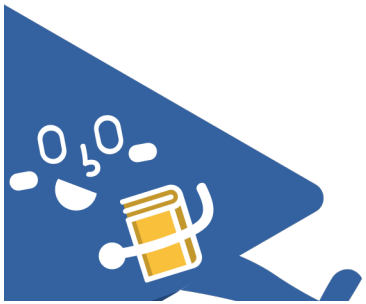
Mouth Articulation-Based Anchoring for Improved Cross-Corpus Speech Emotion Recognition

Authors: Shreya G. Upadhyay, Ali N Salman, Carlos Busso, and Chi-Chun Lee

Outline

- Introduction
  - Background and Motivation
  - Research Goal
- Mouth Articulatory-Gesture Commonality Analyses
- Cross-Domain SER Architecture
- Experiment Results and Analysis
- Conclusion And Future Work

Introduction



Speech Emotion Recognition Applications

- System that recognize the human emotions
- 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

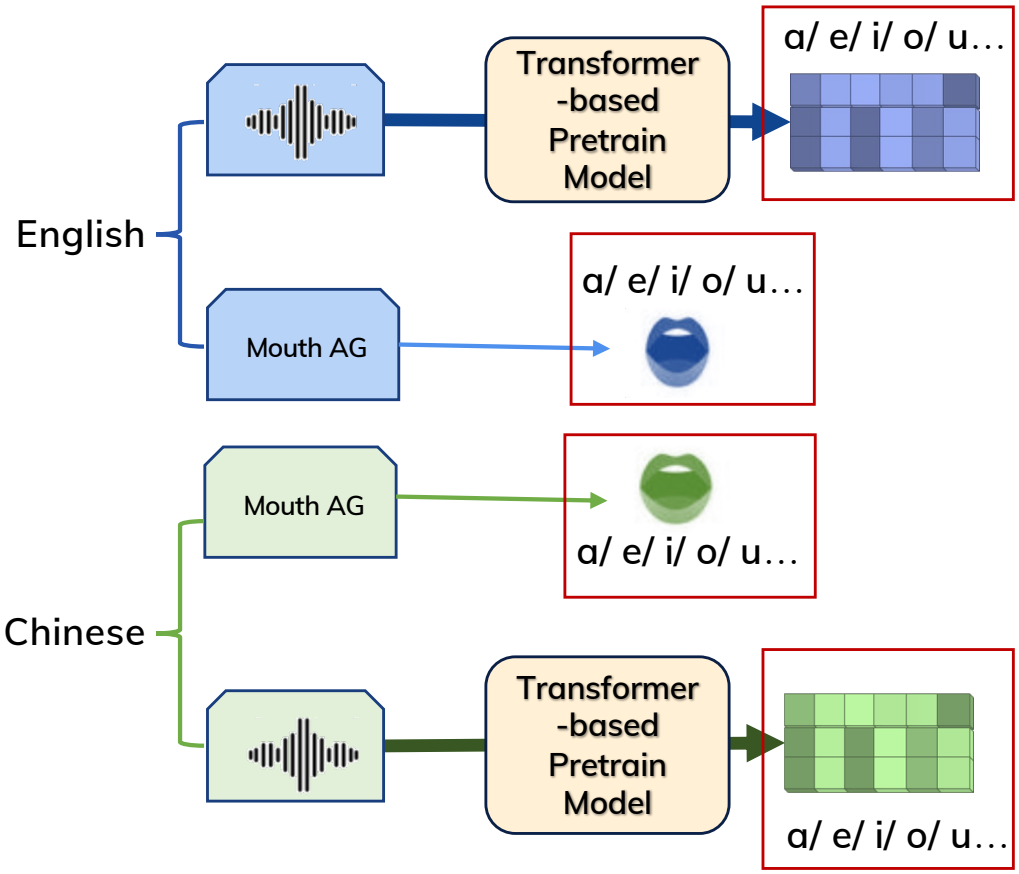
Mouth Articulation Gesture Commonalities
Motivation

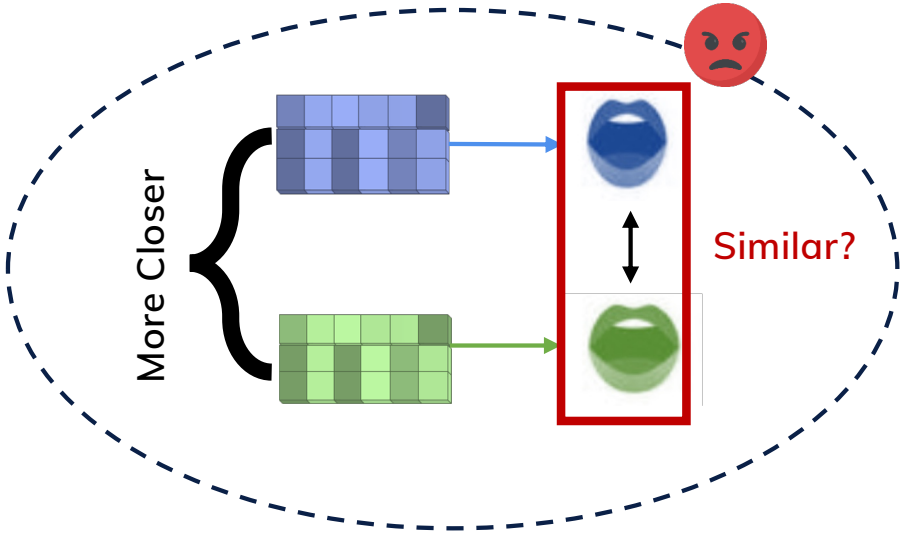
- Literature: works on all acoustic data
- Are they stable? Based on Recording conditions, noise, speaker differences, etc.
- Acoustic signals and articulatory features are intrinsically linked [1]
- Physical mouth shapes influence verbal sound meaning [2]
- Taller mouth shapes impact size and width perception in vowel pronunciation [3]

Mouth Articulation-Based Anchoring for Improved Cross-Corpus Speech Emotion Recognition

[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.

