameters of CGMM are estimated using EM algorithm

Future Work

- Verify the performance improvement in source localisation and ASR accuracy using ideal binary masks.
- Develop a speech model to compute the spectral masks for each time frequency bin
- To setup the entire multichannel based speech recognition pipeline for the Chime Challenge
- Check the improvement in baseline ASR after incorporating the current techniques

References I

Cohen, I., Benesty, J., & Gannot, S. (2009). *Speech processing in modern communication: challenges and perspectives* (Vol. 3). Springer Science & Business Media.

Higuchi, T., Ito, N., Yoshioka, T., & Nakatani, T. (2016). Robust mvdr beamforming using time-frequency masks for online/offline asr in noise. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5210–5214).

Jahn Heymann, L. D., & Haeb-Umbach, R. (2016a). Neural network based spectral mask estimation for acoustic beamforming.

References II

Jahn Heymann, L. D., & Haeb-Umbach, R. (2016b). Wide residual blstm network with discriminative speaker adaptation for robust speech recognition.

Seltzer, M. L. (2003). *Microphone array processing for robust speech recognition* (Unpublished doctoral dissertation). Carnegie Mellon University Pittsburgh, PA.

Souden, M., Benesty, J., & Affes, S. (2010). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on audio, speech, and language processing*, *18*(2), 260–276.

References III

Xiong Xiao,
Z. Z. S. Z. S. S. S. W. L. W. L. X. D. L. J. E. S. C. H. L.,
Chenglin Xu. (2016). A study of learning based beamforming methods for speech recognition.