Student-Teacher Learning for BLSTM Mask-based Speech Enhancement

Aswin Shanmugam Subramanian, Szu-Jui Chen, Shinji Watanabe

Center for Language and Speech Processing, Johns Hopkins University

Abstract

- Neural network mask-based beamforming techniques have improved the performance of multichannel noise robust ASR significantly.
- Spectral masks have not been helpful in the single-channel case.
- We propose a student-teacher learning paradigm for mask estimation to fill out the gap between single-channel and multichannel speech enhancement

BLSTM Masking Network

[Heymann+, 2016]

Layer	Activation	Dimension
Input	-	513
BLSTM	Tanh	256
Feedforward 1	ReLU	513
Feedforward 2	clipped ReLU	513

Table: 1: Masking Network Architecture

- $Y = (\{\|y(t,b)\|\}_{b=1}^{B} | t = 1, \cdots, T)$: sequence of T-length noisy speech magnitude spectra
- $\mathrm{IBM}_\mathrm{X}(t,b) \in \{0,1\}$ and $\mathrm{IBM}_\mathrm{N}(t,b) \in \{0,1\}$ at each time-frequency bin (t,b): ideal binary speech and noise mask target respectively
- $w_\mathrm{X}(t,b) \in [0,1]$ and $w_\mathrm{N}(t,b) \in [0,1]$ at each time-frequency bin (t,b): predicted speech and noise mask respectively
- $\mathrm{loss} = \mathrm{loss}_\mathrm{X} + \mathrm{loss}_\mathrm{N}$
  $\mathrm{loss} = \frac{1}{T*B} \sum_{t,b} \sum_{v \in \{\mathrm{X},\mathrm{N}\}} \mathrm{CE}(\mathrm{IBM}_v(t,b), w_v(t,b))$
  where, $\mathrm{CE}(a, a') \triangleq a \log a' + (1-a) \log(1-a')$

Mask-Based Beamformer

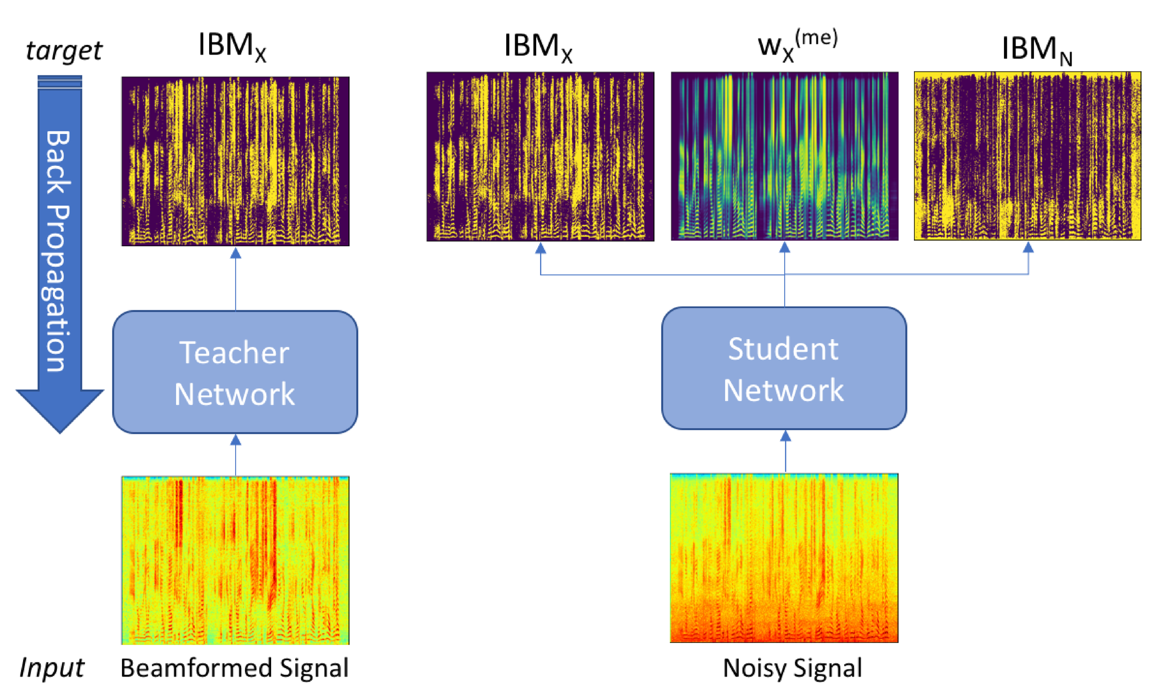
[Heymann+, 2016]

