Speech Dereverberation

Bo Wen, Haiqin Yin & Meiying Chen

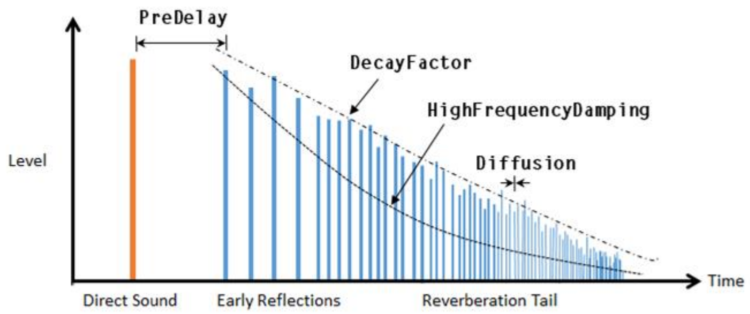
Outline

- Introduction
- Timeline
- Methods
- Summary
- Future work

Introduction - Reverberation

- What's Reverberation
  - Reverberation is the process of multi-path propagation of a sound from its source to one or more receivers
  - Effects on direct speech signal
    - Increase perceived distance
    - Reduce Intelligibility
    - Spectral distortion due to early reflection
- Reverberant signal *x(n)*
  - Anechoic speech signal *s(n)*
  - Acoustic Room Impulse Response (RIR) *h(n)*
  - Additive ambient noise component *v(n)*

$$x_m(n) = \mathbf{h}_m^T(n)\mathbf{s}(n) + \nu_m(n)$$



Introduction

- Problem identification
  - Sound quality & Intelligibility can degrade in reverberant environment
  - Enhance recordings in reverberant environment
- Application
  - Telecommunication
    - Hands -free phone
    - Desktop conference terminal
  - Reverb removal for recording
  - Automatic Speech Recognition

Introduction - Dereverberation

- What’s Dereverberation
    - Dereverberation is the process by which the effects of reverberation are removed from sound
    - Most commonly apply to speech

Timeline - History of Speech Dereverberation

GAUBITH
Spectral Processing
2008

HAN
DNN
2014

SANTOS
RNN
2017

ZHENG
Spectrogram Filtering
2018

ZHAO
RNN
2018

KILIS
Spectral Subtraction
2019

HAN
DNN
2019

2001
LEBART
Spectral Subtraction

2008
NAKATANI
Muti-channel Linear Prediction

2014
WENINGER
Deep BLSTM RNN

2017
WILLIAMSON
DNN

2018
MOHANAN
Non-convolutive NMF Model

2018
ESCUDERO
DNN

2019
ZHAO
DNN

Introduction - Goals

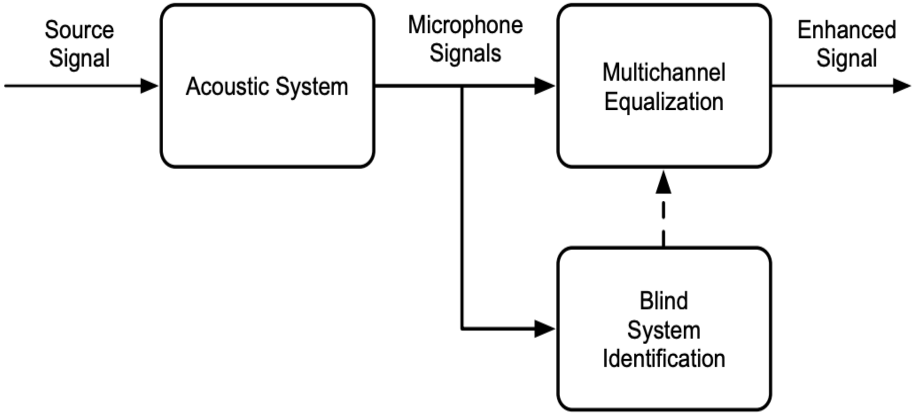
- Ultimate Goal - Complete Dereverberation:
  - Estimation of the anechoic speech signal $s(n)$
- Sufficient Goal - Partial Dereverberation:
  - Estimation of a filter of the anechoic speech signal $s(n)$

