Distant Speech Recognition Using Microphone Arrays

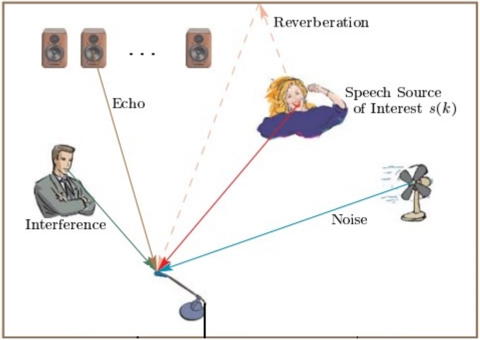
George Jose (153070011)
Guide : Prof. Preeti Rao

Indian Institute of Technology, Bombay

June 29, 2017

Table of Contents

Far Field Speech Recognition : Challenges

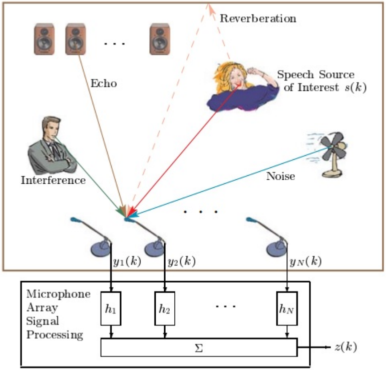


Major Challenges:

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

Table of Contents

System Overview

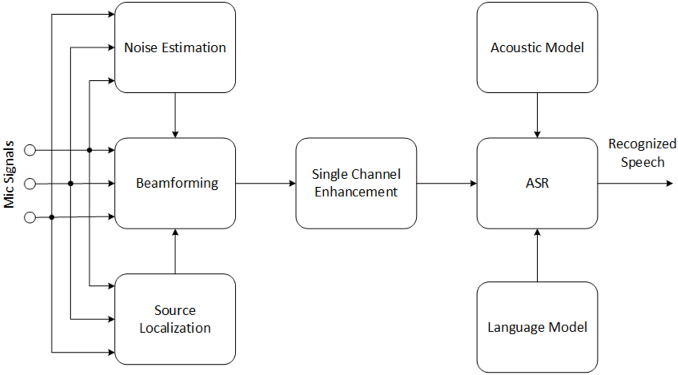


Figure: Overall System Block Diagram

Objective

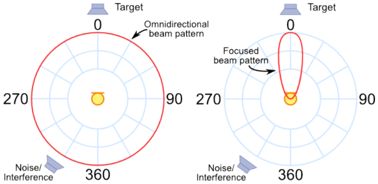


Figure: [1] Omnidirectional response(left) vs Directional response(right)

Two Stage Process

- Source Localization : Identifying the source location
- Spatial Filtering : Steering the response towards source

1

[1] http://www.labbookpages.co.uk/audio/beamforming/delaySum.html

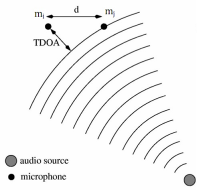
Source Localization

Goal

To find information regarding the position of the source with respect to the microphone array

Approaches broadly classified into 3 categories:

1. Time Delay Of Arrival (TDOA) algorithms
2. Steered Response Power (SRP) algorithms
3. High resolution spectral algorithms



TDOA Algorithms

Cross Correlation Method

Find the time shift which maximizes cross correlation

$$\tau_{12} = \arg\max_{\tau} R_{y_1 y_2}(\tau) = \arg\max_{\tau} E[y_1[n]y_2[n-\tau]]$$

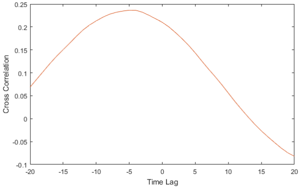


Figure: Cross correlation between 2 signals

In practice, cross correlation computed by:

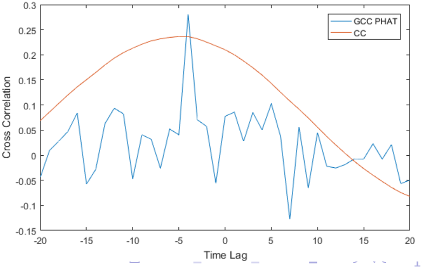
$$R_{y_1 y_2}(\tau) = \text{IFFT}\{G_{y_1 y_2}(f)\} = \text{IFFT}\ \{E[Y_1(f)Y_2^*(f)]\}$$

Generalised Cross Correlation Phase Transform (GCC PHAT) (Knapp & Carter, 1976)

