Multichannel Speech Enhancement Based on Time-frequency Masking Using Subband Long Short-Term Memory

Xiaofei Li and Radu Horaud
Inria Grenoble, France

Introduction

Scenario

- Single speaker
- Ambient noise
- Microphone array

Task

- Speech denoising

Introduction

Conventional unsupervised subband methods

- Single-channel: noise estimation + spectral subtraction [1]
- Multichannel: beamforming (spatial filtering) [2]

[1] Y. Ephraim, et al, "Speech enhancement using a minimum-meansquare error short-time spectral amplitude estimator," *TASP*, 1984.

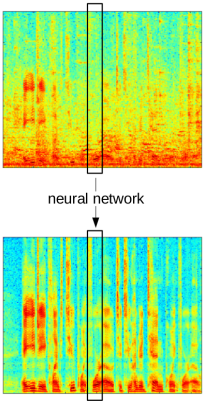
[2] S. Gannot, el al, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *TASLP*, 2017.

Introduction

Deep neural network methods [3]

Single-channel: spectral regression

- Learn using full-band spectra and dynamics (past, present and future frames)
- Sensitive to speaker identities and noise types
- Input/output (noisy speech / clean speech) : high dimensional vectors, therefore a very large network is necessary



[3] D. Wang, et al, "Supervised speech separation based on deep learning: An overview," *TASLP*, 2018.

Introduction

Deep neural network methods

Multichannel

- Single-channel spectral regression + beamforming [4]
- Full-band spatial information learning [5]

[4] C. Boeddeker, et al, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," *ICASSP*, 2018.

[5] S. Chakrabarty, et al, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," *IWAENC*, 2018.

Introduction

Proposed: subband LSTM (Long Short-Term Memory)

Motivation: unsupervised methods (spectral subtraction and beamforming) show that subband-based methods are powerful to discriminate between speech and noise:

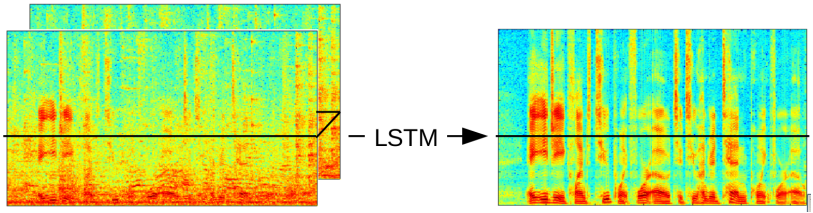
- Speech is temporally nonstationary / noise exhibits more stationarity
- Speech is spatially coherent / noise is diffuse
- Spectral subtraction and spatial filtering are performed in subbands

Objective: train an LSTM network that exploits subband information.

Subband LSTM network

Train one LSTM common to all frequencies

- Learn spatial and temporal information common to all frequencies
- Exclude cross-band spectral information
- Map *multichannel-sequence*-to-*speech-sequence* for each subband



Subband LSTM network

Signal formulation in the STFT domain: $x^i_{t,k} = s^i_{t,k} + u^i_{t,k}$

- Input: STFT Fourier coefficients

$$\mathbf{x}_{t,k} = [\mathcal{R}(x^1_{t,k}), \mathcal{I}(x^1_{t,k}) \dots \mathcal{R}(x^I_{t,k}), \mathcal{I}(x^I_{t,k})]^\top \in \mathbb{R}^{2I}$$

- Target: rectified magnitude ratio mask (MRM)

$$M_{t,k} = \min\left(\frac{|s^1_{t,k}|}{|x^1_{t,k}|}, 1\right) \quad \in [0, 1]$$

- Loss:

$$(\hat{M}(\mathbf{x}_{t,k}, \theta) - M_{t,k})^2 \quad \forall k$$

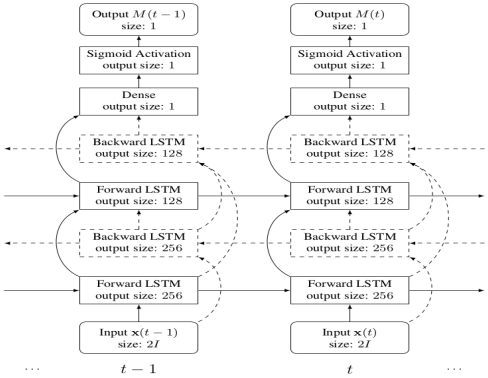
- Test: denoising using the predicted soft mask

$$|\hat{s}_{t,k}| = \hat{M}_{t,k}|x^1_{t,k}|,$$
$$\arg(\hat{s}_{t,k}) = \arg(x^1_{t,k})$$

