# A Layer-Anchoring Strategy for Enhancing Cross-Lingual Speech Emotion Recognition

Shreya G. Upadhyay[1], Carlos Busso[2], Chi-Chun Lee[1]

[1]National Tsing Hua University, Taiwan

[2]University of Texas at Dallas, USA

# Introduction

# Speech Emotion Recognition Applications

- Speech Emotion Recognition (SER) is important for many applications
  - Education [1, 2]
  - Healthcare [3]
  - Call Center [4]
  - Entertainment [5, 6]
  - Social Robotics [7]

SER system's diverse application needs generalization across different domains or languages


[1] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning," in Emotions, technology, design, and learning. Elsevier, 2016, pp. 27–46.
[2] M. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," Smart Learning Environments, vol. 6, no. 1, pp. 1–20, 2019.
[3] Z. Farhoudi and S. Setayeshi, "Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition," Speech Communication, vol. 127, pp. 92–103, 2021.
[4] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in International Conference on Affective Computing and Intelligent Interaction. Springer, 2011, pp. 125–134.
[5] A. Menychtas, M. Galliakis, P. Tsanakas, and I. Maglogiannis, "Real-time integration of emotion analysis into homecare platforms," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019, pp. 3468–3471.
[6] H. Basanta, Y.-P. Huang, and T.-T. Lee, "Assistive design for elderly living ambient using voice and gesture recognition system," in 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2017, pp. 840–845.
[7] E. Polyakov, M. Mazhanov, A. Rolich, L. Voskov, M. Kachalova, and S. Polyakov, "Investigation and development of the intelligent voice assistant for the internet of things using machine learning," in 2018 Moscow Workshop on Electronic and Networking Technologies (MWENT). IEEE, 2018, pp. 1–5.

# Literature Studies

- Common Formulation
  - Mitigate mismatches of Source <--> Target domains
    - Transfer learning, semi-supervised learning, few-shot learning, etc.
  - Optimizing to decrease a distance metric of Source <--> Target features
    - Variations on Generative Adversarial Network (GAN)

Models are useful but come purely from a computational angle

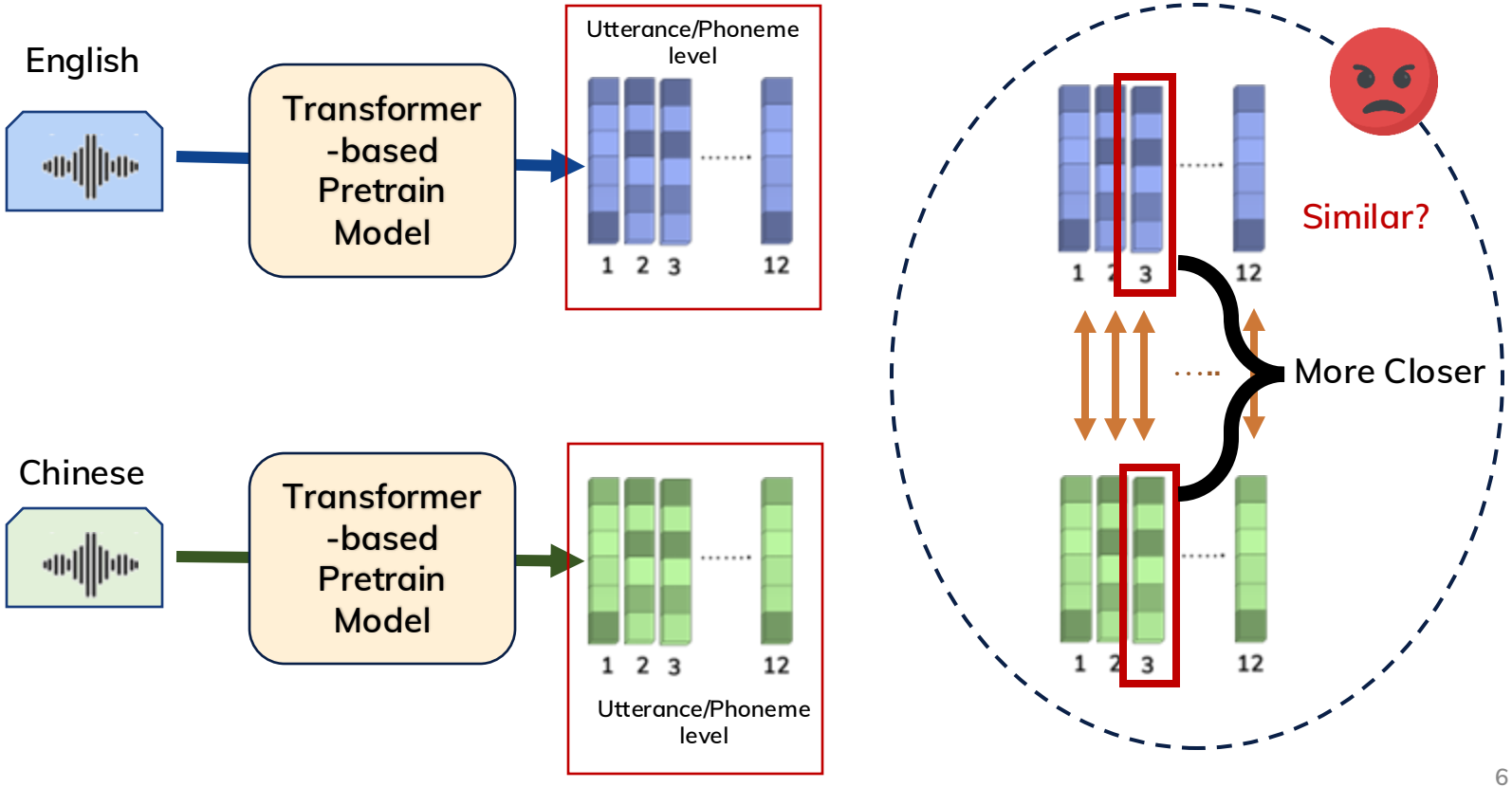
# Pretrained Model's Layer Commonalities
## Motivation

