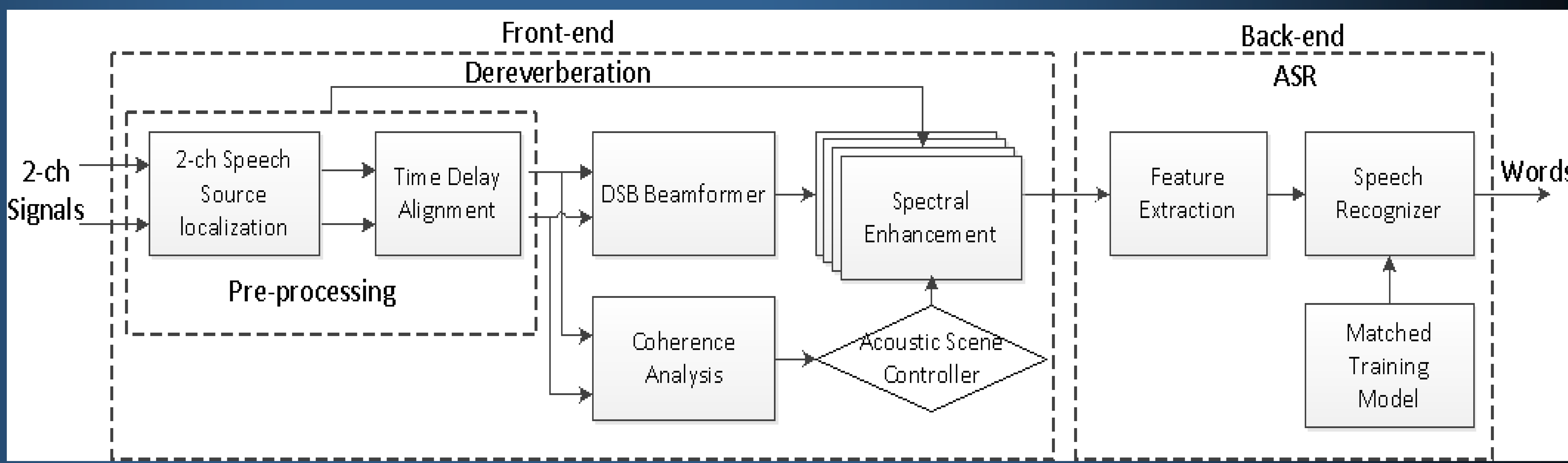
p2.1: Acoustic Scene Aware Dereverberation using 2-channel spectral enhancement for REVERB Challenge

Xiaofei WANG, Yanmeng GUO, Xi YANG, Qiang FU and Yonghong YAN
Institute of Acoustics, Chinese Academy of Sciences

Background

- The reverberation is known to degrade severely the audible quality of speech and performance of automatic speech recognition (ASR).
- The combination of the front-end audio processing with the back-end speech recognition techniques is also effective to improve the ASR performance in reverberant conditions.
- Under certain enhancement method, there is usually a tradeoff between the reverberation or noise suppressing amount and target speech distortion.

Block Diagram of Proposed System



Main Idea

- Acoustic scene classification: based on coherence analysis.
- Selection of spectral enhancement: based on the acoustics scene.
- Eliminate the interference as much as possible while keeping the speech distortion always in a low level.

System Description

Spectral Enhancement

Motivation:

- Robust to both noise and reverberation
- Low calculation complexity

General Form:

$$|\hat{S}(l,k)| = G(l,k)|\hat{X}(l,k)|$$

2-channel Case

Motivation:

- Basic topology of all microphone arrays
- Low requirement for both hardware and software
- It is ideal to fulfill the dereverberation task based on 2 sensors, just like what the human auditory system

Acoustic Scene Awareness

Motivation:

- Reflection condition: high or low
- Speaker-mic distance: near or far
- A controller is needed to better selection of spectral enhancement method since different spectral enhancement methods shows superiority in different kinds of reverberant environment

Pre-processing

Motivation:

- DOA of target signal is unknown
- An alignment filter should be designed for Beamforming

System Description

Back-end ASR

Clean+noEnh: "clean-condition" HMMs without dereverberation

Clean+Enh: "clean-condition" HMMs with dereverberation

Multi+noEnh: "multi-condition" HMMs without dereverberation

Multi+Enh: "multi-condition" HMMs with dereverberation

ReTrn+Enh: re-trained "multi-condition" HMMs with dereverberation

Adapt to the potential distortion in the front-end enhanced signals.

Algorithm

Algorithm 1 Strategy of Spectral enhancement.

if $\hat{\epsilon} > \sigma_1$ **then**
  $G(l,k)(FFT\_bins/3 : FFT\_bins) = 1$
  $G(l,k)(1 : FFT\_bins/3-1) = max(G_{late}, G_{cdr})$
  $\hat{X}(l,k) = X_0$
else if $\hat{\epsilon} > \sigma_2$ **then**
  $G(l,k) = \max(G_{late}, G_{cdr})$
  $\hat{X}(l,k) = X_{DSB}$
else if $\hat{\epsilon} > \sigma_3$ **then**
  $G(l,k) = \min(G_{late}, G_{cdr})$
  $\hat{X}(l,k) = X_{DSB}$
else
  $G(l,k) = G_{late}G_{cdr}$
  $\hat{X}(l,k) = X_{DSB}$
end if

Conclusion and Discussion

- An acoustic scene aware technique is proposed to make dereverberation robust to different conditions
- For SE task, objective indexes illustrate the improvement on speech signal quality
- For ASR task, when it is combined with back-end ASR with matched training, it produces a significant decrease on WER

Results(SE Task)

Table 1. Cepstral distance of test SimData before and after dereverberation.