Mouth Articulation Commonalities

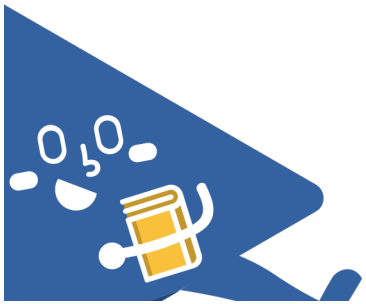




Propose Steps

- A twofold approach
  - 1. Analyze emotion-specific mouth articulation commonalities across languages
  - 2. Leverages these common gestures as a constraining factor to facilitate cross-domain SER

Mouth Articulatory-Gesture Similarity Analysis



Setup

- To implement this idea, the corpora with three modalities are needed
  - 1) Speech
  - 2) Text: to get the phonemes
  - 3) Vision: to track the mouth gestures
- Two different domain corpora:
  - CREMA-D: American English
  - MSP-IMPROV: American English
- Pretrain Model: Wav2vec2.0

Speech Affective Corpora

- CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset)[1]
  - Duration: ~8 hours
  - Emotional Labels: Anger, Disgust, Fear, Happy, Neutral, and Sad
  - Source: Professional actors performing scripted sentences
  - Modalities: Audio and Visual (facial expressions), text
- MSP-IMPROV (Multimodal Spontaneous Emotion Corpus)[2]
  - Duration: ~9 hours
  - Emotional Labels: Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise
  - Source: Dyadic interactions with both improvised and scripted speech
  - Modalities: Audio, Video, and Text

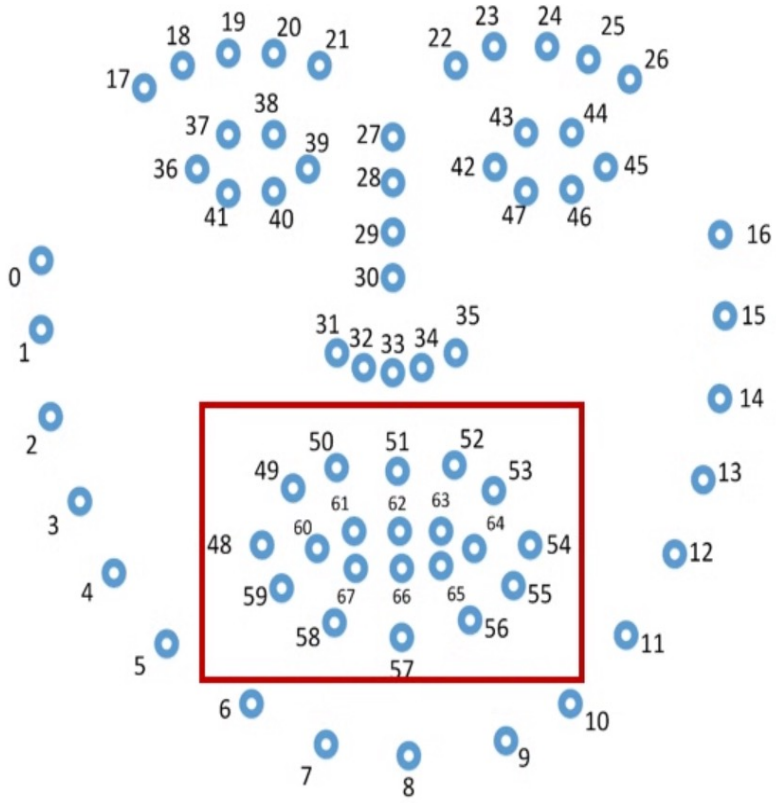
[1] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," IEEE transactions on affective computing, vol. 5, no. 4, pp. 377–390, 2014.
[2] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," IEEE Transactions on Affective Comput- ing, vol. 8, no. 1, pp. 67–80, 2016.

Mouth Articulatory Gestures Commonality

- Steps:
  - 1) Extract articulatory gestures
  - 2) Preprocessing
  - 3) Segmentation
  - 4) Find the commonalities

1) Extract Mouth Articulatory Gestures

- Use OpenFace [1] to detect the face bounding box and 68 2D landmarks (e.g., eyes, chin, lips)
- Focus is on 12 key landmarks (48-59)
  - Define the outer mouth shape



[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, 2016, pp. 1–10.

