The NWPU System for CHiME5 Challenge

Zhiwei Zhao, Jian Wu, Lei Xie

Audio, Speech and Language Processing Lab (ASLP@NWPU)

School of Computer Science, Northwestern Polytechnical University

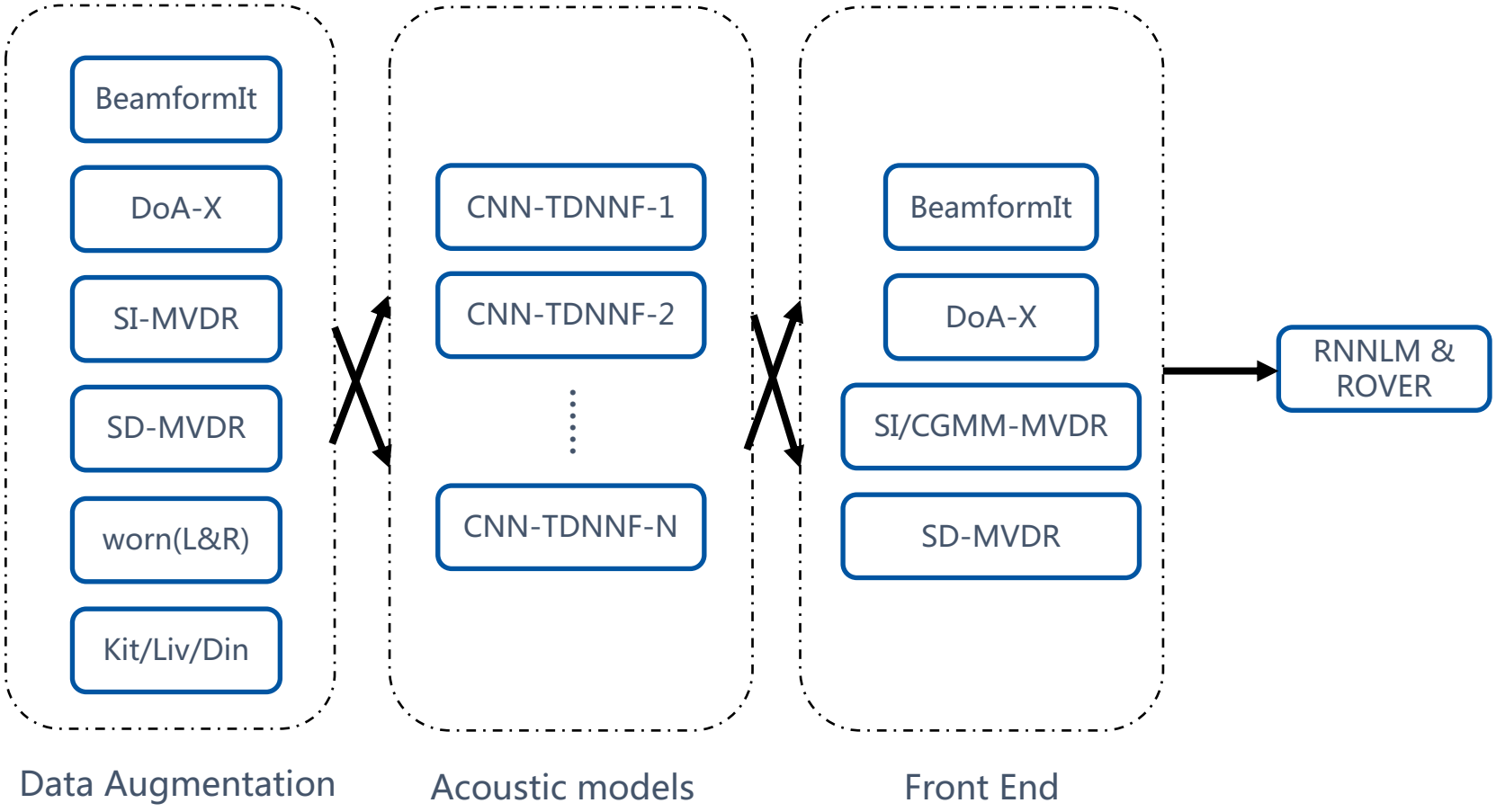
Outline

- System Overview
  - Single track challenge
  - Development set
    - ✓ N-gram 65.17%
    - ✓ RNNLM 63.54%
  - Evaluation set
    - ✓ N-gram 57.85%
    - ✓ RNNLM 56.09%
- Front End
- Acoustic Model
- System Fusion
- Summary