Method

- Three main approaches:
    - Reverberation Cancellation
    - Reverberation Suppression
    - Direct Signal Estimation

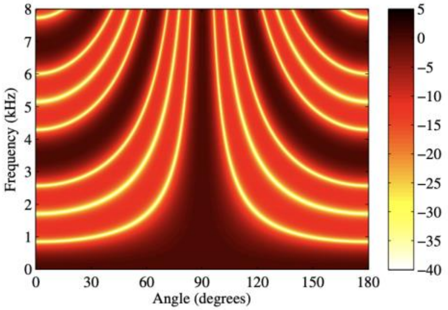
Method - Reverberation Cancellation

- The microphone signal is regarded as a delayed or filtered version of the source signal
    - Estimate of the acoustic impulse response (AIR) is known
    - Ultimate output signal is unknown
- To obtain an estimate of the desired signal:
    - Blindly identify the model parameters of the acoustic system
    - Apply a multichannel equalizer

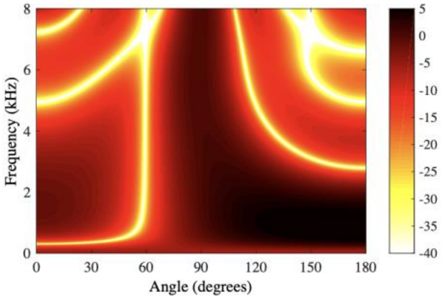


Method - Reverberation Cancellation

- Techniques
  - Inverse Filtering
  - Spatial Filtering
    - DOA (Direction-of-arrival) differ from direct path
    - Spatial filter is used to remove
- Problems:
  - Cause undesired signal coloration
  - High and unknown channel order
  - Hard to adapt to moving sources



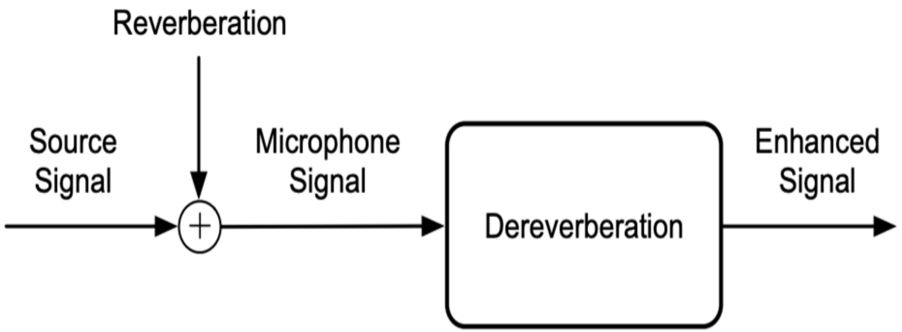
Before



After

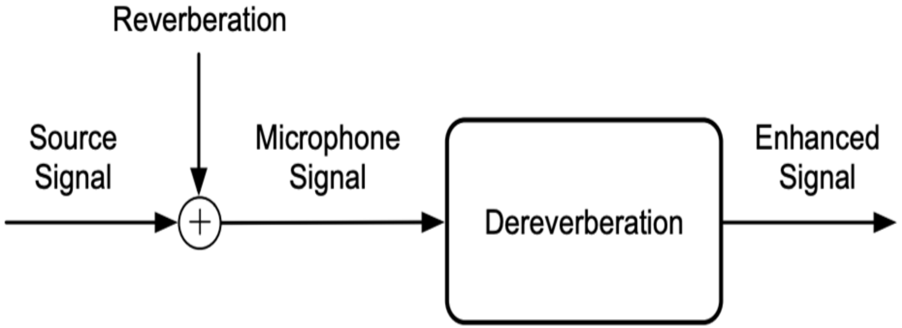
Method - Reverberation Suppression

- Model the reverberation as an additive process based on the assumption that reverberant signal is uncorrelated with direct signal
- Techniques
  - Spatial filtering techniques
  - Spectral enhancement/subtraction techniques