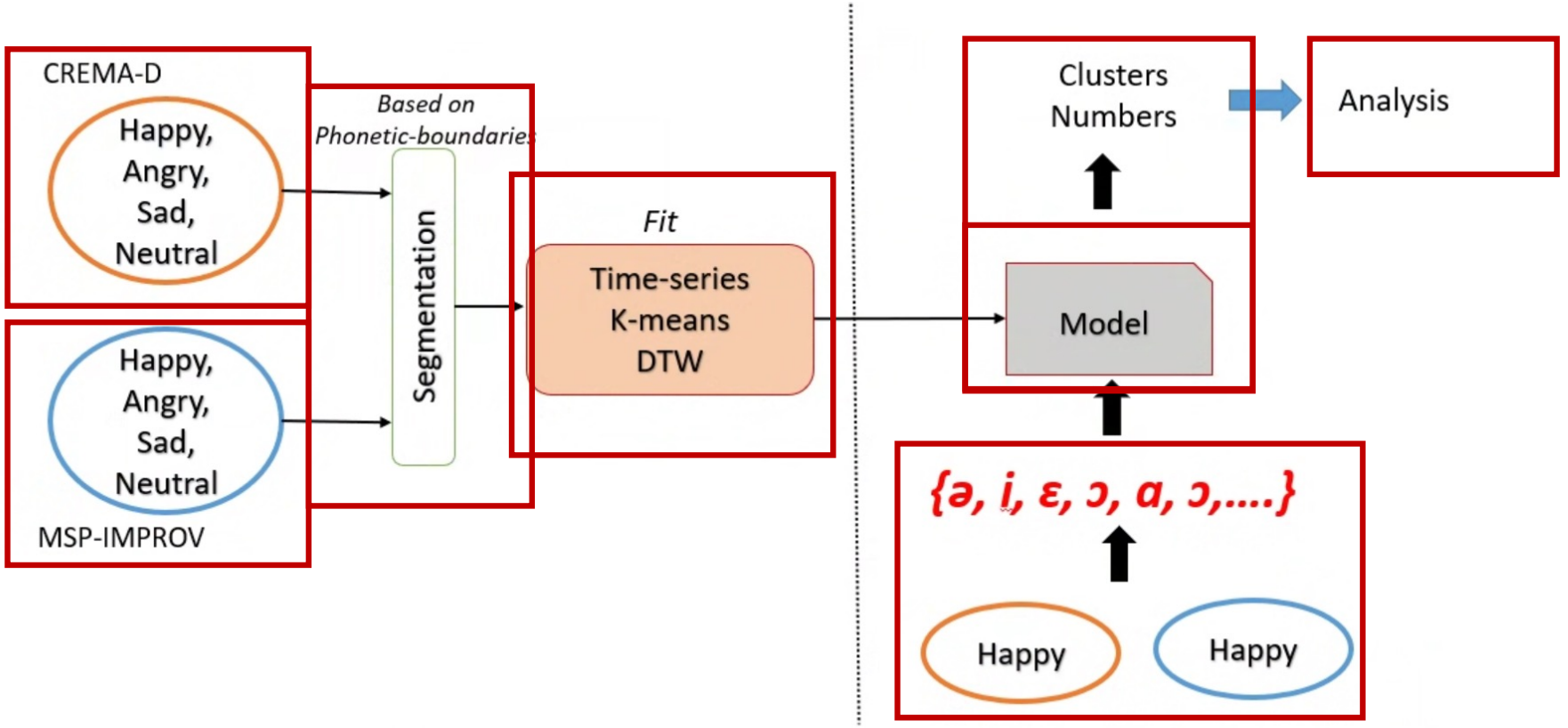
- 2) Preprocessing
  - Align the landmarks with respect to the eye [1,2]
    - Align landmarks by rotating to make the eye distance parallel to the x-axis
    - Normalizing by inter-pupil distance
- 3) Segmentations: Phoneme-based segmentations
  - Mouth gestures are continuous and dynamic: Making hard segmentation difficult without clear boundaries
- 4) Commonality: Clustering approach to cluster the different AGs present
  - Challenge with only Phoneme-Based hard segmentations:
    - It's hard to say articulations are just from the vowels
    - Vowels and consonants influence each other's articulatory gestures

[1] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," IEEE Transactions on Image Processing, vol. 25, no. 3, pp. 1233–1245, 2016.
[2] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2235–2245

Articulatory-Gesture Clustering Steps

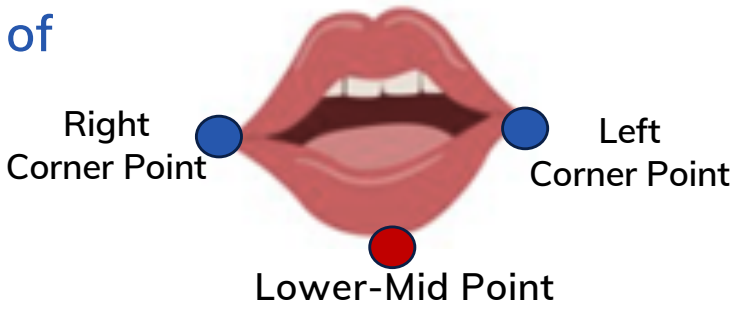
- Use phonetic boundaries to segment landmark sequences
- Apply time-series k-means with Soft-DTW to cluster similar articulatory patterns despite timing variations
- Train the AG cluster model on samples from four emotions across two corpora,
  - Determine 10 as the optimal cluster number using the elbow method

