THE ANDREW & ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

SIPL
Signal and Image Processing Lab

TECHNION
Israel Institute of Technology

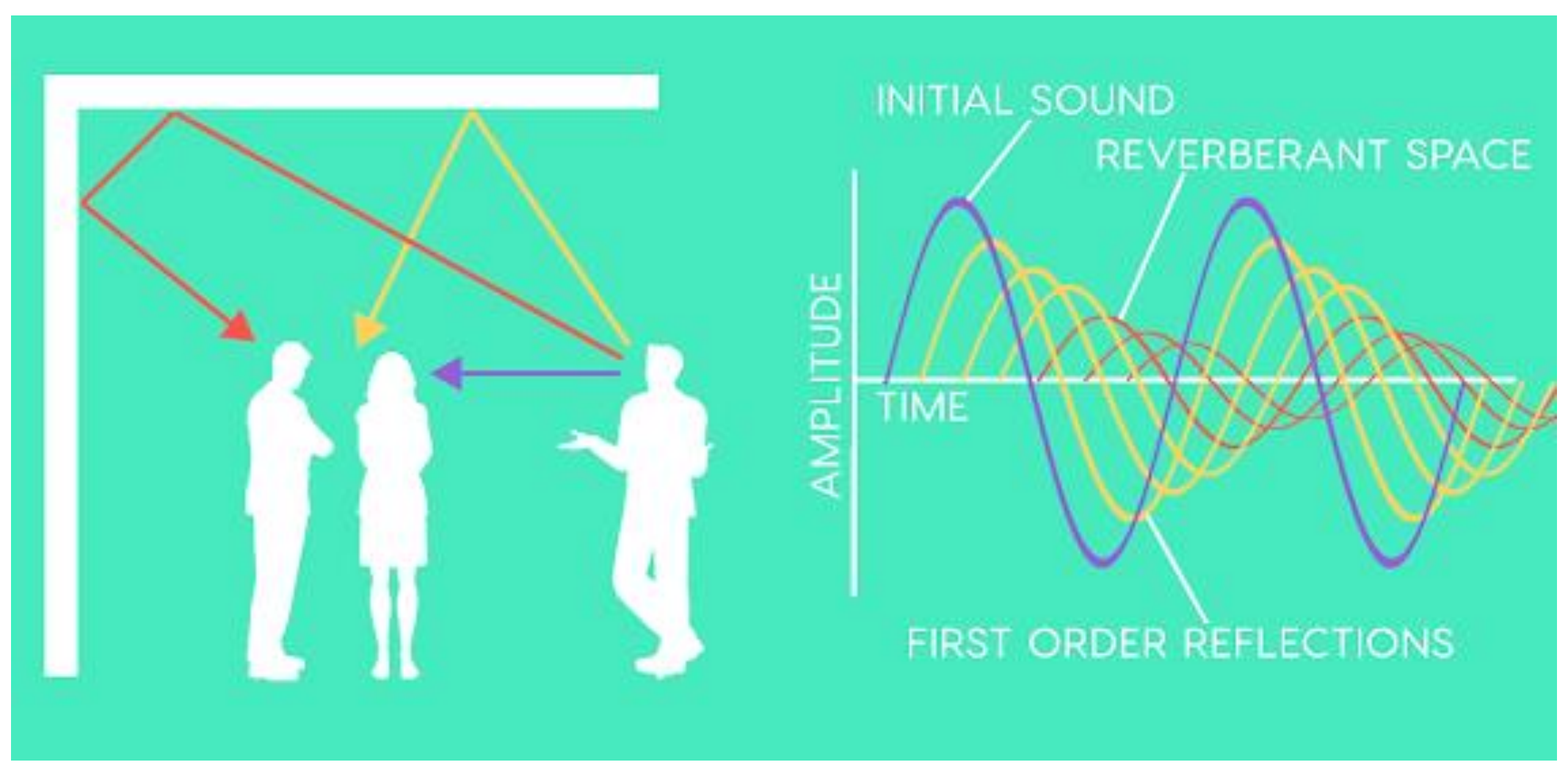
Work in Progress

Deep Learning Based Target Cancellation for Speech Dereverberation

Neriya Golan and Mikhail Klinov
Supervised by Yair Moshe & Baruch Berdugo

Introduction

- Reverberation is the process of multi-path propagation of a sound from its source to a receiver
- Reverberation reduces speech intelligibility, can cause hearing aids to malfunction, impairs performance of speech recognition and more…



The process of reverberation

Project Goals

