ICASSP 2018 paper review

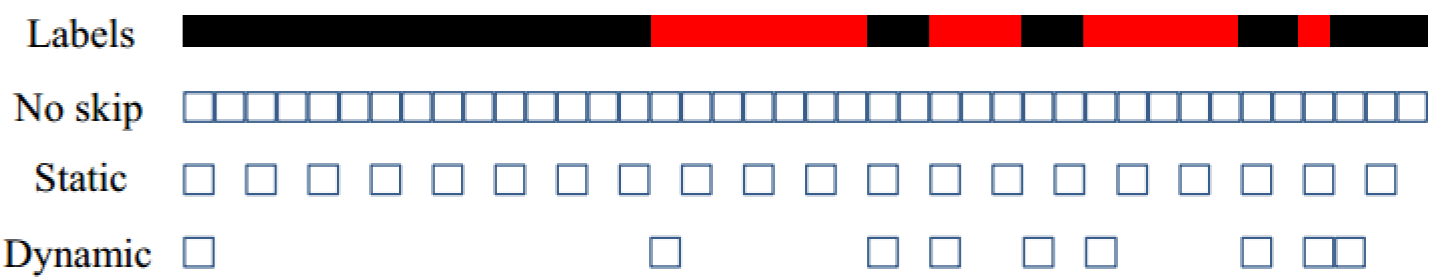
Zhehuai Chen

DYNAMIC FRAME SKIPPING FOR FAST SPEECH RECOGNITION IN RECURRENT NEURAL NETWORK BASED ACOUSTIC MODELS

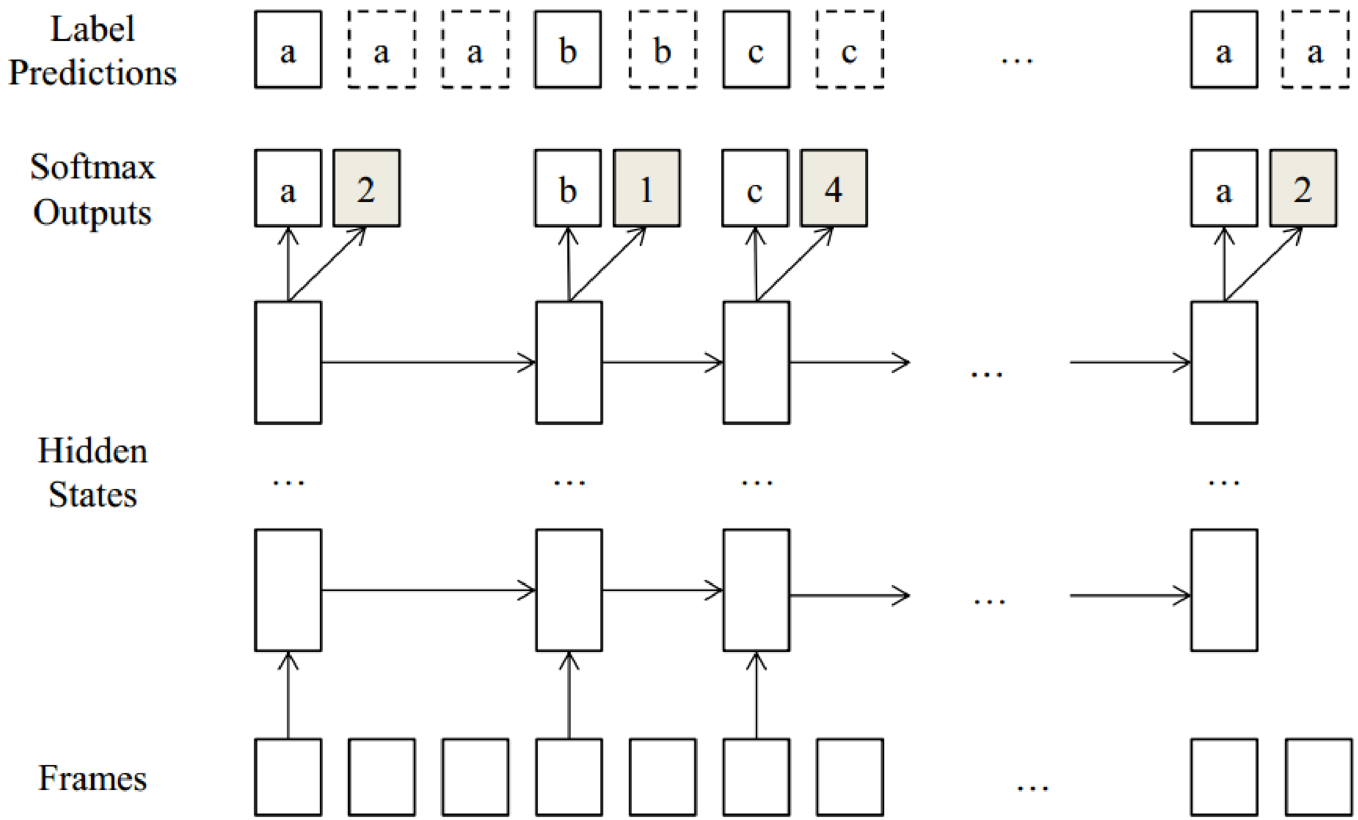
Inchul Song[1], *Junyoung Chung*[2*], *Taesup Kim*[2], *Yoshua Bengio*[2†]

