Multi-channel Processing for Distant Speech Recognition

Preeti Rao and V. Rajbabu

rajbabu, prao@ee.iitb.ac.in
Department of Electrical Engineering
IIT Bombay

24 November 2017

Table of Contents

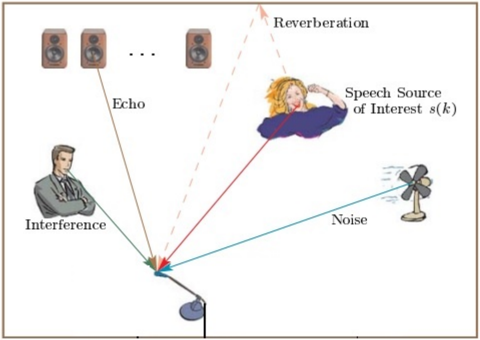
Distant Speech Recognition (DSR): Introduction

Research Outcome
- Multi-channel DSR Framework
- Dereverberation using NMF
- NMF for joint dereverberation and denoising

References

Distant Speech Recognition (DSR): Challenges

Distant speech: source at a relatively further distance from a microphone array compared to the spacing between array elements



Major Challenges:

1. Noise
2. Reverberation
3. Echo
4. Interfering speaker

Image courtesy (Seltzer, 2003)

DSR: Approach

Exploit source separation in spatial domain

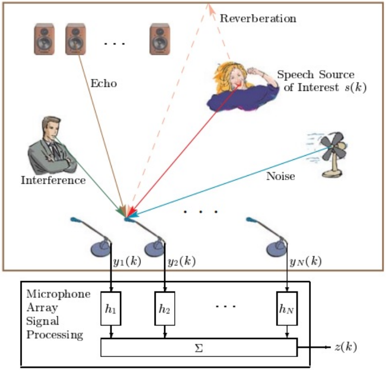


Image courtesy (Seltzer, 2003)

How ?

Use multiple microphones

Why ?

Signals from each source arrive with different delays at each microphone

Table of Contents

Distant Speech Recognition (DSR): Introduction

Research Outcome
- Multi-channel DSR Framework
- Dereverberation using NMF
- NMF for joint dereverberation and denoising

References

Status

Proposed modifications for an improved DSR system

- Framework of multi-channel beamforming followed by single-channel enhancement for improved DSR word error rates (WERs)
- Improved steering vector in MVDR beamforming
- Non-negative matrix factorization (NMF) for improved single-channel dereverberation using appropriate constraints on room impulse response (RIR)
- NMF formulation for joint dereverberation and denoising

Experiments for DSR WERs and speech enhancement measures

Outline

Distant Speech Recognition (DSR): Introduction

Research Outcome
- Multi-channel DSR Framework
- Dereverberation using NMF
- NMF for joint dereverberation and denoising

References

DSR: System Overview

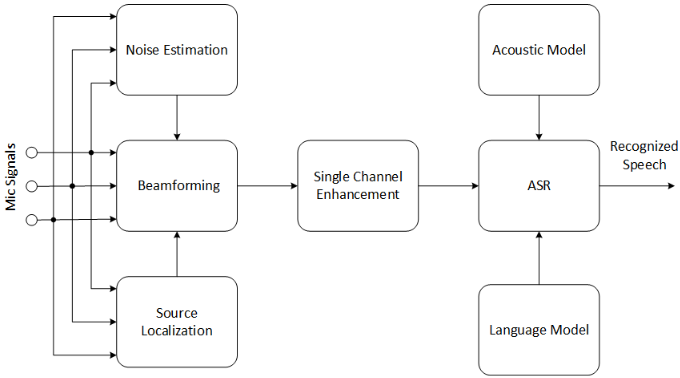


Figure: Block diagram of proposed DSR system

DSR: System Overview

Multiple stages

- Source localization and beamforming: Identifying the source location and performing spatial filtering to enhance the speech signal
- Single-channel enhancement: denoising and dereverberating NMF
- Automatic speech recognition (ASR) to recognise speech - acoustic and language models, training and testing

