AICAS 2022

An Attention-based Neural Network on Multiple Speaker Diarization

Kai Jyun Hung, Shao Wen Cheng,
Hsie Chia Chang, Yen Chin Liao

2022 IEEE International Conference on Artificial Intelligence Circuits and Systems
Virtual & Hybrid Conference

National Yang Ming Chiao Tung University

Outline

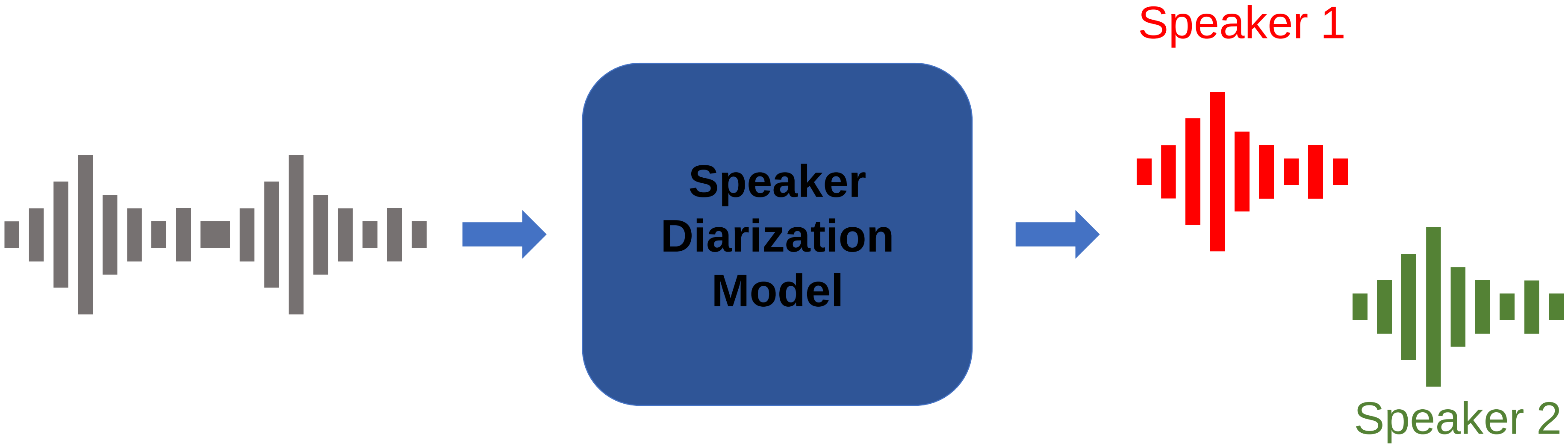
- Introduction
- Background
- Proposed Method
- Experimental Results
- Reference

Outline

- Introduction
- Background
- Proposed Method
- Experimental Results
- Reference

Introduction

- Speaker Diarization
  - Who spoke when?
  - Multiple speakers in each speech utterance



Outline

- Introduction
- Background
- Proposed Method
- Experimental Results
- Reference

Background

- Evaluation metrics - Diarization Error Rate
  - Evaluation metrics in this work

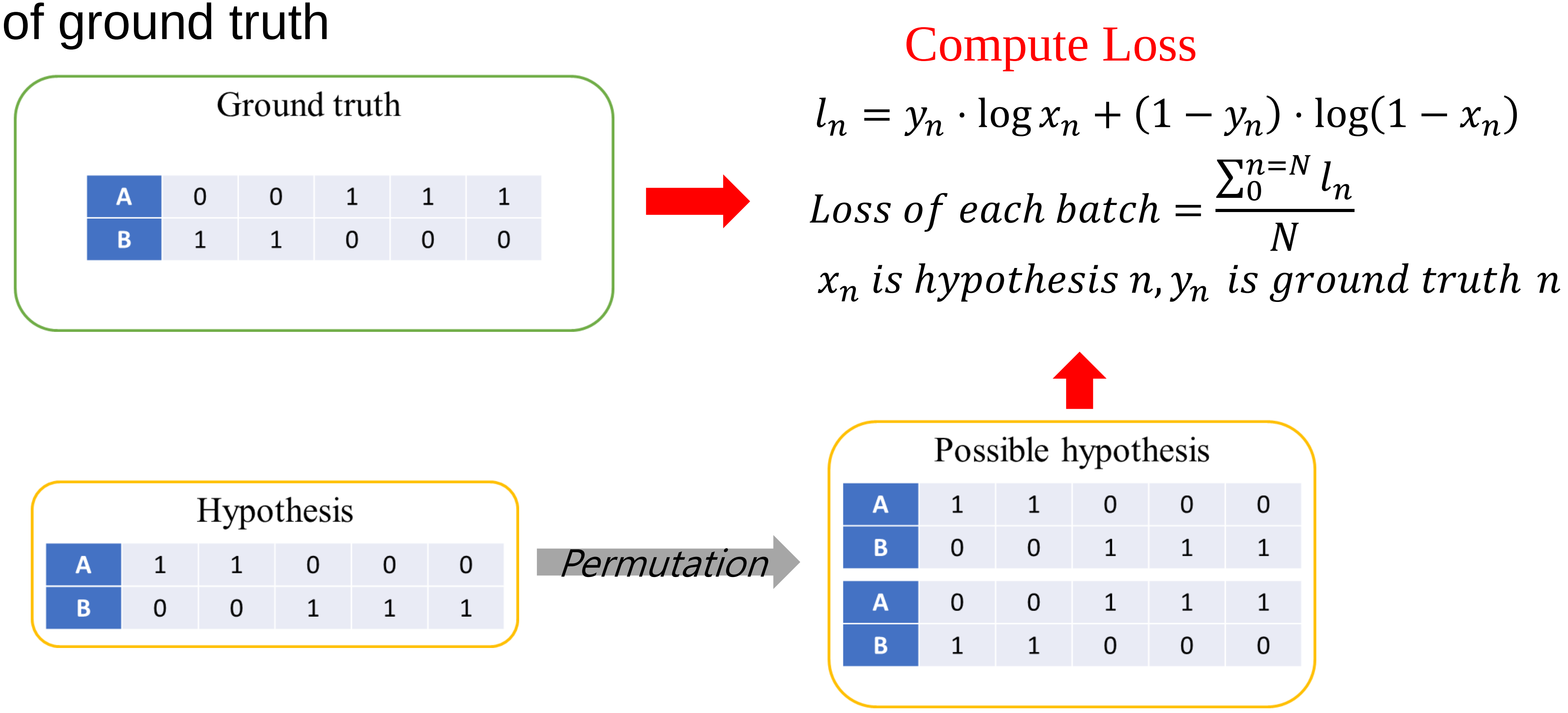
$$DER = \frac{Miss\ Detection + False\ Alarm + Overlap + Confusion}{Reference}$$