- Dereverberation of speech signals using deep learning methods
  - Single channel – start with a solution that works for reverberations recorded by a single microphone
  - Dual channel – expand the solution to two microphones, placed at a certain configuration
  - Real-time – the process of dereverberation should not cause a distinguishable delay

Challenges

- Existing DNN-based solutions are usually complex and lack code, making the results hard to reproduce
- Implementation of an end-to-end DNN-based solution from scratch

Dataset

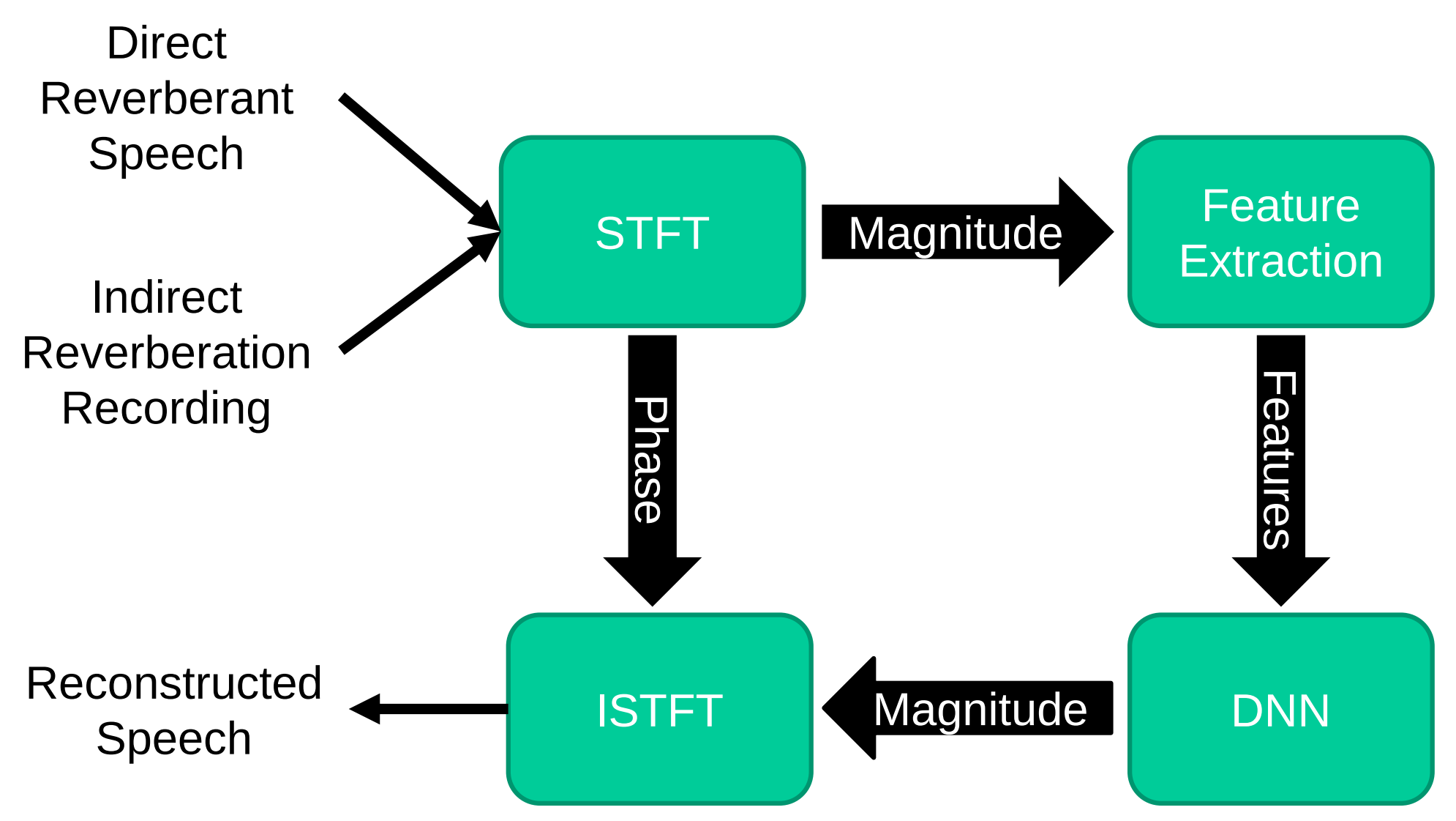
- TIMIT Corpus - 6300 speech recordings carried by 630 speakers from 8 different American dialect regions