System Overview

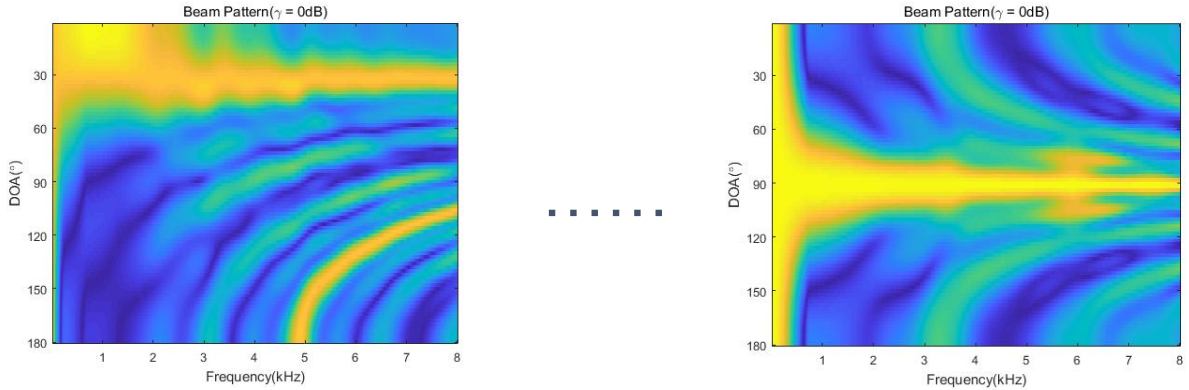
- Front End
  - Fixed Beamforming
  - Adaptive Beamforming with SI/SD masks
- Acoustic Models
  - Deep TDNN-F[D.Povey+2018]
  - Data Augmentation
- System Fusion
  - Location Level
  - Beamforming Level

System Overview



Front End

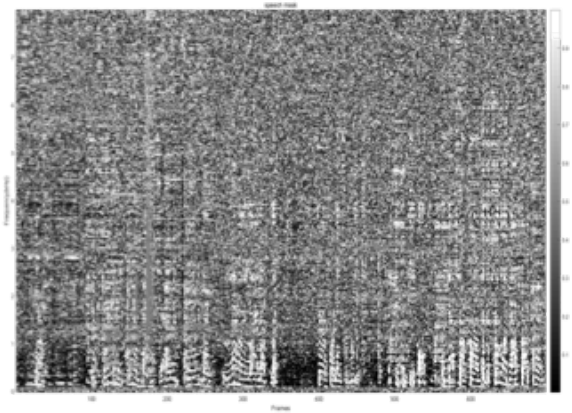
- Fixed Beamforming
  - Design fixed beamformers by sampling DoA (30, 60, 90…)
  - Ensemble methods



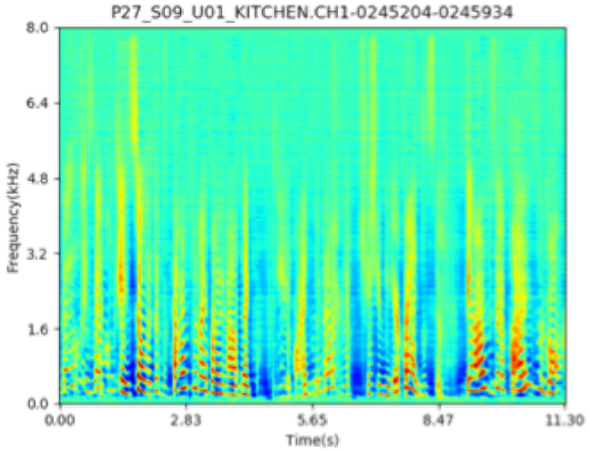
- Adaptive Beamforming
  - **MVDR**/GEV/PMWF……
  - TF-Mask (SI/SD) estimator

Front End

- Adaptive Beamforming (SI)
  - CGMM[NTT+2016] mask estimated on close-talk data as supervision
  - Corresponding far-field speech as input
  - Mask quantization brings little improvements



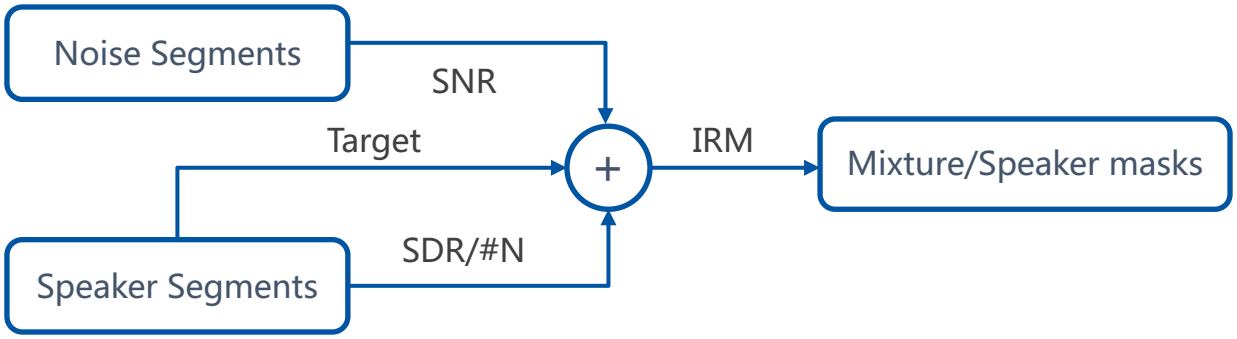
CGMM



NN

Front End

- Adaptive Beamforming (SD)
  - Model training
    - ✓ Split segments which have only one speaker by annotations
    - ✓ Treat non-speaker segments on training set as candidate noise
    - ✓ Exclude silence part from noise set using a simple VAD
    - ✓ Data simulation on above speaker and noise segments
    - ✓ Training IRM estimator on simulated data for each speaker



