Dual system combination approach for various reverberant environments with dereverberation techniques

Yuuki Tachioka, Tomohiro Narita (Mitsubishi Electric)
Felix Weninger, Shinji Watanabe (MERL)

Summary

We have validated the effectiveness of techniques below:

Speech enhancement:

- Single-channel dereverberation method[1)]
  - -reverberation time (RT) estimation
- Eight-channel beam-forming with direction of arrival estimation[2)]

ASR using Kaldi toolkit [3)]:

- Feature transformation and speaker adaptation
  - -LDA, MLLT, basis fMLLR
- Discriminative training and discriminative feature transformation
  - -boosted MMI and feature-space boosted MMI
  - -Deep neural networks

ASR System combination using ROVER[4)]:

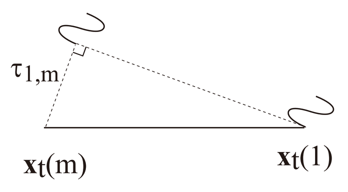
- Discriminative training for system combination
  - -dual system approach[5)]
  - -black box optimization of ROVER parameters[6)]

Speech Enhancement Part

DS beamformer with direction of arrival estimation

DS beam former

$$\tilde{\mathbf{y}}_t = \sum_m \mathbf{x}_t(m) \odot \exp(-\jmath \boldsymbol{\omega} \tau_{1,m})$$



Estimation of direction of arrival

$$\tau_{1,m} = \arg\max \mathcal{S}^{-1}\left[\frac{\mathbf{x}_t(1) \odot \mathbf{x}_t(m)^*}{|\mathbf{x}_t(1)||\mathbf{x}_t(m)|}\right]$$

$\mathcal{S}$: short-time Fourier transform
$\mathbf{x}_t(m)$: m-th mic inputs at t-th frame
$\odot$: element-wise multiplication
$*$: complex conjugate

SS based derev. with RT estimation

$\hat{\mathbf{y}}$: source signal
$\mathbf{n}$: noise signal
w_μ : weight

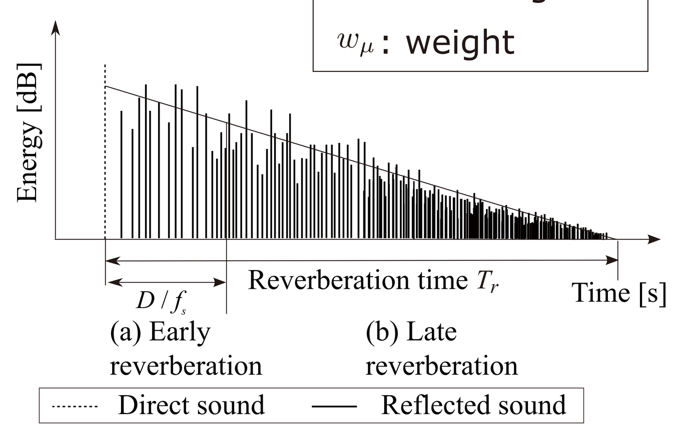
Reverberation model

$$|\mathbf{x}_t|^2 = \sum_{\mu=0}^{t} w_\mu |\hat{\mathbf{y}}_{t-\mu}|^2 + |\mathbf{n}|^2$$

Instantaneous mixture model

$$|\hat{\mathbf{y}}_{t-\mu}|^2 = \eta(T_r)|\mathbf{x}_{t-\mu}|^2 - |\mathbf{n}|^2$$

η: direct sound / total



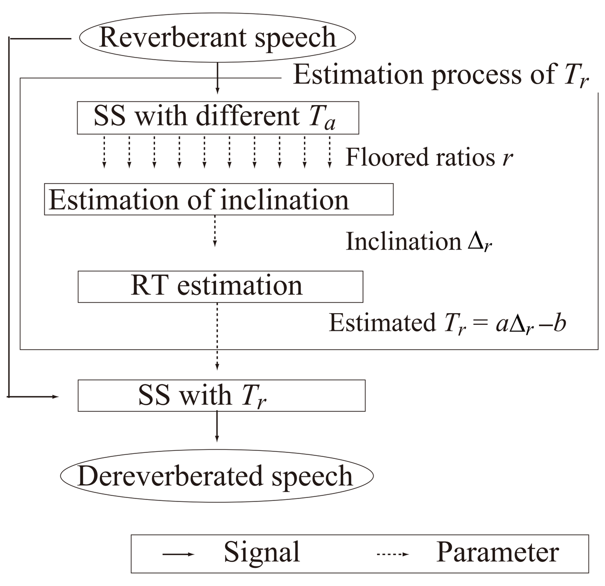
Approx. dereverberation formula (1)

$$|\hat{\mathbf{y}}_t|^2 = |\mathbf{x}_t|^2 - \sum_{\mu=1}^{t} w_\mu \left[\eta(T_r)|\mathbf{x}_{t-\mu}|^2 - |\mathbf{n}|^2\right] - |\mathbf{n}|^2$$