  - Notice that some papers adopt simplified metrics

$$DER\_Simplified = \frac{Confusion}{Reference}$$

Background

- Permutation-Free Objectives with Cross Entropy Loss [1]
    - Find minimum loss between hypothesis and all the possible combinational of ground truth



Compute Loss

$$l_n = y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)$$

$$Loss\ of\ each\ batch = \frac{\sum_0^{n=N} l_n}{N}$$

x_n *is hypothesis* n, y_n *is ground truth* n

Ground truth

A	0	0	1	1	1
B	1	1	0	0	0

Hypothesis

A	1	1	0	0	0
B	0	0	1	1	1

Permutation

Possible hypothesis

A	1	1	0	0	0
B	0	0	1	1	1

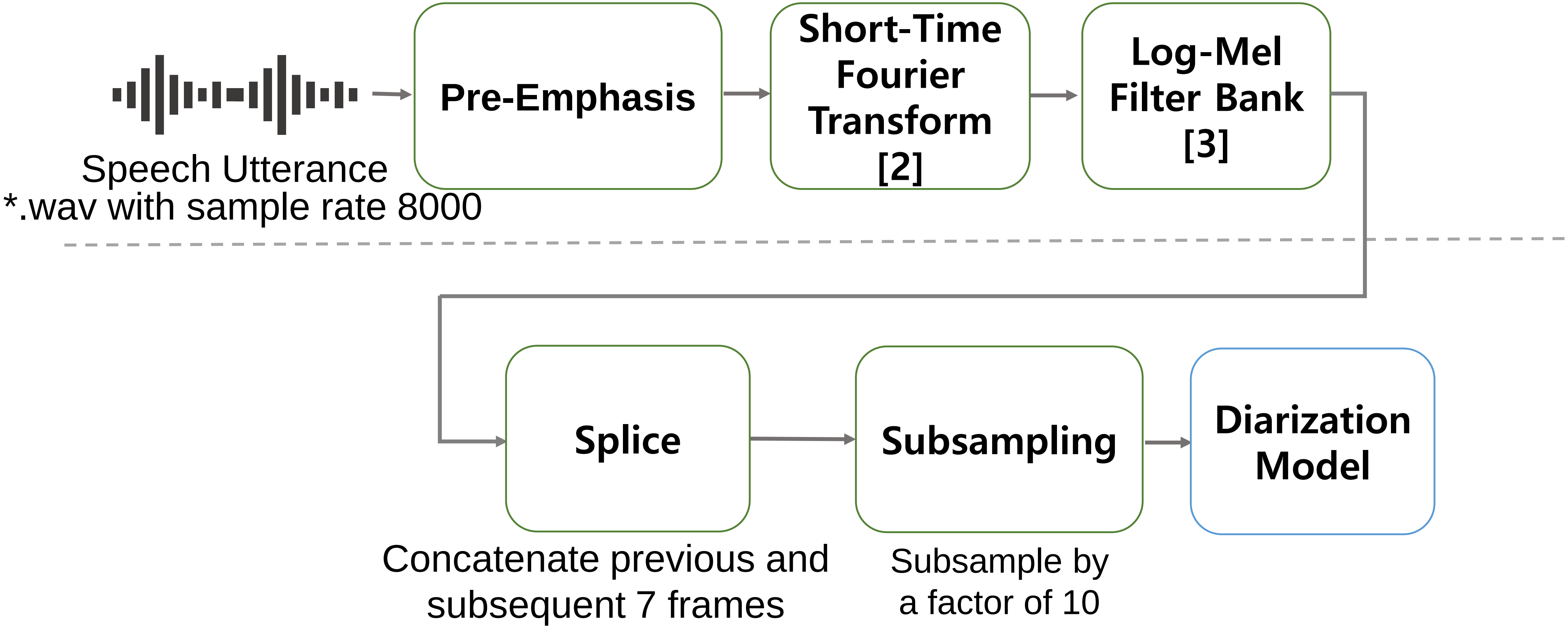
A	0	0	1	1	1
B	1	1	0	0	0

Outline

- Introduction
- Background
- Proposed Method
- Experimental Results
- Reference

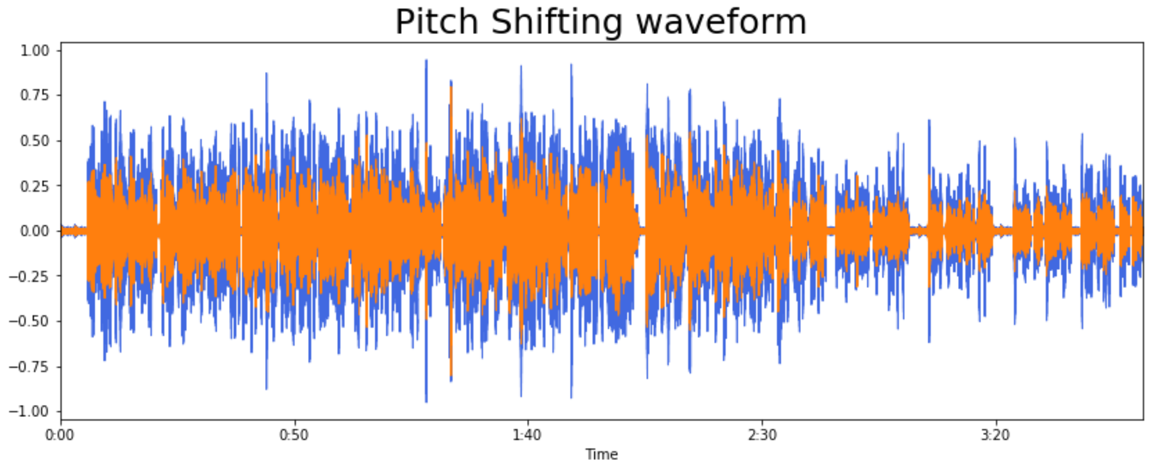
Proposed Method

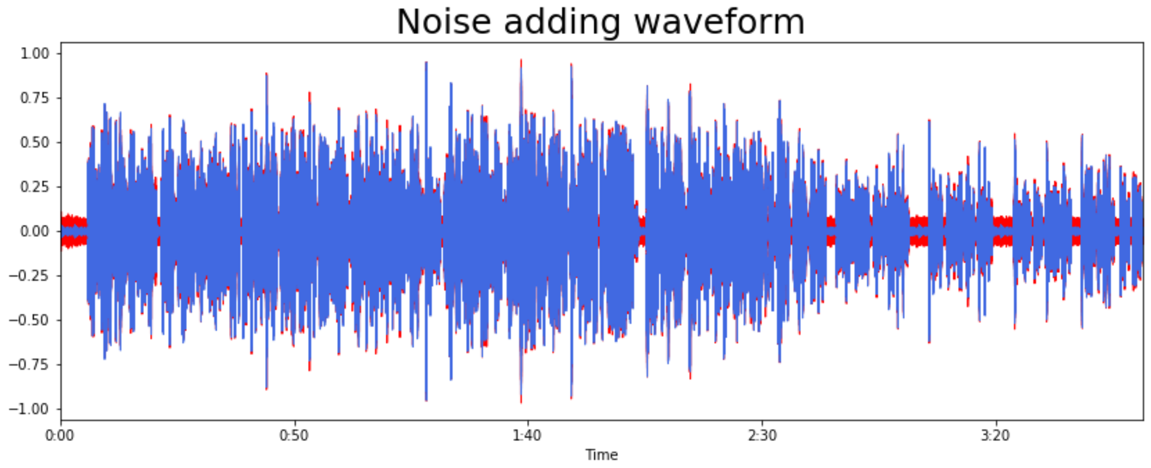
- Data Preprocessing


Speech Utterance
*.wav with sample rate 8000
Pre-Emphasis
Short-Time Fourier Transform [2]
Log-Mel Filter Bank [3]
Splice
Concatenate previous and subsequent 7 frames
Subsampling
Subsample by a factor of 10
Diarization Model

