SPEECH DEREVERBERATION USING NMF WITH REGULARIZED ROOM IMPULSE RESPONSE

Nikhil Mohanan   Rajbabu Velmurugan   Preeti Rao

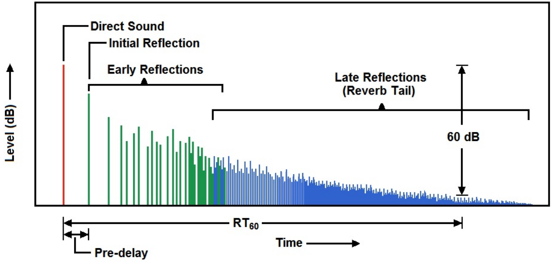
Department of Electrical Engineering
Indian Institute of Technology Bombay

ICASSP 2017
8 March, 2017

Outline

Reverberation

- Source signals undergo multiple reflections in closed surface
- Effects of reverberation depend on
  - room characteristics
  - position of microphone and source
- Typically modeled using room impulse response (RIR)[1]



[1] http://lossenderosstudio.com/newsletter.php?issue=66

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

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 approximation for reverb spectrogram

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

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

$H(n,k)$: spectrogram of RIR

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

Non-negative Matrix Factorizations

Non-negative Matrix Factorization (NMF)

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

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

Convolutive NMF (C-NMF)

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

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

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

Dereverberation using C-NMF (Kameoka, 2009)

- Obtain $\boldsymbol{S}$ and $\boldsymbol{H}$ from $\boldsymbol{Y}$ using C-NMF

Optimization Problem

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

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

- Constraint on $\boldsymbol{H}$ 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 ($\boldsymbol{S} \approx \boldsymbol{WA}$)

Optimization Problem

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

$$H(n,k) \leq H(n-1,k), \;\; \boldsymbol{S} = \boldsymbol{WA}$$

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

Motivation and Objective

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

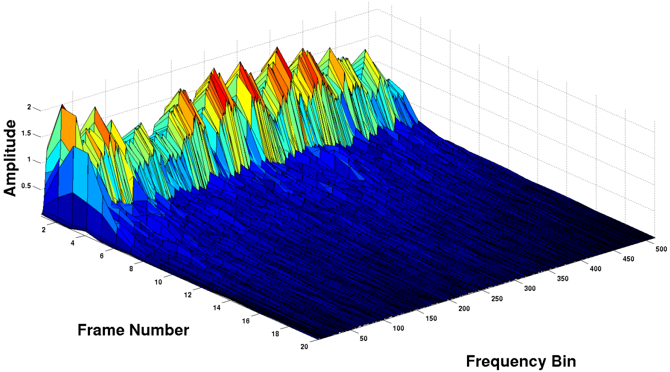
in the NMF formulation

Outline

Method 1: Sparsity on RIR

Enforce sparsity on RIR

- $H(n, k)$ is sparse for larger n



Spectrogram of RIR

Method 1: Sparsity constraint on RIR

Enforce sparsity on RIR

Updated cost function

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

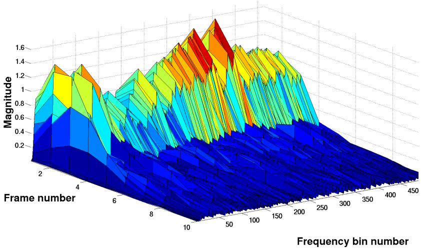
- Second term enforces sparsity on $H(n,k)$
- Referred as $N - CTF + H_{sparse}$

- Extended for $N - CTF + NMF$
- Referred as $N - CTF + NMF + H_{sparse}$

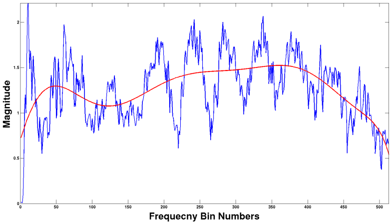
Outline

Method 2: Frequency Envelope of RIR

Given a RIR, frequency envelopes ($g(k)$) corresponding to each frame obtained



Spectrogram of RIR



Spectrum for $n = 2$

Methods 2: Frequency Envelope of RIR

- Frequency envelope $g(k)$ obtained from knowledge of RIR

Updated cost function

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

$$\sum_{n=0}^{L_h-1} H(n,k) = g(k), \forall k$$

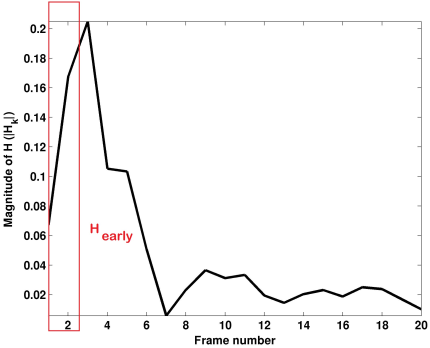
- Referred as $N - CTF + H_{gain}$
- $N - CTF + NMF + H_{gain}$ when extended to $N - CTF + NMF$

Outline

1. Motivation
2. Proposed Dereverberation Methods
   - Sparsity constraint on RIR
   - Frequency envelope on RIR
   - Inclusion of Early RIR
3. Results
   - RIR Estimation
   - Speech Enhancement
4. Summary