- Discards amplitude and uses only phase
- Whitens the cross power spectrum

GCC PHAT :

$$R_{y_1 y_2}(\tau) = IFFT\{\frac{G_{y_1 y_2}(f)}{|G_{y_1 y_2}(f)|}\}$$



SRP PHAT Algorithm (Zhang, 2008)

Limitations of TDOA algorithms

- Do not consider all possible microphone pairs
- Do not use knowledge about microphone positions

SRP PHAT

- Fix the required angular resolution
- Compute TDOA between each microphone pair at each angle
- Evaluate SRP PHAT function for each angle

$$f(\theta) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} R_{y_1 y_2}^{GCC}(\tau_{ij}(\theta))$$

- Find θ which maximizes the SRP PHAT function

Acoustic Beamforming

Objective

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

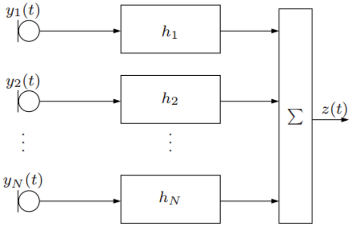


Figure: Beamformer Model (Cohen et al., 2009)

Signal at jth microphone :
$y_j(n) = s_i(n - \tau_{ji}) + v_j(n)$

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

In vector notation,
$\mathbf{Y}(f,k) = \mathbf{d(f)}S_i(f,k) + \mathbf{V}(f,k)$
Steering vector - $\mathbf{d(f)}$

Acoustic Beamforming

Beamformer Output :

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

Beamformer should not distort the speech signal

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

Output power :

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

Noise power at the output should be minimum

MVDR Beamforming

Targets:

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

Optimization Problem

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

Solving the optimization problem gives :

Minimum Variance Distortionless Response (MVDR) Beamformer

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

Delay Sum Beamforming (DSB)

- For spatially uncorrelated noise : $\boldsymbol{R}_v(f) = \sigma_v^2 I$ (σ_v^2 represents the noise PSD)
- DSB maximizes the White Noise Gain (WNG)

Optimization Problem

$$\mathbf{h}(f) = arg\,\underset{\mathbf{h}(f)}{min}\, \sigma_v^2 \mathbf{h}^h(f)\mathbf{h}(f) \quad \text{subject to } \mathbf{h}^h(f)\mathbf{d}(f) = 1$$

Delay Sum Beamformer (DSB)

$$h_{DSB}(k) = \frac{\mathbf{d(f)}}{N}$$

- Phase aligns the signal at different microphones

Super Directive Beamforming (Bitzer & Simmer, 2001)

- Works based on diffuse noise field assumption
- Elements of noise coherence matrix given by :

$$|\Gamma_{diff}(f)|_{ij} = sinc(2\pi f d_{ij}/c)$$

Optimization Problem

$$\mathbf{h}(f) = \underset{\mathbf{h}(f)}{arg\,min}\, \mathbf{h}^h(f)\mathbf{\Gamma}_{diff}(f)\mathbf{h}(f) \quad \text{subject to } \mathbf{h}^h(f)\mathbf{d}(f) = 1$$

Super Directive Beamforming (SDB)

$$\mathbf{h}_{SDB}(f) = \frac{\mathbf{\Gamma}_{diff}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^h(f)\mathbf{\Gamma}_{diff}^{-1}(f)\mathbf{d}(f)}$$

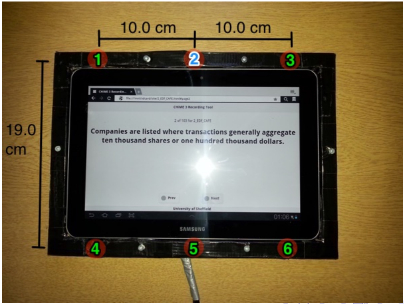
Summary

- Using an array of microphones we can :
  1. Locate the direction of the source using delay information
  2. Steer array response towards the direction of the source
- Depending on the noise conditions we can use :
  1. DSB : For spatially white noises
  2. MVDR : For coherent noise fields
  3. SDB : For diffuse noise fields

Table of Contents

CHiME Challenge Overview

- Distant speech recognition task using microphone arrays
- Six microphones embedded on the frame of a tablet
- Five mics facing upwards and one in backward direction
- Contains real and simulated data from WSJ0 corpus



Source : http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/overview.html

Environments

Cafe

Street

On the bus

Pedestrian area

Source http://spandh.dcs.shef.ac.uk/chime$_{c}hallenge/chime2015/data.html$

Data Overview

Real data recorded from 12 native US takers

Simulated data created by:

- Estimating speaker movements, SNR and noise from real data
- Remixing clean speech with corresponding time-varying delay and same noise signal or other noise signal with same SNR.