- $\bar{w}_v(t,b) = \mathrm{Median}(\{w_{m,v}(t,b)\}_{m=1}^{M})$, where $v \in \{\mathrm{X},\mathrm{N}\}$, $w_{m,\mathrm{X}}(t,b)$ and $w_{m,\mathrm{N}}(t,b)$ are speech and noise mask for each channel m respectively.
- $\mathbf{\Phi}_v(b) = \sum_{t=1}^{T} \bar{w}_v(t,b) \mathbf{y}(t,b) \mathbf{y}(t,b)^\mathrm{H}$, where $v \in \{\mathrm{X},\mathrm{N}\}$, $\mathbf{y}(t,b) \in \mathbb{C}^M$ and $\mathbf{\Phi}_v(b) \in \mathbb{C}^{M \times M}$
- $\mathbf{f}_\mathrm{GEV}(b) = \mathrm{argmax}_{\mathbf{f}(b)} \frac{\mathbf{f}^\mathrm{H}(b) \mathbf{\Phi}_\mathrm{X}(b) \mathbf{f}(b)}{\mathbf{f}^\mathrm{H}(b) \mathbf{\Phi}_\mathrm{N}(b) \mathbf{f}(b)}$

Mask-Based Beamformer Cnt'd

- $(\mathbf{\Phi}_\mathrm{N}(b))^{-1} \mathbf{\Phi}_\mathrm{X}(b) \mathbf{f}(b) = \lambda \mathbf{f}(b)$
- $x^{(\mathrm{me})}(t,b) = \mathbf{f}_\mathrm{GEV}^\mathrm{H}(b) \mathbf{y}(t,b)$, where $\mathbf{f}_\mathrm{GEV}(b)$ is the beamforming filter and $x^{(\mathrm{me})}(t,b)$ is the multichannel enhanced signal
- single-channel enhanced signal, $x^{(\mathrm{se})}(t,b) = w_\mathrm{X}(t,b) y(t,b)$

Student-Teacher Model



Teacher Model:

$$\mathrm{loss} = \frac{1}{T*B} \sum_{t,b} \mathrm{CE}(\mathrm{IBM}_\mathrm{X}(t,b), w_\mathrm{X}^{(me)}(t,b)))$$

Student Model (Additional Loss Term):

$$\mathrm{loss}_\mathrm{st} = \frac{1}{T*B} \sum_{t,b} \mathrm{CE}(w_\mathrm{X}^{(me)}(t,b)), w_\mathrm{X}^{(se)}(t,b))).$$

Student Model with Real Data:

$$\mathrm{loss} = \begin{cases} \mathrm{loss}_\mathrm{st} & \text{for real} \\ \lambda_1 \mathrm{loss}_\mathrm{st} + \lambda_2 \mathrm{loss}_\mathrm{X} + \lambda_3 \mathrm{loss}_\mathrm{N} & \text{for simulation} \end{cases}$$

Experiments

- Dataset: 1 channel track in CHiME-4
  - Training: 1600 (real) + 7138 (simulated) - use all 6ch data
  - Dev & Test: 3280 & 2640 respectively - equal real and simulated noisy utterances.
  - HMM-GMM ASR system - Kaldi CHiME4 recipe

