THE NTU-ADSC SYSTEMS FOR REVERBERATION CHALLENGE 2014

presented by

<u>Xiong Xiao</u>[1], *Shengkui Zhao*[2], *Duc Hoang Ha Nguyen*[3], *Xionghu Zhong*[3], *Douglas L. Jones*[2], *Eng Siong Chng*[1,3], *Haizhou Li*[1,3,4]

[1]Temasek Lab@NTU, Nanyang Technological University, Singapore.
[2]Advanced Digital Sciences Center, Singapore.
[3]School of Computer Engineering, Nanyang Technological University, Singapore.
[4]Department of Human Language Technology, Institute for Infocomm Research, Singapore.

Outline

- System Highlights
- Speech Enhancement
  - Delay and Sum + spectral subtraction
  - MVDR + DNN spectrogram enhancement
- Speech Recognition
  - Multi condition training
  - Clean condition training
- Summary

System Highlights

- Beamforming
  - Delay and Sum, MVDR
  - Classic method, always works!

- **DNN feature mapping**
  - Mapping reverberant spectrogram to clean spectrogram for enhancement
  - Mapping reverberant MFCC features to clean features for ASR

- DNN acoustic modeling for ASR
  - Discriminative feature learning and modeling in a single framework.

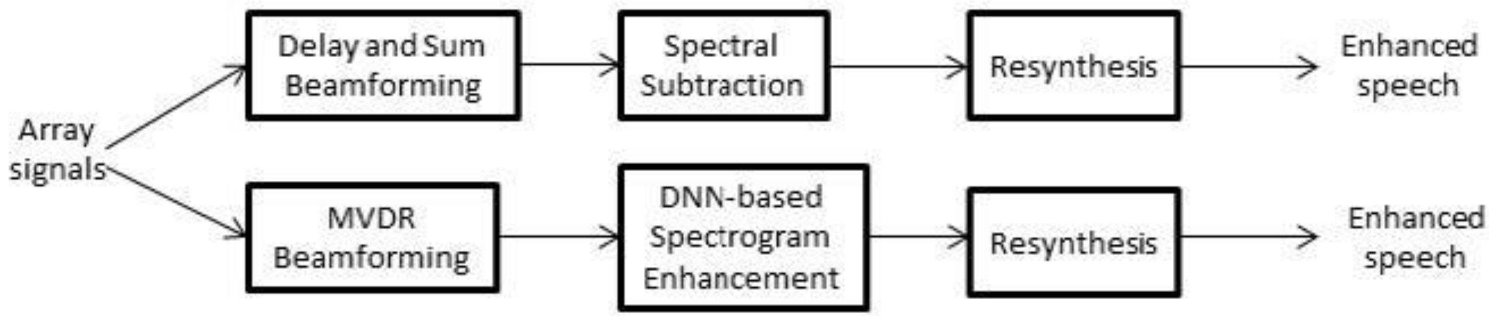
- **Feature adaptation (Cross-transform) for ASR**
  - a generalization of temporal filter and fMLLR transform.
  - explicitly use the correlation between feature frames to counter distortions that have effects over many frames.

Outline

- System Highlights
- Speech Enhancement
  - Delay and Sum + spectral subtraction
  - MVDR + DNN spectrogram enhancement
- Speech Recognition
  - Multi condition training
  - Clean condition training
- Summary

NANYANG TECHNOLOGICAL UNIVERSITY

Speech Enhancement Systems

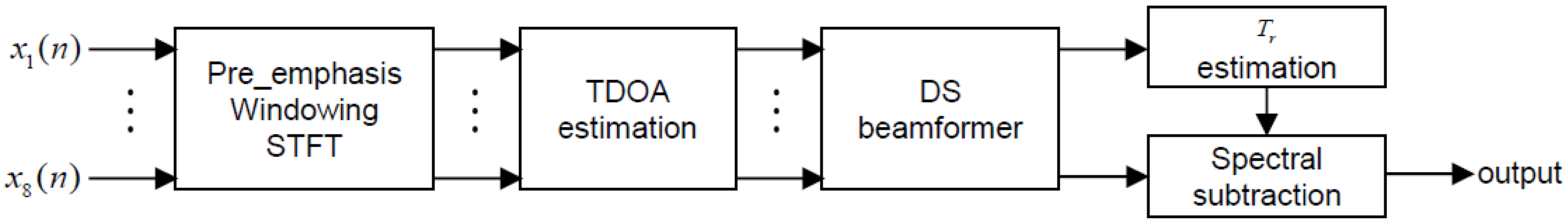


Two speech enhancement systems are considered:

- **DS beamforming + spectral subtraction (DS+SS);**
- **MVDR beamforming + DNN based spectrogram enhancement (MVDR + DNN).**

NANYANG TECHNOLOGICAL UNIVERSITY

Speech Enhancement – DS + Spectral Susbtraction



- ❑ **DS beamforming**
  - ➢ **Windowing STFT: 64ms Hanning window,**
  - ➢ **GCC-PHAT for TDOA estimation,**
  - ➢ **Multi-channel phase alignment and sum.**

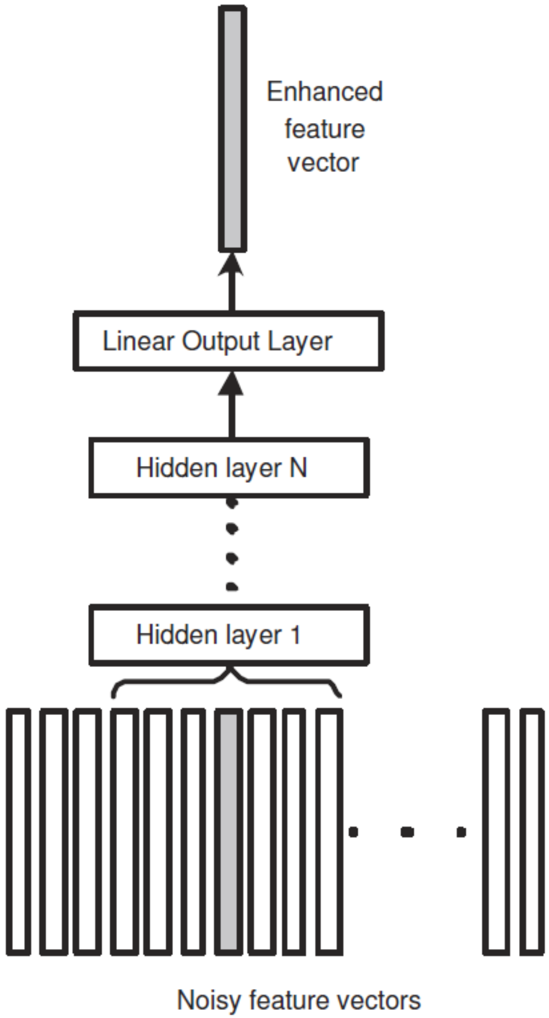
  75% frame overlap, 1024 point STFT.

- ❑ **Spectral Subtraction**
  - ➢ **Reverberation time estimation: ML method.**
  - ➢ **Amplitude spectral subtraction.**

NANYANG TECHNOLOGICAL UNIVERSITY

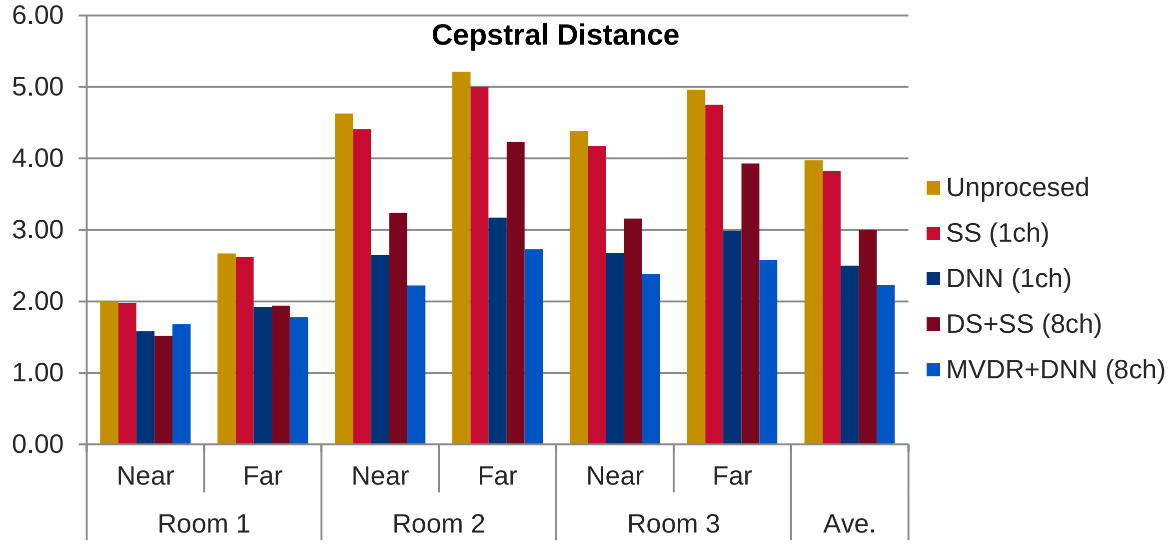
Speech Enhancement – MVDR + DNN feature mapping