Simulated data doesnot contain echoes, reverberation, mic failures

Dataset		# speakers	# utterances
Training	real	4	1600
	simu	83	7138
Devel	real	4	410
	simu	4	410
Test	real	4	330
	simu	4	330

Chime Enhancement Baselines

BeamformIt (Anguera, Wooters, & Hernando, 2007)

- Source localization : GCC PHAT
- TDOA Post Processing : Viterbi Algorithm
- Channel Selection : Cross Correlation based
- Beamforming : Weighted Delay and Sum Beamforming

MVDR beamforming

- Source localization : GCC PHAT
- TDOA Post Processing : Viterbi Algorithm
- Channel Selection : Power Thresholding
- Noise Estimation : 500ms context prior to utterance

Baselines

GMM-HMM Baseline :

- Input Vector : 40-D MFCC Vector obtained after applying LDA to 91-D vector (13x7)
- Architecture : Total of 2500 GMMs with 6 Gaussians each

DNN-HMM Baseline :

- Input Vector : 440-D filter bank features (40x11)
- Architecture : 7 Hidden layer with 2048 neurons in each layer
- Cost function : Minimum Bayesian Risk (MBR) function

LM : 3-gram LM with 5-gram and RNN LM for lattice rescoring

ASR Results

Using a GMM-HMM and trigram LM

Method	Real	Simu	Average
None	22.16	24.44	23.3
DSB	12.71	13.73	13.22
SDB	12.76	13.57	13.17
BeamformIt	12.99	14.30	13.64
MVDR	17.12	10.67	13.92

Table: WER (%) obtained on Chime Challenge development set using a GMM-HMM model trained on noisy data with a trigram language model

- MVDR has best performance in simulated data and worst performance in real data !!

Table of Contents

Multi Channel Alignment (MCA) Beamforming (Stolbov & Aleinik, 2015)

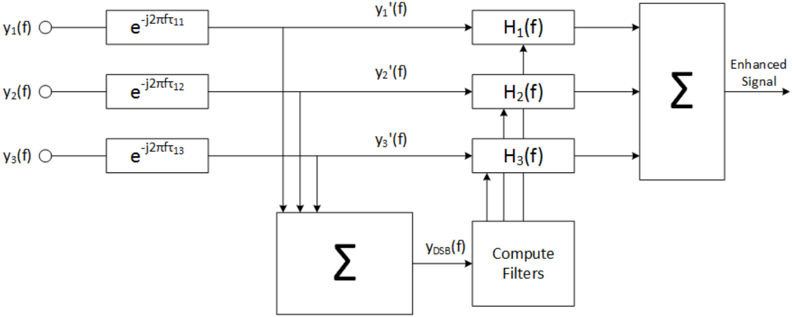


Figure: Multi Channel Alignment Beamforming

MCA Algorithm

MCA Algorithm

1. Compute the TDOAs using source localization algorithm
2. Phase align speech signals using the estimated TDOAs
3. Perform DSB to compute reference signal for filter estimation
4. Apply the filters and sum the filtered signals

Filter Estimation

$$H_i(f,k) = \frac{|E\{y_i'(f,k)y_{DSB}^*(f,k)\}|}{E\{y_i'(f,k)y_i'^*(f,k)\}}$$

This is equivalent to a Weiner filter !!

Proposed Approach

- Combines Weiner filtering with MVDR beamforming
- Constraint the filters to take the form of a Weiner filter
- Modify steering vector by adding gains to each element

Modified Steering Vector

$$\mathbf{d}(f,k) = [g_1(f,k)e^{-j2\pi f\tau_{11}}\ g_2(f,k)e^{-j2\pi f\tau_{12}}\ \ldots\ g_N(f,k)e^{-j2\pi f\tau_{1N}}]^T$$

$$g_i(f,k) = \frac{1}{H_i(f,k)} = \frac{E\{y_i'(f,k)y_i'^*(f,k)\}}{|E\{y_i'(f,k)y_{DSB}^*(f,k)\}|}$$

- Optimization constraint : $\mathbf{d}$(f,k)$\mathbf{h}^H$(f,k)=1
- Ensures each filter take the form of a Weiner filter

Table of Contents

Objective Measures on Real Data

Method	CD	f-SNR	SRMR
None	3.88	-1.26	2.06
DSB	3.37	2.97	2.29
Gain-DSB + DSB	3.36	6.02	2.36
MVDR	3.52	-0.63	2.52
Gain-DSB + MVDR	3.52	5.08	2.69