Articulatory-Gesture Clustering

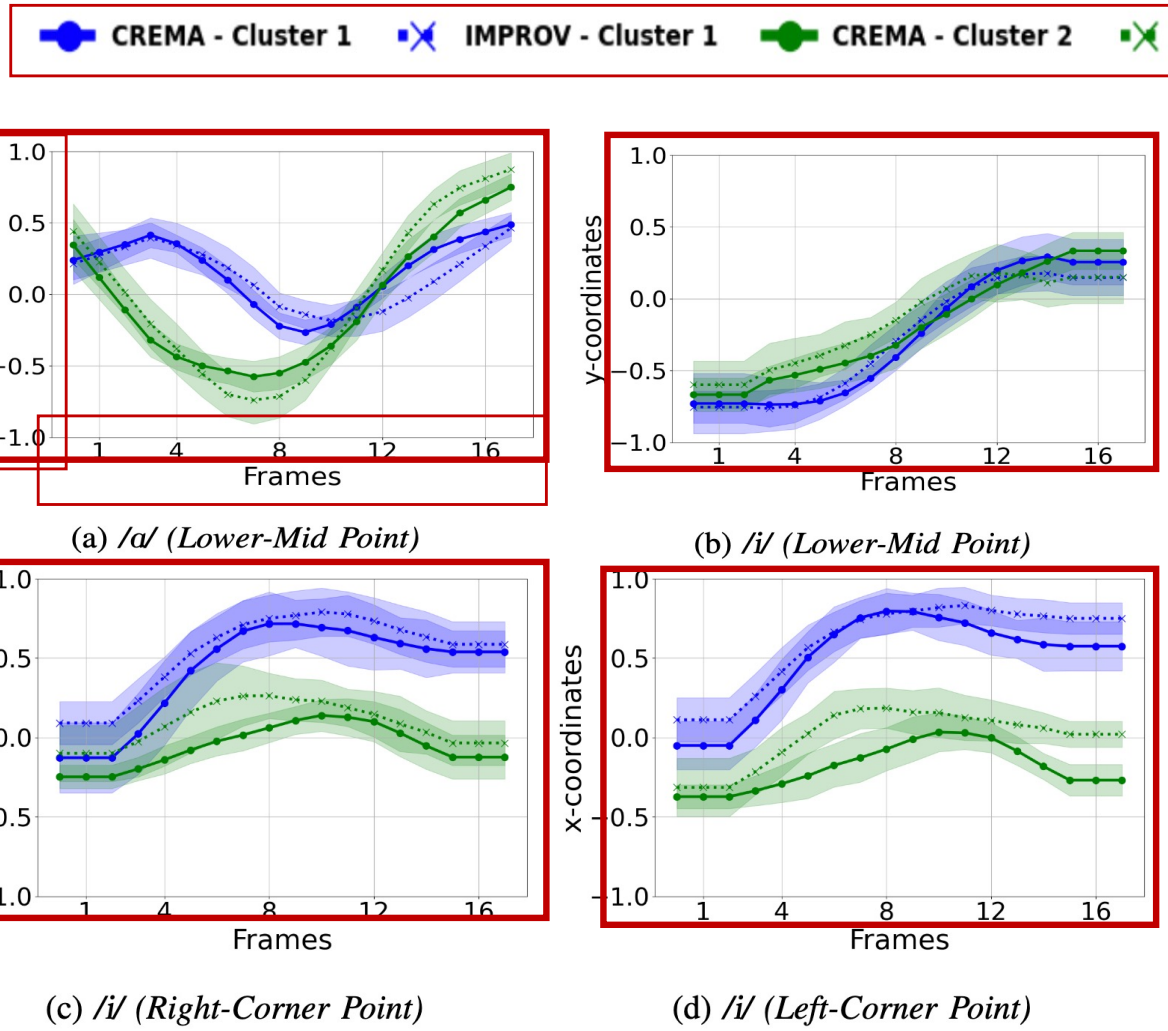


Clustering Model Sanity Check

- Visualize whether two clusters have different shapes of the same phoneme or not
- Steps:
  - Get the vowel segments coordinates of different clusters
  - Plot them over frames
  - Focus is on three points



CREMA→ CREMA-D
IMPROV→ MSP-IMPROV



(a) */ɑ/ (Lower-Mid Point)*

(b) */i/ (Lower-Mid Point)*

(c) */i/ (Right-Corner Point)*

(d) */i/ (Left-Corner Point)*

- Reveal distinct AG patterns across two clusters
- But shows similar AG patterns from different corpora within the same cluster

1) Cross-Corpus AG Cluster Overlap Analysis

- Identify overlap in AG clusters between two corpora
- Passed each corpora sample through an AG cluster model
  - Determine their cluster ID
- Calculate the percentage of samples that are similarly clustered

$$Sim = \frac{1}{K_{c,t}} \sum_{i=1}^{K_{c,t}} \min\left(\frac{N_{1,i}}{N_1}, \frac{N_{2,i}}{N_2}\right) \times 100$$

- Corpus 1 has clusters 1 and 2
- Corpus 2 includes clusters 1, 2, and 3, we focus on the common clusters
- for similarity evaluation → 1, 2
- To account for potential discrepancies,
  - we use a 25% threshold to exclude very less common clusters

Cross-Corpus AG Cluster Overlap Analysis

- A single phoneme is scattered over more than one cluster
- Some phonemes having a common gesture

AG Cluster	/ɑ/	/ə/	/ɛ/	/i/	/æ/	/u/
Cluster_1	75.3	2.1	60.6	71.1	78.9	0
Cluster_2	70.4	79.4	3.6	79.9	75.3	75.5
Cluster_3	61.2	64.5	77.1	0	76.7	10.8
Cluster_4	0	0	78.2	69.3	0	76.2
Cluster_5	0	0	0	5.7	15.5	0
Cluster_6	0	66.8	0	0	12.3	0
Cluster_7	68.4	0	18.4	2.5	0	74.5
Cluster_8	7.1	0	0	0	0	0
Cluster_9	0	12.9	0	0	0	0
Cluster_10	0	0	13.6	0	7.4	0