Proposed Method

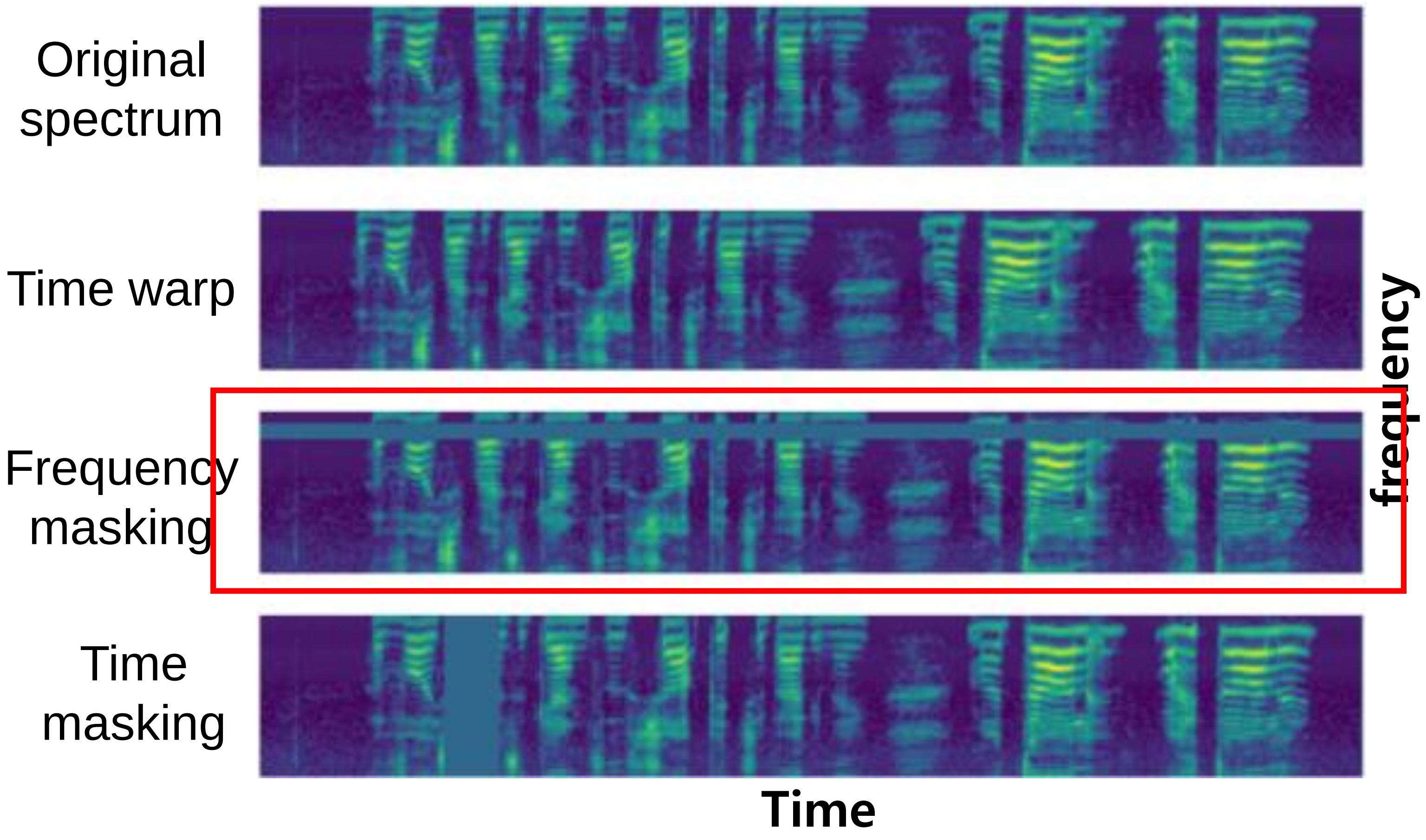
- Data Augmentation
  - Pitch shifting
  - Noise adding
  - SpecAugment [4]