Table: Objective measures on Chime Challenge development set

Method	CD	f-SNR	SRMR
None	3.17	1.89	1.73
DSB	3.01	5.85	1.94
Gain-DSB + DSB	3.22	5.99	2.03
MVDR	3.05	3.06	2.24
Gain-DSB + MVDR	3.46	6.67	2.38

Comparison of WERs

Using a GMM-HMM acoustic model and trigram LM

Method	**Real**	**Simu**	**Average**
BeamformIt	12.99	14.30	13.64
DSB	12.71	13.73	13.22
Gain-DSB + DSB	**12.04**	12.05	12.04
MVDR	17.12	10.67	13.92
Gain-DSB + MVDR	12.75	**10.48**	**11.62**

Table: WER (%) obtained on Chime Challenge development set using a GMM-HMM model trained on noisy data with a trigram language model

NMF Based Postprocessing

Post processing	Real	Simu	Average
None	12.75	10.48	11.62
CNMF	15.25	12.87	14.06
CNMF + NMF	14.26	12.08	13.17

Table: WER (%) obtained with NMF based post processing methods to Gain-DSB + MVDR

- NMF based postprocessing techniques increases the WER
- Designed to reduce the amount of reverberation
- Presence of residual noise degrades the performance

Effect of DNN-HMM Acoustic Model on WERs

Using a DNN-HMM acoustic model and trigram LM

Method	**Real**	**Simu**	**Average**
BeamformIt	8.14	9.03	8.59
DSB	8.08	8.29	8.18
Gain-DSB + DSB	7.87	7.73	7.80
MVDR	12.38	6.25	9.31
Gain-DSB + MVDR	8.71	6.60	7.66

Table: WER (%) obtained on Chime Challenge development set using a DNN-HMM model trained on noisy data with a trigram language model

Effect of Lattice Rescoring on WERs

Lattice Rescoring using a 5-gram LM

Method	**Real**	**Simu**	**Average**
BeamformIt	6.85	7.75	7.30
DSB	6.59	7.29	6.94
Gain-DSB + DSB	6.39	6.66	6.52
MVDR	10.93	5.29	8.11
Gain-DSB + MVDR	7.39	5.50	6.44

Table: WER obtained on Chime Challenge development set using a DNN-HMM model trained on noisy data after lattice rescoring with 5-gram language model

Effect of Lattice Rescoring on WERs

Lattice Rescoring using a RNN LM

Method	**Real**	**Simu**	**Average**
BeamformIt	5.76	6.77	6.27
DSB	5.55	6.27	5.90
Gain-DSB + DSB	5.35	5.69	5.52
MVDR	9.85	4.51	7.18
Gain-DSB + MVDR	6.57	4.75	5.66

Table: WER obtained on Chime Challenge development set using a DNN-HMM model trained on noisy data after lattice rescoring with RNN language model

Objective Measure on TCS Data

Method	CD	f-SNR	SRMR
None	2.57	4.69	6.65
DSB	2.28	8.80	8.28
Gain-DSB + DSB	2.42	9.36	8.81
MVDR	2.33	6.23	6.77
Gain-DSB + MVDR	2.53	9.60	8.54

Table: Objective measures on TCS Data

References I

Anguera, X., Wooters, C., & Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2011–2022.

Benesty, J., Chen, J., & Huang, Y. (2008). *Microphone array signal processing* (Vol. 1). Springer Science & Business Media.

Bitzer, J., & Simmer, K. U. (2001). Superdirective microphone arrays. In *Microphone arrays* (pp. 19–38). Springer.

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

References II

Habets, E. (2008). *Room impulse response (rir) generator.* https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

Johnson, D. H., & Dudgeon, D. E. (1992). *Array signal processing: concepts and techniques*. Simon & Schuster.

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(4), 320–327.

Kumatani, K., Arakawa, T., Yamamoto, K., McDonough, J., Raj, B., Singh, R., & Tashev, I. (2012). Microphone array processing for distant speech recognition: Towards real-world deployment. In *Signal & information processing association annual summit and conference (apsipa asc), 2012 asia-pacific* (pp. 1–10).

References III

Kumatani, K., McDonough, J., & Raj, B. (2012). Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *Signal Processing Magazine, IEEE*, *29*(6), 127–140.

Perez-Lorenzo, J., Viciana-Abad, R., Reche-Lopez, P., Rivas, F., & Escolano, J. (2012). Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. *Applied Acoustics*, *73*(8), 698–712.

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

References IV

Stolbov, M. B., & Aleinik, S. V. (2015). Improvement of microphone array characteristics for speech capturing. *Modern Applied Science*, *9*(6), 310.

Zhang, C., Florêncio, D., Ba, D. E., & Zhang, Z. (2008). Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *IEEE Transactions on Multimedia*, *10*(3), 538–548.