Polack model

$$w_\mu = \begin{cases} 0 & (1 \le \mu \le D) \\ \frac{\alpha_s}{\eta(T_r)} e^{-2\Delta\varphi\mu} & (D < \mu) \end{cases}$$

α_s: sub. param.
φ : frame shift
Δ : constants

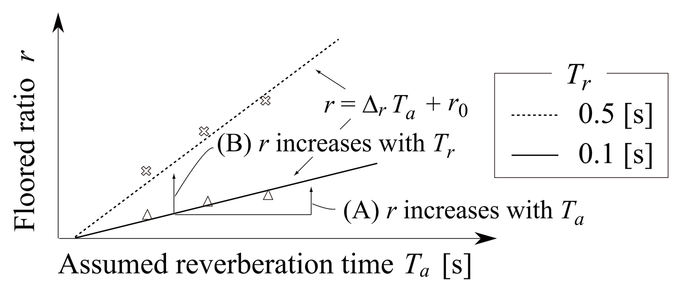


To estimate RT, floored ratio of Eq.(1) is calculated for assumed RT

Two obserbations:

- r increases with T_a (assumed RT)
- r increases with T_r (actual RT)

Using these, RT can be estimated from the floored ratio



ASR part

MMI discriminative training of acoustic models

MMI objective function to optimize λ and λ_c

$$\mathcal{F}_\lambda^{\mathrm{MMI}}(\omega_r) = \ln \frac{P_\lambda(\omega_r, \mathbf{X})}{\sum_\omega P_\lambda(\omega, \mathbf{X})} = \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_\lambda(s_r, \mathbf{X})^\kappa p_L(\omega_r)}{\sum_\omega \sum_{s \in \mathcal{S}_\omega} p_\lambda(s, \mathbf{X})^\kappa p_L(\omega)}$$

b-MMI objective function

$$\mathcal{F}_\lambda^{\mathrm{bMMI}}(\omega_r) = \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_\lambda(s_r, \mathbf{X})^\kappa p_L(\omega_r)}{\sum_\omega \sum_{s \in \mathcal{S}_\omega} p_\lambda(s, \mathbf{X})^\kappa p_L(\omega) e^{-bA(s,s_r)}}$$

κ : acoustic scale
p_λ : acoustic score with HMM state sequence s
p_L : language model score
s_r : reference state sequence
$\mathbf{X} = \{\mathbf{x}_t | t = 1, \cdots, T\}$: feature vector sequence (T-frame)
$\mathcal{S}_{\omega_r}, \mathcal{S}_\omega$: set of HMM state sequences which output ω_r and ω, respectively
$A(s,s_r)$: state/phoneme/word accuracy calculated from the HMM state sequences of s for a reference s_r
b : boosting factor
b_1 : boosting factor for complementary system

Discriminative training for system combination

Discriminative training principle:

MI between ref., 1-best of base system, and hypotheses of comp. system

Proposed objective function $\mathcal{F}_\varphi^c$:

$$\mathcal{F}_\varphi^c(\omega_r, \omega_{q,1}) = (1+\alpha)\mathcal{F}_\varphi(\omega_r) - \frac{\alpha}{Q}\sum_{q=1}^{Q} \mathcal{F}_\varphi(\omega_{q,1})$$