[1]Samsung Advanced Institute of Technology, Republic of Korea
[2]MILA, Université de Montréal, Canada
inchul2.song@samsung.com, {junyoung.chung,taesup.kim,yoshua.bengio}@umontreal.ca

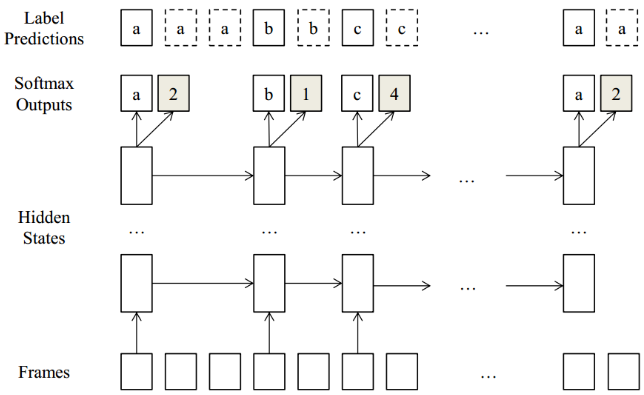
- Motivation



- Motivation
- Framework
  - Using state alignment



- Motivation
- Framework
- Method
  - Policy gradients (for future decision)



$$J(\theta_s) = \mathbb{E}_{\pi_{\theta_s}}\left[\sum_{i=1}^{N} \gamma^{i-1} r_i\right]$$

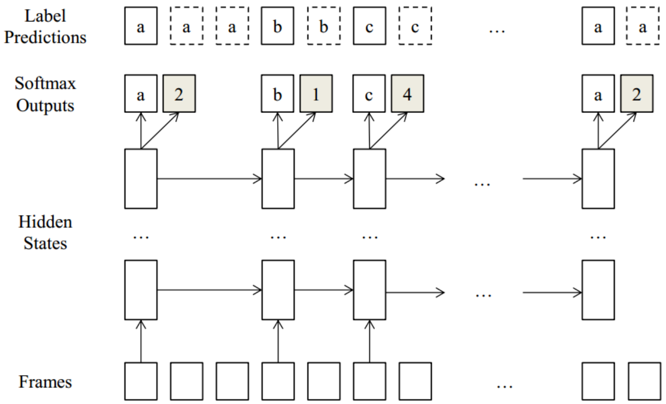
$$r_i = -|s_i^* - s_i|$$

$$\nabla_{\theta_s} J(\theta_s) = \mathbb{E}_{\pi_{\theta_s}}\left[\sum_{i=1}^{N} \nabla_{\theta_s} \log \pi_{\theta_s}(s_i|h_i) R_i\right]$$

$$R_i = \sum_{k=i}^{N} \gamma^{k-i} r_k$$

SHANGHAI JIAO TONG UNIVERSITY

- Motivation
- Framework
- Method: Policy gradients
- Result
  - Bad
- TODO
  - Stack feature?
  - Better using alignment? (label delay)
  - Better criteria?
  - Sequence (long temporal) criteria?