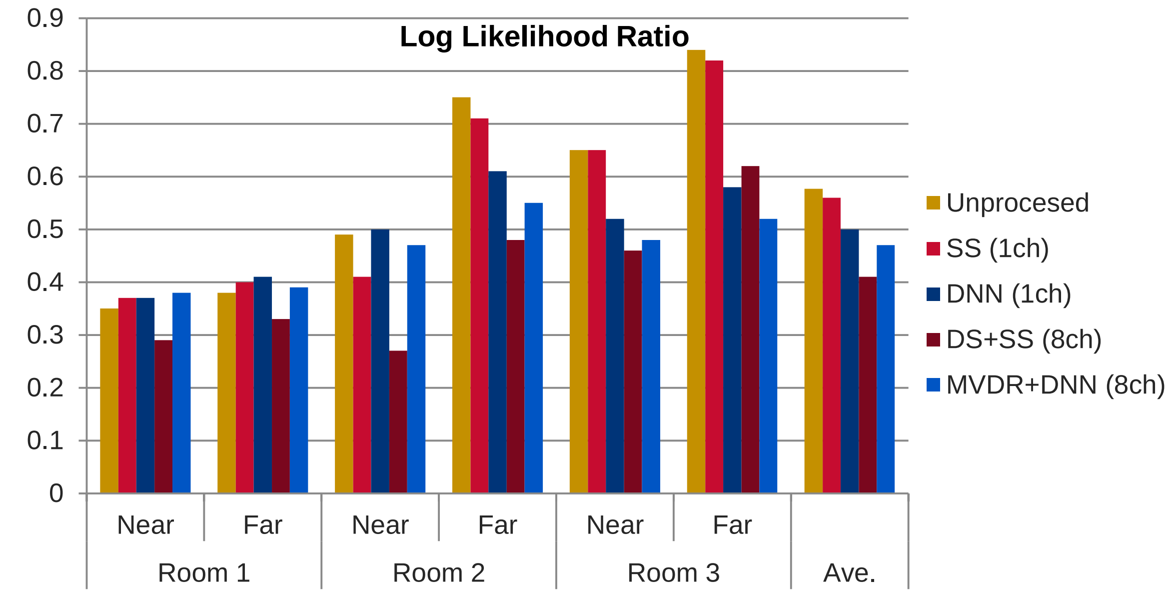
- Use DNN to map a window of reverberant feature vectors to a (central) clean feature vector.
- Let DNN learn to do dereverberation.
- For speech enhancement, input and output are spectrum vectors.
- For ASR, input and out are MFCC feature vectors.
- Training data: frame aligned clean and multi-condition data.
- **DNN size: 2827– 3x3072 – 771**
  - Predict both static and dynamic spectrum, then merge them to produce smoothed static spectrum.



