The THU-SPMI CHiME-4 system : Lightweight design with advanced multi-channel processing, feature enhancement, and language modeling

Hongyu Xiang, Bin Wang, Zhijian Ou

Speech Processing and Machine Intelligence (SPMI) Lab,

Tsinghua University, Beijing, China

xianghy12@163.com, wb.th08@gmail.com, ozj@tsinghua.edu.cn

Our submission and contributions

For time constraint, we only submit results for 6-channel track.

1. For multi-channel processing, we explore a bundle of beamforming techniques and compare them with the same back-end.
   - Most effective: MVDR beamformer with steering vector being estimated by time-frequency masks.
2. Propose to applying histogram equalization (HEQ) to compensate for the residual noise in beamformed speech.
3. Interpolated Trans-dimensional Random Field (TRF) language model with LSTM-RNN language model.

Beamforming

The signal model is

$$\mathbf{X}(t,\omega) = S(t,\omega)\mathbf{d}(\omega) + \mathbf{N}(t,\omega) = \mathbf{G}(t,\omega) + \mathbf{N}(t,\omega)$$

Beamfomer	Description
GSC	$\min_{\mathbf{R}} \left\| \mathbf{b} - \mathbf{R}^{\mathbf{H}}\mathbf{z} \right\|^2$
MVDR	$\min_{\mathbf{W}} \mathbf{E}\left\| \mathbf{W}^{\mathbf{H}}\mathbf{X} \right\|^2 \ \text{s.t.}\ \mathbf{W}^{\mathbf{H}}\mathbf{d} = 1$
MVDR-RTF	$\min_{\mathbf{W}} \mathbf{E}\left\| \mathbf{W}^{\mathbf{H}}\mathbf{X} \right\|^2 \ \text{s.t.}\ \mathbf{W}^{\mathbf{H}}\mathbf{d} = d_1$
PMWF	$\min_{\mathbf{W}} \mathbf{E}\left(\left\| \mathbf{W}^{\mathbf{H}}\mathbf{X} - G_1 \right\|^2 + \beta \left\| \mathbf{W}^{\mathbf{H}}\mathbf{N} \right\|^2 \right)$
WDAS	$y(t) = \sum_i \omega_i x_i(t - \tau_i)$

$\mathbf{b}$ is the output of GSC fixed beamformer.

$\mathbf{R}$ is the adaptive filter of GSC. $\mathbf{z}$ is the output of GSC block matrix.

$\mathbf{W}$ is the filter-and-sum coefficients.

$\mathbf{d}$ is the steering vector. d_1 is the first (reference) element of $\mathbf{d}$.

Abbrev.

GSC	Generalized Sidelobe Canceller
MVDR	Minimum Variance Distortionless Response
MVDR-RTF	MVDR with Relative Transform Function
PMWF	Parameterized Multi-channel Wiener Filter
WDAS	Weighted Delay And Sum filter

Notations for Table 3

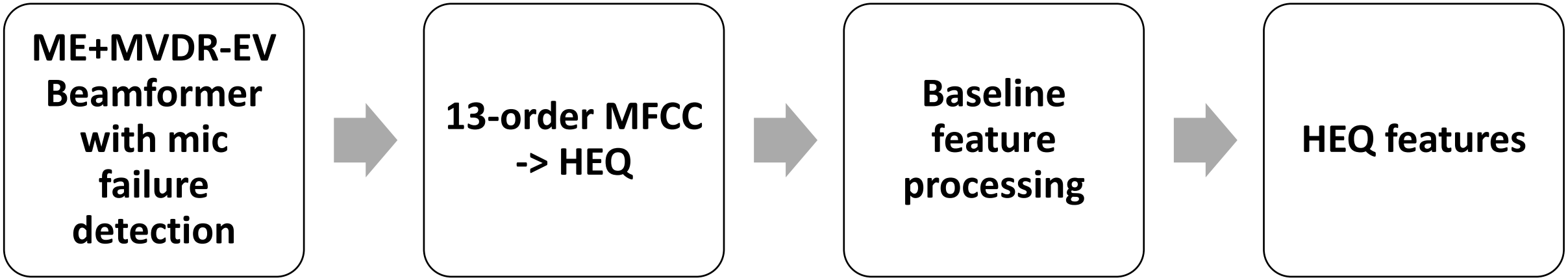
WDAS	BeamformIt in the baseline
MVDR	The one released at CHiME-3
GSC	Standard one with fixed beamformer and fixed block matrix
GSC+WDAS	Use WDAS to replace the beamformer block of GSC
MVDR-RTF	MVDR with relative transform function
MVDR-EV	Obtain the steering vector from the principal eigenvector of the estimated spatial correlation matrix of clean signal.
ME	Complex GMM based spectral mask estimation (ME), directly to separate the source ("ME+direct") or to estimate the spatial corelation matrices (the last three rows)

Beamforming

Table 3: Average WER (%) of different beamformers with the baseline back-end (without RNN).