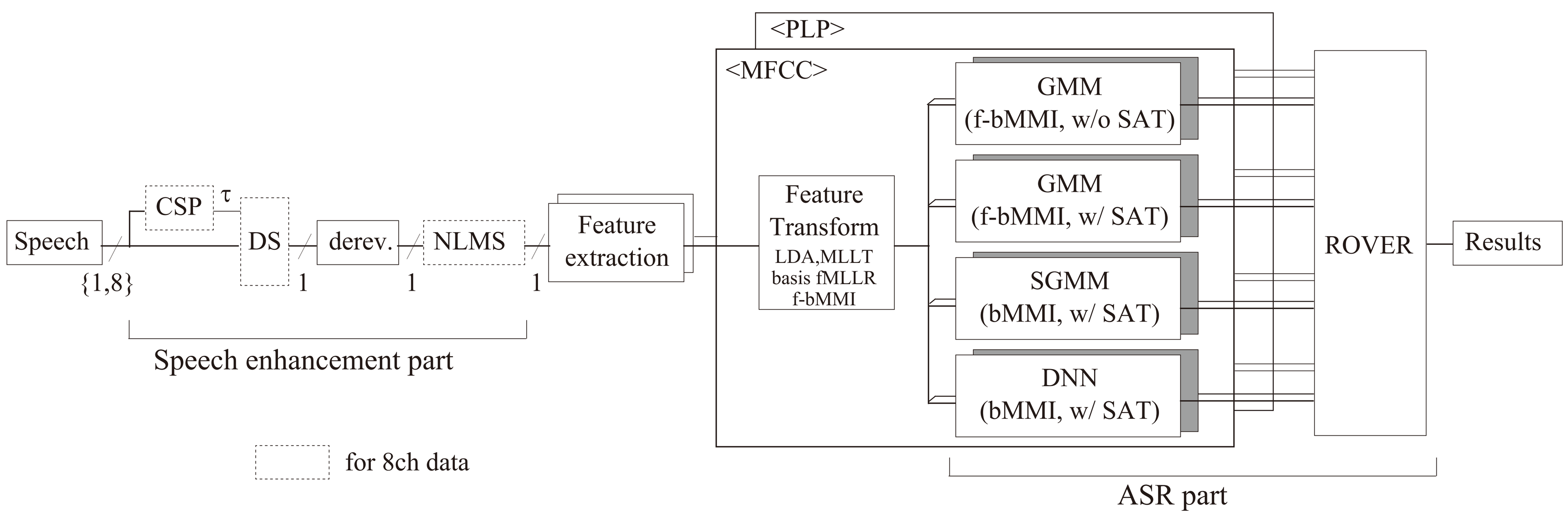
Reference ω_r
max. MI
max. MI
Q Base system $\omega_{q,1}$ ↔ Comp. system
min. MI

φ : set of model parameters of complementary system to be optimized
α : scaling factor

$$\mathcal{F}_{\lambda_c}^{c}(\omega_r, \omega_1) = \mathcal{F}_{\lambda_c}^{\mathrm{MMI}}(\omega_r) + \alpha \ln \frac{P_{\lambda_c}(\omega_r, \mathbf{X})}{P_{\lambda_c}(\omega_1, \mathbf{X})}$$

$$\mathcal{F}_{\lambda_c}^{c}(\omega_r, \omega_1) = \mathcal{F}_{\lambda_c}^{\mathrm{bMMI}}(\omega_r) + \alpha \ln \frac{\sum_{s_r \in \mathcal{S}_{\omega_r}} p_\lambda(s_r, \mathbf{X})^\kappa p_L(\omega_r)}{\sum_{s_1 \in \mathcal{S}_{\omega_1}} p_\lambda(s_1, \mathbf{X})^\kappa p_L(\omega_1) e^{b_1 A(s_1, s_r)}}$$

System overview



Experiments

Task description and setup

A middle-size vocabulary continuous speech recognition task

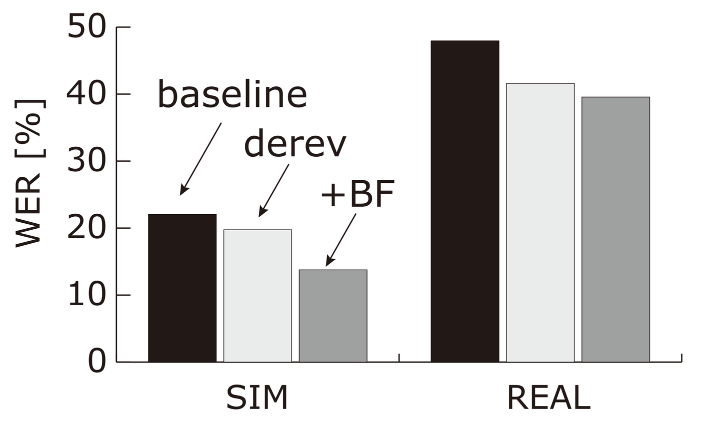
8 different reverberant environments:

- -3 rooms with near/far mic settings for SIMulated data
- -1 room with near/far mic settings for REAL data with noise