NANYANG TECHNOLOGICAL UNIVERSITY

Objective measures – CD and LLR



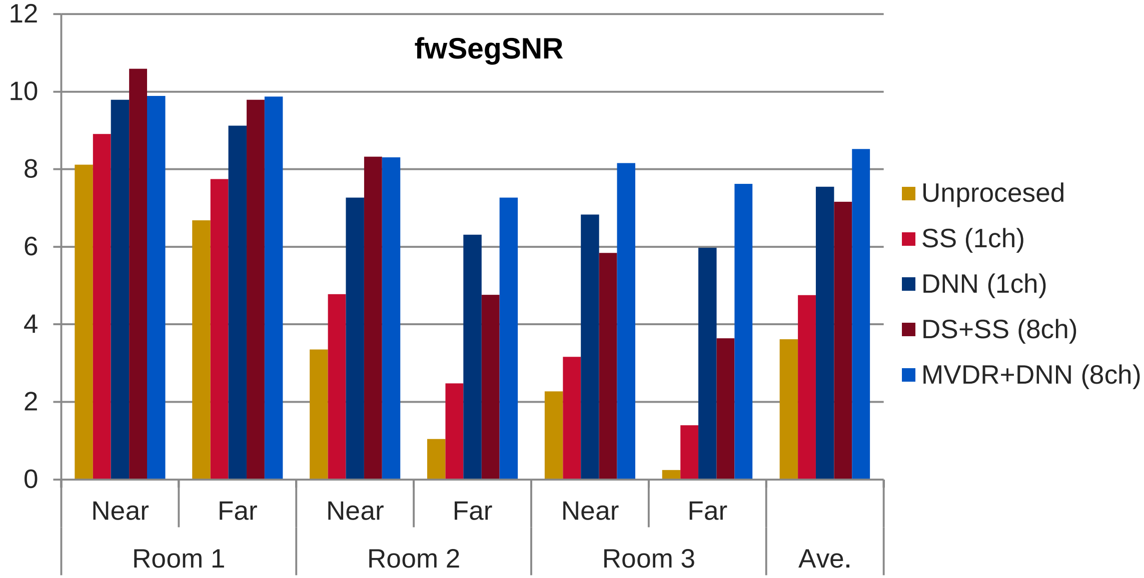


Both DS+SS and MVDR+DNN reduces cepstral distances and LLR significantly, especially for high reverberation cases.

DNN degrades LLR significantly for 8-ch low reverberation cases.

NANYANG TECHNOLOGICAL UNIVERSITY

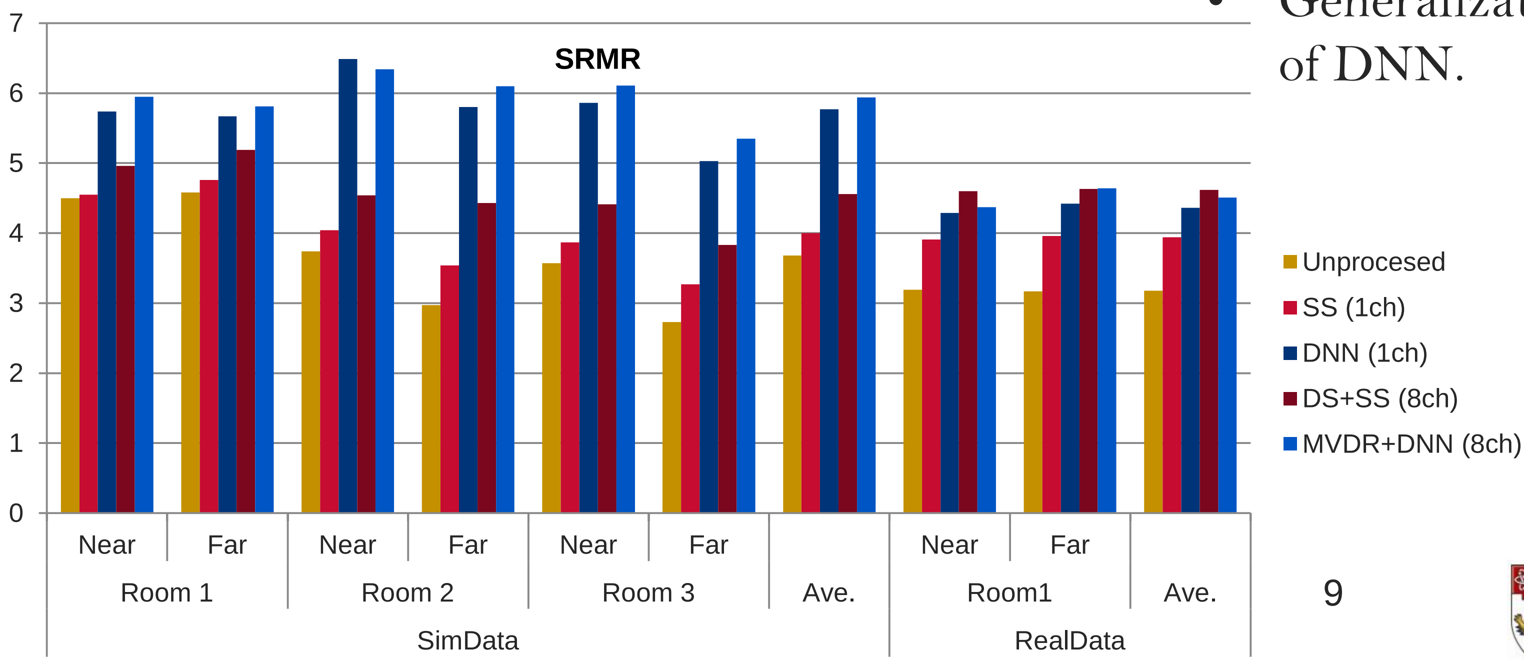
Objective measures – fwSegSNR and SRMR



DNN improves fwSegSNR for most cases.

