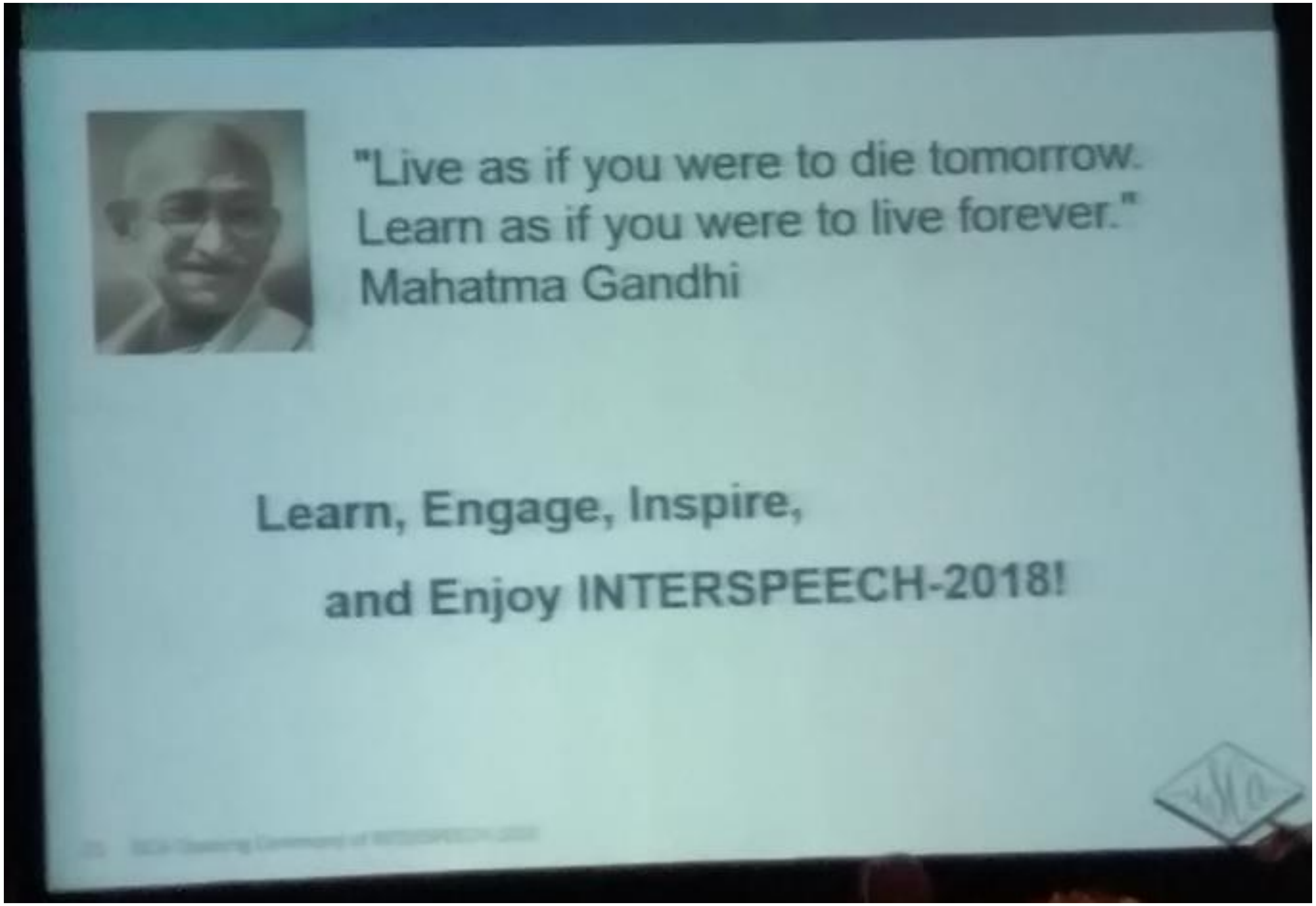
Interspeech 2018 paper review

Zhehuai Chen

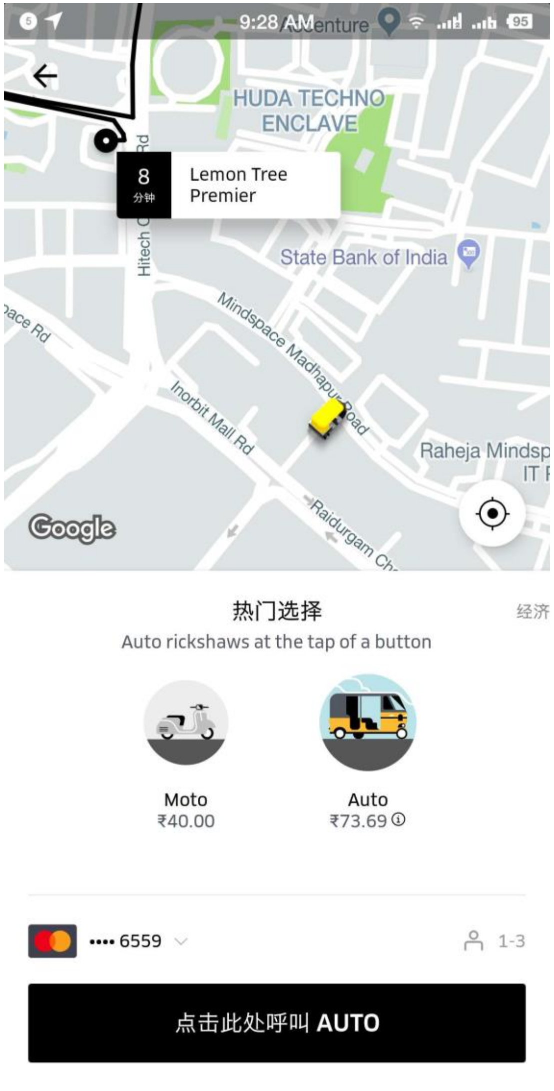
Opening



- Video

Life in India

- Video



Machine Speech Chain with One-shot Speaker Adaptation