Sentence Type	#Sentences	#Speakers	Total	#Sentences per Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342		6300	10

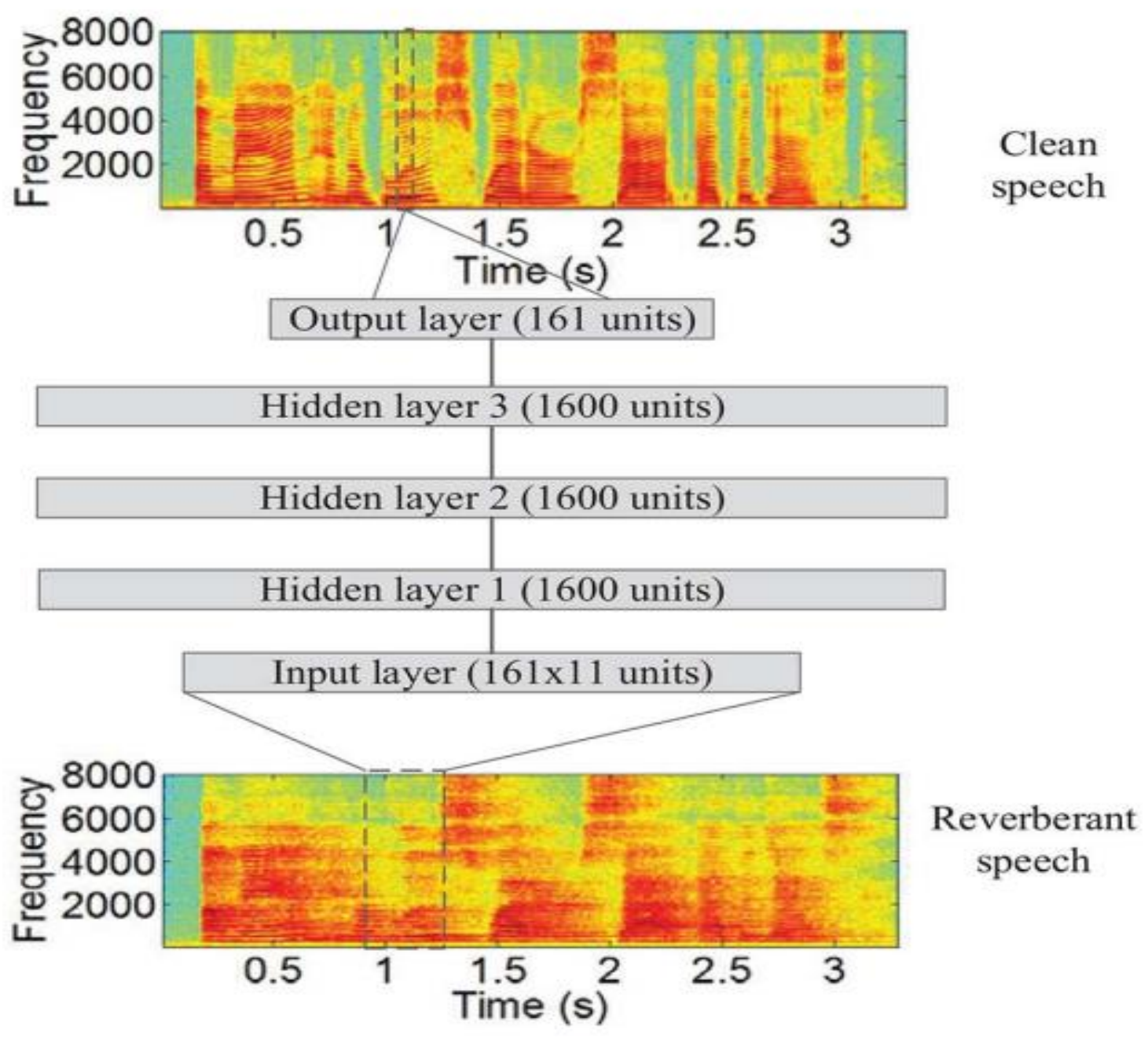
Reverberant Speech

- Simulated using Room Impulse Response Generator by [Jarrett et al., 2012]
- Simulation of 2 different rooms and 3 different reverberation times (T60) – 0.3, 0.6, 0.9 sec for training and 0.2, 0.5, 0.8 sec for testing