DNN has smaller improvements in SRMR for real data.

- Generalization problem of DNN.



NANYANG TECHNOLOGICAL UNIVERSITY

Subjective measures

Amont of Reverberation Score						
			Mean			
			Simulated Room 2		RealData Room1	
			Near	Far	Near	Far
1ch		Unprocessed	41.5	31.0	28.9	21.5
	SS	Processed	52.6	42.7	37.8	38.6
		Improvement	***11.1***	***11.7***	***8.9***	***17.2***
	DNN	Processed	59.3	51.7	63.9	63.5
		Improvement	**17.8**	**20.7**	**35.0**	**42.0**
8ch		Unprocessed	21.5	18.9	14.6	16.6
	DS+SS	Processed	47.4	42.1	42.2	30.7
		Improvement	***25.9***	***23.2***	***27.6***	***14.1***
	MVDR+DNN	Processed	83.3	50.1	50.2	29.4
		Improvement	**61.8**	**31.2**	**35.6**	**12.9**

Overall Quality Score						
			Mean			
			Simulated Room 2		RealData Room1	
			Near	Far	Near	Far
1ch		Unprocessed	36.7	46.3	51.9	42.9
	SS	Processed	47.9	47.4	45.6	50.2
		Improvement	***11.2***	***1.1***	***-6.3***	***7.3***
	DNN	Processed	19.6	16.6	16.7	16.4
		Improvement	**-17.1**	**-29.7**	**-35.3**	**-26.5**
8ch		Unprocessed	37.0	33.8	30.6	25.3
	DS+SS	Processed	57.8	55.8	52.0	43.9
		Improvement	***20.8***	***22.0***	***21.4***	***18.6***
	MVDR+DNN	Processed	31.9	20.7	15.5	9.3
		Improvement	**-5.1**	**-13.2**	**-15.1**	**-16.0**

MVDR+DNN generally removes more reverberation than DS+SS.

But it also introduces more speech distortion and results in poorer quality.

Reasons?

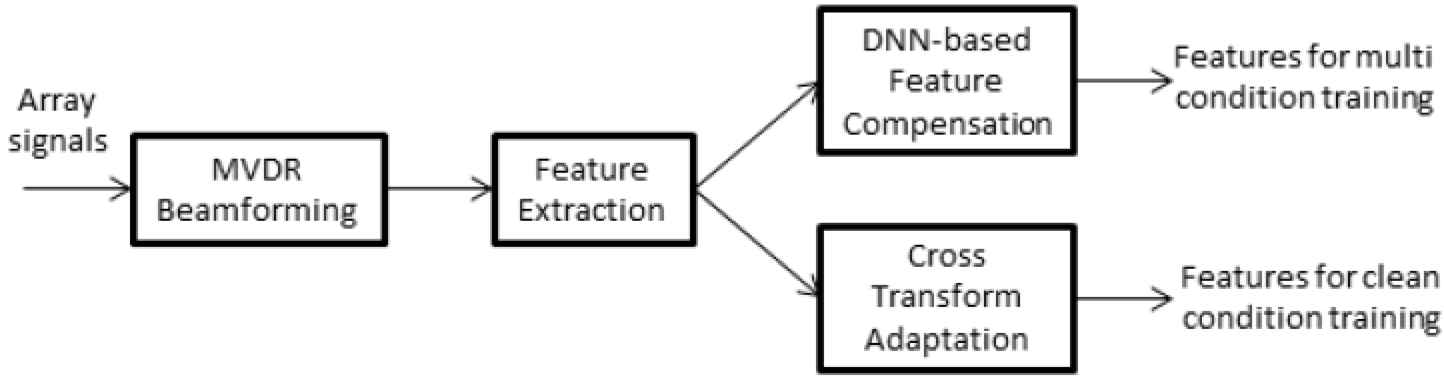
- Frame-by-frame processing of DNN.
- DNN reduces mean square errors between predicted log spectrum and clean log spectrum, not a perceptually meaningful error.