Track	System	Dev		Test	
		real	simu	real	simu
6ch	WDAS	8.19	9.40	15.59	15.61
	MVDR	14.31	5.97	25.89	6.99
	GSC	10.99	15.77	19.79	24.17
	GSC+WDAS	9.46	11.73	16.61	19.00
	PMWF	10.82	9.90	19.58	14.18
	ME+direct	8.75	6.98	15.05	7.74
	ME+PMWF	8.87	6.51	15.52	7.33
	ME+MVDR-EV	8.04	6.07	13.59	7.32
	ME+MVDR-RTF	11.07	6.90	18.99	8.53

Histogram equalization (HEQ)



WER comparison after using HEQ. Back-end is DNN (sMBR) and KN5+RNN LM.

Track	System	Dev		Test	
		real	simu	real	simu
6ch	ME+MVDR-EV	5.16	4.70	8.21	5.79
	Single HEQ	4.97	4.75	7.27	6.47
	Stack HEQ	4.64	4.22	7.15	5.51

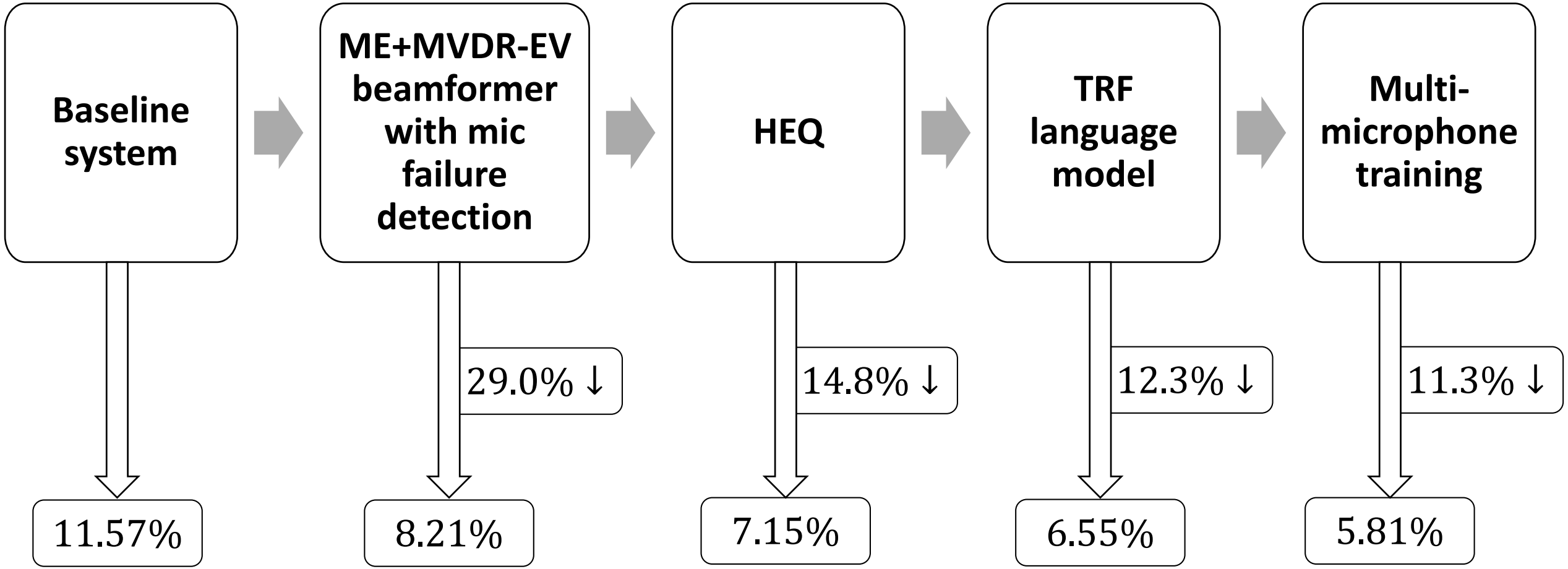
Trans-dimensional Random Field (TRF) LMs

$$p(l, x^l; \lambda) = \pi_l \cdot \frac{1}{Z_l(\lambda)} \cdot \exp\left[\lambda^T f(x^l)\right], \qquad l = 1, \cdots, m$$

Track	System	Dev		Test	
		real	simu	real	simu
6ch	KN5	5.57	5.11	8.25	6.42
	RNN	5.24	4.82	7.92	5.91
	TRF	5.09	4.56	7.92	6.04
	KN5+RNN	4.64	4.22	7.15	5.51
	TRF+RNN	4.48	4.06	6.96	5.26
	TRF+LSTM	4.58	3.78	6.55	4.95

- B. Wang, Z. Ou, and Z. Tan, "Trans-dimensional random fields for language modeling," in ACL, 2015.
- B. Wang, Z. Ou, Y. He, and A. Kawamura, "Improving and scaling trans-dimensional random field language models," arXiv 2016.

Summary



WERs on the real data evaluation test set