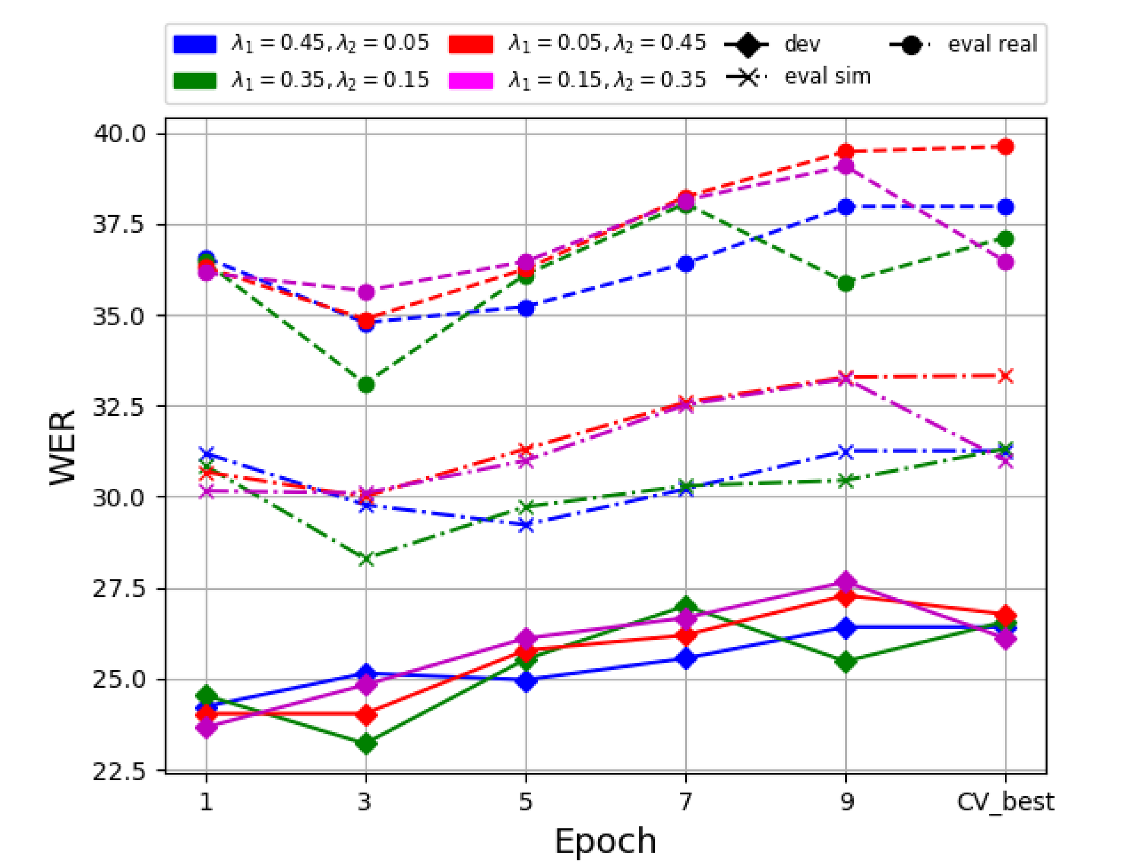


Figure: WER vs Epoch for Different Parameter Combinations

- Best validation loss - not necessarily gives best word error rate (WER).
- Choosing the epoch based on the WER of the development data seems to be a better criterion.