Chosen Solution

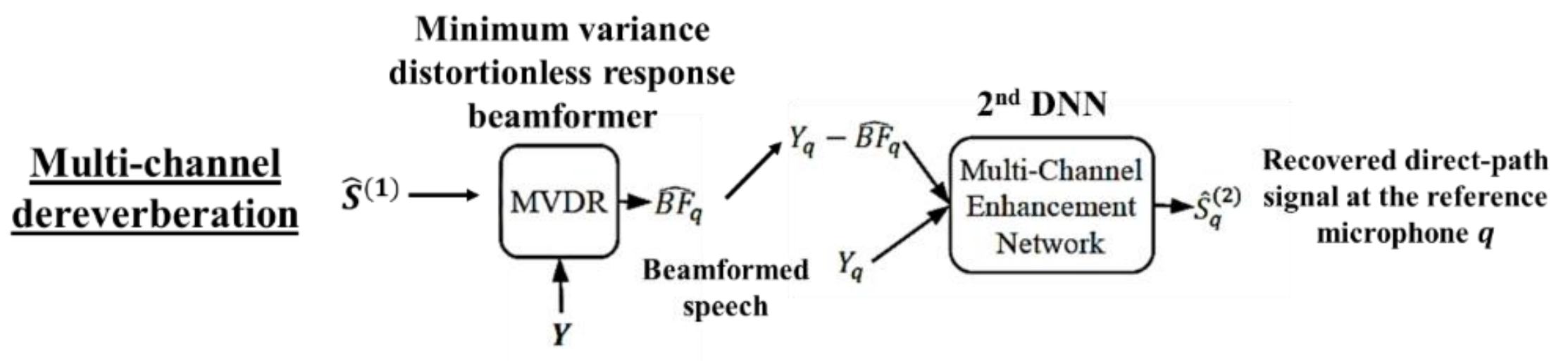


- STFT domain input-output
- Feed Forward DNN based on [Han et al.,2015]


Frequency
8000
6000
4000
2000
0.5
1
1.5
2
2.5
3
Time (s)
Clean speech
Output layer (161 units)
Hidden layer 3 (1600 units)
Hidden layer 2 (1600 units)
Hidden layer 1 (1600 units)
Input layer (161x11 units)
Reverberant speech