NANYANG TECHNOLOGICAL UNIVERSITY

Outline

- System Highlights
- Speech Enhancement
  - Delay and Sum + spectral subtraction
  - MVDR + DNN spectrogram enhancement
- Speech Recognition
  - Multi condition training
  - Clean condition training
- Summary

Speech Recognition Systems

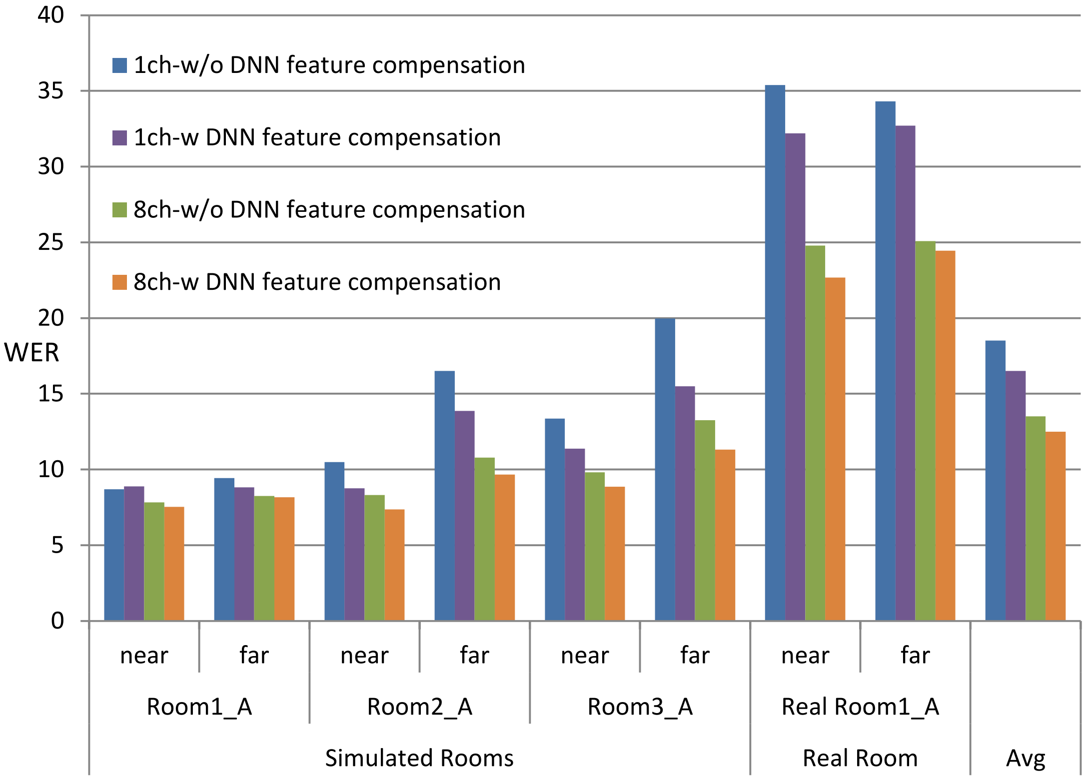


- MVDR beamforming for 2ch and 8ch.
- Clean condition training scheme
  - Cross Transform Adaptation
  - CMLLR (256 class) model adaptation.
  - HMM/GMM model (the challenge baseline settings)
- Multi condition training scheme
  - DNN based feature compensation
  - DNN based acoustic modeling

NANYANG TECHNOLOGICAL UNIVERSITY

ASR - Multi-condition training – results

- DNN feature mapping (585-3x2048-39)
- DNN acoustic modeling (351-7x2048-3500, RBM pretraining + CrossEntropy + SMBR)



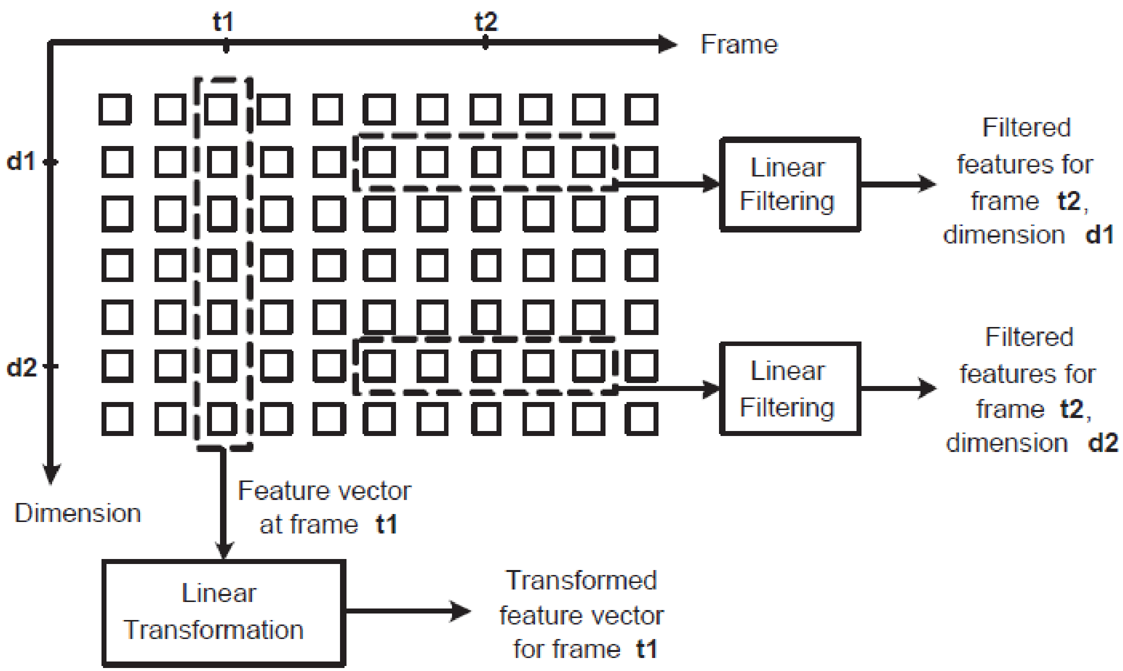
DNN feature compensation and DNN acoustic model are complementary.