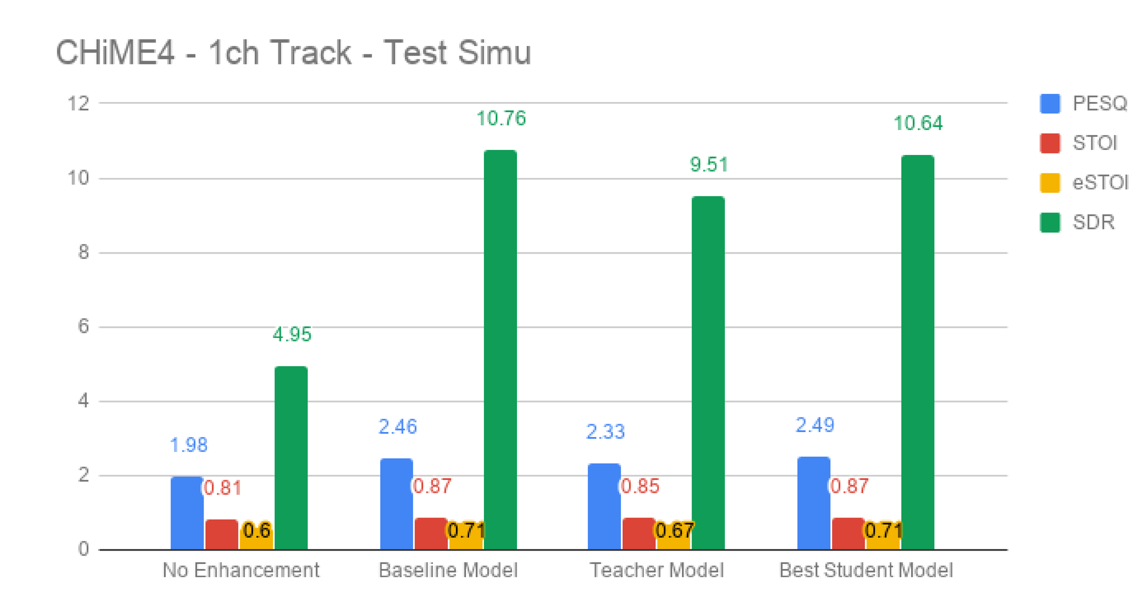
Table 2: WER of HMM-GMM ASR System

	Parameters						WER Dev (%)		WER Test (%)	
	λ_1	λ_2	λ_3	epoch	Train data (ASR)	BLSTM Mask	real	simu	real	simu
1	-	-	-	-	all 6ch noisy	-	**21.40**	**23.22**	**35.63**	**31.98**
2	-	-	-	14	all 6ch noisy	Baseline	28.99	28.05	40.98	35.50
3	-	-	-	7	all 6ch noisy	Teacher	24.91	26.00	40.26	35.73
4	1/3	1/3	1/3	6	all 6ch noisy	Student	25.95	24.66	35.50	29.98
5	0.25	0.25	0.5	12	all 6ch noisy	Student	26.56	26.19	36.33	31.36
6	0.35	0.15	0.50	3	all 6ch noisy	Student	**23.34**	**23.11**	33.11	**28.30**
7	0.35	0.15	0.50	3	all 6ch noisy	Student with real	23.42	23.55	**32.64**	28.88
8	-	-	-	-	all 6ch noisy + 5th ch enhanced data from baseline	Baseline	22.07	23.37	34.02	30.41
9	0.35	0.15	0.50	3	all 6ch noisy + 5th ch enhanced data from baseline	Student	**19.78**	**20.76**	30.66	**26.60**
10	0.35	0.15	0.50	3	all 6ch noisy + 5th ch enhanced data from baseline	Student with real	19.79	20.85	**29.80**	26.66

Discussion

- Table 2 (rows 4-7):
  - The training data for ASR not enhanced
  - Student models performed better than both the teacher model and baseline
  - Student models don't perform better than the non-enhanced noisy speech.
- Table 2 (rows 8-10):
  - 5^{th} channel data enhanced using the baseline masking model is included as part of ASR training
  - The performance improved significantly compared to using the original noisy data in the all conditions when the development and evaluation data was enhanced using our best student models (rows 9 and 10).
  - WER improvement for the real test set was observed when real training data was included while training the mask (row 10).

Speech Enhancement Scores



- Masking gives significantly better scores in all four metrics.
- No considerable difference in the scores amongst the masking models.

Conclusion

- The proposed student-teacher paradigm improved the performance of a GMM-HMM ASR system from both original noisy speech and the baseline masking.
- Our preliminary experiments on a strong ASR backend improved performance over the baseline masking but not the original noisy data.

Mail: {aswin, schen146, shinjiw}@jhu.edu

WWW: http://www.clsp.jhu.edu