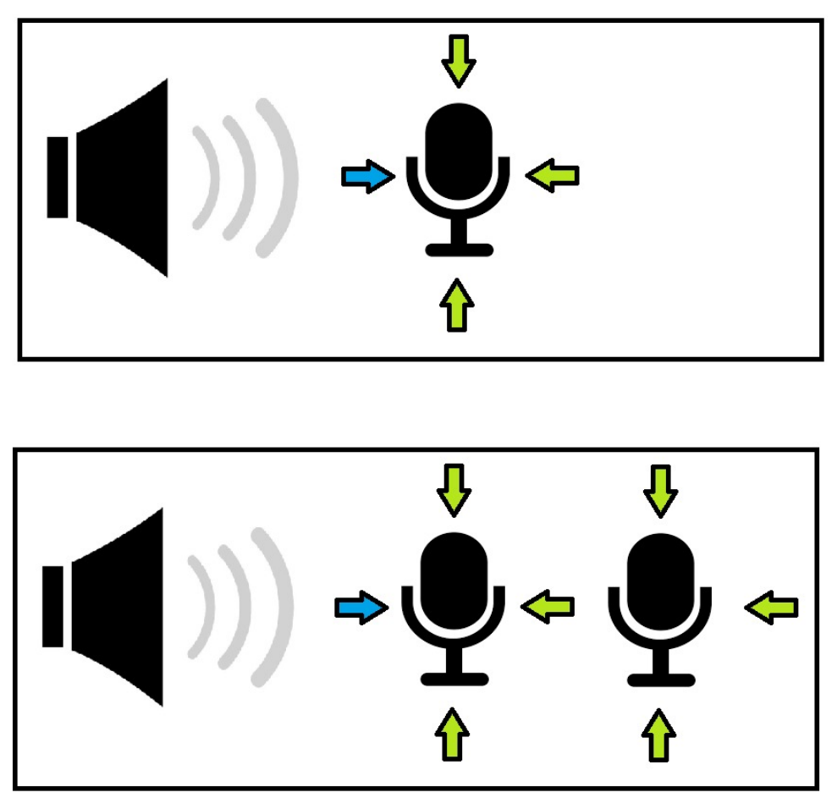
Dual Channel Approach

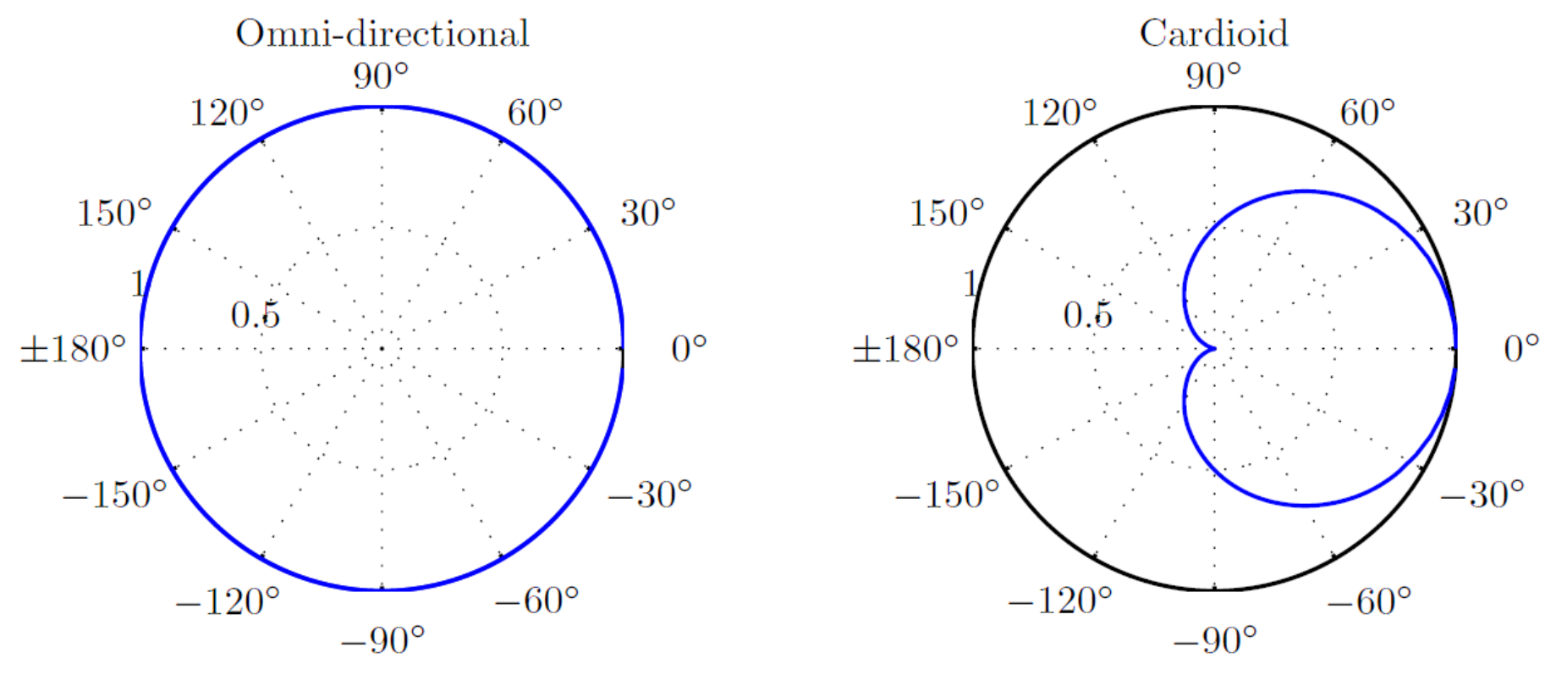
- It is possible to enhance performance by feeding the DNN a 2nd input – consisting of reverberations only (without the direct path signal)
- This is usually done by a beamformer which utilizes a microphone array to derive the reverberations


Multi-channel dereverberation
$\hat{S}^{(1)}$
Minimum variance distortionless response beamformer
MVDR
Y
$\widehat{BF}_q$
Beamformed speech
$Y_q - \widehat{BF}_q$
Y_q
2nd DNN
Multi-Channel Enhancement Network
$\hat{S}_q^{(2)}$
Recovered direct-path signal at the reference microphone q