Andros Tjandra[1,2], *Sakriani Sakti*[1,2], *Satoshi Nakamura*[1,2]

[1]Nara Institute of Science and Technology, Graduate School of Information Science, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

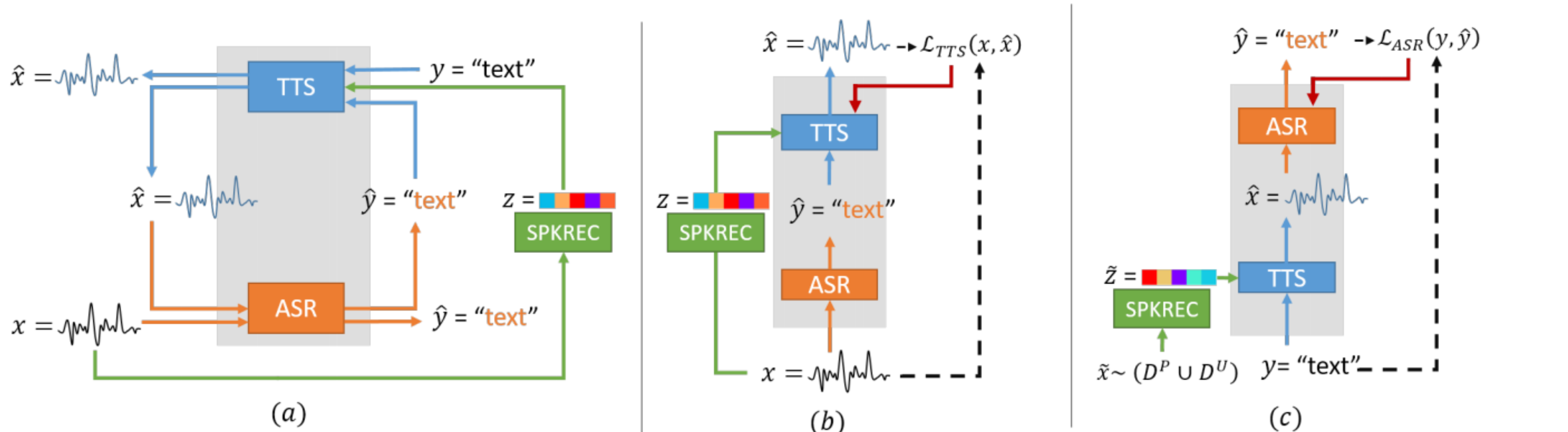
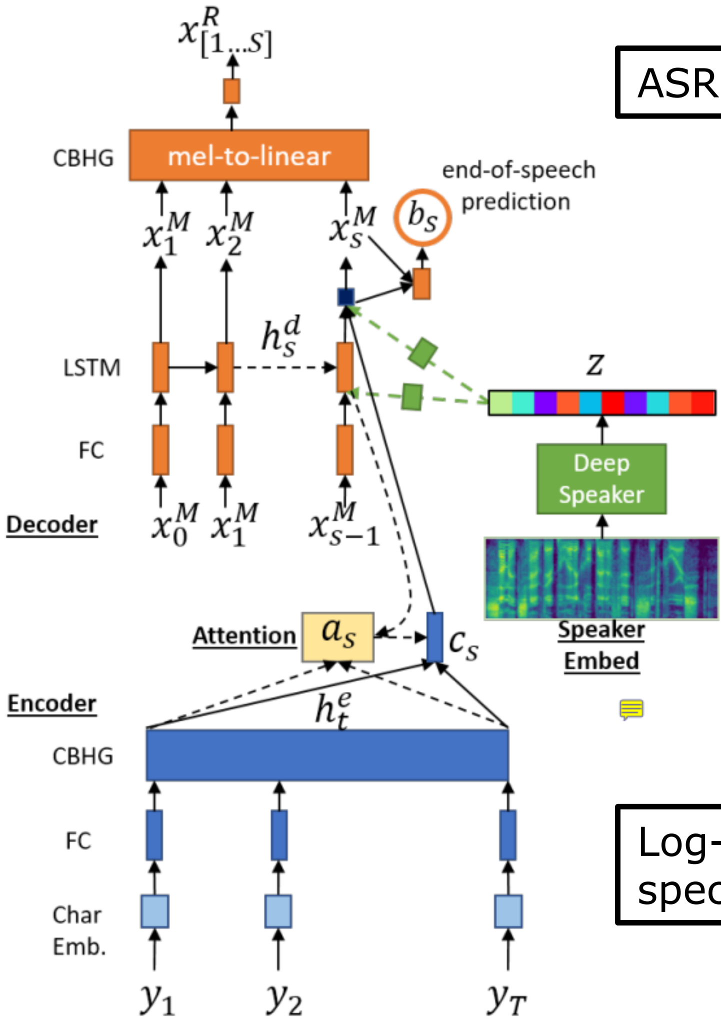