- Pretrained Model Dependency [1]:
  - Current research heavily relies on large pretrained models
  - Often focusing only on the final layer
- Task-Specific Nature and Hierarchical Structure [2, 3]:
  - Suggests that different layers encapsulate varying levels of information

We introduce a layer-anchoring mechanism to leverage multi-layer information for effective cross-lingual emotion transfer

[1] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan, "Emotion recognition based on phoneme classes," in 8th International Conference on Spoken Language Processing (ICSLP 04), Jeju Island, Korea, October 2004, pp. 889–892.
[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern recognition, vol. 44, no. 3, pp. 572–587, 2011.
[3] H. Li, X. Zhang, S. Duan, and H. Liang, "Speech emotion recognition based on bi-directional acoustic-articulatory conversion," Knowledge- Based Systems, p. 112123, 2024.

# Pretrain Layer Linguistic-Commonality



# Propose Steps

- A twofold approach
  - 1. Analyze emotion-specific layer commonalities across languages
    - Commonality Analysis → Anchoring Candidates
  - 2. Leverages these anchoring units as a constraining factor to facilitate cross-domain/lingual SER
    - Architecture
    - Experiment results
    - Insights

# Layer Similarity Analysis

# Speech Affective Corpora

- Two Naturalistic Corpora
  - MSP-Podcast [1]: American English (200 hrs.)
  - BIIC-Podcast [2]: Taiwanese Mandarin (140 hrs.)

[1] Lotfian, Reza, and Carlos Busso. "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings." *IEEE Transactions on Affective Computing* 10.4 (2017): 471-483.
[2] Upadhyay, Shreya G., et al. "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection." *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023.

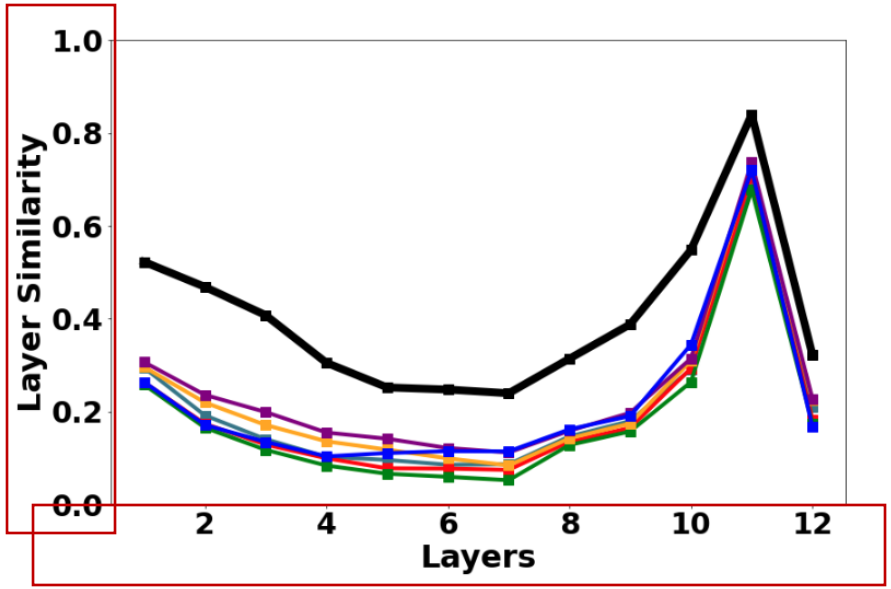
# Setup : Pretrain Models

- WavLM [1]:
  - Self-Supervised Learning (SSL) for Speech
  - Designed for Multiple Tasks
  - Robust to Noisy & Overlapped Speech
- Whisper [2]:
  - Weakly Supervised Learning
  - Multilingual & Multitask
  - Robust in Noisy Environments

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., "Wavlm: Large-scale self- supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.
[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in International Conference on Machine Learning. PMLR, 2023, pp. 28 492–28 518.