Percentage overlap of AG clusters across corpora

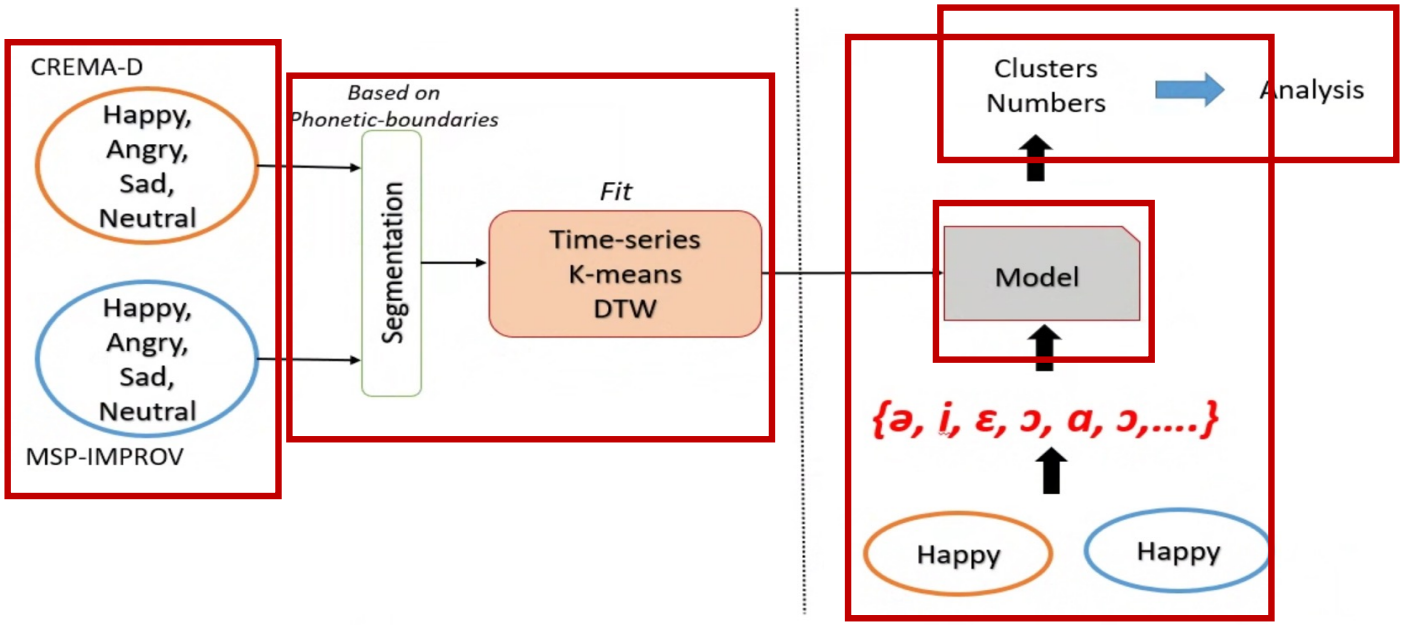
Emotion-Specific Average Overlap

AG Cluster	/ɑ/	/ə/	/ɛ/	/i/	/æ/	/u/
All	68.8	69.3	74.5	73.4	76.6	75.0
Neutral	**75.3**	68.4	**80.2**	66.7	72.5	60.3
Happiness	**80.1**	74.3	70.9	**78.2**	68.6	61.4
Anger	**73.4**	70.6	66.1	**75.6**	71.2	59.3
Sadness	66.2	65.7	**68.3**	63.8	67.9	**72.5**

- Some phonemes have more overlap under different emotions

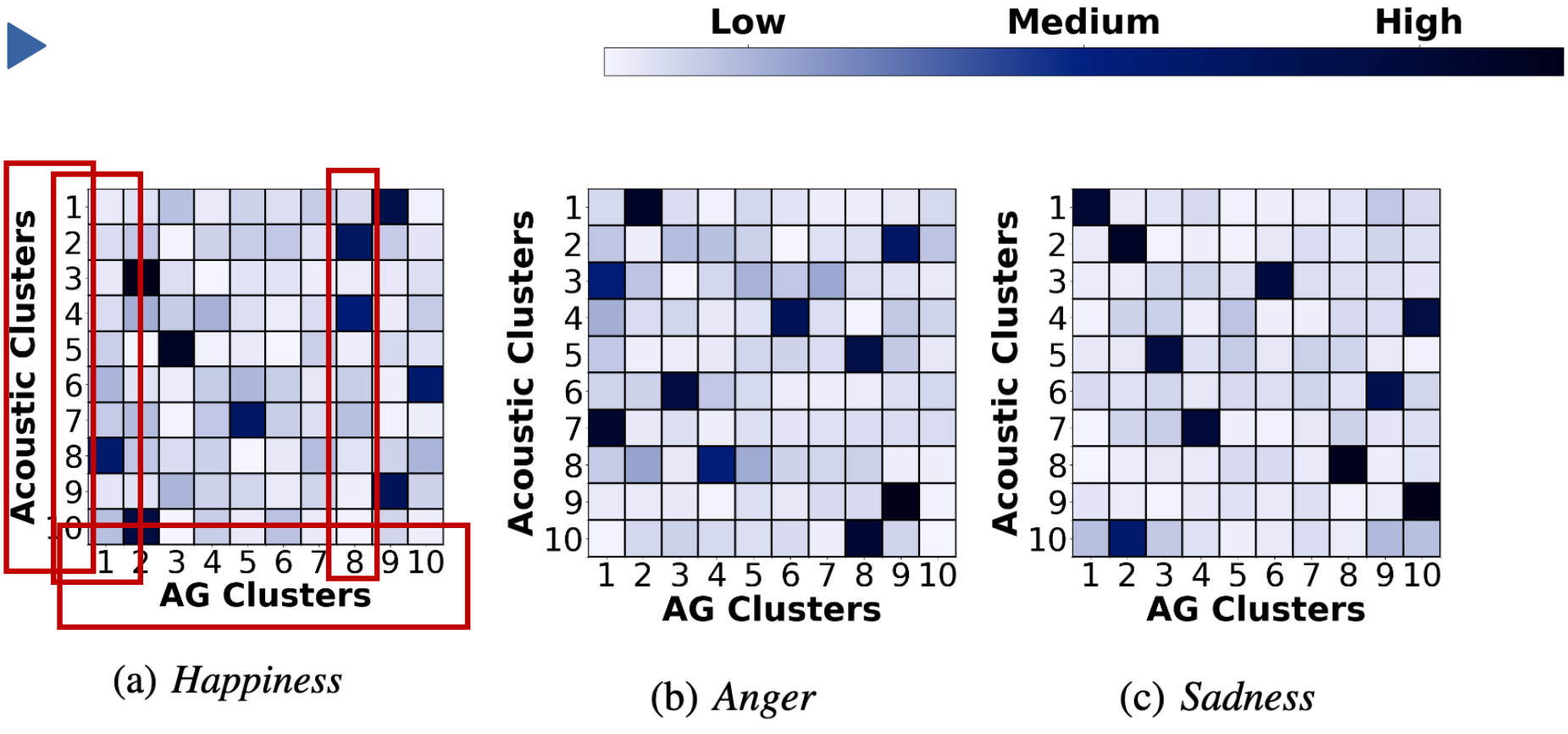
2) AG-Acoustic Association Analysis

- Assess how well samples grouped by AG clusters align with their corresponding acoustic features
- Model an acoustic cluster system similar to the AG cluster model