Front End

- Adaptive Beamforming (SD)
  - Model details
    - ✓ 3 layer BLSTMs
    - ✓ log filter bank feature as input
  - Results for each speaker (absolute WER reduction)
    - ✓ P05 2.91%+
    - ✓ P06 2.52%+
    - ✓ P07 3.00%+
    - ✓ P08 **3.62%+**
    - ✓ P25 1.86%+
    - ✓ P26 2.17%+
    - ✓ P27 *0.39%+*
    - ✓ P28 2.39%+

Front End

- Results on baseline AM

Methods	WER %
Beamformit	82.40
CGMM-MVDR	81.08
SI-MVDR	80.82
SD-MVDR	80.02
DoA-X joint decoding	80.69
SD-MVDR + DoA-X joint decoding	80.05

- Joint decoding means state-level posterior average
- In submitted transcripts, we use ROVER [J.G.Fiscus+1997] instead of posterior average

Acoustic Models

- Factored form of TDNN
- Location dependent training
- Data augmentation (MCT)

Acoustic Models

- Deep TDNN-F
  - Front end: official beamformer
  - Training data: same as baseline
  - TDNN-F brings **7% absolute reduction** on WER

Acoustic model	Feature	Dev (WER%)
Official GMM-HMM	MFCC	91.83
Official TDNN	MFCC	80.11
Official TDNN	LMFB	79.91
TDNN-F (11 layers)	LMFB	75.34
1CNN+TDNN-F (15 layers)	LMFB	73.16
1CNN + TDNN-F (19 layers)	LMFB	72.61

  - For TDNN-F, scaled case is better than floating case

Acoustic Models

- Data Augmentation
  - Basic data augmentation brings **2% absolute reduction** on WER

ID	Training Data	Dev(WER%)
Baseline	1+3	72.61
Model 1	1+3+5	70.49
Model 2	1+3+4	71.32

- ✓ 1 → All close talk data
- ✓ 3 → Far-field data used in baseline setup
- ✓ 4 → DoA-90 enhanced speech (25k, 60h)
- ✓ 5 → WS (weight and sum/Beamformit) on far-field data (460h)

Acoustic Models

- Data Augmentation
  - Further data augmentation could bring **3% absolute WER reduction**
    - ✓ Fix beamforming {DoA-60,90,120}
    - ✓ SI/SD-MVDR beamforming on far-field training data
    - ✓ Single channel speech enhancement
  - 5k utterances per kinect, 25k utterances in total for each method
  - Results on dining room
    - ✓ Before 305h 71.83%
    - ✓ After 605h 68.95%
  - Not fully finished until submission

System Fusion

- ROVER on different AMs & Front Ends
  - ROVER brings **3% absolute WER reduction**
  - Baseline N-gram LM

Beamformer	Model1	Model 2	Model{3~5}
CGMM-MVDR	70.05%	70.80%	71.22%
DoA-105	69.67%	70.40%	70.72%
DoA-90	69.87%	70.62%	71.15%
DoA-60	69.66%	70.63%	70.94%
Beamformit	70.21%	71.07%	71.55%
SI-MVDR	68.83%	69.53%	70.07%
SD-MVDR	68.66 %	69.06%	69.57%
ROVER	65.17%		

  - Model 3~5: TDNN-F trained on dining/living/kitchen rooms

System Fusion

- ROVER on different AMs & Front Ends
  - RNNLM brings **2% absolute WER reduction**
    - ✓ LSTM-TDNN structure & Kaldi-RNNLM toolkit
    - ✓ 3-order pruned lattice-rescoring algorithm

Beamformer	Model1	Model 2	Model{3∼5}
CGMM-MVDR	68.25%	68.68%	69.08%
DoA-105	68.13%	68.43%	68.65%
DoA-90	68.13%	68.65%	69.25%
DoA-60	68.09%	68.64%	69.09%
Beamformit	68.50%	69.28%	69.48%
SI-MVDR	66.98%	67.39%	68.07%
SD-MVDR	66.91%	67.00%	67.63%
ROVER	63.54%		

Summary

• Tune baseline	2%+	
• Front End	2%+	
• Acoustic Model	7%+	
• Partial Augmentation	2%+	
• System fusion	3%+	16%+ total
• RNNLM	2%+	18%+ total

Zhiwei Zhao, Jian Wu, Lei Xie
ASLP@NWPU
URL : www.npu-aslp.org

Thank You