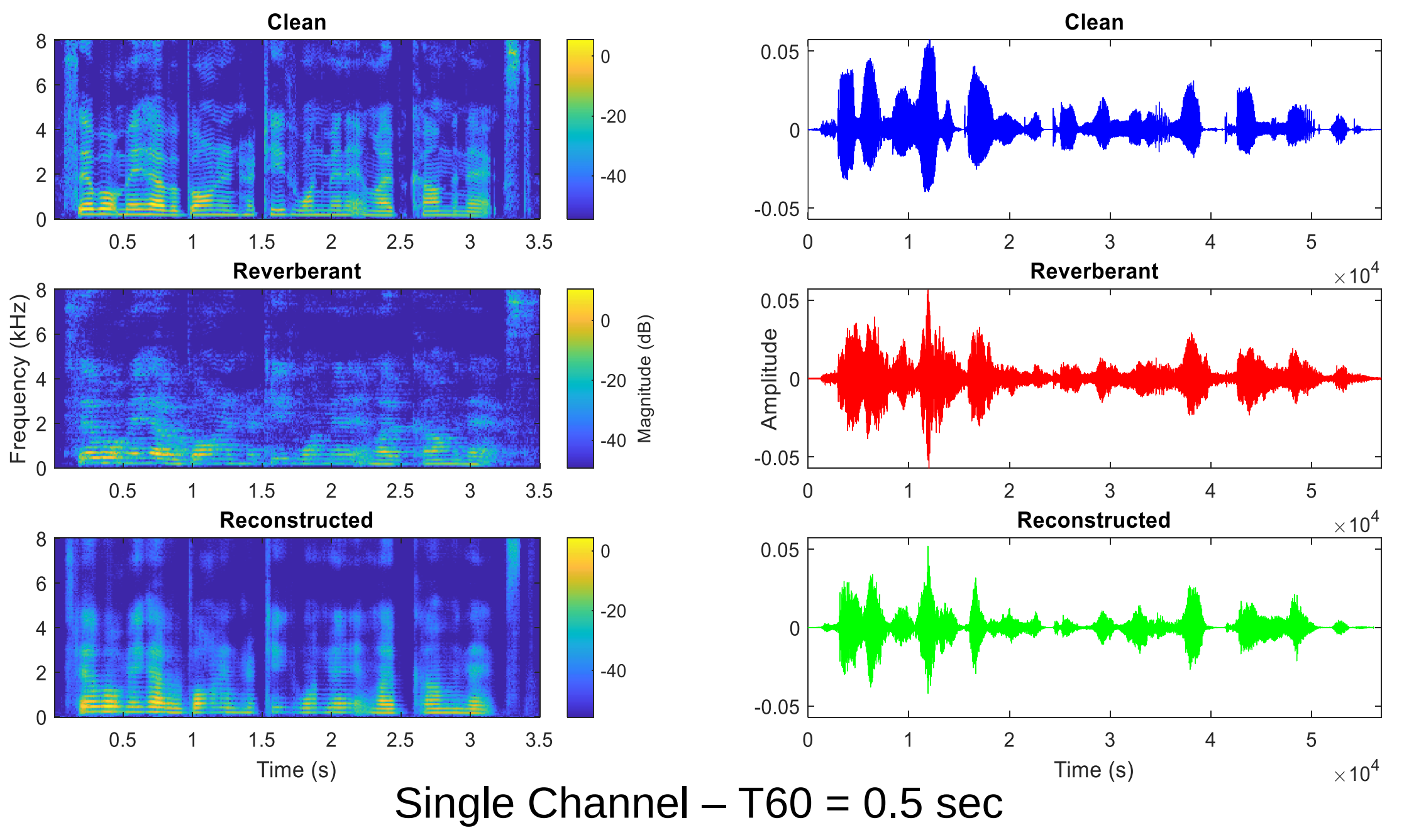
Solution by [Zhong-Qiu Wang et al., 2020]

- We propose to replace the microphone array by adding a single microphone which is directed at the opposite direction of the speaker, hence “hears” only reverberations