AG-Acoustic Features Association Analysis

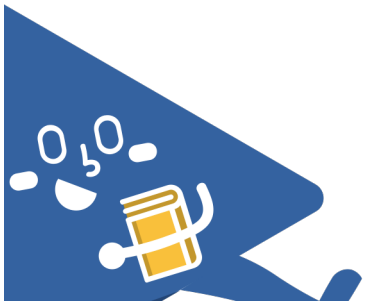
- 2 clustered models:
  - Acoustic and AG cluster model
- First, we cluster the samples using the AG cluster model
- For each AG cluster,
  - We extracted the acoustic features of the grouped samples and
  - Passed them through the acoustic cluster model
- The goal is to assess how many acoustic clusters form within each AG cluster to evaluate the association



(a) *Happiness* (b) *Anger* (c) *Sadness*

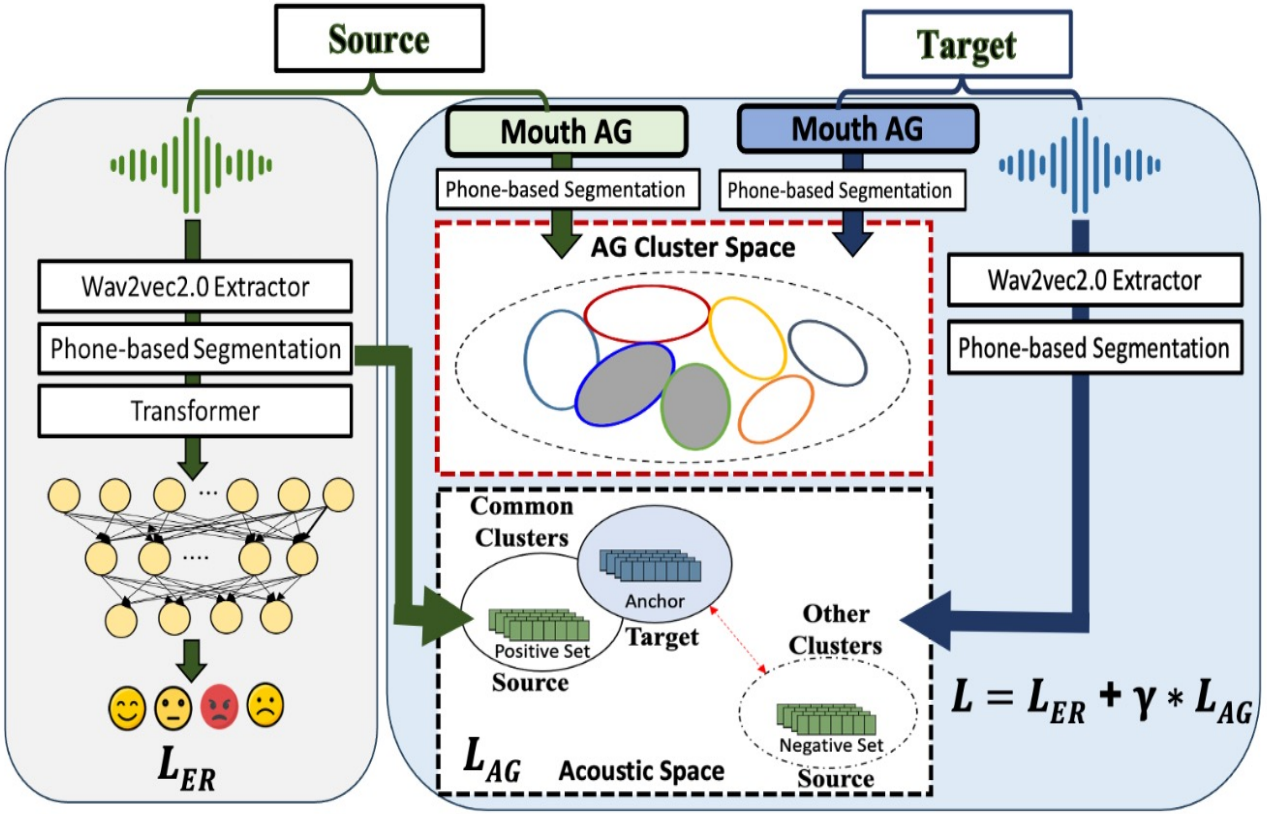
- Acoustic embeddings often cluster in fewer clusters within each AG cluster
- Reflecting that there is a correlation between AG gestures and their acoustic counterparts