Pitch Shifting waveform


Noise adding waveform

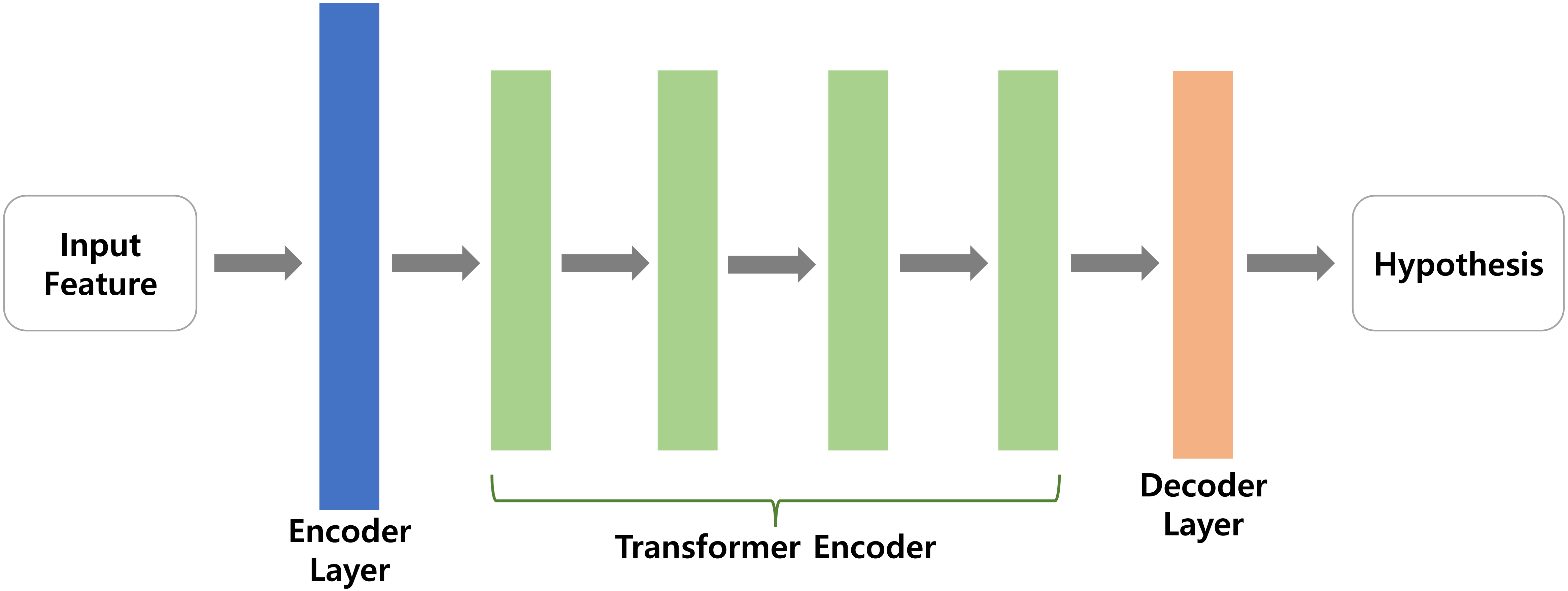
Proposed Method

- Data Augmentation
  - Pitch shifting
  - Noise adding
  - SpecAugment [4]



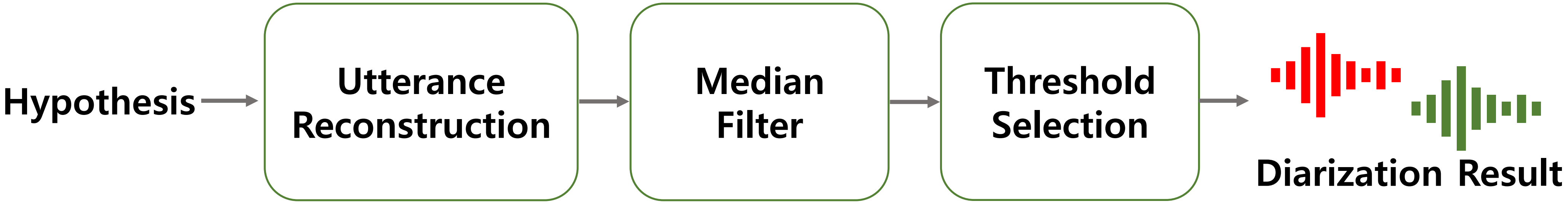
Proposed Method

- Model Structure [5]