SEQUENCE-TO-SEQUENCE ASR OPTIMIZATION VIA REINFORCEMENT LEARNING

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

[1] Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2] RIKEN, Center for Advanced Intelligence Project AIP, Japan
{andros.tjandra.ai6, ssakti, s-nakamura}@is.naist.jp

- Motivation
  - Sequence level criterion deriving from RL
- Method
  - Agent: S2S model;
  - State: context & hidden state in S2S;
  - Action: output label set

$$\pi_\theta(a_t|s_t) = P(y_t|h_t^{D(n)}, c_t^{(n)}; \theta) = P(y_t|\mathbf{y}_{<t}, \mathbf{x}^{(n)}; \theta)$$

  - Reward: WER variants

$$\nabla_\theta E_{\mathbf{y}} \left[R^{(n)} | \pi_\theta \right] = \nabla_\theta \int P(\mathbf{y}|\mathbf{x}^{(n)}; \theta) R^{(n)} \, d\mathbf{y}$$

- Motivation: sequence level criterion deriving from RL
- Method
  - Agent: S2S model;
  - State: context & hidden state in S2S;
  - Action: output label set
  - Reward: WER variants
    - change to temporal distributed reward

$$\nabla_\theta E_{\mathbf{y}}\left[R^{(n)}|\pi_\theta\right] = \nabla_\theta \int P(\mathbf{y}|\mathbf{x}^{(n)};\theta)R^{(n)}\,d\mathbf{y}$$

$$\downarrow$$

$$\nabla_\theta E_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t^{(n)}|\pi_\theta\right]$$

$$= E_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t^{(n)} \sum_{t=1}^{T} \nabla_\theta \log P(y_t|\mathbf{y}_{<t},\mathbf{x}^{(n)};\theta)\right]$$

- Motivation: sequence level criterion deriving from RL
- Method
  - Agent: S2S model;
  - State: context & hidden state in S2S;
  - Action: output label set
  - Reward: WER variants
    - change to temporal distributed reward: whether becomes worse

$$\nabla_\theta E_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t^{(n)} | \pi_\theta\right]$$

$$= E_{\mathbf{y}}\left[\sum_{t=1}^{T} r_t^{(n)} \sum_{t=1}^{T} \nabla_\theta \log P(y_t | \mathbf{y}_{<t}, \mathbf{x}^{(n)}; \theta)\right]$$

$$r_t^{(n)} = -(ED(\mathbf{y}_{1:t}, \mathbf{y}^{(\mathbf{n})}) - ED(\mathbf{y}_{1:t-1}, \mathbf{y}^{(\mathbf{n})}))$$

- Motivation: sequence level criterion deriving from RL
- Method
- Comparison with "Minimum Risk Training for Neural Machine Translation"

$$\nabla_\theta E_{\mathbf{y}}\left[R^{(n)}|\pi_\theta\right] = \nabla_\theta \int P(\mathbf{y}|\mathbf{x}^{(n)};\theta)R^{(n)}\,d\mathbf{y}$$

$$\mathcal{L}_{\text{werr}}(\mathbf{x},\mathbf{y}^*) = \mathbb{E}[\mathcal{W}(\mathbf{y},\mathbf{y}^*)] = \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})\mathcal{W}(\mathbf{y},\mathbf{y}^*)$$

- Sampling method
- Reward construction

$$\sum_{\mathbf{y}_i \in \text{Beam}(\mathbf{x},N)}$$

$$\left[\mathcal{W}(\mathbf{y}_i,\mathbf{y}^*) - \widehat{W}\right]$$

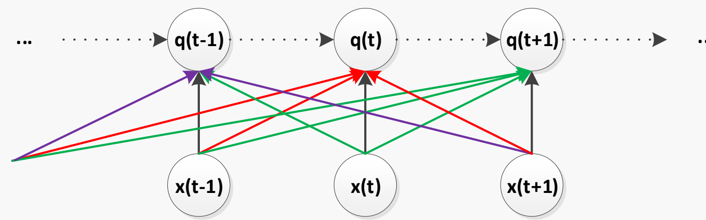
$$r_t^{(n)} = -(ED(\mathbf{y}_{1:t},\mathbf{y}^{(\mathbf{n})}) - ED(\mathbf{y}_{1:t-1},\mathbf{y}^{(\mathbf{n})}))$$