Method 3: Retaining Early Part of RIR

- First few frames of RIR (50 ms) constitute the early part



Early part of RIR

Method 3: Retaining Early Part of RIR

- Retaining early part enhances speech quality
- Enhanced spectrogram obtained as

$$\hat{S}_{est}(n,k) = S_{est}(n,k) * H_{early}(n,k)$$

$S_{est}(n,k)$, $H(n,k)$ obtained using any existing NMF-based methods

- Referred as $N - CTF + H_{early}$
- $N - CTF + NMF + H_{early}$ when extended to N-CTF+NMF

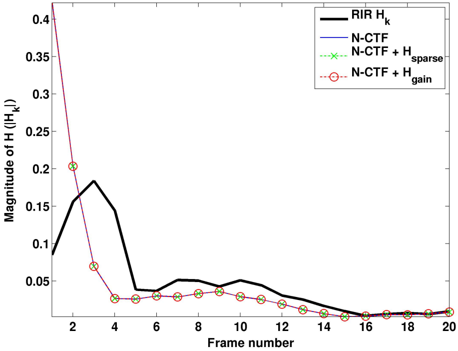
Experiment Setup

- 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
  - intrusive methods (comparison with clean speech)
    - PESQ
    - Cepstral distance (CD)
  - non-intrusive method
    - speech to reverberation modulation energy ratio (SRMR)

Outline

Results: RIR Estimation without Speech Model

- Constraints did not improve RIR estimate



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

RIR Comparison without Speech Model

- Comparison using Mean square error (MSE)
- MSE defined as

$$MSE = ||\boldsymbol{H_{act}} - \boldsymbol{H_{est}}||_F$$

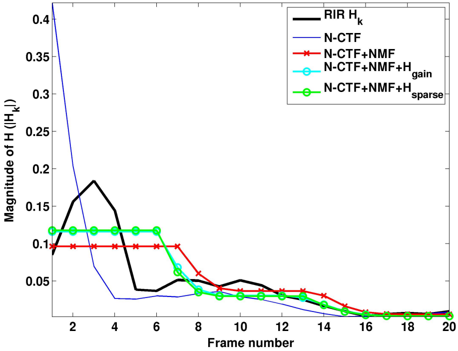
$\boldsymbol{H_{act}}$, $\boldsymbol{H_{est}}$: normalised actual and estimated RIR spectrogram

Methods	MSE
N-CTF	6.26
N-CTF + H_{sparse}	6.26
N-CTF + H_{gain}	6.26

- Constraints tend to be equivalent, no change in MSE

Results: RIR Estimation with Speech Model

- Proposed constraints improved RIR estimate



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

RIR Comparison with Speech Model

Methods	*MSE*
N-CTF	6.26
N-CTF + NMF	4.31
N-CTF + NMF + H_{gain}	4.26
N-CTF + NMF + H_{sparse}	4.33

- N-CTF+NMF case
  - no significant MSE change

Outline

1. Motivation
2. Proposed Dereverberation Methods
   - Sparsity constraint on RIR
   - Frequency envelope on RIR
   - Inclusion of Early RIR
3. Results
   - RIR Estimation
   - Speech Enhancement
4. Summary

Results: Dereverberation without Speech Model

Methods	$\Delta PESQ$	ΔCD	$\Delta SRMR$
N-CTF	0.27	0.71	1.48
N-CTF + H_{sparse}	0.27	0.71	1.48
N-CTF + H_{gain}	0.28	0.71	1.18
N-CTF + H_{early}	0.33	0.77	1.92

- Significant variations in SRMR
- Sparsity and frequency envelope constraints did not help
- H constraints
  - $N - CTF$ and $N - CTF + H_{sparse}$ are equivalent
  - $N - CTF$ and $N - CTF + H_{gain}$ differ in scaling factor ($g(k)$)
- Retaining early part of RIR helped

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

Summary

- Developed an improved NMF frame work for dereverberation
- Constraints on RIR
  - sparsity on RIR
  - frequency envelope of RIR
  - retaining early part of RIR
- Enhancement without speech model
  - improvement with inclusion of early part
  - no improvement with sparsity and frequency envelope
- Enhancement with speech model
  - improved performance with inclusion of early part
  - improvement with frequency envelope constraint
  - marginal improvement with sparsity

Future Work

- Effects of constraints on ASR performance
- Performance of constraints in an NMF supervised setting
- Relax inequality constraint $H(n,k) \leq H(n-1,k)$ in $N - CTF + NMF$

Acknowledgement

Thanks,

- Council of Scientific & Industrial Research (CSIR), India
- Tata Consultancy Services (TCS), India

Reference I

Anguera, X., Wooters, C., & Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. , *15*(7), 2011–2022.

Avargel, Y., & Cohen, I. (2007). System identification in the short-time fourier transform domain with crossband filtering. , *15*(4), 1305–1319.

Baby, D., & hamme, H. V. (2016). Supervised speech dereverberation in noisy environments using exemplar-based sparse representations. In *Proc. of ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 156–160).