Proposed Method

- Data Postprocessing



Hypothesis

Utterance Reconstruction

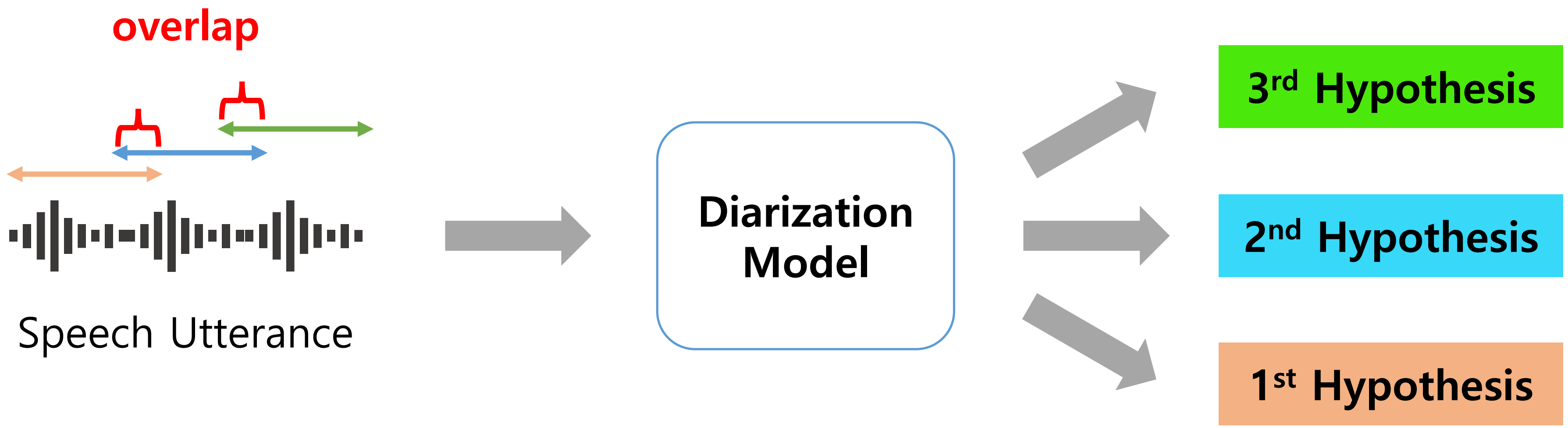
Median Filter

Threshold Selection

Diarization Result

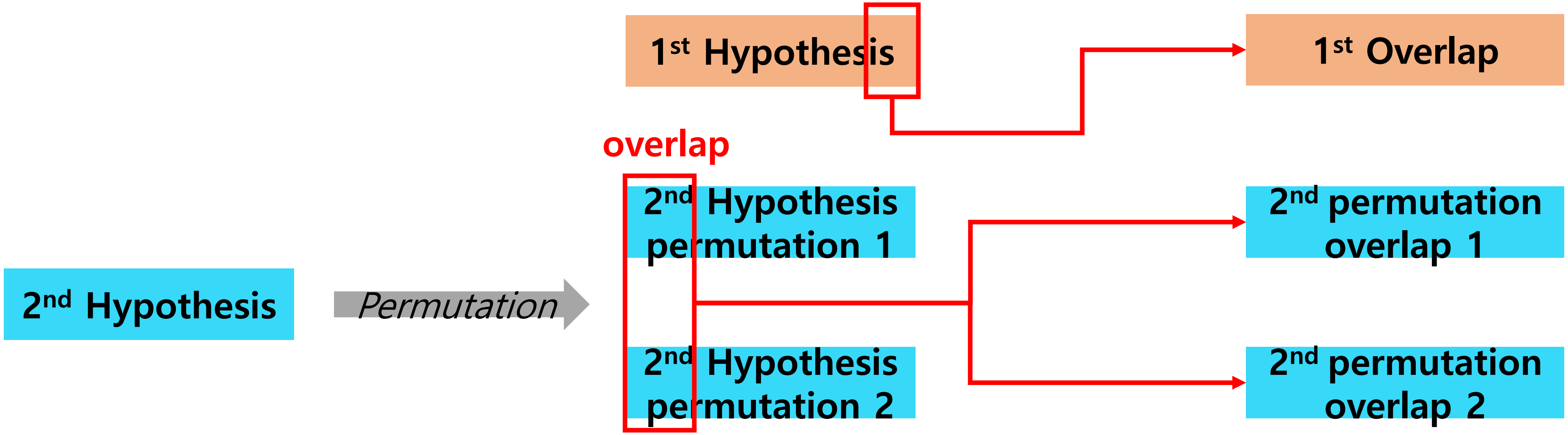
Proposed Method

- Utterance Reconstruction (1/4)
  - Split the speech utterance before inference due to length limited. Then co ncatenate the result by computing the loss of overlapping part.



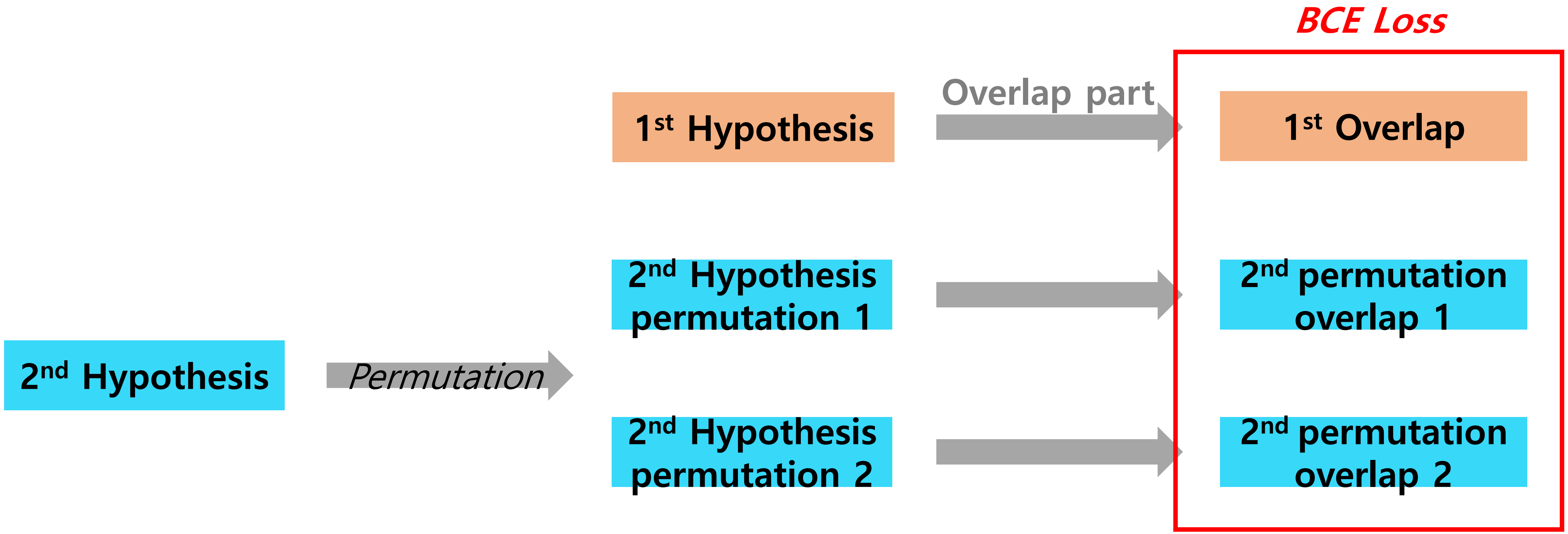
Proposed Method

- Utterance Reconstruction (2/4)
  - Split the speech utterance before inference due to length limited. Then co ncatenate the result by computing the loss of overlapping part.