Speech enhancement

derev improves the performance

BF improves it further



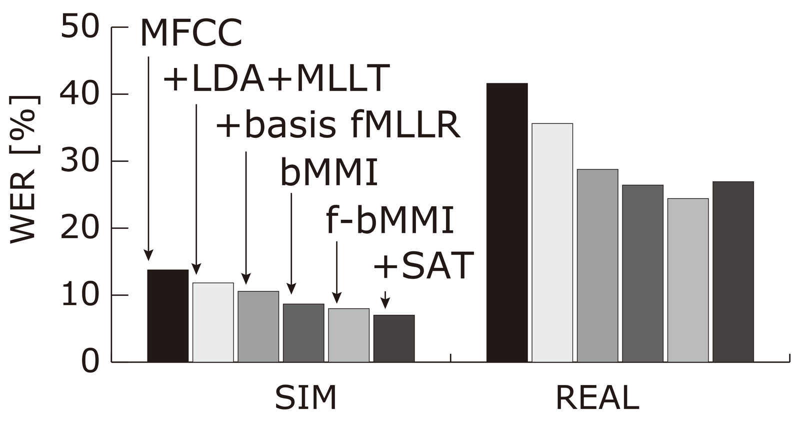
Feature transformation and discriminative methods

LDA improves the performance due to the use of long context

basis fMLLR is effective

f-bMMI is effective

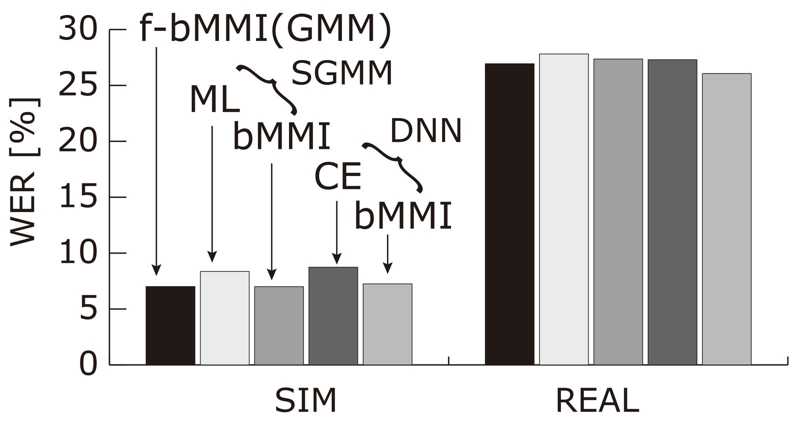
SAT is unstable



Subspace GMM and DNN

bMMI is effective

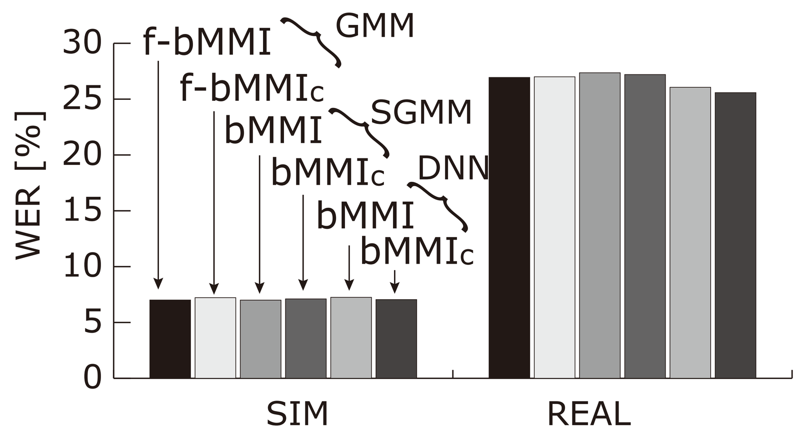
Best performer is different for each environment



Complementary systems

Performance is moderate

Output tendendies are different



System combination

System combination improves the accuracies for all the cases

Proposed method is effective for all the environment

Combination of different types of systems is effective

		Number of systems				SimData							RealData		
						Room 1		Room 2		Room 3		Avg	Room 1		Avg
	ID	GMM	SAT-GMM	SGMM	DNN	near	far	near	far	near	far		near	far	
8ch	0)		1			5.01	6.76	5.96	9.07	5.84	9.40	7.01	24.27	29.60	26.94
	1)		2			4.72	5.83	5.96	8.92	5.37	8.75	6.59	23.27	28.30	25.79
	2)	2	2			4.72	6.02	5.72	8.26	5.14	8.56	6.40	22.27	26.59	24.43
	3)	4	4			4.72	5.83	5.77	8.21	5.19	8.38	6.35	22.52	26.52	24.52
	4)	4	4	4		**4.08**	5.16	5.62	7.79	**4.80**	8.38	5.97	22.40	27.00	24.70
	5)	4	4	4	4	4.18	**5.11**	**5.50**	**7.74**	4.85	**8.23**	**5.94**	21.90	26.52	**24.21**
	6)	3	1	4	2	4.18	5.51	**5.50**	**7.74**	4.97	8.43	6.06	**21.58**	**26.32**	*23.95*