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

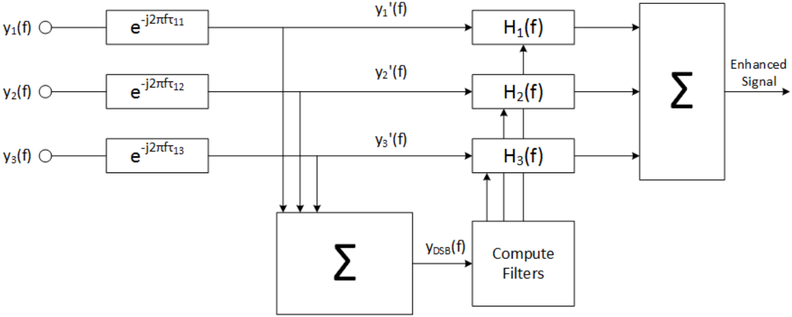


Figure: Multi Channel Alignment Beamforming

MCA Algorithm

1. Compute time-difference of arrivals (TDOAs) - source localization
2. Phase align speech signals using the estimated TDOAs
3. Delay-sum beamforming (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 MVDR Beamforming

- Combines Weiner filtering with minimum variance distortionless (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

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

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

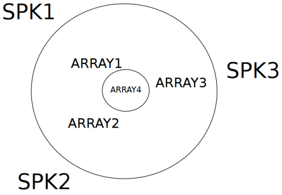
- Optimization constraint : $\mathbf{d}$(f,k)$\mathbf{h}^H$(f,k)=1
- Ensures each filter take the form of a Weiner filter

Dataset: In-house Microphone Array Setup

Data collection setup was created with additional support from TCS

- Multi-channel sound-card with 8 individual channel mics and pre-amplifiers
- 4 mics for the array and 4 mics used as close talking microphones (CTM) or lapel mics

- Number of speakers : 3
- Distance from speaker to mic : 50cm
- Array Diameter : 20cm



Speech scenarios

- Non Overlapping
- Partial Overlapping
- Complete Overlapping

In-house Data: Speech Enhancement Measures

Degraded data objective measures:
CD = 2.57 f-SNR = 4.69 SRMR = 6.65

Method	CD	f-SNR	SRMR
DSB	0.29	4.11	1.63
Gain-DSB + DSB	0.14	4.69	2.16
MVDR	0.24	1.54	0.12
Gain-DSB + MVDR	-0.04	4.91	1.89

Table: Objective measures on in-house non-overlapping data

- Increase in cepstral distance due to distortions
- Increase in f-SNR and SRMR objective measures
- Trends consistent with those observed in CHiME data

DSR Framework: Summary

- Beamforming based enhancement improved speech recognition accuracies
- Proposed modification to MVDR beamforming has improved performance in CHiME real-data
- Superior noise reduction is achieved at the cost of some speech distortion

Outline

Distant Speech Recognition (DSR): Introduction

Research Outcome
- Multi-channel DSR Framework
- Dereverberation using NMF
- NMF for joint dereverberation and denoising

References

Reverberation Models

- Time domain model

$$y(n) = s(n) * h(n) = \sum_{k=0}^{L-1} h(k)s(n-k)$$

$y(n)$: reverb speech, $s(n)$: clean speech

$h(n)$: RIR, L : length of RIR

- Spectrogram model -smooth approx. for reverb spectrogram ($Y(n,k)$)

$$Y(n,k) \approx S(n,k) * H(n,k) = \sum_{m=0}^{L_h-1} H(m,k)S(n-m,k)$$

$S(n,k)$: magnitude spectrogram of clean, RIR

L_h : Number of frames in H(n,k)

Our focus is on NMF based single channel dereverberation

Non-negative Matrix Factorization (NMF)

NMF model

- Factorizes non-negative matrix $\mathbf{S}$
  $\mathbf{S} \approx \mathbf{WA}$, where $\mathbf{W} \geq 0$, $\mathbf{A} \geq 0$
- $\mathbf{W}$: set of basis vectors, $\mathbf{A}$: corresponding activations

- Clean speech $S(n,k)$ can be decomposed using NMF

Convolutive NMF (C-NMF)

$$\mathbf{Y} \approx \sum_{m=0}^{L_h-1} \mathbf{H}_m \overset{m\rightarrow}{\mathbf{S}},$$

where, $\mathbf{H}_m = diag(\mathbf{H(m,0)}, \mathbf{H(m,1)}, ..., \mathbf{H(m,K-1)})$

- Reverb speech $Y(n,k)$ modeled using C-NMF

Dereverberation using C-NMF (Kameoka, 2009)