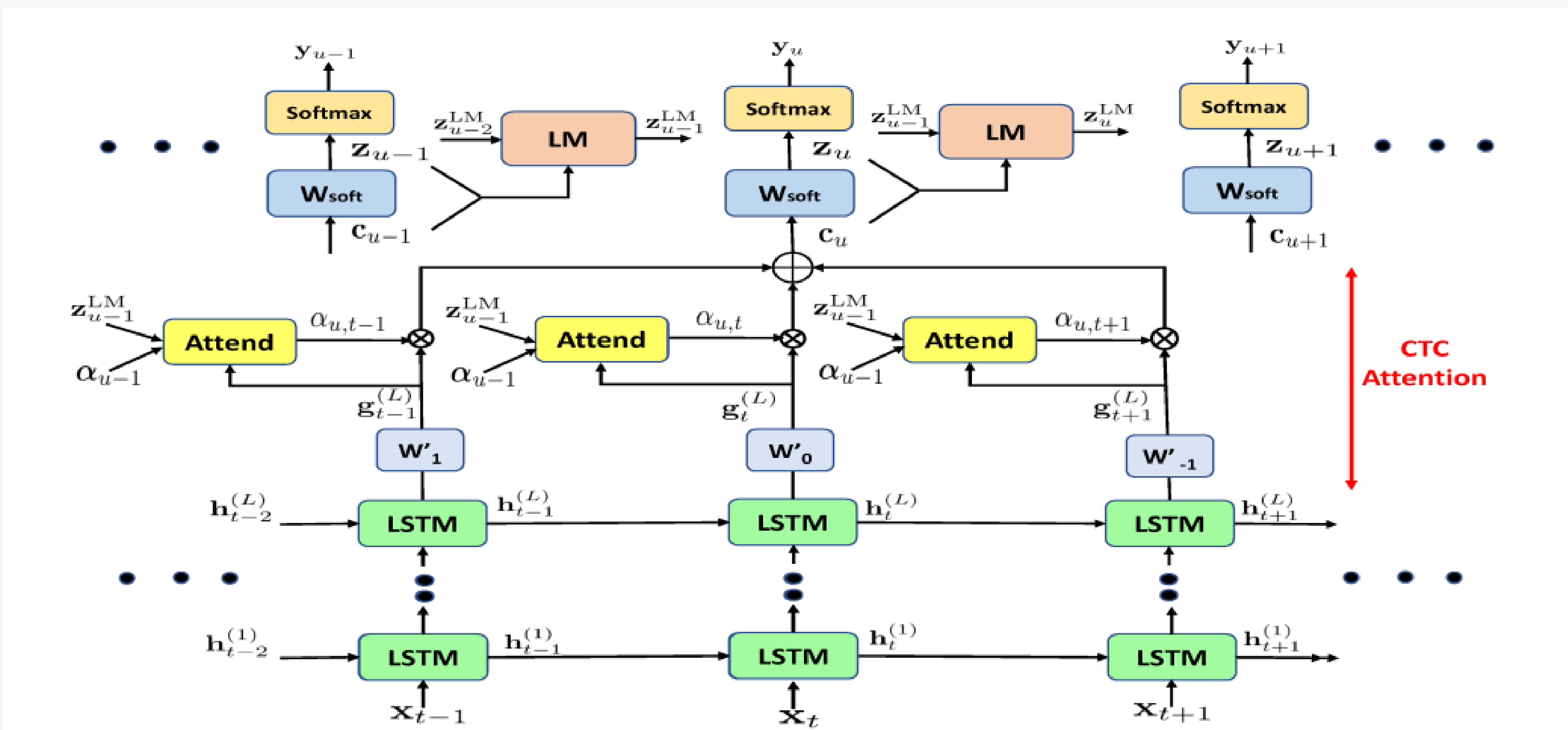
ADVANCING CONNECTIONIST TEMPORAL CLASSIFICATION WITH ATTENTION MODELING

Amit Das, Jinyu Li, Rui Zhao, Yifan Gong*

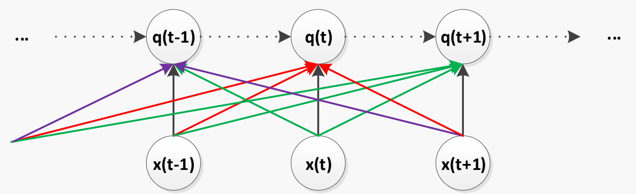
Microsoft AI and Research, One Microsoft Way, Redmond, WA 98052