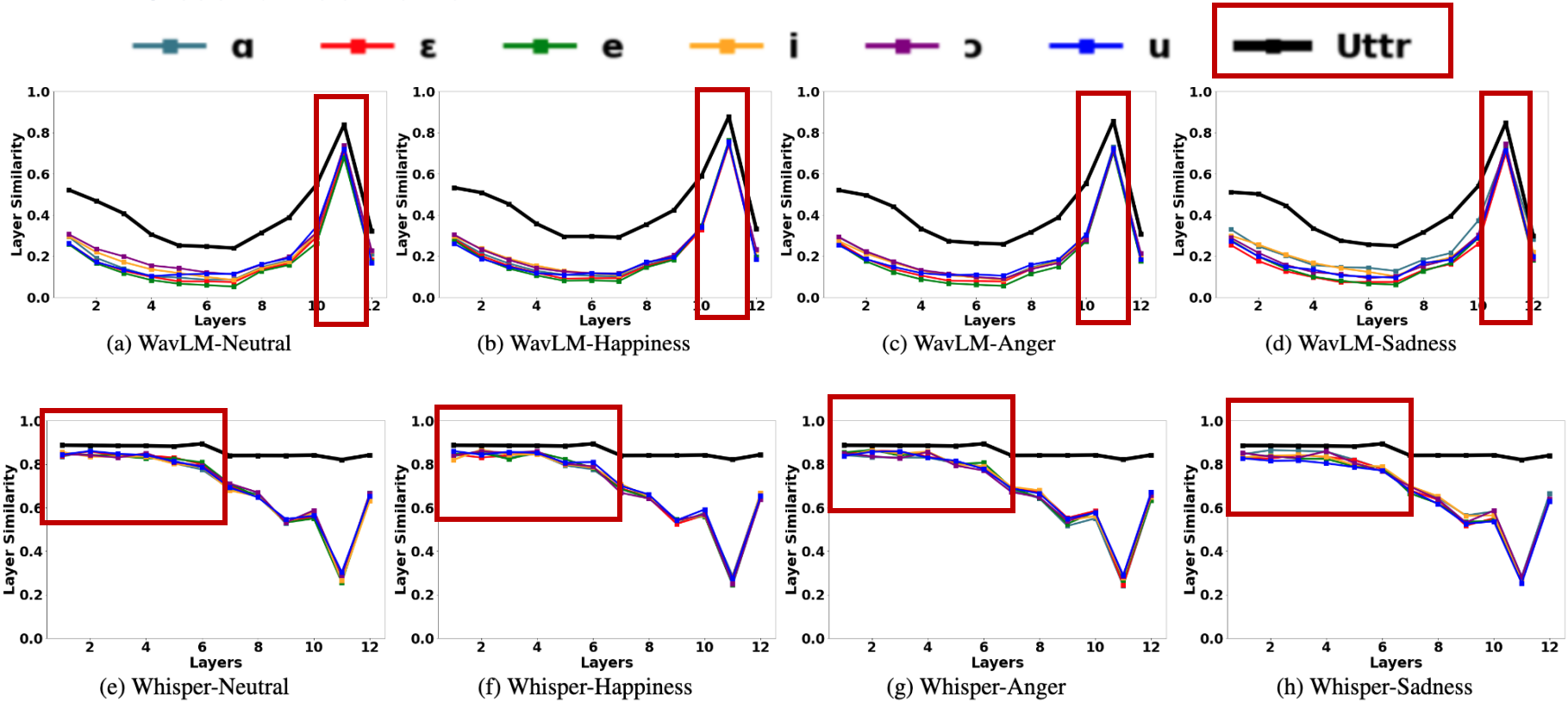
# Layer Similarity


- Layer similarity is estimated using Cosine Similarity
  - Over both corpora
- Over the 12 layers (Base models)
- 2 level similarities
  - Utterance level similarities
  - Phoneme level similarities