1st Hypothesis
1st Overlap
overlap
2nd Hypothesis permutation 1
2nd permutation overlap 1
2nd Hypothesis
Permutation
2nd Hypothesis permutation 2
2nd permutation overlap 2

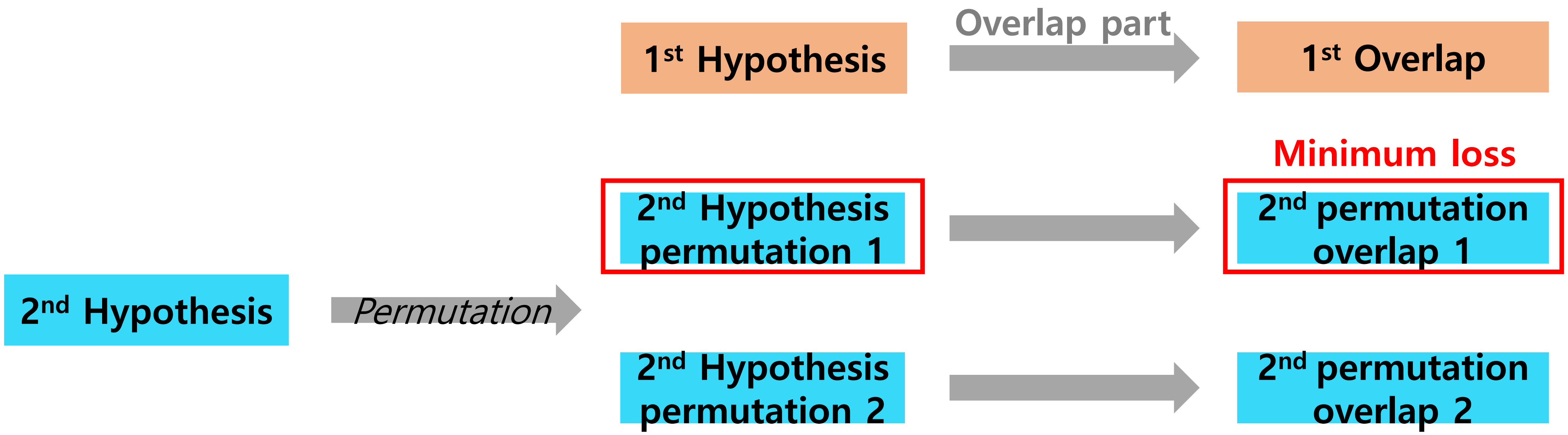
Proposed Method

- Utterance Reconstruction (3/4)
  - Split the speech utterance before inference due to length limited. Then co ncatenate the result by computing the loss of overlapping part.


BCE Loss
1st Hypothesis
Overlap part
1st Overlap
2nd Hypothesis permutation 1
2nd permutation overlap 1
2nd Hypothesis
Permutation
2nd Hypothesis permutation 2
2nd permutation overlap 2

Proposed Method

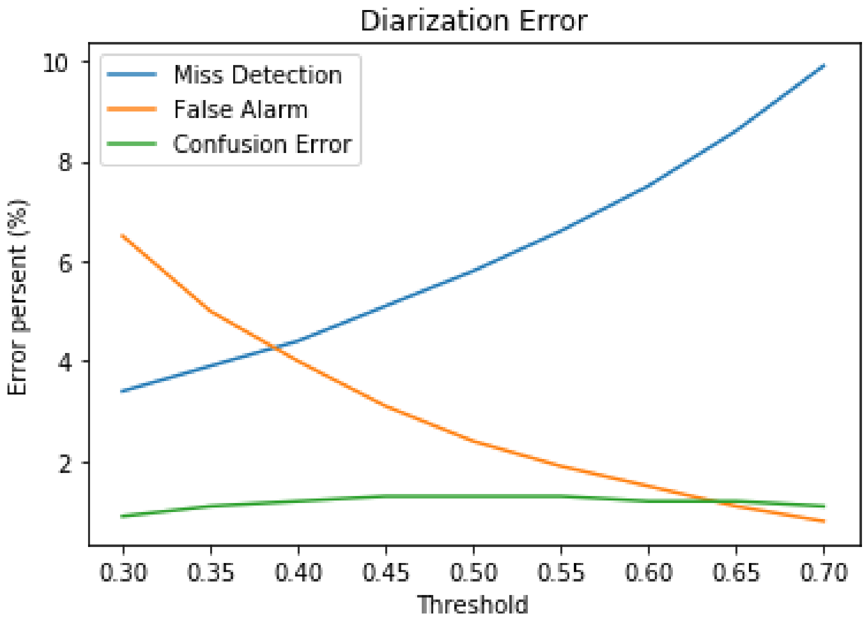
- Utterance Reconstruction (4/4)
  - Split the speech utterance before inference due to length limited. Then co ncatenate the result by computing the loss of overlapping part.