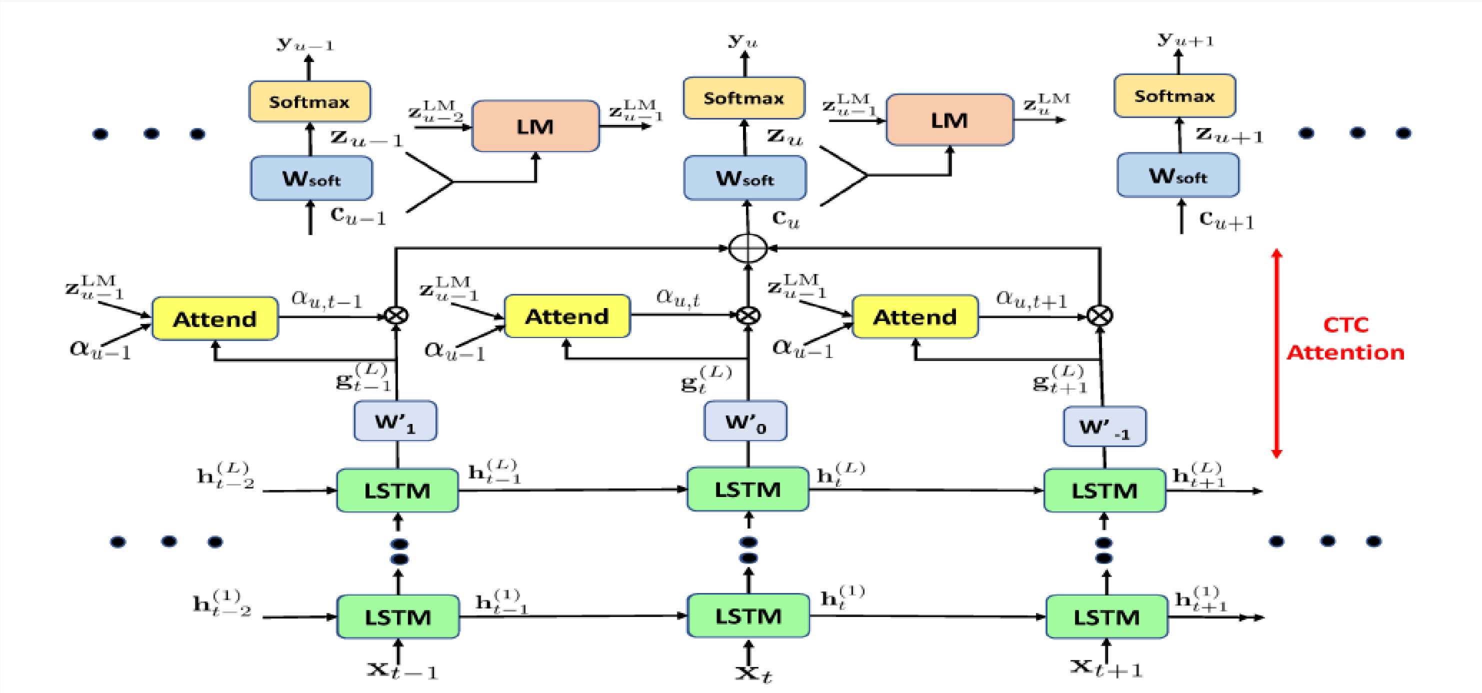
- Motivation:
  - 1. hard align -> soft align
  - 2. change modeling but not criterion





- Method
  - 1. Chunk based
  - 2. time convolution to obtain g_t
  - 3. output z_u to replace h_u in obtaining attention weight \alpha
  - 4. diff weight \alpha for diff dimension of g_t
  - 5. Add language model as a “decoder”





- WER Results
  - Result 1: single letter CTC: 23.29 → 18.49
  - Result 2:

E2E Model	WER (	
	Vanilla	Attention
single-letter	17.54	14.30
double-letter	15.37	12.16
triple-letter	13.28	11.36

  - Result 3:

mixed (OOV: word + triple-letter) CTC	9.32
mixed (OOV: word + triple-letter) attention CTC	8.65