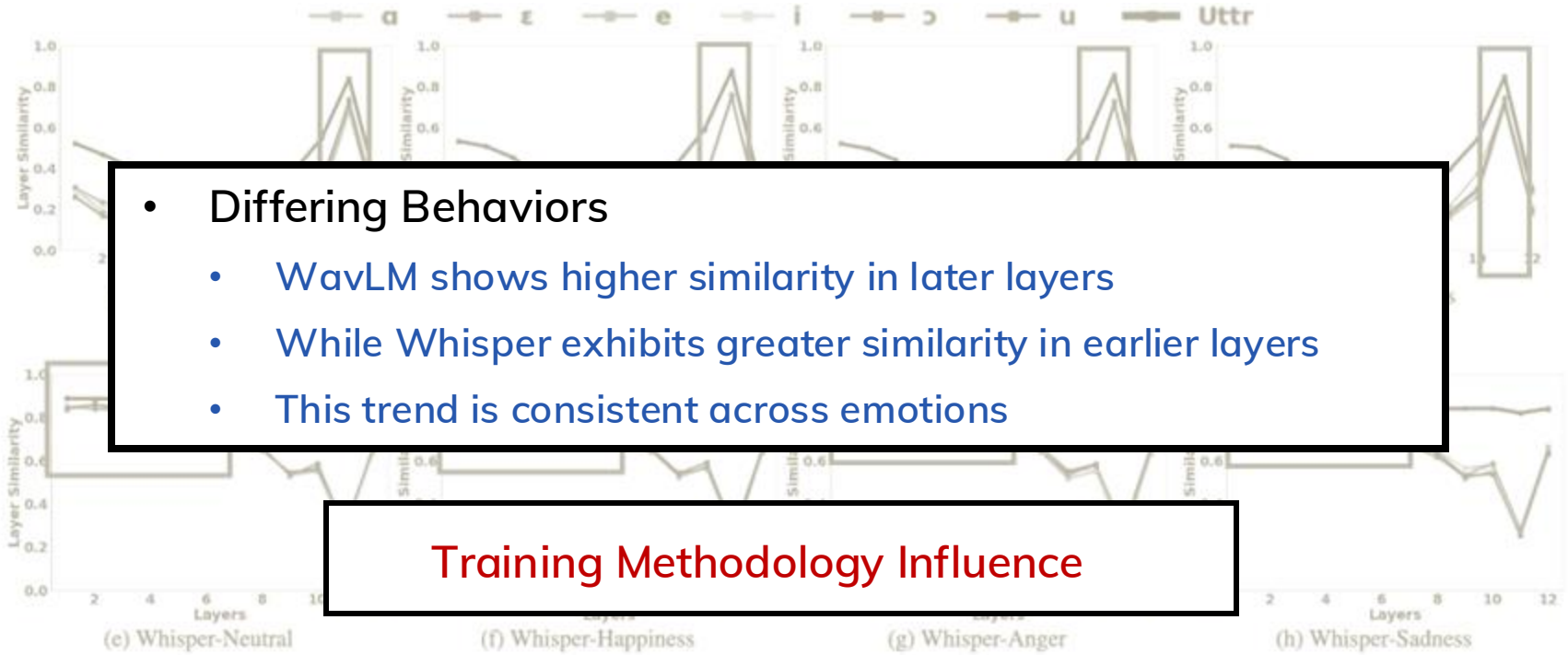


# Emotion-specific Layer similarity

## Utterance-level



# Emotion-specific Layer similarity

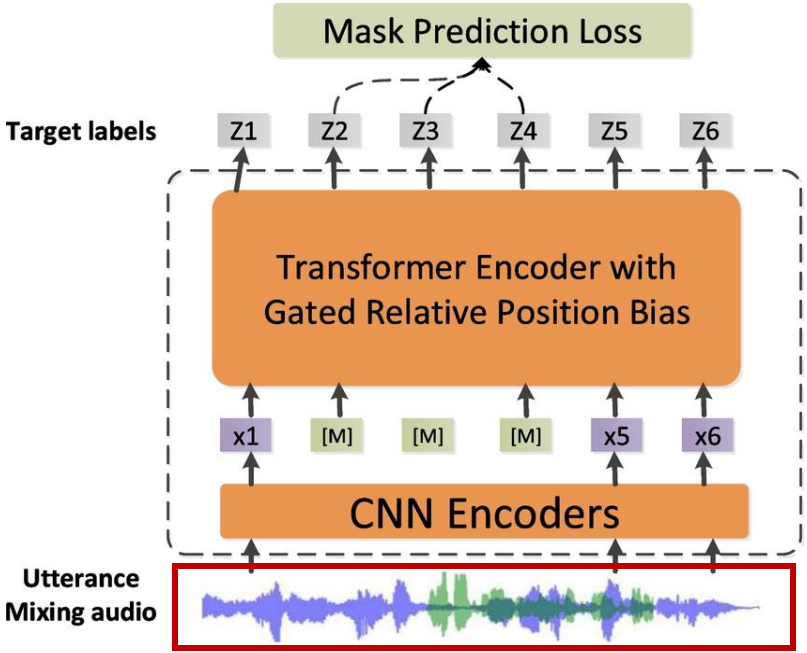



- Differing Behaviors
  - WavLM shows higher similarity in later layers
  - While Whisper exhibits greater similarity in earlier layers
  - This trend is consistent across emotions

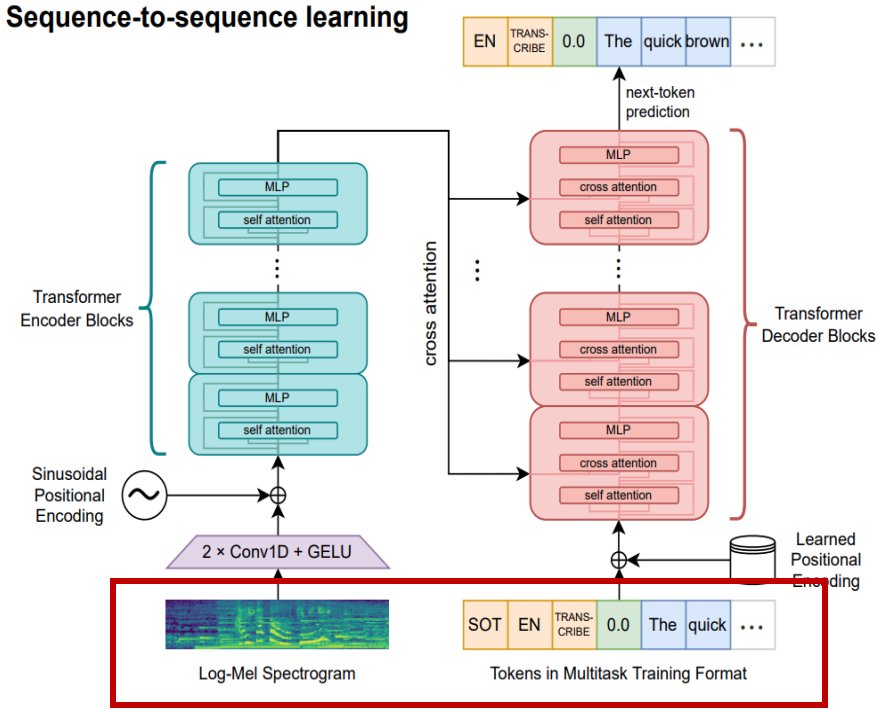
Training Methodology Influence

# Training Methodology

## WavLM [1]



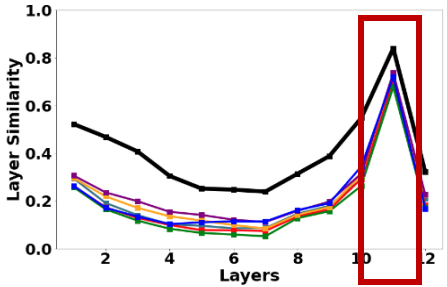

## Whisper [2]



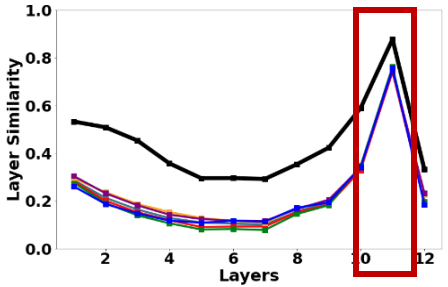


[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., "Wavlm: Large-scale self- supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.
[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in International Conference on Machine Learning. PMLR, 2023, pp. 28 492–28 518.