- Obtain **S** and **H** from **Y** using C-NMF

Optimization Problem

$$\min_{\mathbf{H},\mathbf{S}} \sum_{n,k} KL(Y(n,k)||S(n,k) * H(n,k))$$

$$s.t. \quad \sum_{n=0}^{L_h-1} H(n,k) = 1, \forall k,\ S \geq 0,\ H_m \geq 0$$

- Constraint on **H** to avoid gain uncertainty
- Referred as Non-negative Convolutive Transfer Function ($N - CTF$)

Derevberation using C-NMF with Speech Model (Mohammadiha, 2015)

- Additional NMF model for clean speech ($\mathbf{S} \approx \mathbf{WA}$)

Optimization Problem

$$\min_{H,W,A} \sum_{n,k} KL(Y(n,k) || S(n,k) * H(n,k))$$
$$H(n,k) \leq H(n-1,k), \ \mathbf{S} = \mathbf{WA}$$

- Constraints on $\mathbf{H}$ to avoid distortions
- Referred as N-CTF+NMF

Proposed NMF: Constrained RIR

Motivation

Current NMF based methods

- do not use appropriate prior on RIR
- do not focus on RIR estimation for speech dereverberation

Objective

Use appropriate constraints on RIR to obtain improved

- RIR estimates
- speech dereverberation

in the NMF formulation

NMF Dereverberation: Experiment Setup

- Clean speech
  - 16 TIMIT sentences spoken by different speakers
- RIR
  - REVERB 2014 challenge
  - T_{60}=700ms, d = 2m
- STFT parameters
  - 64ms window, 16ms hop size
  - square root of Hanning window
- RIR estimate

- Objective measures for speech enhancement
  - Perceptual Evaluation of Speech Quality (PESQ)
  - Cepstral Distance (CD)
  - Speech to Reverberation Modulation energy Ratio (SRMR)

Results: RIR Estimation with Speech Model

- Proposed constraints improved RIR estimate

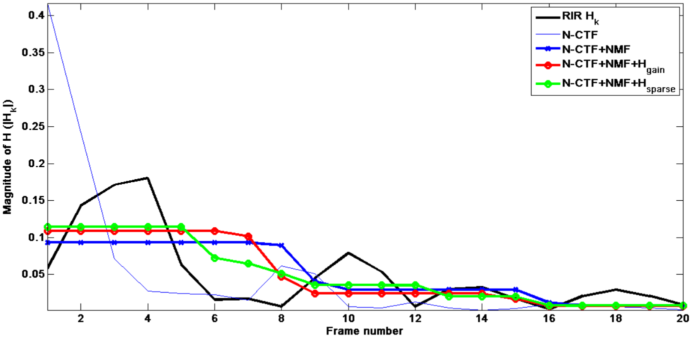


Figure: Normalized RIR estimates for a specific RIR with $T_{60} = 700$ ms and frequency band ($k = 218$).

Results: Dereverberation with Speech Model

Methods	$\Delta PESQ$	ΔCD	$\Delta SRMR$
N-CTF	0.27	0.71	1.48
N-CTF + NMF	0.54	0.92	1.65
N-CTF + NMF + H_{sparse}	0.54	0.92	1.65
N-CTF + NMF + H_{gain}	0.54	0.94	2.14
N-CTF + NMF + H_{early}	0.49	0.93	2.22

- Sparsity on RIR marginally improved results
  - better RIR estimate did not lead to better clean speech estimate
- Frequency envelope constraint improved performance
- Retaining early part of RIR helped

NMF Dereverberation: Summary

- Developed an improved NMF frame work for dereverberation
- Constraints on RIR
  - sparsity, frequency envelope, retaining early part of RIR
- Enhancement without speech model
  - improvement with inclusion of early part
- Enhancement with speech model
  - improved performance with sparsity, frequency envelope and early part of RIR

Outline

Distant Speech Recognition (DSR): Introduction

Research Outcome
- Multi-channel DSR Framework
- Dereverberation using NMF
- NMF for joint dereverberation and denoising

References

Joint Dereverberation and Denosing using NMF [Deepak Baby, ICASSP 2016]

- Supervised approach
  - Clean basis (5000) from WSJ0 training data
  - Noise basis ($2500 + 1000$ *sniffed basis*) from CHiME2 background noise
- Two models proposed to perform joint dereverberztion and denoising