Mouth AG-Anchored SER Architecture

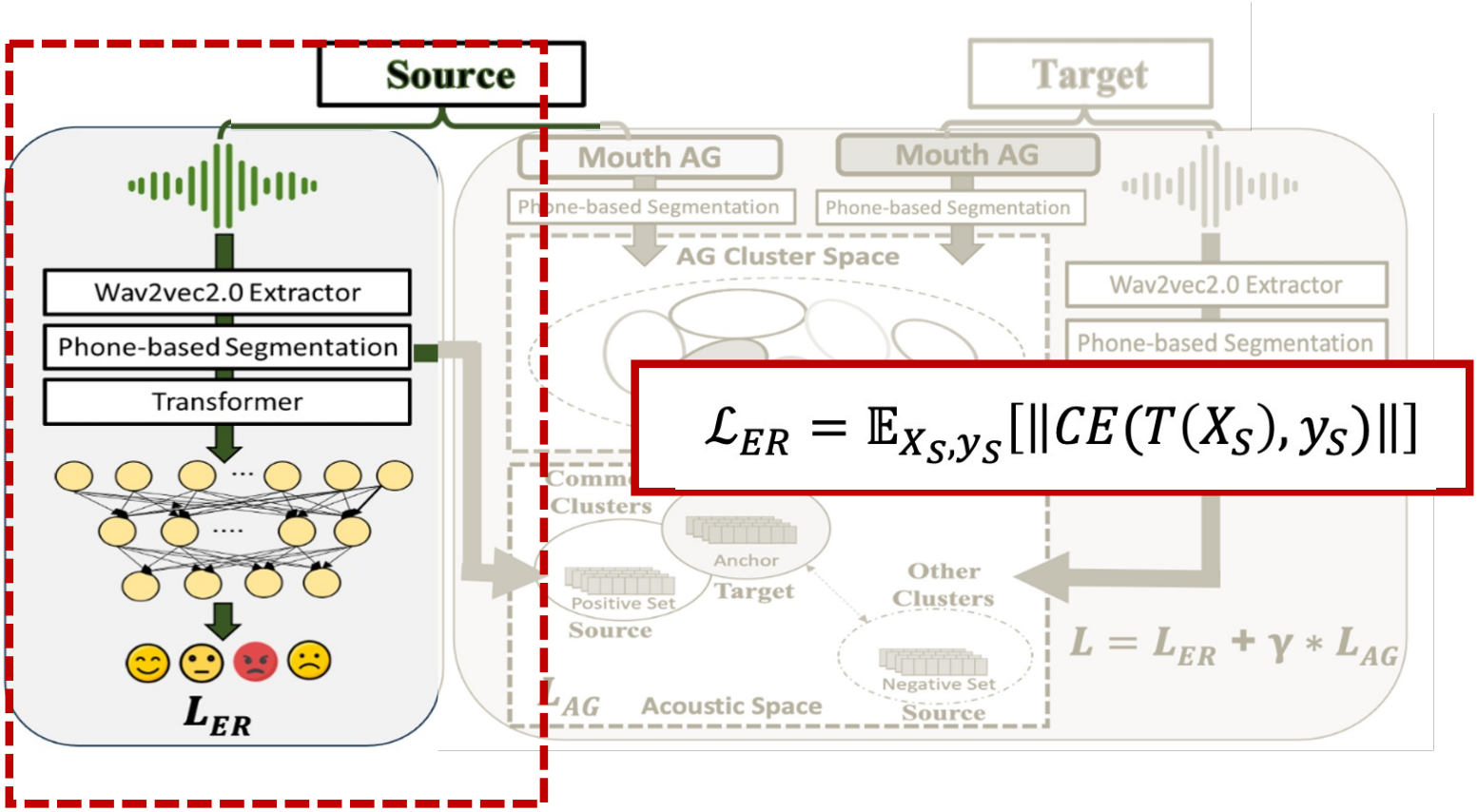


AG-ANCHORED CROSS-CORPUS SER

- 2 Branches:
  - 1st SER branch
  - 2nd AG-anchoring branch

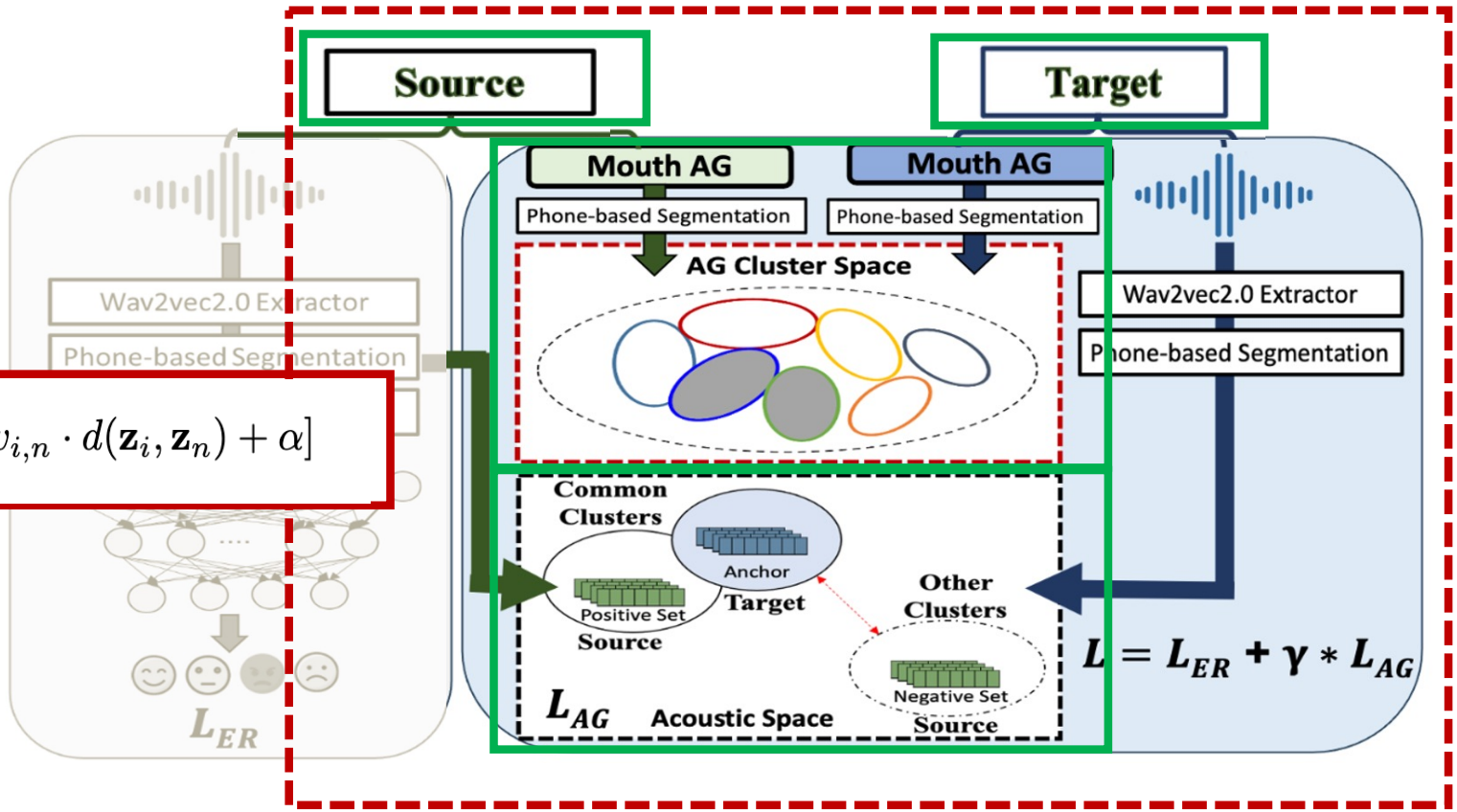