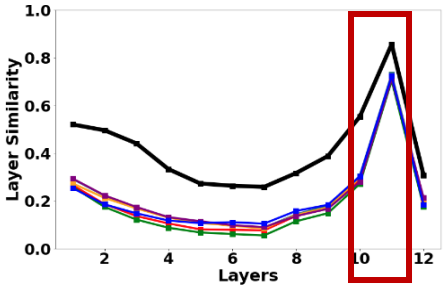
# Phonetic-Level Layer Similarity

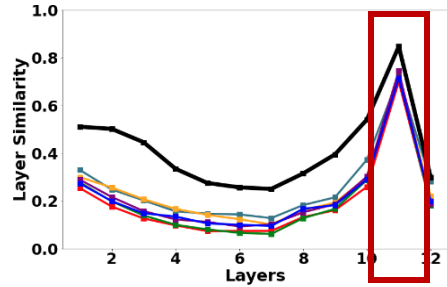

(a) WavLM-Neutral



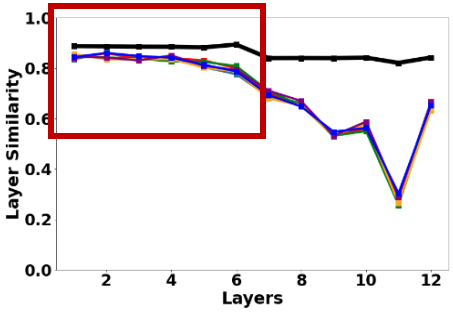

(b) WavLM-Happiness



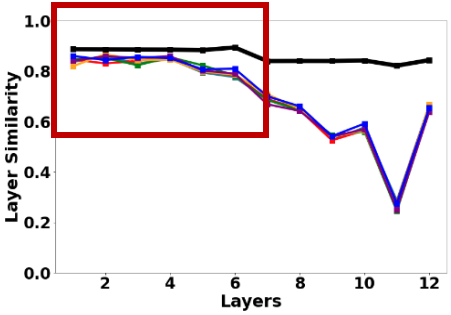

(c) WavLM-Anger



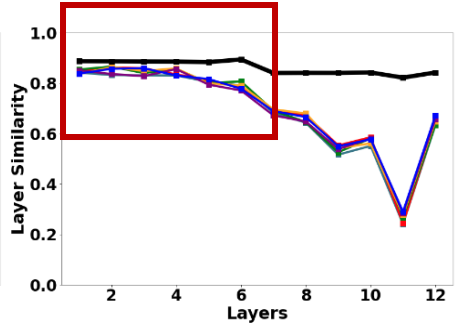

(d) WavLM-Sadness



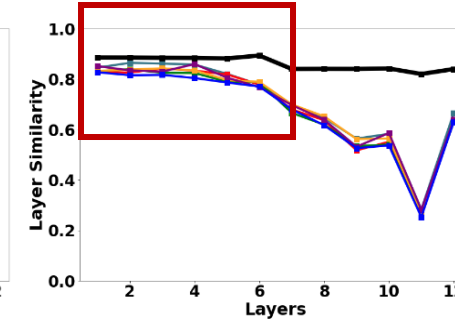

(e) Whisper-Neutral




(f) Whisper-Happiness

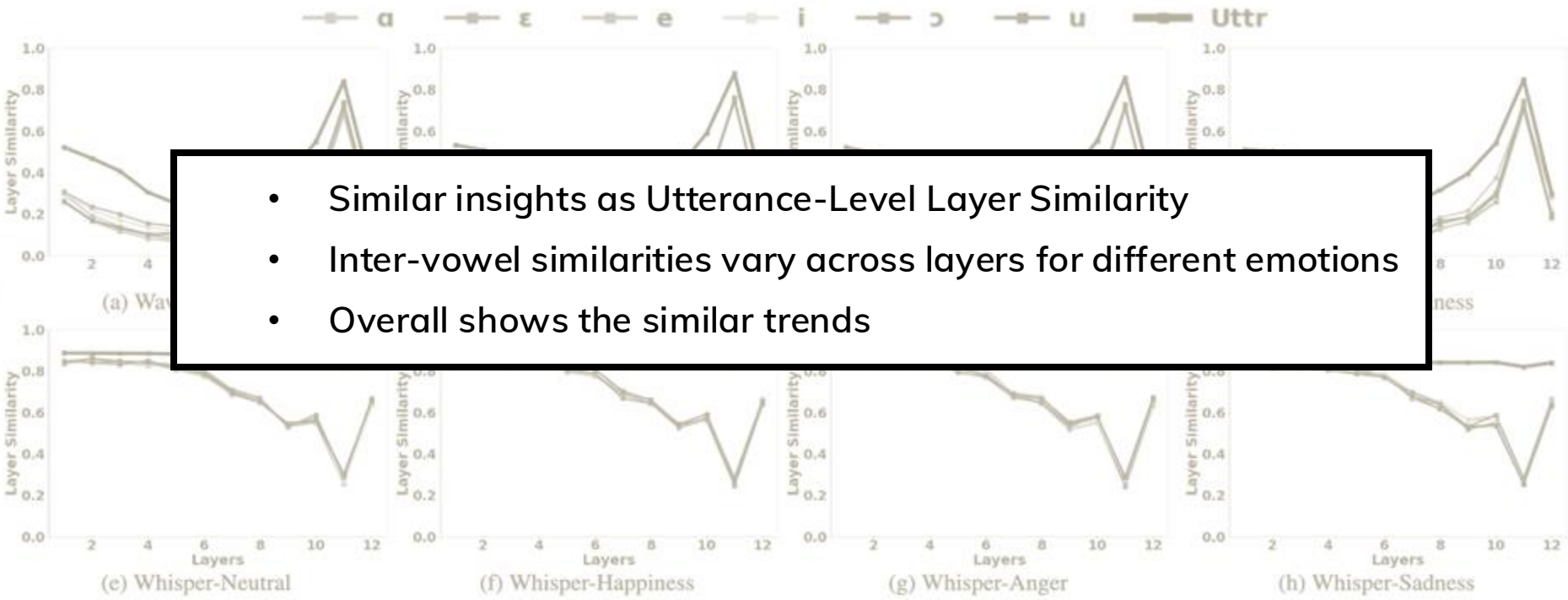



(g) Whisper-Anger




(h) Whisper-Sadness

# Phonetic-Level Layer Similarity




- Similar insights as Utterance-Level Layer Similarity
- Inter-vowel similarities vary across layers for different emotions
- Overall shows the similar trends