NMF model for clean speech and noise

$$Y = H * (W_{speech} X_{speech}) + W_{noise} X_{noise}$$

Convolutive NMF model for clean speech and noise

$$Y = H * (W_{speech} * X_{speech}) + W_{noise} * X_{noise}$$

- Multiplicative update for estimating X_{speech}, X_{noise} and H
- Shows objective measure improvement for CHiME2 challenge

Proposed Joint Dereverberation and Denoising using NMF

- Representation using NMF model for clean speech and noise

$$
\begin{aligned}
Y(n,k) &= \sum_{l=0}^{L_h-1} H(k,l)S(k,n-l) + W_{noise}X_{noise} \\
&= \sum_{l=0}^{L_h-1} H(k,l) \sum_{r=1}^{R} W_{speech}(k,r)X_{speech}(r,n-l) + W_{noise}X_{noise} \\
&= \sum_{r=1}^{R} W_{speech}(k,r) \sum_{l=0}^{L_h-1} H(k,l)X_{speech}(r,n-l) + W_{noise}X_{noise}
\end{aligned}
$$

Proposed Joint Dereverberation and Denoising using NMF

- Assumption

  $H(k,l)$ independent of k, i.e. $H(k,l) = H(l)\ \forall k \in \{1, 2, ..., K-1\}$

$$
\begin{aligned}
Y(n,k) &= \sum_{r=1}^{R} W_{speech}(k,r) \sum_{l=0}^{L_h-1} H(l) X_{speech}(r, n-l) + W_{noise} X_{noise} \\
&= W_{speech} X_{reverb} + W_{noise} X_{noise} \\
&= [W_{speech}\ \ W_{reverb}][X_{reverb}^T\ \ X_{noise}^T]^T
\end{aligned}
$$

where X_{reverb} is defined as

$$X_{reverb} = X_{clean} * H(l)$$

- X_{clean} is estimated from X_{reverb}

Effect of Reverberation on Activation

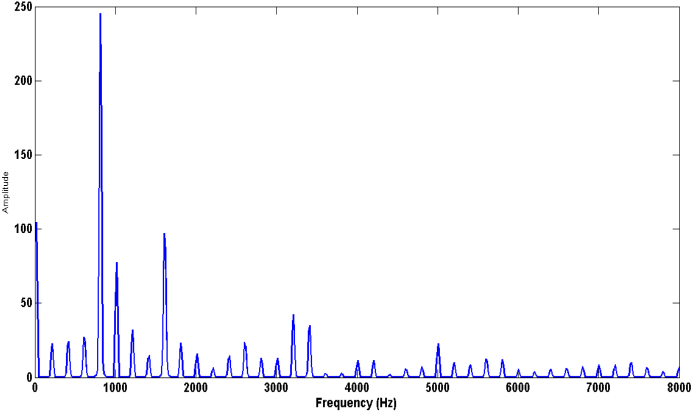


Figure: Basis learned for the vowel 'a'

Effect of Reverberation on Activation

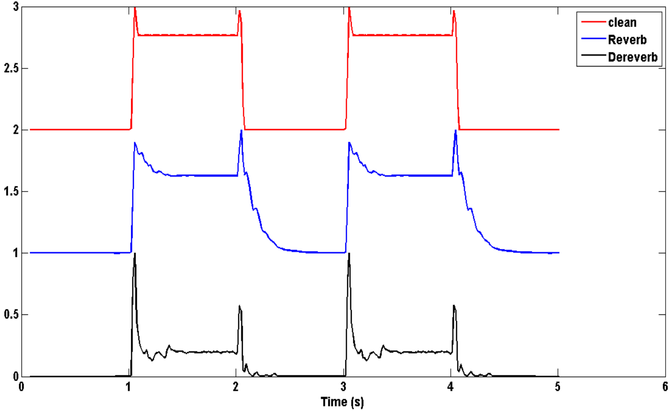


Figure: Effect of Reverberation and Dereverberation algorithm on Activation of a Vowel