1st Hypothesis
Overlap part
1st Overlap
Minimum loss
2nd Hypothesis permutation 1
2nd permutation overlap 1
2nd Hypothesis
Permutation
2nd Hypothesis permutation 2
2nd permutation overlap 2

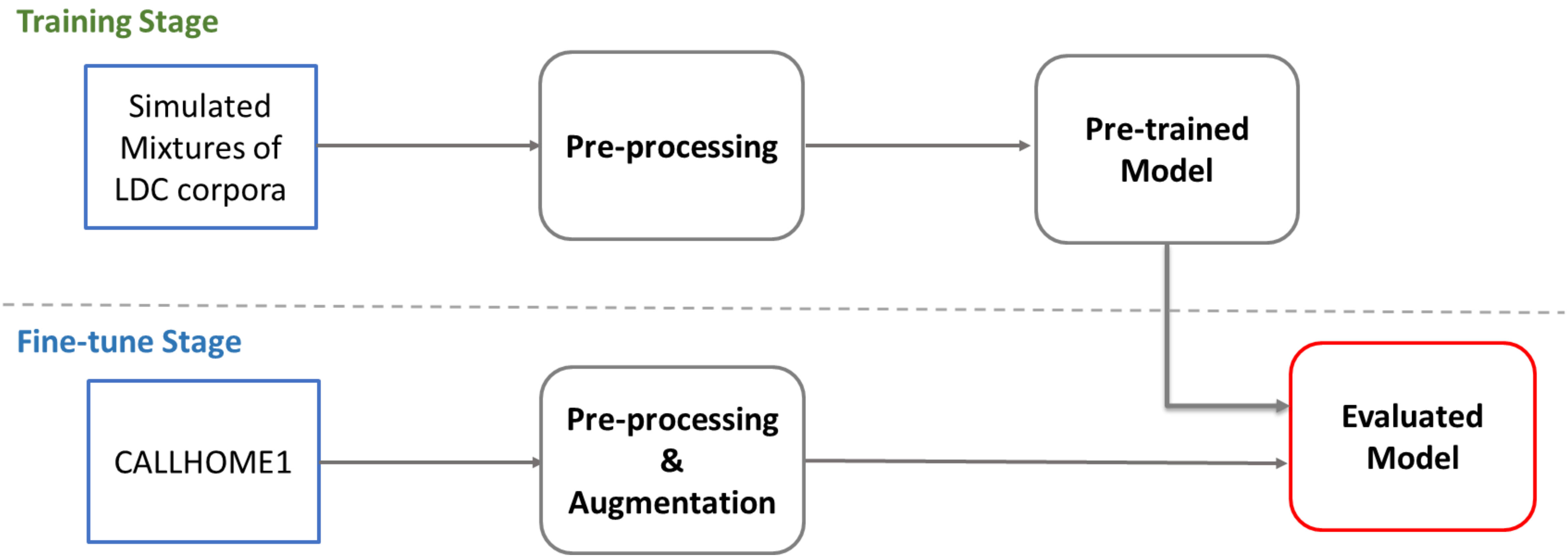
Proposed Method

- Threshold Choosing
  - Select different threshold with binary search.
    - The lower the threshold, the higher the false alarm.
    - The higher the threshold, the higher the miss detection.
    - The relationship between threshold and confusion error is small.



Proposed Method

- Training of Diarization System



Training Stage

Simulated Mixtures of LDC corpora

Pre-processing

Pre-trained Model

Fine-tune Stage

CALLHOME1

Pre-processing & Augmentation

Evaluated Model

Outline

- Introduction
- Background
- Proposed Method
- Experimental Results
- Reference

Experimental Results

- Experimental Data
  - Training set:

    ➢Simulated Mixtures [6] of LDC Corpora [7]

  - Fine-tune set:

    ➢Subset of 2000 NIST Speaker Recognition Evaluation (CALLHOME) [8]

  - Fine-tune set:

    ➢Subset of 2000 NIST Speaker Recognition Evaluation (CALLHOME)

Experimental Results

- Experimental Results

Model	Pre-trained data	DER
Proposed convolution model	100000 simulated mixtures	9.47%
Proposed linear model		9.12%
SA-EEND [5]		9.54%
SA-EEND-EDA [9]	400000 simulated mixtures	8.07%

Outline

- Introduction
- Background
- Proposed Method
- Experimental Results
- Reference

Reference

- [1] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant trainingof deep models for speaker-independent multi-talker speech separation,” inIEEEInternational Conference on Acoustics, Speech and Signal Processing (ICASSP),2017, pp. 241–245.
- [2] J. Allen, “Short term spectral analysis, synthesis, and modification by discretefourier transform,”IEEE Transactions on Acoustics, Speech, and Signal Process-ing, vol. 25, no. 3, pp. 235–238, 1977.
- [3] S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement ofthe psychological magnitude pitch,”Journal of the Acoustical Society of America,vol. 8, pp. 185–190, 1937.
- [4] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk and Q. V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. 2019.
- [5] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, and K. Nagamatsu, “End-to-endneural diarization: Reformulating speaker diarization as simple multi-label classi-fication,”ArXiv, vol. abs/2003.02966, 2020.

Reference

- [6] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, “End-to-end neural speaker diarization with permutation-free objectives,” in Proc. Interspeech, 2019, pp. 4300–4304.
- [7] Linguistic Data Consortium. [Online]. Available: https://www.ldc.upenn.edu/
- [8] Przybocki, Mark, and Alvin Martin, “2000 NIST Speaker Recognition Evaluation LDC2001S97,” Web Download, 2001, Philadelphia: Linguistic Data Consortium.
- [9] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, “End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors,” ArXiv, 2020.