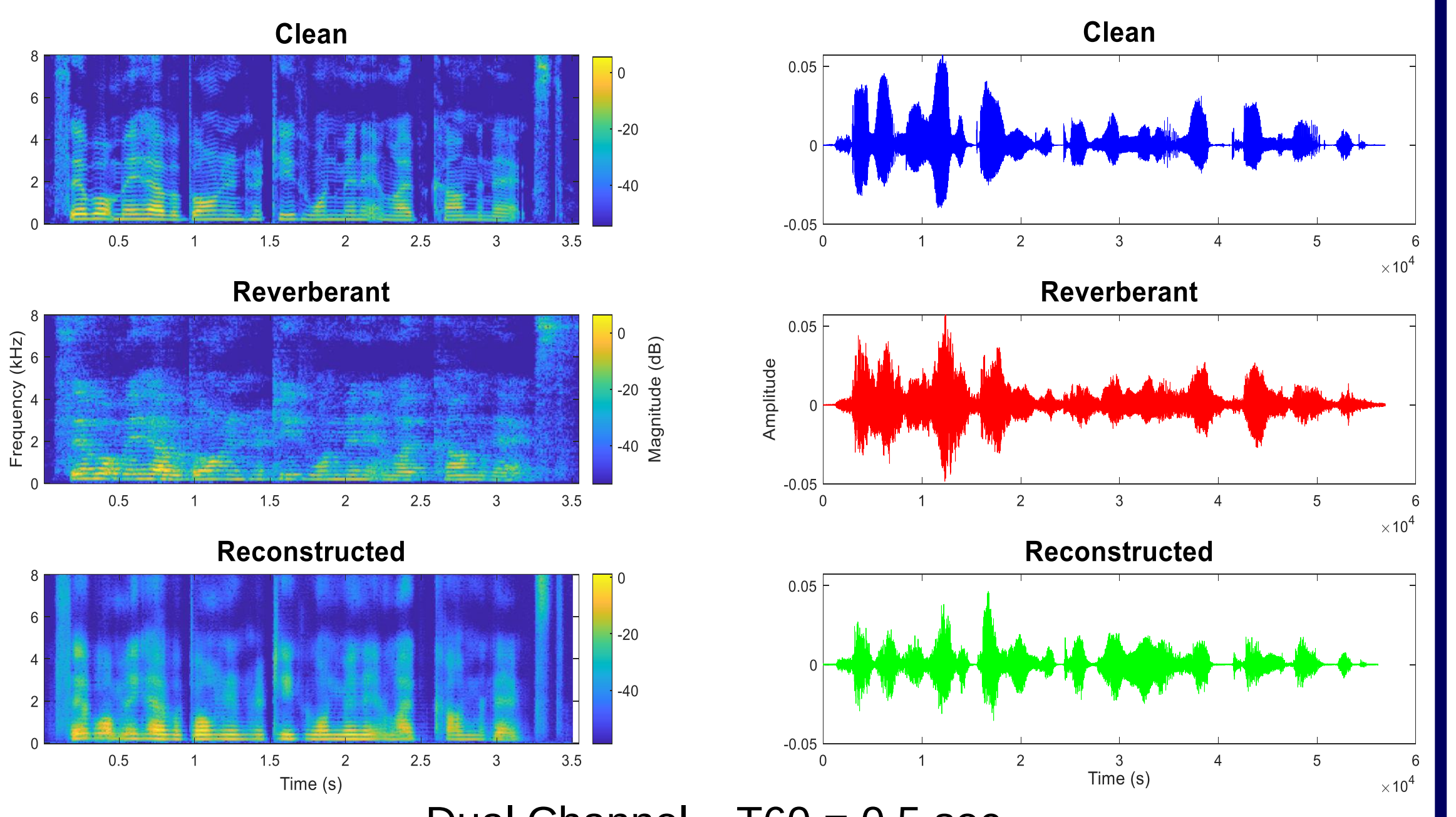




Results



Single Channel – T60 = 0.5 sec



Dual Channel – T60 = 0.5 sec

Speech Quality

Clean		T60 = 0.2 sec	T60 = 0.5 sec	T60 = 0.8 sec
3.77	Rev	3.71	3.10	3.03
	Rec - Single	3.06	2.62	2.65
	Rec – Dual	2.99	2.58	2.53

DNSMOS Score – Speech Quality Metric – Higher is better

Spectral Distance

	T60 = 0.2 sec	T60 = 0.5 sec	T60 = 0.8 sec
Rev	3.29	4.56	5.10
Rec - Single	2.97	3.47	3.93
Rec – Dual	3.09	3.24	3.34

Mean distance between spectrograms – Lower is better

Conclusions

- Successful removal of reverberations, at a cost of decreasing speech quality caused by phase – magnitude inconsistency
- The dual channel approach:
  - Insufficient results
  - Should be tested in more complex architectures
- Work in progress:
  - Pseudo-phase sensitive loss functions for training
  - Advanced STFT reconstruction – HiFi-GAN [Kong et al., 2020]

June 2021