Subband LSTM network

Bidirectional LSTM network

- Sequence-to-sequence
- Small-sized network: two BLSTM layers, 1.19 M learnable parameters



Experiments

CHiME3 dataset [6], six microphones, four noise types

- BUS: bus
- CAF: coffee shop
- PED: pedestrian area
- STR: street junction

Data generation: mutichannel *booth speech* with added backgroud noise

- Four speakers for training, one for validation, three for test
- Training data: about 11 hours generated data, cut into sequences with length of 192 frames (LSTM learns 192 time steps of memory)
- Test data: original CHiME utterances with varying lengths, SNR is 0 dB

[6] J. Barker, et al, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *ASRU*, 2015.

Experiments

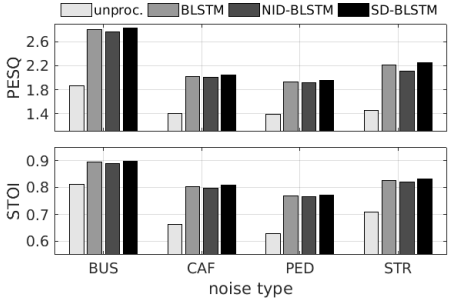
Performance metrics

- PESQ (perceptual evaluation of speech quality): speech quality
- STOI (short-time objective intelligibility): speech intelligibility

Experiments

Network generalization performance in terms of speaker and noise type

- BLSTM: speaker-independent, noise-type-dependent
- NID-BLSTM: speaker-independent, noise-type-independent
- SD-BLSTM: speaker-dependent, noise-type-dependent
- Results: two microphones



Experiments

Baseline methods

- BeamformIt: unsupervised filter-and-sum beamforming [7]
- Multichannel full-band CNN [8]
  - Input: multichannel, single-frame, full-band STFT coefficients
  - Output: full-band time-frequency mask
  - CNN network

[7] X. Anguera, et al, "Acoustic beamforming for speaker diarization of meetings," *TASLP*, 2007.

[8] S. Chakrabarty, et al, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," *IWAENC*, 2018.

Experiments

Audio examples [9] (four microphones)

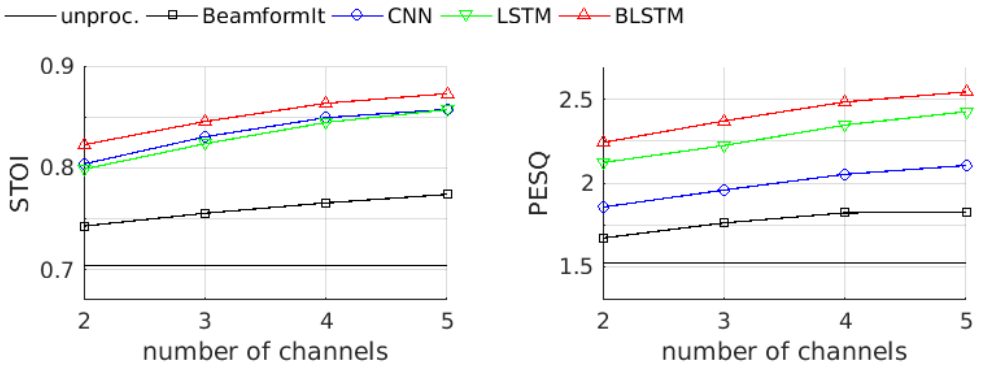
	unproc.	BeamformIt	CNN	BLSTM
CAF	audio	audio	audio	audio
STR	audio	audio	audio	audio

[9] https://team.inria.fr/perception/research/mse-lstm/

Experiments

Quantitative results (averaged over all noise types)

- The proposed achieves better speech quality



Conclusions

Multichannel subband LSTM

- Generalizes well to unseen speakers and noise types
- Complementary to full-band spectral regression methods

In the future we plan to experiment with other targets, such as

- *complex ideal ratio mask* or *Fourier coefficients* to predict both the module and the phase,
- *spatial filtering* to mimic beamforming.

Website

https://team.inria.fr/perception/research/mse-lstm/