Effect of Reverberation on Activation

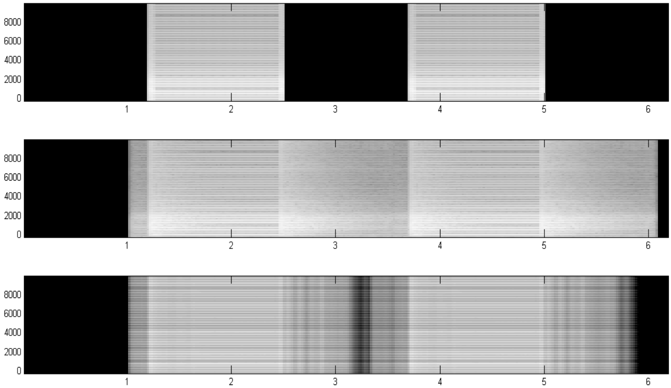


Figure: Spectrogram of (a)clean, (b) reverb and (c)dereverb

Enhancement Result for Speaker Specific Model

- Clean basis
  - Learned from 8 TIMIT sentences uttered by speaker
  - 100 basis for each speaker
- Noise basis
  - Factory-2 noise
  - 100 basis obtained from 24 s noise data
- Degraded condition
  - Reverb challenge RIR - $T_{60} = 700\ ms$ and $d = 2\ m$
  - Factory-2 noise with $SNR = 0dB$

	$PESQ$	CD
Degraded speech	0.93	4.53
	$\Delta PESQ$	ΔCD
Reference method	1.53	0.80
Proposed method	**3.18**	**3.60**

Table: Enhancement Result

Enhancement for CHiME Challenge Noise

RIR : Reverb Challenge, 700 ms, 2 m
Noise type: Cafe
SNR = 0dB

	PESQ	*CD*	*fSNR*	*LLR*
Degraded speech	0.71	5.49	−2.40	1.23
	$\Delta PESQ$	ΔCD	$\Delta fSNR$	ΔLLR
Reference Method	0.38	0.10	2.93	−0.14
Proposed method	0.34	0.31	2.31	0.03

Table: Enhancement Result

Joint Dereverberation and Denoising: Summary

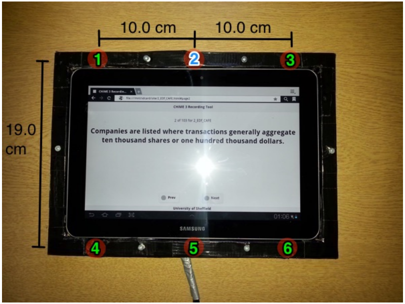
- Supervised approach to jointly handle reverberation and noise
- Two step approach - less numbers of parameters to be estimated and better estimates for clean speech
- Significant improvement in enhancement measures - need to consider ASR measures

Thanks

Questions

Dataset: CHiME Challenge

- DSR 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

Dataset: Environments

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

CHiME Speech Data

Real data recorded from 12 native US talkers

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 does not 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 Data: Enhancement Measures

Degraded data measures:
CD = 3.17   f-SNR = 1.89   SRMR = 1.73

Method	**Δ CD**	**Δ f-SNR**	**Δ SRMR**
DSB	0.16	3.96	0.21
Proposed + DSB	-0.05	4.10	0.30
MVDR	0.12	1.17	0.51
Proposed + MVDR	-0.29	4.78	0.65

Table: Objective measures on Chime Challenge evaluation set

- Consistent improvements in f-SNR and SRMR
- Speech distortions are introduced due to addition of gains

CHiME Data: GMM-HMM Acoustic Model

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
Proposed + DSB	**12.04**	12.05	12.04
MVDR	17.12	10.67	13.92
Proposed + 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

- Proposed steering vector has improved WERs
- Significant improvements in real data

CHiME Data: NMF based Post-processing

Convolutive NMF (CNMF) for dereverberation

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

CHiME Data: DNN-HMM Acoustic Model

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
Proposed + DSB	**7.87**	7.73	7.80
MVDR	12.38	**6.25**	9.31
Proposed + 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

- 4 % decrease in WER compared to GMM-HMM model
- Relative improvements independent of acoustic model

CHiME Data: Lattice Rescoring

Lattice rescoring using a RNN language model

Method	Real	Simu	Average
BeamformIt	5.76	6.77	6.27
DSB	5.55	6.27	5.90
Proposed + DSB	**5.35**	5.69	**5.52**
MVDR	9.85	**4.51**	7.18
Proposed + 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

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.

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).
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.
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.