Multiple Beamformers with ROVER for the CHiME-5 Challenge

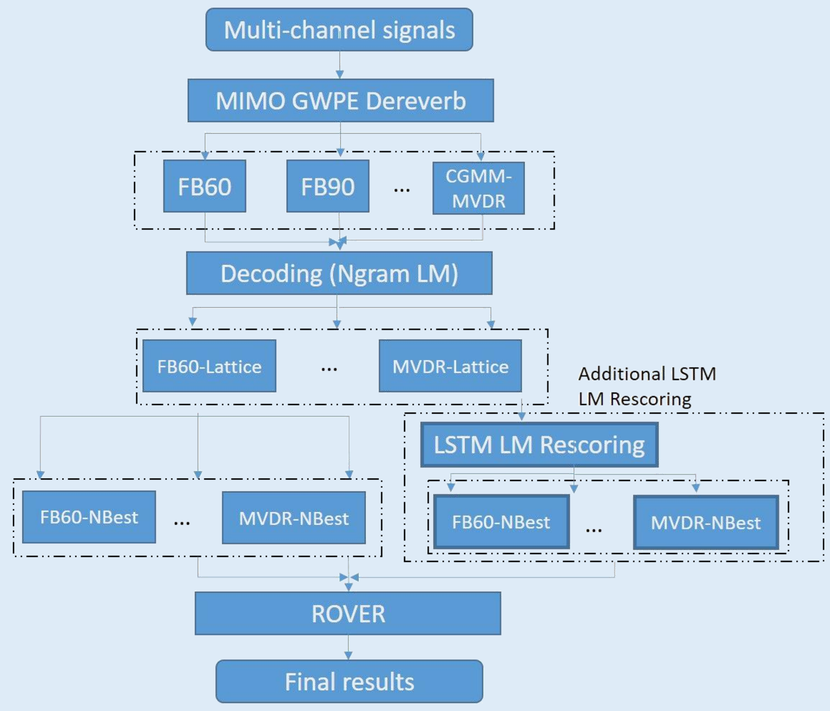
Sining Sun [1] , Yangyang Shi [2] , Ching-Feng Yeh [2] ,Suliang Bu [3],
Mei-Yuh Hwang [2] , Lei Xie [1]
[1] Northwestern Polytechnical University [1]
[2] Mobvoi AI Lab, Seattle, USA [2]
[3] Dept. of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA

Introduction

- **We only focus on single array track**
- **Main work**
  - Multi-channel generalized weighted prediction error (GWPE) dereverberation
  - Multiple beamformers (fixed beamformers and CGMM-MVDR) with Nbest lists ROVER
  - Data augmentation
  - CNN-TDNN-F acoustic model
  - LSTM language model

Details

- **GWPE dereverberation**
  - GWPE is Multi-Input and Multi-Output (MIMO) algorithm
  - GWPE can keep DOA information
  - Our matlab code is available at https://github.com/snsun/gwpe-speech-dereverb
- **Multiple beamformers with Nbest lists ROVER**
  - Because it is very difficult to get the accurate DOA estimation, we design several fixed beamformers (FB) and each of them only focuses on one specific direction
  - CGMM-MVDR beamformer is also used because of its great performance
  - Get Nbest lists for beamformed signals and the final results achieved by ROVER
- **Acoustic model**
  - One CNN layer before factorized time delay neural network (CNN-TDNN-F)
  - CNN-TDNN-F outperforms official TDNN model
- **LSTM language model**
  - For every beamformed speech, LSTM language model is used to do lattice rescoring and generate additional Nbest lists for ROVER



Experimental Evaluation

- **Acoustic models**
  - TDNN AM vs CNN-TDNN-F AM on dev set, using official training data

Training data	Beamforming	Acoustic model	WER(%)
Official (worn+100K far-filed)	BeamFormit	TDNN	81.30
		CNN-TDNN-F	75.91

- **Dereverberation and multiple beamformers with Nbest lists ROVER**
  - CNN-TDNN-F is used as AM
  - GWPE and our proposed beamforming strategy are used

Training data	Acoustic model	Dereverb	Beamforming	WER(%)
Official (worn+100K far-filed)	CNN-TDNN-F	No	Beamformit	75.91
		No	Multi beamformers with ROVER	72.54
		Yes		71.56

- **Data augmentation and LSTM LM rescoring**
  - Only 22K utterances are selected randomly to do dereverb and beamforming
  - Augment official training data using another enhanced 44K (CGMM-MVDR and 90-degree beamformer) utterances
  - LSTM LM Nbest lists are used during ROVER

Training data	Dereverb	Beamforming	Nbest	WER(%)
Offical	Yes	Multi beamformers with ROVER	3-gram	71.56
Augmented (Offical + 44K)			3-gram	70.68
			Additional LSTM LM	69.57

Summary

- On the front-end signal processing, combining GWPE dereverberaton and multiple beamformers with N-Best ROVER gave significant improvement
- On acoustic models, CNN-TDNN-F significantly improved over the TDNN backend
- On language models, LSTM language model rescoring led to a small WER reduction