Figure 1: *(a) Overview of proposed machine speech chain architecture with speaker recognition; (b) Unrolled process with only speech utterances and no text transcription (speech → **[ASR,SPKREC]** → [text + speaker vector] → **TTS** → speech); (c) Unrolled process with only text but no corresponding speech utterance ([text + speaker vector by sampling **SPKREC**] → **TTS** → speech → **ASR** → text). Note: grayed box is the original speech chain mechanism.*

- “闭环学习”
- **Sample speaker vector to do multi-style training**



ASR

Model	CER (%)
Supervised training: WSJ *train_si84* (paired) → Baseline	
Att Enc-Dec [19]	17.01
Att Enc-Dec [20]	17.68
Att Enc-Dec (ours)	17.35
Supervised training: WSJ *train_si284* (paired) → Upperbound	
Att Enc-Dec [19]	8.17
Att Enc-Dec [20]	7.69
Att Enc-Dec (ours)	7.12
Semi-supervised training: WSJ *train_si84* (paired) + *train_si200* (unpaired)	
Label propagation (greedy)	17.52
Label propagation (beam=5)	14.58
Proposed speech chain (Sec. 2)	**9.86**

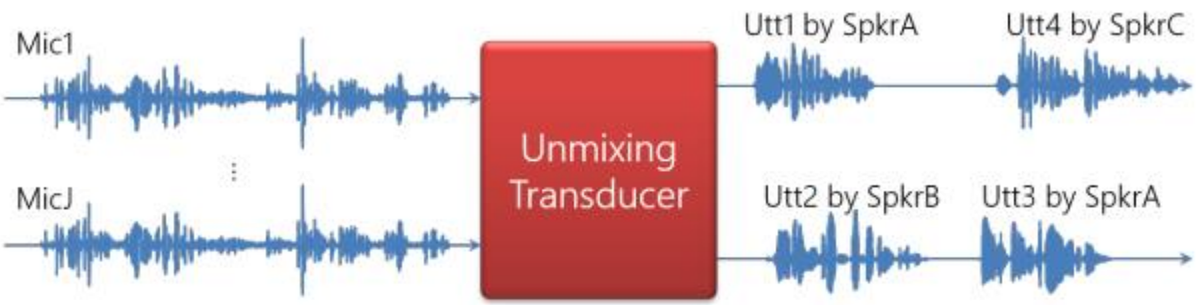
Log-mel spectrogram

Model	L2-norm^2
Supervised training: WSJ *train_si84* (paired) → Baseline	
Proposed Tacotron (Sec. 4) (ours)	1.036
Supervised training: WSJ *train_si284* (paired) → Upperbound	
Proposed Tacotron (Sec. 4) (ours)	0.836
Semi-supervised training: WSJ *train_si84* (paired) + *train_si200* (unpaired)	
Proposed speech chain (Sec. 2 + Sec. 4)	**0.886**