# Unified Layer Selection

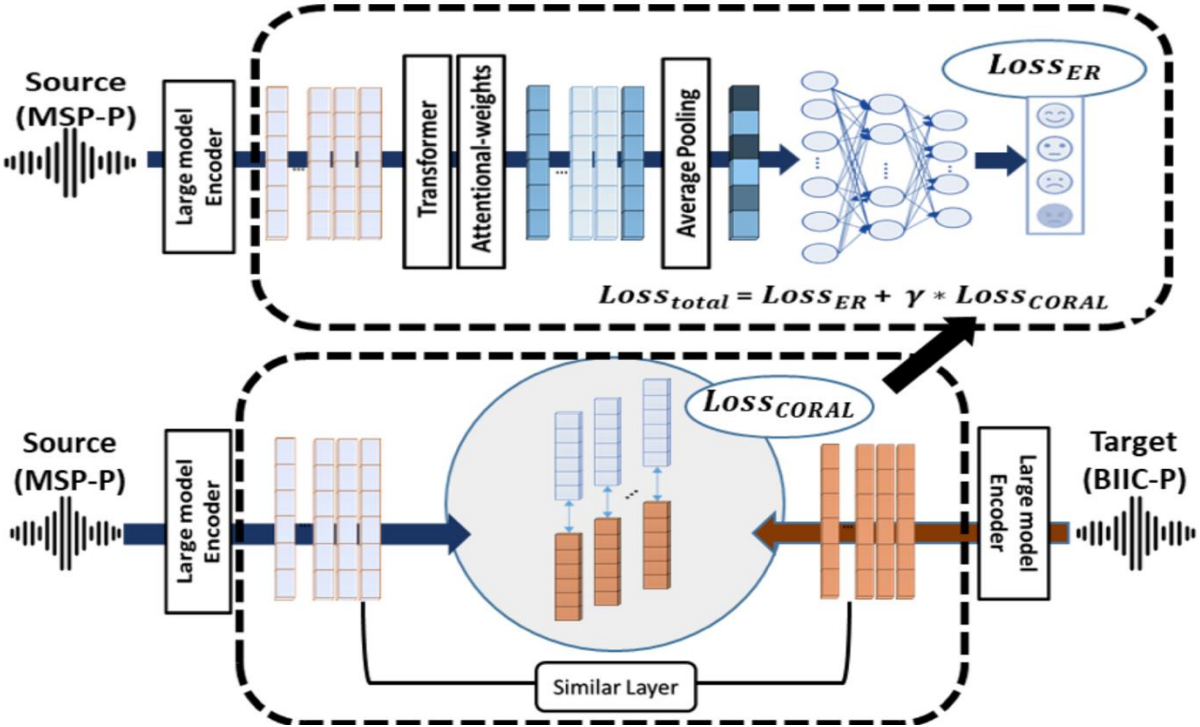
- Group Layer:
  - Top 3 most similar layers
- Best Layer:
  - The best similar layer
- Worst Layer:
  - 3 least similar layer

| | WavLM | Whisper |
|---|---|---|
| Group-Layers (GL) | [8, 9, 11] | [1, 2, 3] |
| Best-Layer (BL) | [11] | [2] |
| Worst-Layers (WL) | [5, 6, 7] | [7, 10, 11] |

# Layer-Anchored SER Architecture

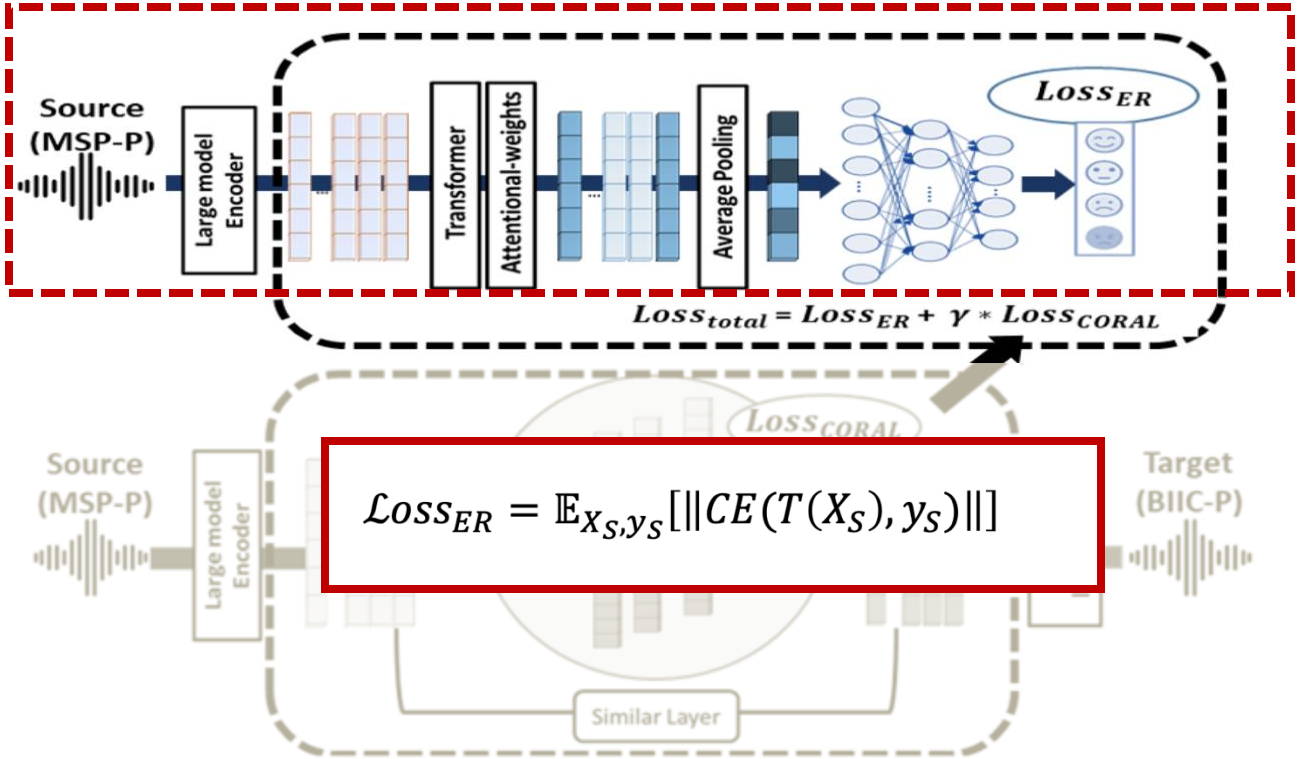
# Architecture: Layer Anchored Cross-Lingual SER

