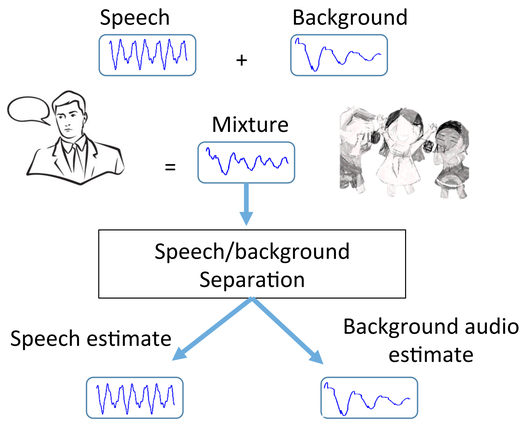
PHASE-SENSITIVE AND RECOGNITION-BOOSTED SPEECH SEPARATION USING DEEP RECURRENT NEURAL NETWORKS

HAKAN ERDOGAN[1,2], JOHN R. HERSHEY[1], JONATHAN LE ROUX[1], SHINJI WATANABE[1]
[1] MITSUBISHI ELECTRIC RESEARCH LABORATORIES, CAMBRIDGE MA 02309, USA
[2] SABANCI UNIVERSITY, ISTANBUL, 34956, TURKEY
haerdogan@sabanciuniv.edu, {hershey, leroux, watanabe}@merl.com

PROBLEM

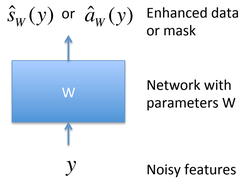
- Goal: Separate speech signal from background noise given a single channel recording of both
- Assumption: available training data with ground truths



- In the time domain $y(\tau) = s(\tau) + n(\tau)$
- In the STFT domain: $y_{t,f} = s_{t,f} + n_{t,f}$
- Often used approximation: $|y_{t,f}| \approx |s_{t,f}| + |n_{t,f}|$
- Problem: Given mixed STFT $\boldsymbol{y}$ and given *training data*, find an estimate of speech STFT and use it to reconstruct speech signal in the time domain.

NEURAL NET WITH MASK PREDICTION

- Direct magnitude prediction $\hat{s}_W(y)$ versus mask or adaptive filter prediction $\hat{a}_W(y)$
- For mask prediction, obtain speech estimate by $\hat{s}(y) = \hat{a}_W(y)y$



Predicting masks may be better since

- Logistic sigmoid outputs have a suitable range $[0, 1]$
- Easier to learn for the network since the mask is slowly changing
- No global variance issues in the predictions
- Pass through input directly when no noise

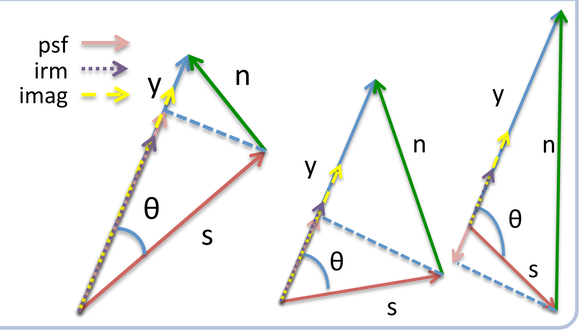
WHAT KIND OF NEURAL NETWORK?

- Long short-term memory (**LSTM**) recurrent networks were shown to be better than DNNs for this problem [1]
- The best performing input features are log-mel-filterbank energies with 100 mel filters (mfb) [1]
- We show that bidirectional LSTM (**BLSTM**) works better than unidirectional LSTM

IDEAL MASKS

target mask/filter	formula	optimality principle
IBM:	$a^{\text{ibm}} = \delta(\lvert s\rvert > \lvert n\rvert)$,	max SNR given $a \in \{0, 1\}$
IRM:	$a^{\text{irm}} = \frac{\lvert s\rvert}{\lvert s\rvert + \lvert n\rvert}$,	max SNR given $\theta_{\text{s}} = \theta_{\text{n}}$
"Wiener like":	$a^{\text{wf}} = \frac{\lvert s\rvert^2}{\lvert s\rvert^2 + \lvert n\rvert^2}$,	max SNR, expected power
ideal amplitude:	$a^{\text{iaf}} = \lvert s\rvert/\lvert y\rvert$,	exact $\lvert\hat{s}\rvert$, max SNR $\theta_{\text{s}} = \theta_y$
phase-sensitive:	$a^{\text{psf}} = \frac{\lvert s\rvert}{\lvert y\rvert}\cos(\theta)$,	max SNR given $a \in \mathbb{R}$
ideal complex:	$a^{\text{icf}} = s/y$,	max SNR given $a \in \mathbb{C}$

IDEAL MASKS IN COMPLEX DOMAIN



IDEAL MASKS CHIME-2 DEV SET SDR (IN DB)

dt	-6 dB	9 dB	Avg
IBM	14.56	20.89	17.59
IRM	14.13	20.69	17.29
"Wiener-like"	15.20	21.49	18.21
ideal amplitude	13.97	21.35	17.52
phase sensitive filter	17.74	24.09	**20.76**
truncated PSF	16.13	22.49	**19.17**