Recognizing Overlapped Speech in Meetings: A Multichannel Separation Approach Using Neural Networks

Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva

Microsoft AI and Research, One Microsoft Way, Redmond, WA, USA



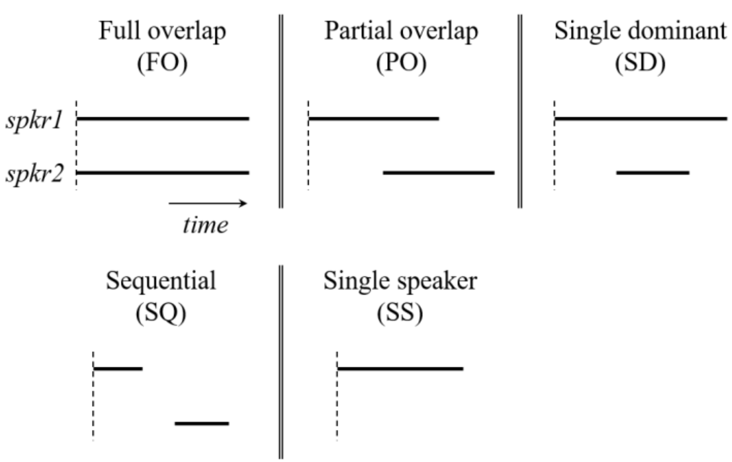
- **REAL conference task**

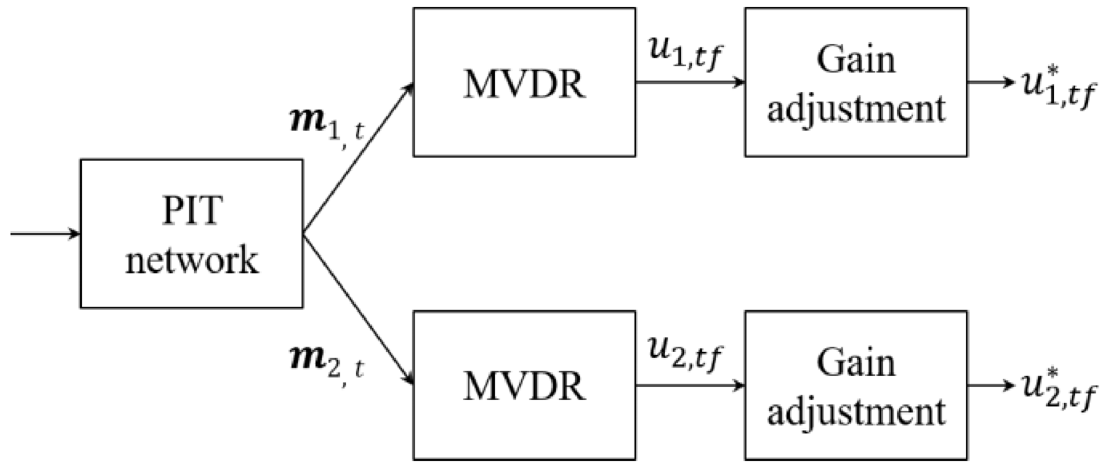
MULTI-MICROPHONE NEURAL SPEECH SEPARATION FOR FAR-FIELD MULTI-TALKER SPEECH RECOGNITION

Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Fil Alleva

Microsoft AI and Research, One Microsoft Way, Redmond, WA

- Spectral and spatial inputs:
  - The magnitude spectra
  - Inter-microphone phrase diff (IPD) to the first one
- Mask-driven beamforming outputs (separate ASR)
  - Mask-driven MVDR beamforming
  - Gain adjustment
- Test 5 cases