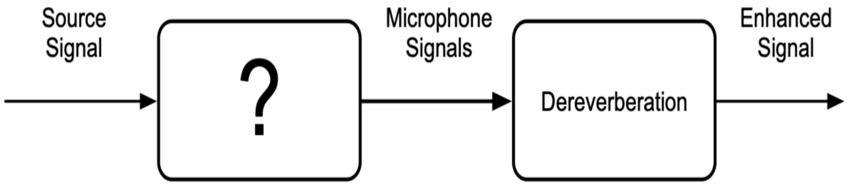
Method - Reverberation Suppression

- Advantages:
  - Effective for light reverberant speech signal
- Problems:
  - Only partial dereverberation is possibl
  - Require prior knowledge of source and channel
  - Introduce speech distortion



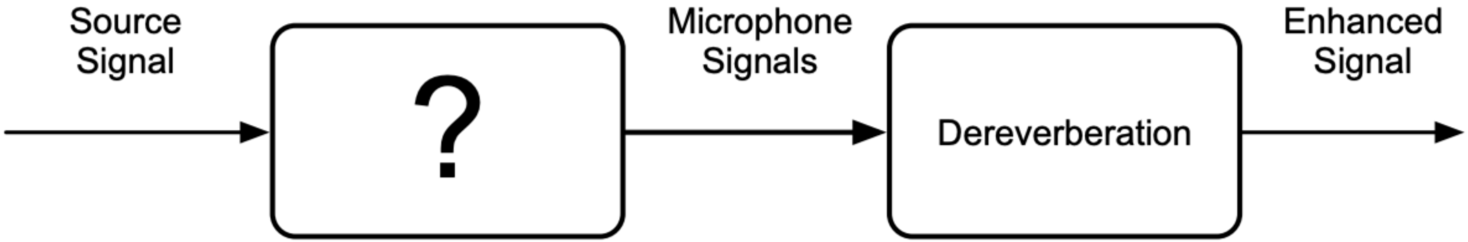
Method - Direct Signal Estimation

- Directly estimate the source signal from the microphone signals by regarding the acoustic system as unknown
- Techniques
  - Linear Prediction/LP residual processing
  - Temporal Envelope Processing
  - NMF - Non-negative Matrix Factorization
  - Deep Learning - Ideal Binary Mask (IBM)
    - CNN
    - RNN



Method - Direct Signal Estimation

- Problems:
  - Hard to train and generalize
  - Missing contextual information



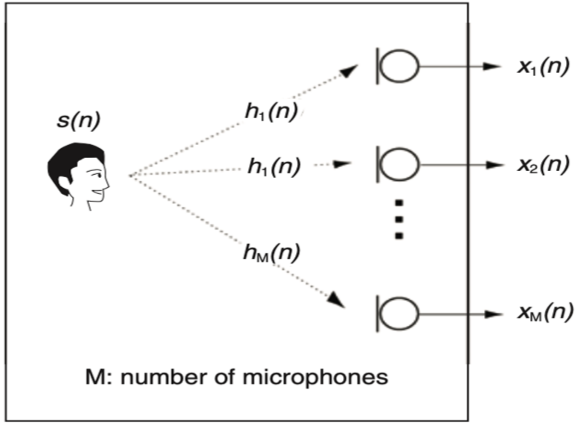
Spatiotemporal & spectral processing

Nikolay Gaubitch, Emanuel Habets & Patrick Naylor   IEEE 2018

- Reverberation:

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n) + \nu_m(n)$$

- Two-stage multi-microphone method
    - Stage I: Spatio-temporal Averaging Method
        - Early reflection
    - Stage II: Spectral subtraction
        - Late reverberation



Stage I - Spatio-temporal Averaging (SMERSH)

- Spatially averaged speech
- Compensate for the propagation time

$$\bar{x}(n) = \frac{1}{M} \sum_{m=1}^{M} x_m(n - \tau_m),$$

Stage I - Spatio-temporal Averaging (SMERSH)

Speech signal = linear predictor(LP) =

LP coefficient

Prediction residual

GCIs (glottal closure instance)

Periodic peak

Random peak

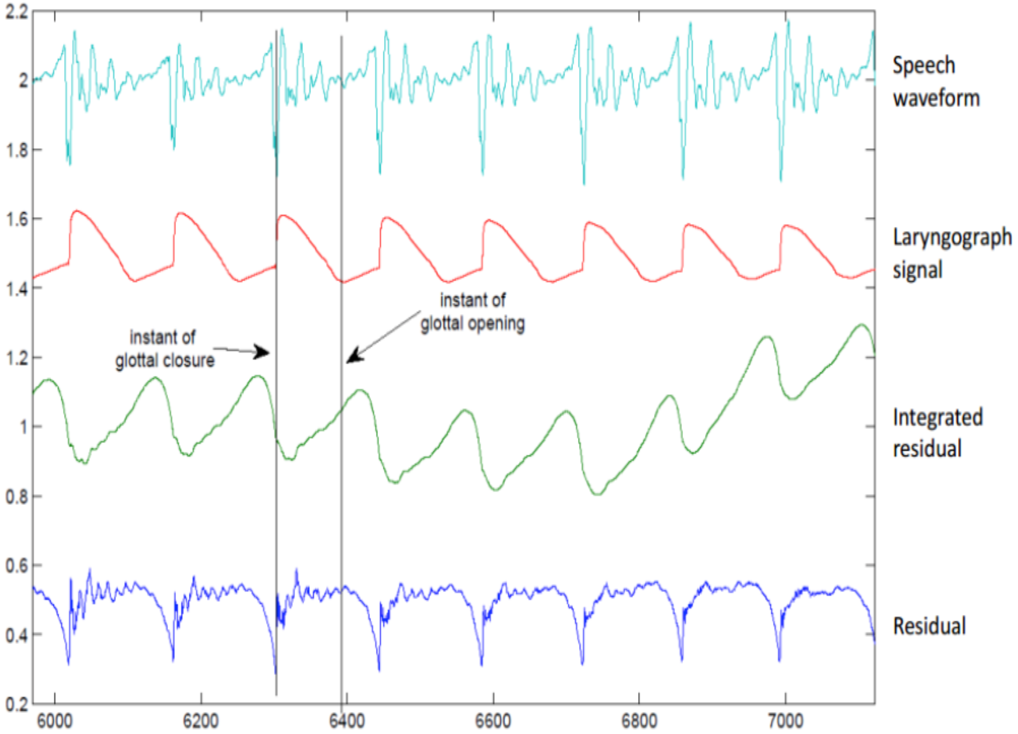
Similar strength

Affects

**Reverberatio
n**

What is GCIs

Glottal closure instants (GCIs) (also marks or epochs) refer to peaks in the speech signal that correspond glottal closure, a significant excitat tract



Stage I - Spatio-temporal Averaging (SMERSH)

- Apply a weight function to excluded the GCIs
- Averaging process
- Add an L-tap FIR filter to address unvoiced speech

Periodic peak

Random peak

Stage II - Spectral Subtraction

- Spectral subtraction assumes a statistical model of Room Impulse response (RIR), which is given by:
  - b(n) is a stationary zero mean white Gaussian noise sequence
  - δ is the room damping constant
  - T60 is the reverberation time
  - Fs is the sampling rate

$$h_n = \begin{cases} b(n)e^{-\delta n} & \text{for } n \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