SUPERVISED NOISE REDUCTION FOR MULTICHANNEL KEYWORD SPOTTING

Yiteng (Arden) Huang, Thad Hughes, Turaj Z. Shabestary, Taylor Applebaum

Google Inc., USA

{ardenhuang, thadh, turajs, applebaum}@google.com

- Motivation
  - Baseline
    - per-channel energy normalization (PCEN)
    - 2-channel adaptive noise cancellation (ANC)

$$\mathcal{E}(j\omega, m) \triangleq X_1(j\omega, m) - \mathbf{h}^H(j\omega, m)\mathbf{x}_2(j\omega, m)$$

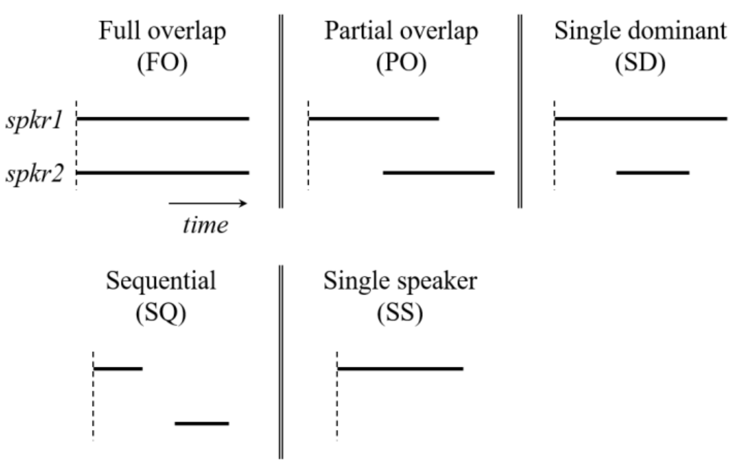
  - Decide noise or speech based on 1-pass KWS result
  - If speech → ANC succeeds to get clean speech → do not change filter coefficients
  - If noise → ANC fails to clean the speech → change coefficients and double check KWS

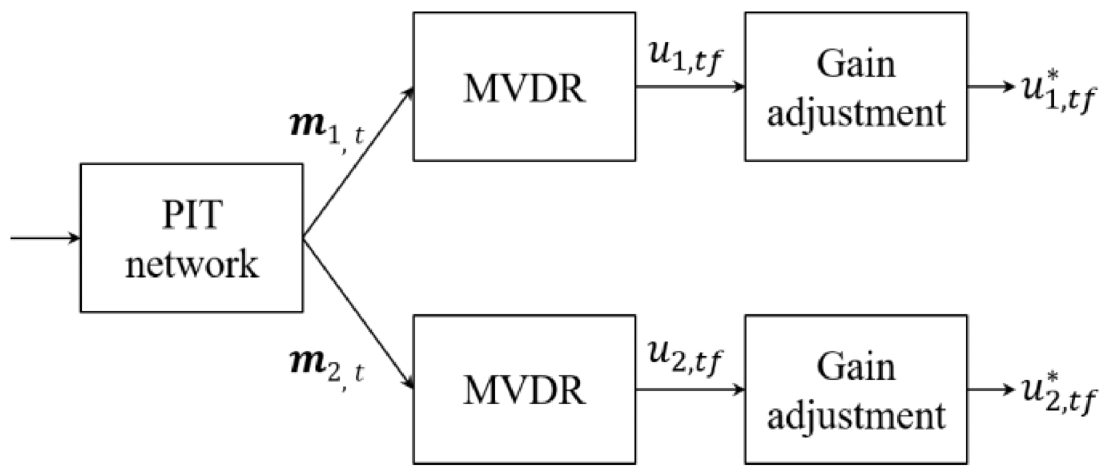
MULTI-MICROPHONE NEURAL SPEECH SEPARATION FOR FAR-FIELD MULTI-TALKER SPEECH RECOGNITION

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





EFFICIENT INTEGRATION OF FIXED BEAMFORMERS AND SPEECH SEPARATION NETWORKS FOR MULTI-CHANNEL FAR-FIELD SPEECH SEPARATION

Zhuo Chen, Takuya Yoshioka, Xiong Xiao, Jinyu Li, Michael L. Seltzer, Yifan Gong

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

- Beam prediction
  - the best beam is related to both target and interfering speakers (cannot directly use DOA information)
  - CE between N-hot selection vector and prediction (N spks)
- Multi-view PIT