- 2 Branches:
  - 1) Conventional SER
  - 2) Layer Anchoring



# 1st Branch: Conventional SER

Loss: Cross-Entropy loss

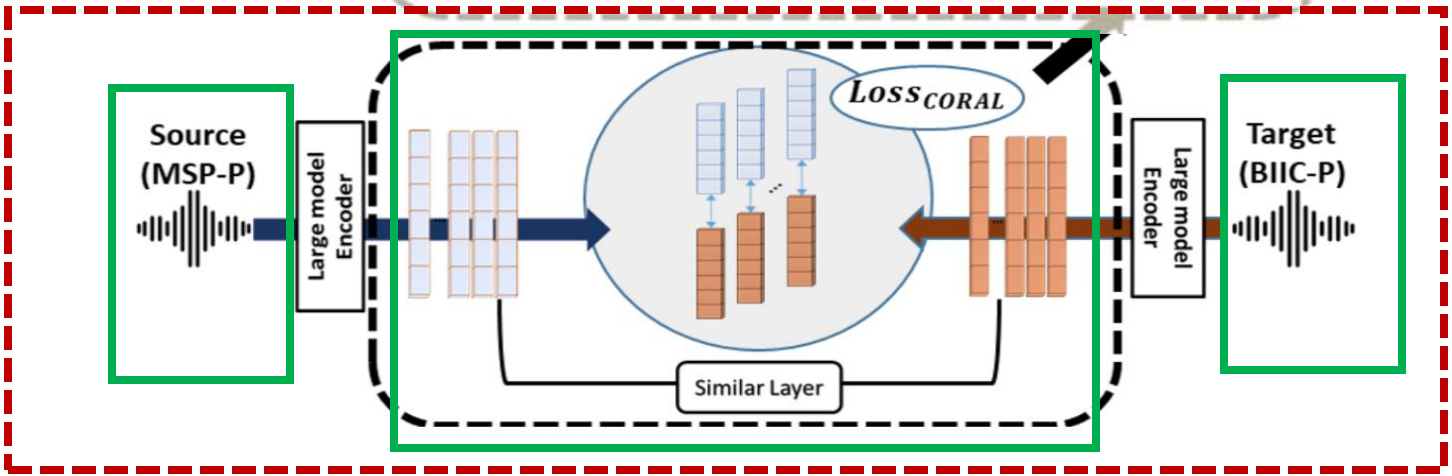



$$\mathcal{L}oss_{ER} = \mathbb{E}_{X_S, y_S}[\|CE(T(X_S), y_S)\|]$$

# 2nd Branch: Layer-Anchoring

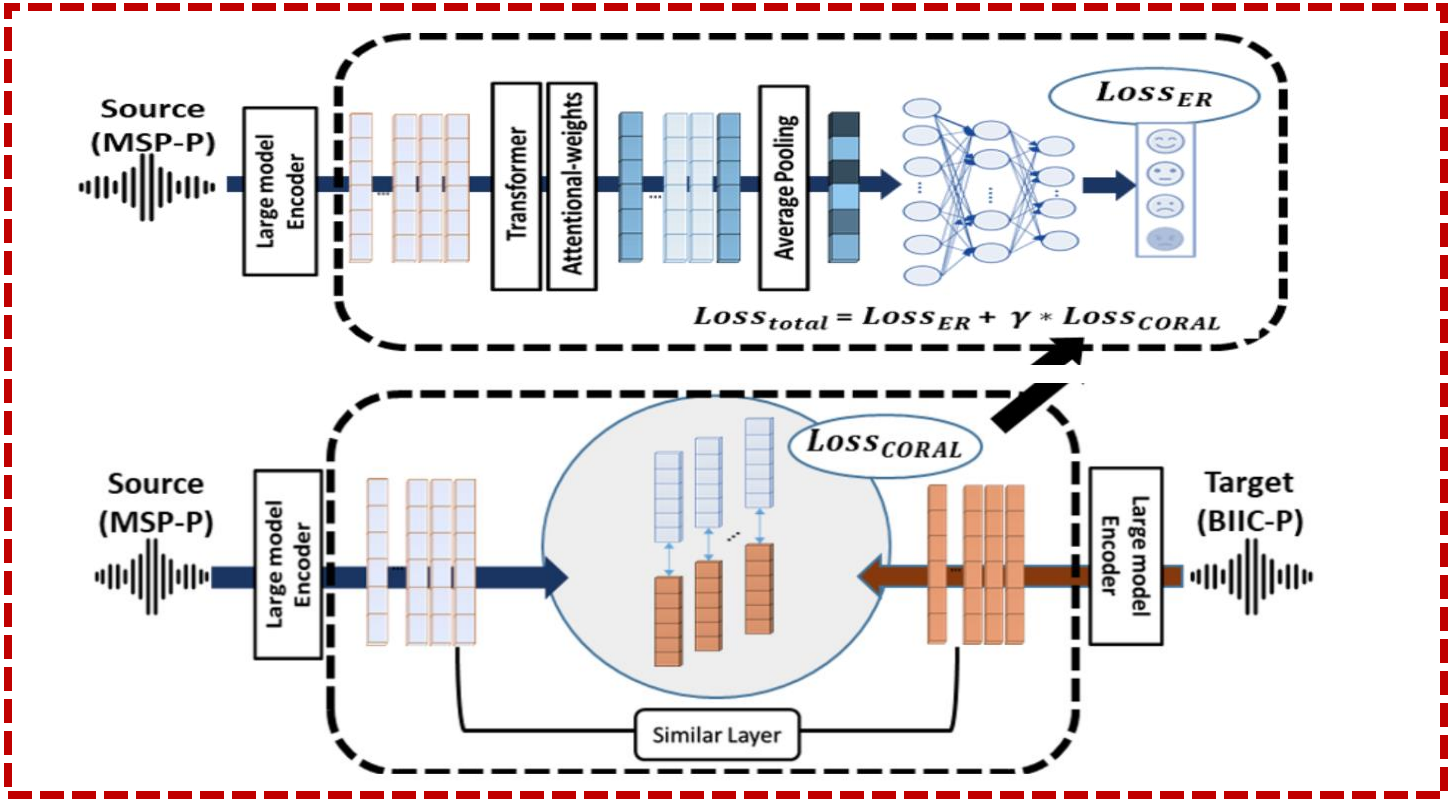
Loss: Coral Loss

$$Loss_{CORAL} = \sum_{i}^{N} \|Cov(L_{src}^{(i)}) - Cov(L_{tar}^{(i)})\|_F$$



# Overall Loss

$$L_{total} = Loss_{ER} + \gamma * Loss_{CORAL}$$



# Experimental Results

# Baseline Models

- 1) Ensemble Learning [1]
- 2) Few-Shot Learning [2]
- 3) Phonetic-Anchored Learning (PA) [3]
  - From the last section proposed the Phonetic-Anchoring Idea

[1] Wisha Zehra, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu, “Cross corpus multi-lingual speech emotion recognition using ensemble learning,” Complex & Intelligent Systems, vol. 7, no. 4, pp. 1845–1854, 2021.
[2] Youngdo Ahn, Sung Joo Lee, and Jong Won Shin, “Cross-corpus speech emotion recognition based on few-shot learning and do- main adaptation,” IEEE Signal Processing Letters, vol. 28, pp. 1190– 1194, 2021.
[3] Shreya G Upadhyay, Luz Martinez-Lucas, Bo-Hao Su, Wei-Cheng Lin, Woan-Shiuan Chien, Ya-Tse Wu, William Katz, Carlos Busso, and Chi-Chun Lee, “Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition,” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

# Baseline Performance Comparison

| | | MSP-P → BIIC-P | | BIIC-P → MSP-P | |
|---|---|---|---|---|---|
| | | WavLM | Whisper | WavLM | Whisper |
| Top Layer | CC | 52.01 | 51.87 | 48.39 | 49.01 |
| | Ensemble [1] | 52.18 | 52.03 | 51.75 | 51.98 |
| | Few-shot [2] | 53.62 | 52.74 | 50.59 | 51.75 |
| | PA [3] | 58.14 | 57.83 | 55.35 | 54.93 |
| w/ Layer | PA-Avg | 58.06 | 58.01 | 55.24 | 54.32 |
| | PA-Atn | 58.83 | 58.92 | 55.64 | 56.10 |
| | **LA** | **60.21** | **59.65** | **56.68** | **56.37** |

CC: Cross-corpus (Direct Test)
PA: Phonetic-Anchored Model
LA: Layer-Anchored Model

LA model achieves higher UAR with both WavLM and Whisper features