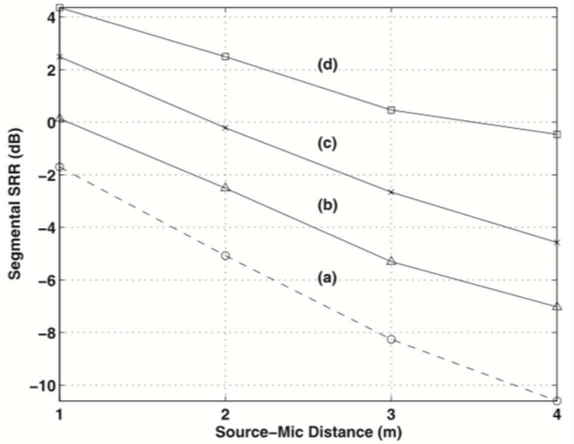
$$\delta = 3\ln(10)/(T_{60}f_s)$$

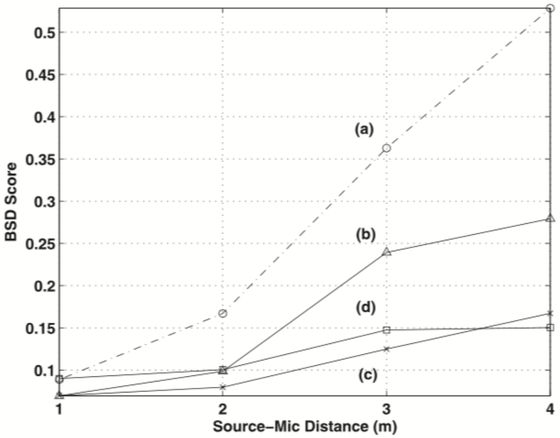
Stage II - Spectral Subtraction

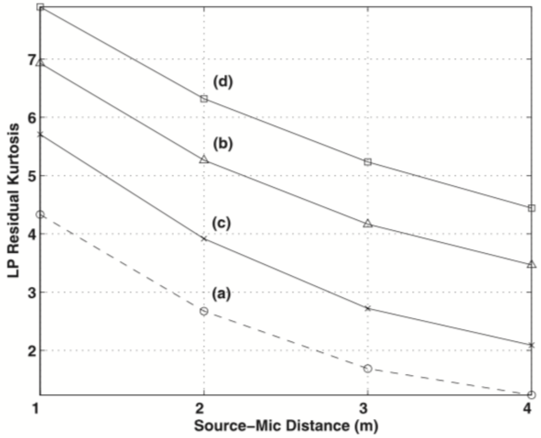
- Power spectral density(PSD) has additive property
- The direct component can be obtained by estimating and subtracting the late reverberant short-term power spectral density (PSD)

$$h_n = \begin{cases} b(n)e^{-\delta n} & \text{for } n \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

Evaluation







(a) reverberant (unprocessed) speech,
(b) speech processed with SMERSH,
(c) speech processed with spectral subtraction (using only one microphone)
(d) the combination of SMERSH and spectral subtraction.