Reason?

- DNN feature compensation uses parallel corpus and wider context.
- Good to have a two concatenated DNN architecture than a big DNN?

ASR - Clean-condition training

- Use cross transform for feature compensation
- Use CMLLR for model adaptation (challenge script)
- HMM/GMM system (challenge script)



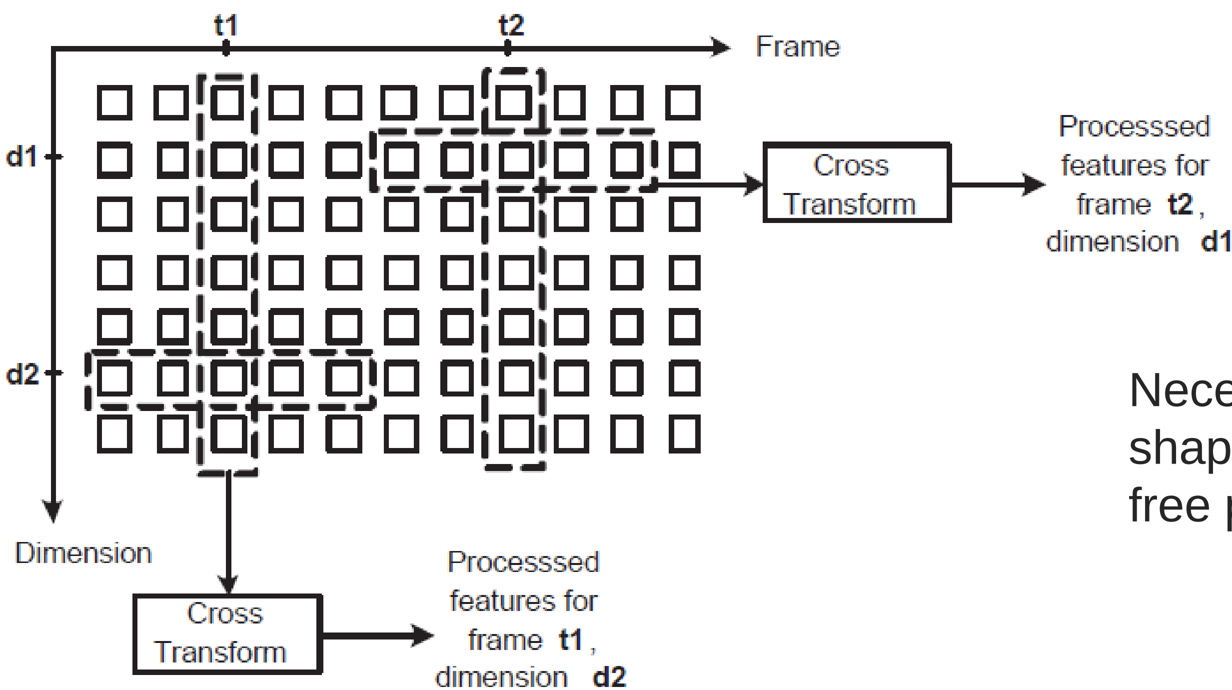
Temporal filtering processes the feature trajectories.

Linear transform processes feature vectors.

How about combine them?