PHASE-SENSITIVE APPROXIMATION LOSS

Loss function for training the network:

$$\mathcal{L}(W) = \sum_{t,f} D(\hat{a}_{t,f})$$

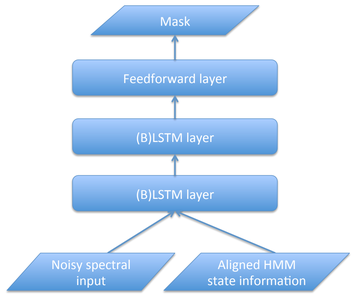
Distortion measures:

- Mask approximation (**MA**): $D_{\text{ma}}(\hat{a}) = |\hat{a} - a^*|^2$
- Magnitude spectrum approximation (**MSA**): $D_{\text{msa}}(\hat{a}) = (\hat{a}|y| - |s|)^2$
- Phase-sensitive spectrum approximation (**PSA**): $D_{\text{psa}}(\hat{a}) = |\hat{a}y - s|^2$
- **PSA** is equivalent to: $D_{\text{psa}}(\hat{a}) = (\hat{a}|y| - |s|\cos(\theta))^2$

USING ASR INFORMATION

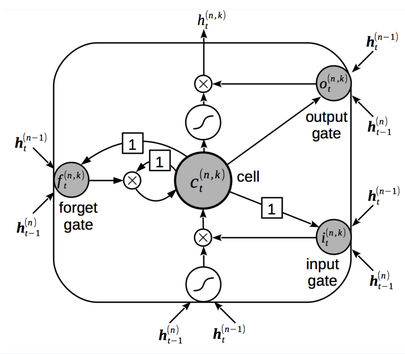
- BLSTM neural network is only trained on acoustic data and does not directly use language model information for the target speech for long term context
- How to provide this information to the neural network?
- Use ASR alignment features as additional inputs to the network
- Obtain an alignment of HMM states to the one-best decoding of an ASR system
- How to encode the alignment features?
  - One-hot state-alignment vector
  - Average/power-mean features corresponding to the aligned state in training data

EXPERIMENTAL SETUP



- Apply speech separation methods for speech enhancement of CHiME-2 data
- 2-layer (B)LSTMs followed by one feedforward layer and sigmoid nonlinearity
- Training: Input Gaussian noise, backprop, stochastic gradient, init from earlier net if possible
- First train with MA objective to predict a mel-mask, then add one more layer to expand to full fft spectrum and train with the MSA/PSA objective
- Evaluate using SDR/SIR on CHiME-2 dev and eval sets

LSTM CELL



CHIME-2 DEV SET SDR (IN DB)

Network	Loss	Input	-6 dB	9 dB	Avg
LSTM 2x256	MA	mfb	8.77	16.71	12.76
BLSTM 2x256	MA	mfb	8.92	16.83	12.90
BLSTM 2x384	MA	mfb	9.39	16.97	13.19
LSTM 2x256	MSA	mfb	9.24	16.93	13.03
BLSTM 2x384	MSA	mfb	9.76	17.28	13.45
LSTM 2x256	PSA	mfb	9.71	17.09	13.36
BLSTM 2x384	PSA	mfb	10.21	17.43	13.76
LSTM 2x256	MSA	mfb+align-avg	9.64	17.92	13.36
LSTM 2x256	MSA	mfb+align-pm	9.59	17.15	13.33
BLSTM 2x384	PSA	mfb+align-avg	10.50	17.56	**13.97**

CHIME-2 EVAL SET SDR/SIR (IN DB)

Network	Loss	Input	Avg-SDR	Avg-SIR
LSTM 2x256	MSA	mfb	13.83	17.53
BLSTM 2x384	MSA	mfb	14.22	18.24
LSTM 2x256	PSA	mfb	14.14	19.20
BLSTM 2x384	PSA	mfb	14.51	19.78
BLSTM 2x384	PSA	mfb+align-avg	**14.75**	**20.46**

CHIME-2 DEV/EVAL SETS WER - NEW [2]

Enhancement	WER (dev) Avg	WER (eval) Input SNR [dB] -6	9	Avg
None	29.39	40.31	13.86	23.41
NMF-MSA	28.38	37.57	12.63	22.02
LSTM-MSA	23.99	30.92	11.68	18.63
LSTM-PSA	23.72	30.90	11.34	18.31
BLSTM-PSA	22.87	29.20	11.26	17.74
BLSTM+align+PSA	**21.54**	**28.04**	**10.97**	**16.58**

- State-of-the-art sequence-discriminatively trained DNN-HMM hybrid ASR system, 15.1 hr train, 4.6 hr dev, 4 hr test
- Speech enhancement reduces WER about 7% absolute

CONCLUSIONS AND FUTURE WORK

- Improved speech enhancement results by:
  - bidirectionality of the recurrent network
  - phase-sensitive spectrum approximation
  - incorporating speech recognition alignment information within the LSTM-DRNN framework
- Future work may be: prediction of the target phase, phase consistency, preserve uncertainty in speech estimates, tighter integration of language model information

REFERENCES

[1] F. J. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *GlobalSIP Machine Learning Applications in Speech Processing Symposium*, 2014.

[2] H. Erdogan, S. Watanabe, J. R. Hershey, and J. Le Roux, "Noise-robust speech recognition with channel adaptive training of recurrent speech enhancement neural networks," 2015, submitted to Interspeech.

The first author was supported by TUBITAK BIDEB-2219 program.