Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline

Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, Shinji Watanabe

Center for Language and Speech Processing, Johns Hopkins University

Objectives

- State-of-the-art system with a simplified single system comparable to the complicated top systems in the challenge
- Publicly available and reproducible recipe in the Kaldi (https://github.com/Szu-JuiChen/kaldi/tree/lmrescore)
- Incorporate computation of four different speech enhancement measures:
  - Perceptual evaluation of speech quality (PESQ)
  - Short-time objective intelligibility measure (STOI)
  - extended STOI (eSTOI)
  - Speech distortion ratio (SDR)

Proposed system

- Bidirectional long short-term memory (BLSTM) mask based beamformer
- Sub-sampled time delay neural network (TDNN) with the lattice-free version of the maximum mutual information (LF-MMI)
- LSTM language model (LSTMLM)

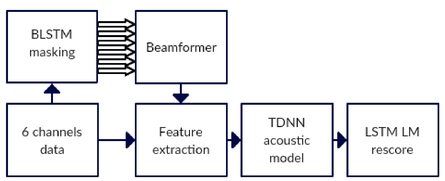

Figure: 1: Diagram of speech recognition system

Formulation

- Data augmentation:
  $\mathbf{O}^{\text{enh}} = (\mathbf{o}^{\text{enh}(t)} \in \mathbb{R}^D | t = 1, \ldots, T)$ (enhanced data) is included with the multichannel data $\mathbf{O}$:

$$\hat{\Theta} = \arg\max_{\Theta} \mathcal{L}(\{\mathbf{O}, \mathbf{O}^{\text{enh}}\}), \tag{1}$$

- LSTM language modeling:

$$\mathcal{L}_{\text{LM}} = z_j + 1 - \sum_i \exp(z_i) \tag{2}$$

where z is a pre-activation vector in the layer of neural network before the final softmax operation and j is an index for the correct word.

Formulation cnt'd

- BLSTM mask based beamformer:
  The PSD matrices of speech components $\mathbf{\Phi}_{\text{speech}}(b) \in \mathbb{C}^{M \times M}$ at frequency bin b, and that of noise components $\mathbf{\Phi}_{\text{noise}}(b) \in \mathbb{C}^{M \times M}$ can be estimated as follows:

$$\mathbf{\Phi}_v(b) = \sum_{t=1}^{T} w_v(t,b)\mathbf{y}(t,b)\mathbf{y}(t,b)^{\text{H}} \tag{3}$$

where $v \in \{\text{speech}, \text{noise}\}$ and $\mathbf{y}(t,b) \in \mathbb{C}^M$ is an M-dimensional complex spectrum at time (frame) t in frequency bin b. $\mathbf{y}^{\text{H}}$ denotes the conjugate transpose. $w_v(t,b) \in [0,1]$ is the mask value.

- TDNN with lattice-free MMI:
  The LF-MMI objective function is shown below

$$\mathcal{L}_{\text{MMI}} = \sum_{n=1}^{N} \log \frac{p(\mathbf{O}^n|S^n)^{\mathcal{C}} P(L^n)}{\sum_L p(\mathbf{O}^n|S^L)^{\mathcal{C}} P(L)} \tag{4}$$

where $p(\mathbf{O}^n|S^L)$ is the likelihood function of a speech feature sequence $\mathbf{O}^n$ given the state sequence S^L at n'th utterance. $P(L)$ is the phoneme language model probability and $\mathcal{C}$ is the probability scale.

Experiments

Table: 1: Speech Enhancement Scores

		Dev (Simu)				Test (Simu)			
Track	Enhancement Method	PESQ	STOI	eSTOI	SDR	PESQ	STOI	eSTOI	SDR
1ch	No Enhancement	2.01	0.82	0.61	3.92	1.98	0.81	0.60	4.95
1ch	BLSTM Mask	**2.52**	**0.88**	**0.73**	**9.26**	**2.46**	**0.87**	**0.71**	**10.76**
2ch	BeamformIt	**2.15**	0.85	0.65	**4.61**	2.07	0.83	0.62	**5.60**
2ch	BLSTM Gev	2.13	**0.87**	**0.69**	2.86	**2.12**	**0.87**	**0.69**	3.10
6ch	BeamformIt	2.31	0.88	0.70	**5.52**	2.20	0.86	0.65	**6.30**
6ch	BLSTM Gev	**2.45**	0.88	**0.75**	3.57	**2.46**	**0.87**	**0.73**	2.92

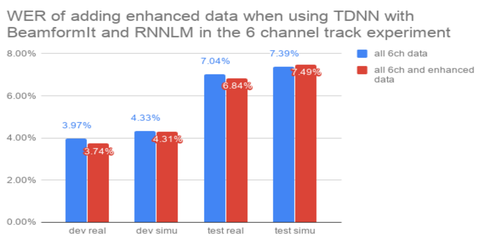

Figure: 2: WER of adding enhanced data

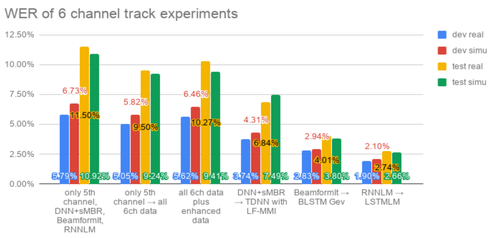

Figure: 3: WER of 6 channel track experiments

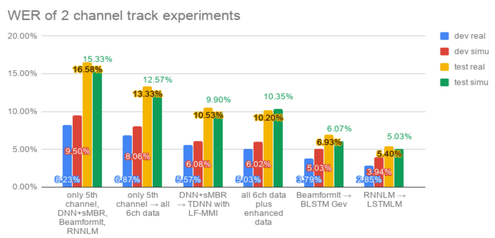

Figure: 4: WER of 2 channel track experiments

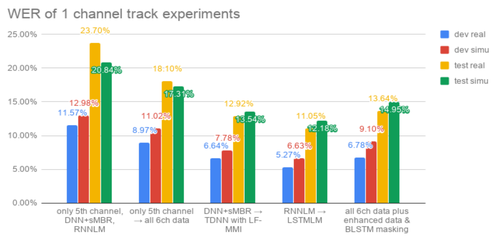

Figure: 5: WER of 1 channel track experiments

Discussion

- Table 1:
  - BLSTM-based speech enhancement shows improvement in most of conditions except for the case of the multichannel SDR metric.
  - For the 1ch track, the BLSTM mask gives significantly better scores in all four metrics. But this is contrary to the ASR results.
- Figure 2 shows the effectiveness of the data augmentation for the system. Improvement confirmed by adding enhanced data in almost all cases except for the simulation test data.
- Figure 3 and Figure 4:
  - Experimental condition is changed incrementally. In most of the cases, every method improved the WER steadily.
  - The performance was degraded if we applied enhanced data on the system using DNN+sMBR.
  - There always seems to be a negative correlation between the ASR performance and the SDR scores.
- Figure 5 shows BLSTM masking was not effective if we only used one microphone although it scores better in terms of all four speech enhancement metrics.

Final Results

Table: 2: Final WER comparison for the real test set.

System	# systems	WER (%)
CHiME-4 baseline	1	11.51
Proposed system	1	2.74
USTC-iFlytek	5	2.24
RWTH/UPB/FORTH	5	2.91
MERL	6	2.98

Conclusion

The system finally achieved 2.74% WER, which outperforms the 2nd place result in the challenge and is publicly available through the Kaldi toolkit.

Contact Information

- Email:
  {schen146,asubra13,hxu31,shinjiw}@jhu.edu