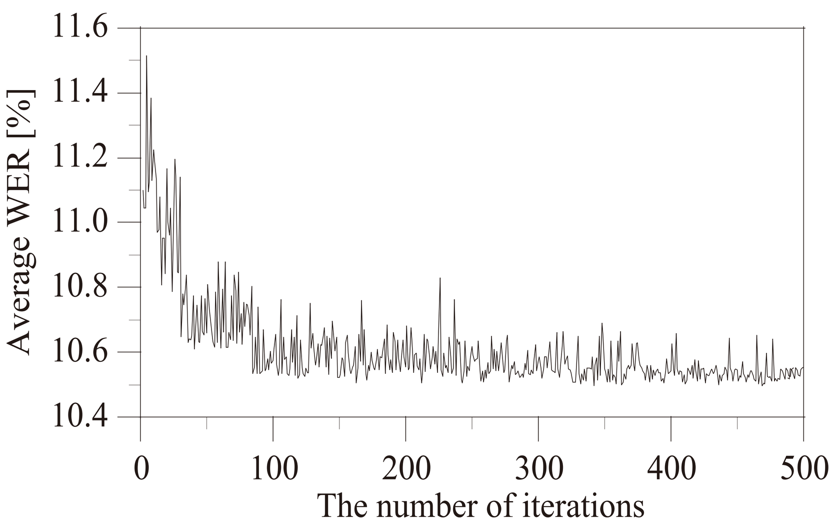
Black box optimization on ROVER parameters

Selection of combined system

ROVER parameter

WER improves monotonically

100 iterations are enough



Evaluation set

		SimData							RealData		
		Room 1		Room 2		Room 3		Avg	Room 1		Avg
		near	far	near	far	near	far		near	far	
1ch	Kaldi baseline	13.23	14.13	15.54	29.69	20.06	37.44	21.68	50.62	45.98	48.30
	derev.	12.50	13.43	14.61	24.71	17.09	32.62	19.16	44.75	43.32	44.04
	f-bMMI	7.27	8.17	8.82	14.11	10.54	18.76	11.28	28.65	29.54	29.10
	SAT+f-bMMI	6.44	7.22	7.57	13.97	9.52	18.44	10.53	28.87	29.78	29.33
	SGMM+bMMI	5.81	6.54	7.22	13.84	8.70	18.17	10.05	27.75	28.36	28.06
	DNN+bMMI	5.90	6.84	7.35	12.57	9.40	16.55	9.77	25.97	25.69	25.83
	ROVER	5.30	5.61	6.30	11.16	7.76	14.95	**8.51**	23.79	23.60	**23.70**
8ch	CSP+BF+derev.	10.94	11.69	10.98	16.33	12.79	21.39	14.02	34.33	36.93	35.63
	f-bMMI	6.57	6.93	6.80	9.93	7.47	12.76	8.41	20.22	23.19	21.71
	SAT+f-bMMI	6.17	6.64	6.51	10.13	7.40	13.15	8.33	20.63	23.67	22.15
	SGMM+bMMI	5.86	6.44	6.29	9.23	6.96	12.83	7.94	20.66	23.50	22.08
	DNN+bMMI	5.64	6.18	6.16	9.29	7.08	12.40	7.79	19.35	22.28	20.82
	ROVER	4.96	5.62	5.58	8.18	5.73	10.47	**6.76**	16.90	20.29	**18.60**

1)Y. Tachioka, T. Hanazawa, and T. Iwasaki, Dereverberation method with reverberation time estimation using floored ratio of spectral subtraction, Acoustical Science and Technology, vol. 34, pp. 212-215, 2013.
2)Y. Tachioka, T. Narita, and T. Iwasaki, Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information, Acoustical Science and Technology, vol. 33, pp. 68-71, 1 2012.
3)D. Povey, et al., The Kaldi speech recognition toolkit, in Proc. of ASRU, 2011.
4) J.G. Fiscus, A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER), in Proc. of ASRU, 1997, pp. 347-354.
5) Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, A generalized framework of discriminative training for system combination, in Proc. of ASRU, 2013.
6) S. Watanabe and J. Le Roux, Black box optimization for automatic speech recognition, in Proc. of ICASSP, 2014.