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

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., ... others (2014). Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge. In *Reverb workshop.*

Reference II

Falk, T. H., Zheng, C., & Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. , *18*(7), 1766–1774.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus. *Linguistic data consortium, Philadelphia*, *33*.

Kallasjoki, H., Gemmeke, J. F., Palomaki, K. J., Beeston, A. V., & Brown, G. J. (2014, May). Recognition of reverberant speech by missing data imputation and NMF feature enhancement. In *Proc. reverb workshop.*

Kameoka, H., Nakatani, T., & Yoshioka, T. (2009). Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In *Proc. of ieee international conference on acoustics, speech and signal processing (ICASSP)* (pp. 45–48).

Reference III

Kumar, K., Singh, R., Raj, B., & Stern, R. (2011). Gammatone sub-band magnitude-domain dereverberation for ASR. In *Proc. of ieee international conference on acoustics, speech and signal processing (ICASSP).*

Kumatani, K., McDonough, J., & Raj, B. (2012, Nov.). Microphone array processing for distant speech recognition. , 127–140.

Kuttruff, H. (2009). *Room acoustics*. Spon Press.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, *401*, 788–791.

Mirsamadi, S., & Hansen, J. H. L. (2014). Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications. In *Proc. of fifteenth annual conference of the international speech communication association (INTERSPEECH).*

Reference IV

Mohammadiha, N., & Doclo, S. (2016). Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling. , *24*(2), 276–289.

Mohammadiha, N., Smaragdis, P., & Doclo, S. (2015). Joint acoustic and spectral modeling for speech dereverberation using non-negative representations. In *Proc. of ieee international conference on acoustics, speech and signal processing (ICASSP).*

Patrick, N., & Nikolay, G. (2010). *Speech dereverberation*. New York: Springer.

REVERB 2014. (n.d.). `http://reverb2014.dereverberation.com/workshop/proceedings.html`. (Online accessed: 2016-03-23)

Reverberation. (n.d.). `http://http://lossenderosstudio.com/newsletter.php?issue=66`. (Online accessed: 2016-08-16)

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

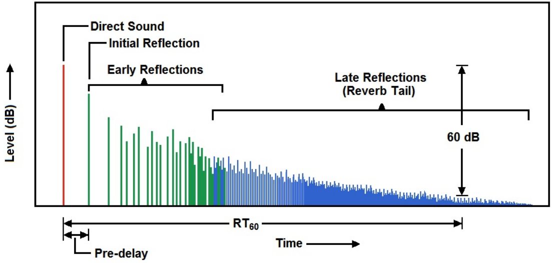
Reference V

Smaragdis, P. (2004, September 22). Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent component analysis and blind signal separation: Fifth international conference* (Vol. 3195, p. 494+). Springer-Verlag GmbH.

Yu, M. (2012). *Multi-channel speech enhancement by regularized optimization* (Unpublished doctoral dissertation). University of California, Irvine.

Yu, M., & Soong, F. K. (2014, May). Speech dereverberation by constrained and regularized multi-channel spectral decomposition: evaluated on REVERB challenge. In *Proc. reverb workshop.*

Room Impulse Response (RIR)



- Divided into two regions
  1. Early reflections - 50 ms after direct path
  2. Late reflections (reverberation tail) - beyond 50 ms
- Exponentially decaying envelope
- RIR parameters
  - Reverberation time (T_{60})
  - Source to microphone distance (d)[2]

[2] http://lossenderosstudio.com/newsletter.php?issue=66

Dereverberation

Problem

Given reverberated speech $y(n)$ or $Y(n,k)$ estimate clean speech $\hat{s}(n)$ or $\hat{S}(n,k)$

Challenges

- Unknown RIR and clean speech - blind deconvolution, ill-posed
- Requires meaningful constraints

Possible approaches

- Reverberation cancellation - estimate RIR and apply deconvolution
  - Multichannel equalization
  - NMF based approaches
- Reverberation suppression - suppress late reflections
  - Spectral subtraction
  - Linear prediction based methods

SRMR improvement for Different RIRs

Table: (T_{60}, d) of $(700ms, 2m)$, $(700ms, 0.5m)$, $(600ms, 2m)$, $(600ms, 0.5m)$

Methods	RIR_1	RIR_2	RIR_3	RIR_4
N-CTF	1.4844	1.7406	1.4180	1.4625
N-CTF + $\boldsymbol{H}_{sparse}$	1.4844	1.7406	1.4180	1.4625
N-CTF + $\boldsymbol{H}_{gain}$	1.1799	1.1144	1.2541	0.9284
N-CTF + $\boldsymbol{H}_{early}$	1.9197	2.1331	1.8982	1.8443
N-CTF + NMF	1.6489	1.6670	1.6754	1.4382
N-CTF + NMF + $\boldsymbol{H}_{gain}$	2.1398	1.4496	2.2688	1.2094
N-CTF + NMF + $\boldsymbol{H}_{sparse}$	1.6547	1.6436	1.6781	1.4297
N-CTF + NMF + $\boldsymbol{H}_{early}$	2.2210	1.8693	2.2526	1.6523