[1] Wisha Zehra, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," Complex & Intelligent Systems, vol. 7, no. 4, pp. 1845–1854, 2021.
[2] Youngdo Ahn, Sung Joo Lee, and Jong Won Shin, "Cross-corpus speech emotion recognition based on few-shot learning and do- main adaptation," IEEE Signal Processing Letters, vol. 28, pp. 1190– 1194, 2021.
[3] Shreya G Upadhyay, Luz Martinez-Lucas, Bo-Hao Su, Wei-Cheng Lin, Woan-Shiuan Chien, Ya-Tse Wu, William Katz, Carlos Busso, and Chi-Chun Lee, "Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

# Layer Selection

| | WavLM | Whisper |
|---|---|---|
| Group-Layers (GL) | [8, 9, 11] | [1, 2, 3] |
| Best-Layer (BL) | [11] | [2] |
| Worst-Layers (WL) | [5, 6, 7] | [7, 10, 11] |

| | MSP-P → BIIC-P | | BIIC-P → MSP-P | |
|---|---|---|---|---|
| | WavLM | Whisper | WavLM | Whisper |
| **LA (Group Layers)** | **60.21** | **59.65** | **56.68** | **56.37** |
| LA (All Layers) | 58.54 | 57.97 | 55.39 | 54.91 |
| LA (Best Layer) | 59.16 | 58.11 | 55.75 | 54.21 |
| LA (Worst Layers) | 58.01 | 57.72 | 54.64 | 53.77 |
| LA (Random Layers-1 ) | 58.94 | 56.24 | 54.93 | 54.29 |
| LA (Random Layers-2) | 58.55 | 57.39 | 53.85 | 53.38 |
| LA (Random Layrers-3) | 57.23 | 57.84 | 54.20 | 54.43 |

LA: Layer-Anchored Model

- Precise layer selection in LA improves results
- Random or non-optimal layers do not enhance results

# Insights

# Insights — Layer-based Anchoring

- Presents Layer-Anchoring strategy for cross-lingual SER
  - Effectively aligns the phonetic characteristics and mitigates the discrepancies across languages
- Analysis shows that layer selection varies by the task and the training methodology

# Future works

- Enhance Generalization
  - Integrate anchoring with advanced domain adaptation techniques
- Leverage Large Pretrained Models
  - Evaluate different encoders to maximize performance
- Expand Emotional Categories for a more inclusive SER model

# Thanks!

Carnegie Mellon University MSP Multimodal Speech Processing Laboratory

biic National Tsing Hua University

NSF

NSTC 國科會

NSTC 國科會