- Beamforming with noise estimation

$$y_{i,tf} = \boldsymbol{w}^{H}_{c,i,f}\boldsymbol{x}_{tf}$$

target

$$\boldsymbol{w}_{c,i,f} = \boldsymbol{\Psi}^{-1}_{c,i,f}\boldsymbol{\Phi}_{c,i,f}\boldsymbol{e}/\rho_{c,i,f}$$

$$\boldsymbol{\Psi}_{c,i,f} = \boldsymbol{\Phi}_{c,\bar{i},f} + \boxed{\boldsymbol{\Phi}_{c,\mathrm{N},f}}$$

interfering

noise

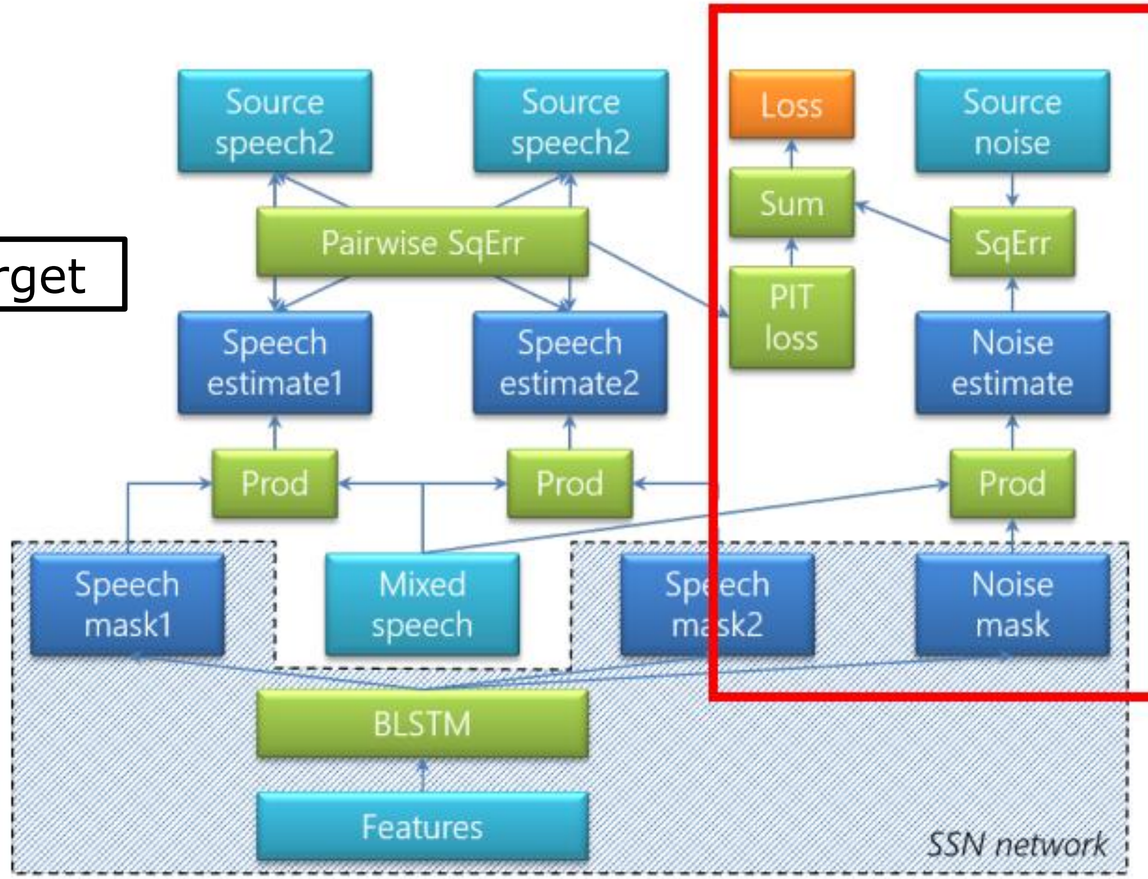


Figure 3: *SSN model and the network for training it.*

- DOA differences to decide num. of speakers

Table 1: *%WER of different front-ends.*

System	Overlapped segments Included	Overlapped segments Excluded
No processing (mic0)	44.6	40.9
Dereverb. [17]	42.1	38.7
+BeamformIt [29]	43.2	40.6
+MaskBF [23]	37.9	32.8
+Unmix. Trans. (proposed)	**33.8**	**30.4**
+UT trained only on WSJMix	34.2	30.8
+UT without noise channel	36.8	34.5

14% overlap real test data

Half data

Without noise estimation

Improve both non-overlapped & overlapped

[23] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, accepted.