Room	Cepstral distance in dB			
	mean		median	
	org	enh	org	enh
room1_near	1.99	1.96	1.68	1.69
room1_far	2.67	2.78	2.38	2.65
room2_near	4.63	3.52	4.24	3.35
room2_far	5.21	4.51	5.04	4.25
room3_near	4.38	3.57	4.04	3.43
room3_far	4.96	4.42	4.73	4.18
average	**3.97**	**3.46**	**3.69**	**3.26**

Table 2 . SRMR of test SimData before and after dereverberation.

Room	SRMR (only mean used)			
	mean		median	
	org	enh	org	enh
room1_near	4.50	4.13	_	_
room1_far	4.58	4.53	_	_
room2_near	3.74	3.88	_	_
room2_far	2.97	4.25	_	_
room3_near	3.57	3.80	_	_
room3_far	2.73	3.84	_	_
average	**3.68**	**4.07**	_	_

Table 3 . Log likelihood ratio of test SimData before and after dereverberation.

Room	Log likelihood ratio			
	mean		median	
	org	enh	org	enh
room1_near	0.35	0.35	0.33	0.33
room1_far	0.38	0.45	0.35	0.42
room2_near	0.49	0.56	0.40	0.49
room2_far	0.75	0.78	0.63	0.71
room3_near	0.65	0.65	0.59	0.60
room3_far	0.84	0.80	0.76	0.75
average	**0.58**	**0.60**	**0.51**	**0.55**

Table 4. Frequency-weighted segmental SNR of test SimData before and after dereverberation.

Room	FWSegSNR in dB			
	mean		median	
	org	enh	org	enh
room1_near	8.12	9.86	10.72	10.99
room1_far	6.68	8.56	9.24	8.72
room2_near	3.35	7.19	5.52	8.76
room2_far	1.04	4.29	1.77	6.43
room3_near	2.27	5.59	4.21	6.82
room3_far	0.24	3.04	0.89	4.78
average	**3.62**	**6.42**	**5.39**	**7.75**

Table 5. PESQ of test SimData before and after dereverberation.

Room	PESQ (only mean used)			
	mean		median	
	org	enh	org	enh
room1_near	2.14	2.09	_	_
room1_far	1.61	1.64	_	_
room2_near	1.40	1.69	_	_
room2_far	1.19	1.36	_	_
room3_near	1.37	1.53	_	_
room3_far	1.17	1.23	_	_
average	**1.48**	**1.59**	_	_

Table 6. SRMR of test RealData before and after enhancement

Room	SRMR (only mean used)			
	mean		median	
	org	enh	org	enh
room1_near	3.17	4.44	_	_
room1_far	3.19	4.67	_	_
average	**3.18**	**4.55**	_	_

Results(ASR Task)

Test Data		Word error rate(%)													
		SimData											RealData		
		Room 1,2,3			Ave.	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
		Clean				Near	Far	Near	Far	Near	Far		Near	Far	
Clean+ noEnh	nocmllr	12.84	12.49	12.13	12.48	18.06	25.38	42.98	82.20	53.54	88.04	51.68	89.72	87.34	88.53
	cmllr	_	_	_	_	14.81	18.86	24.63	64.58	33.77	78.42	39.16	82.31	80.76	81.53
Clean+ Enh	nocmllr	_	_	_	_	17.43	25.25	27.85	49.48	36.51	65.94	37.06	73.91	71.34	72.62
	cmllr	_	_	_	_	14.47	19.47	21.19	34.86	27.16	50.50	27.93	62.66	61.58	62.12
Multi+ noEnh	nocmllr	30.29	30.07	30.11	30.15	20.60	21.15	23.70	38.72	28.08	44.86	29.51	58.45	55.44	56.94
	cmllr	15.99	15.52	15.70	15.73	16.23	18.71	20.50	32.47	24.76	38.88	25.25	50.14	47.57	48.85
Multi+ Enh	nocmllr	_	_	_	_	23.64	36.46	27.72	37.69	34.00	45.85	34.22	59.95	59.49	59.72
	cmllr	_	_	_	_	16.93	20.04	19.91	26.84	23.95	34.33	23.66	44.87	45.81	45.34
ReTrn+ Enh	nocmllr	16.59	15.84	16.41	16.27	15.64	18.76	19.79	28.56	24.01	35.15	23.64	49.50	49.49	49.49
	cmllr	13.76	13.43	13.62	13.60	**14.76**	**16.52**	**18.23**	**24.79**	**21.09**	**31.50**	**21.14**	**42.10**	**45.17**	**43.63**

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).

Presented at the REVERB Workshop, 10 May, 2014. Florence, Italy.