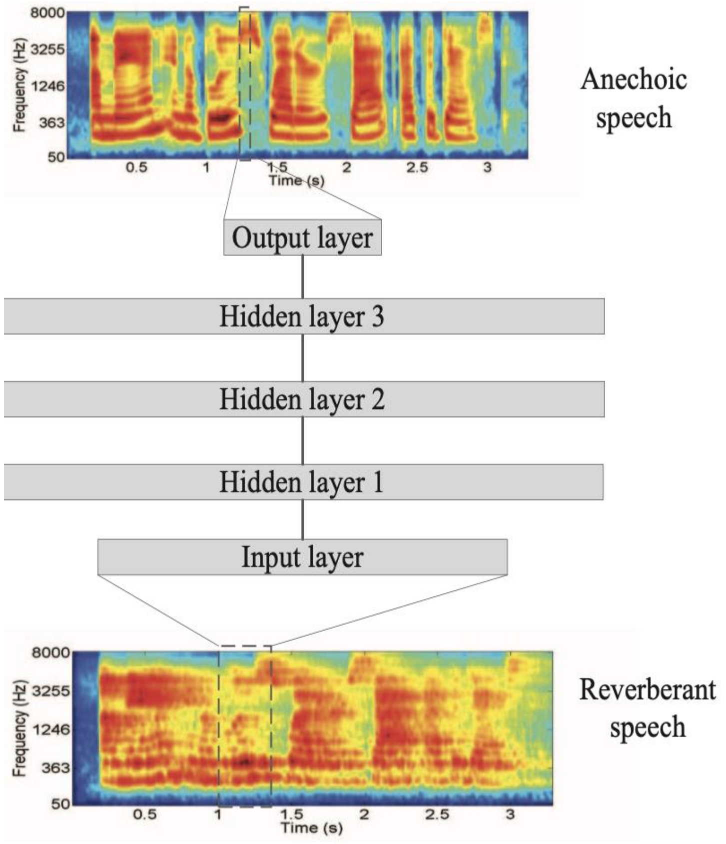
CNN - Learning Spectral Mapping

Kun Han, Yuxuan Wang, Deliang Wang ICASSP 2014

- The magnitude relationship between anechoic and reverberant signal is consistent, especially within the same room
- Learning a spectral mapping from the reverberant speech to regenerate the anechoic speech
- Methodology:
  - Ideal binary mask:
    - target -> direct sound + early reflections
    - mask -> the late reflection

Model Design

- Spectral features
  - Input:
    - Gammatone filterbank + framing
    - Neighboring frames are also considered
    - $\tilde{x}(m) = [x(m-d), \ldots, x(m), \ldots, x(m+d)]^T$
  - Output: 64d vector y(m)
- CNN based spectral mapping
  - Training : Pre-train with RBM + fine tuning (+ two regularizations)



Evaluation

- Traditional Models
  - Inverse filter must be estimated, which is not a trivial problem
  - Assumes that the RIR function is a minimum-phase function that is often not satisfied in practice
- CNN
  - Simple and efficiency, became new SOTA
  - With good generalization ability

Evaluation

- Use synthetic signals to train and test, and the dataset is small (200)
- Neighboring frames issue
  - Previous or succeeding frames
  - Number of neighboring frames chosen
  - Importance of neighboring frames varies at different evaluation location

Summary

- Reverberation Cancellation
- Reverberation Suppression
- **Direct Signal Estimation**
    - **Linear Prediction (by wikipedia)**

- Evaluation Methods
    - DRR - Direct to Reverberation Ratio
    - SRMR - Speech to Reverberation Modulation Energy Ratio
    - STOI - Short-Time Objective Intelligibility
    - BSD - Bark Spectral Distortion (incorporate psychoacoustic response)
    - LP Residual Kurtosis
    - PESQ - Perceptual Evaluation of Speech Quality

Futurework

- Performance Evaluation Metrics
- Reduce early reflection
- Lower DRR & SNRs
- Binaural dereverberation

Reference

Gaubitch, N. D., Habets, E. A. P., & Naylor, P. A. (2008). Multimicrophone speech dereverberation using spatiotemporal and spectral processing. *2008 IEEE International Symposium on Circuits and Systems*. doi: 10.1109/iscas.2008.4542144

Han, K., Wang, Y., & Wang, D. (2014). Learning spectral mapping for speech dereverberation. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/icassp.2014.6854479

M, N., Velmurugan, R., & Rao, P. (2018). A Non-convolutive NMF Model for Speech Dereverberation. *Interspeech 2018*. doi: 10.21437/interspeech.2018-1834

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., & Juang, B.-H. (2008). Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. doi: 10.1109/icassp.2008.4517552

Q&A

Thank you!