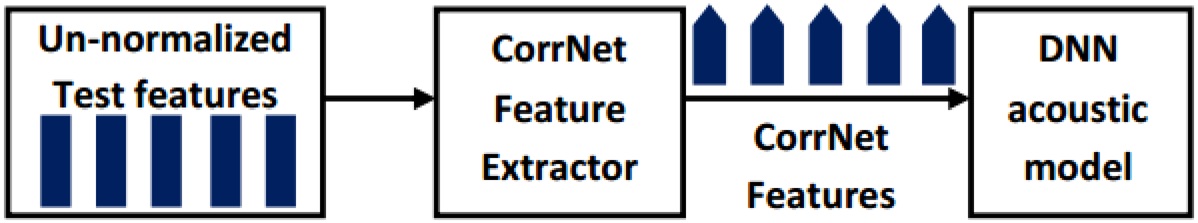
Correlation Networks for Speaker Normalization in Automatic Speech Recognition

Rini Sharon A, Sandeep Reddy Kothinti, Srinivasan Umesh

Indian Institute of Technology Madras, India

- Motivation:
  - Fmllr needs 2-pass decoding
  - i-vector needs long utt.s

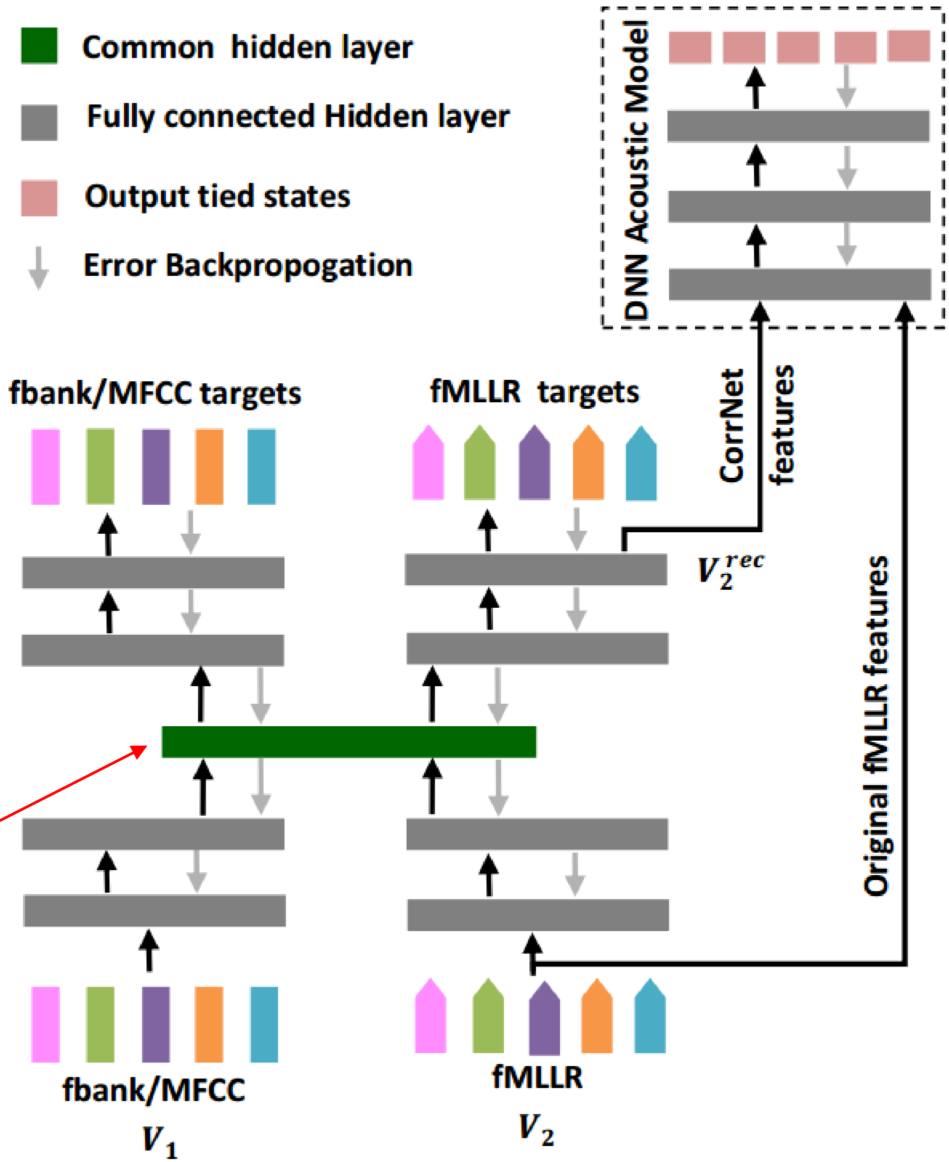
- Proposed decoding pipeline:

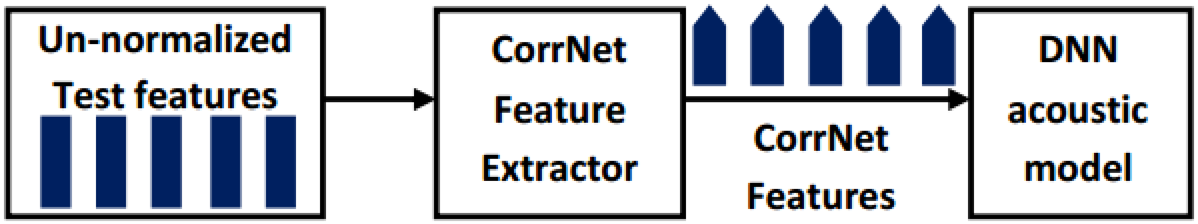


- Training
  - Reconstruct fmllr from itself
  - Reconstruct fmllr from fbank
  - Reconstruct fmllr from fmllr+fbank
  - Maximize the correlation between fmllr and fbank

$$
\begin{aligned}
Loss = & L_{mse}([none, V_2], V_2^{rec}) \\
& + L_{mse}([V_1, none], V_2^{rec}) \\
& + L_{mse}([V_1, V_2], V_2^{rec}) \\
& - \lambda \times L_{corr}(\mathbb{P}(V_1), \mathbb{P}(V_2))
\end{aligned}
$$

$$
L_{corr}(A, B) = \frac{\sum_{i=1}^{N}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{N}(A_i - \overline{A})^2 \sum_{i=1}^{N}(B_i - \overline{B})^2}}
$$





Input features to DNN Acoustic Model	TIMIT				SWBD-33				WSJ-84	
	spk-norm		*utt-norm*		*spk-norm eval2000*		*utt-norm eval2000*		*spk-norm*	*utt-norm*
	test	*dev*	*test*	*dev*	*swbd*	*callhm*	*swbd*	*callhm*	eval	eval
MFCC	20.3	18.9	21.4	20.0	23.7	35.5	24.5	35.7	14.26	15.86
MFCC + i-vectors	20.2	18.3	20.9	19.6	23.4	35.2	24.1	35.5	14.0	15.61
Filterbank	20.0	18.4	21.4	20.7	22.8	34.6	24.3	34.6	13.74	14.96
Filterbank + i-vectors	19.5	17.9	21.6	19.3	22.04	33.6	23.6	34.8	13.57	14.19
fMLLR	18.3	17.4	25.4	24.8	**20.8**	**31.4**	25.0	40.1	11.56	18.24
fMLLR + i-vectors	**18.1**	**17.1**	25.1	23.8	21.02	31.4	24.7	40.1	**11.36**	18.04
CorrNet Models										
CorrNet (Recon fM←fb)	19.6	18.0	19.7	18.4	21.9	33.7	21.8	35.0	12.75	13.39
CorrNet (All loss)	19.4	17.9	19.0	18.3	21.53	32.5	21.9	34.5	12.64	13.29
CorrNet (Weighted loss)	19.4	17.8	19.0	18.3	21.5	32.6	21.7	34.5	12.58	13.23
Combined Scoring	18.8	17.7	**18.9**	**18.2**	21.1	32.5	**21.3**	**34.1**	12.52	**13.17**

- BUT, why does it work?

A Novel Approach for Effective Recognition of the Code-Switched Data on Monolingual Language Model

Ganji Sreeram, Rohit Sinha

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati - 781039, India

Homophone Identification and Merging for Code-switched Speech Recognition

Brij Mohan Lal Srivastava and Sunayana Sitaram

Microsoft Research India

Mandarin-English Code-switching Speech Recognition

Haihua Xu[1], Van Tung Pham[1,2], Zin Tun Kyaw[2], Zhi Hao Lim[1], Eng Siong Chng[1,2], Haizhou Li[3]

[1]Temasek Laboratories, Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Study of Semi-supervised Approaches to Improving English-Mandarin Code-Switching Speech Recognition

Pengcheng Guo[1,2], Haihua Xu[2], Lei Xie[1,], Eng Siong Chng[2,3]*

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2] Temasek Laboratories, Nanyang Technological University, Singapore
[3] School of Computer Science and Engineering, Nanyang Technological University, Singapore