ASR – Cross-transform

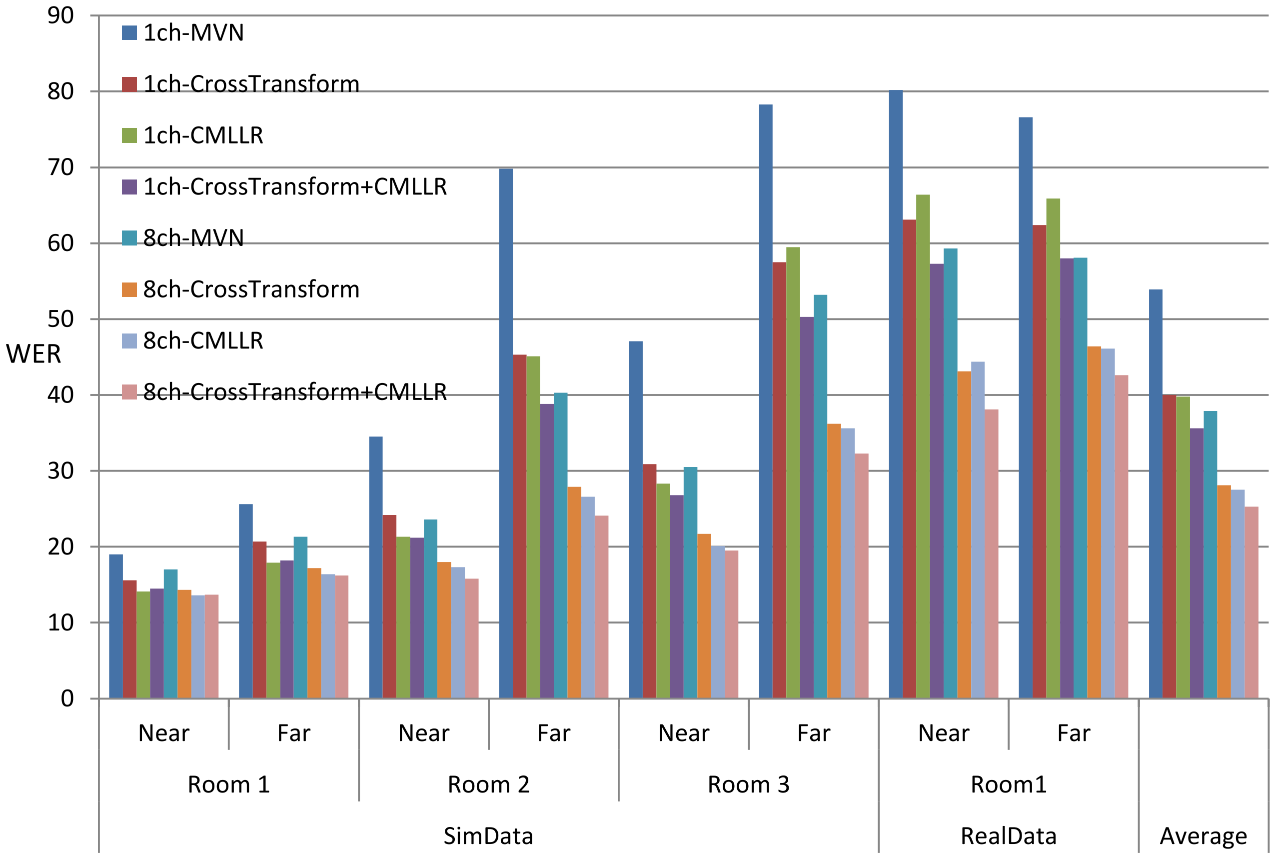
- Cross-transform is a generalization of both temporal filtering and linear transform.
- To adapt the features at a time-frequency location, both the feature vector and feature trajectory that contains the location are used in the regression.



Necessary to take the cross-shape to reduce the number of free parameters.

ASR - Clean-condition training – Results

- Cross-transform (33 frame window size, batch mode)
- CMLLR (256 class, batch mode)
- HMM/GMM system (Challenge scripts)



Cross-transform and CMLLR model adaptation are complementary.

Reason:

- Cross-transform uses longer context size.
- Multi-class CMLLR is more flexible: different transform for different classes.

Summary

- Traditional beamforming works well for both speech enhancement and recognition.

- DNN reduces reverberation significantly, but also introduces high distortion especially in high reverberation cases.

- Cross-transform adapts features using both long term temporal information and spectral information. Complementary to CMLLR.

- Future directions
  - Analyze why DNN produces distortions to speech signal and propose solution.
  - Apply cross-transform to adaptive training of DNN based acoustic model in multi-condition training scheme.

Thank you!