- Text normalization
  - Numbers, temperatures, etc..

  e.g.那家酒店的顾客好评率在4.2分; he gave 4.2 points for his score
  - Homophones identification & merging
    - Clustering based on pronunciation of the words

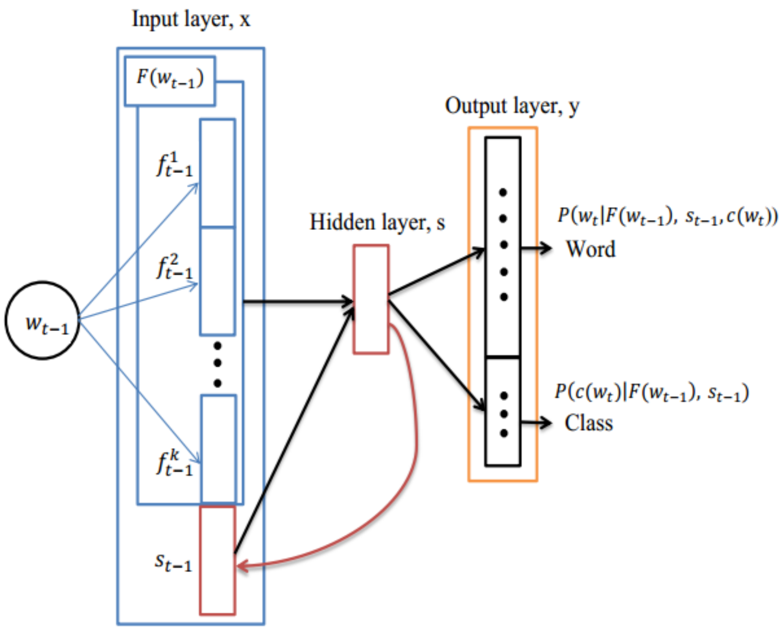
    e.g. 酷 k u4

    & cool k u4 r
- Lexicon learning
  - Collect alternative pronunciations from lexicon, G2P and phonetic decoding
  - Prune alternative pronunciations based on a data likelihood based criterion
  - Use new lexicon, change to a semi-supervised problem (re-decode & re-train)

- Get word-pair in 2 languages
  - Do translation
  - Get word-pairs from the translation alignment

  e.g. 好酷 赞 棒呆 good cool brilliant perfect

- Cluster the low freq. words → group them together as a class
- Use word-pairs all as input of NNLM (Factored LM)



- Get word-pair in 2 languages
  - Do translation
  - Get word-pairs from the translation alignment

  e.g. 好酷 赞 棒呆 good cool brilliant perfect

- Cluster the low freq. words → group them together as a class
- Use word-pairs all as input of NNLM (Factored LM)

- Solve the sparsity in code-switching LM
  - Class-LM based on word pairs & word class
  - Generate text to train the normal LM
  - Do some singleton phrase substitution for all above LMs

  e.g. 我 要 <u>收听</u> Taylor Swift 的 歌

  & 我 要 <u>欣赏</u> Taylor Swift 的 歌

Some papers for engineering

1. Improved Training of End-to-end Attention Models for Speech Recognition
2. End-to-end Speech Recognition Using Lattice-free MMI
3. Compression of End-to-End Models
4. Robust TDOA Estimation Based on Time-Frequency Masking and Deep Neural Networks
5. Comparison of an End-to-end Trainable Dialogue System with a Modular Statistical Dialogue System
6. Improving Attention Based Sequence-to-Sequence Models for End-to-End English Conversational Speech Recognition
7. Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning
8. A Multistage Training Framework for Acoustic-to-Word Model
9. Compressing End-to-end ASR Networks by Tensor-Train Decomposition
10. Phase-locked Loop Based Phase Estimation in Single Channel Speech Enhancement
11. Cycle-Consistent Speech Enhancement
12. Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition
13. Acoustic Modeling with Densely Connected Residual Network for Multichannel Speech Recognition
14. Attention-based End-to-End Models for Small-Footprint Keyword Spotting
15. Automatic Speech Recognition System Development in the "Wild"
16. An Investigation of Mixup Training Strategies for Acoustic Models in ASR
17. A Probability Weighted Beamformer for Noise Robust ASR
18. Investigations on Data Augmentation and Loss Functions for Deep Learning Based Speech-Background Separation