1st Branch: Conventional SER



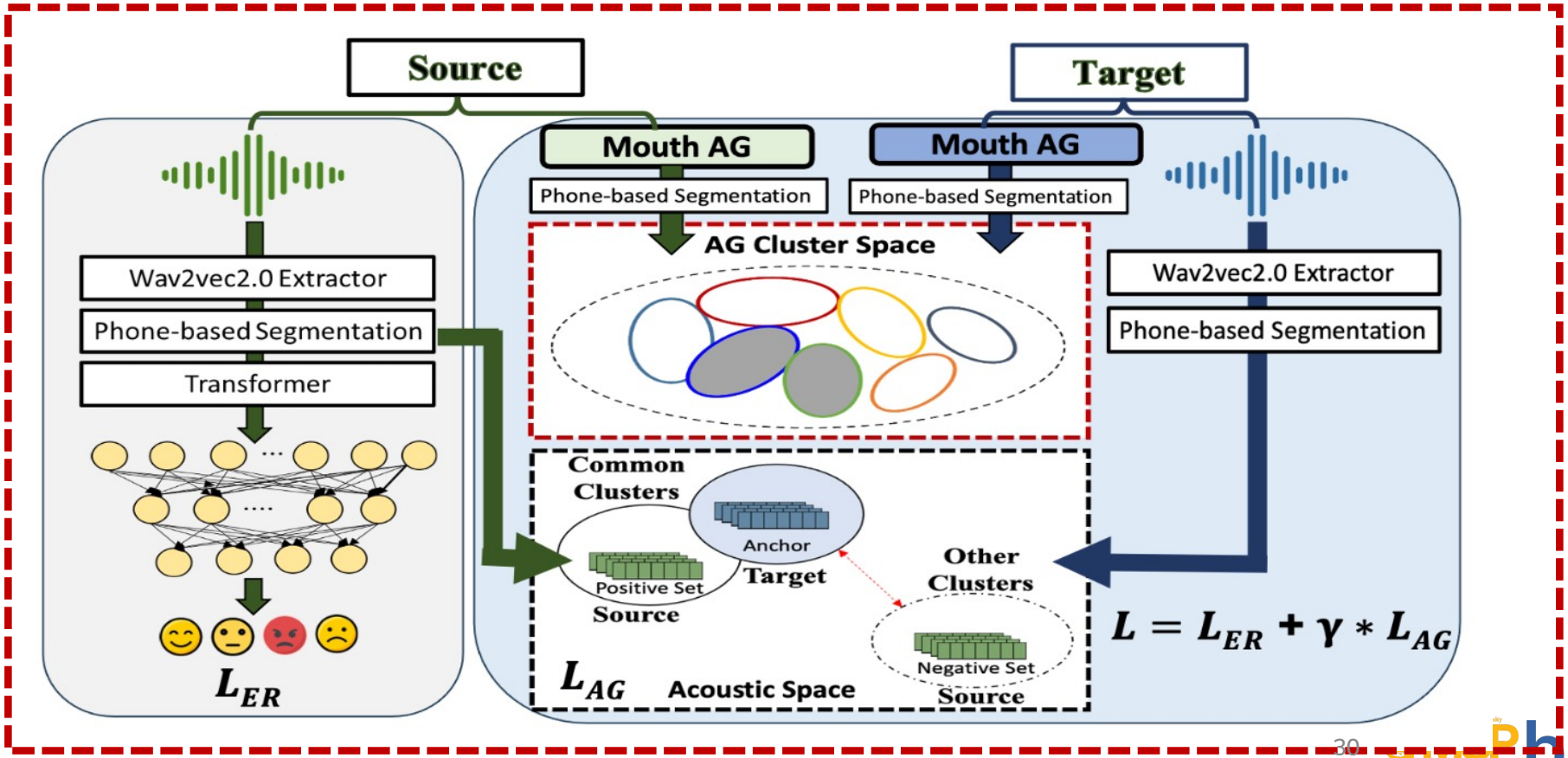
2nd Branch: AG-Anchoring

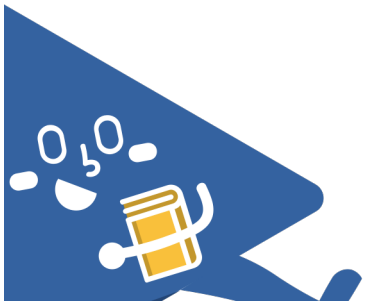
$$\mathcal{L}_{\text{AG}} = \sum_{i} [d(\mathbf{z}_i, \mathbf{z}_p) - w_{i,n} \cdot d(\mathbf{z}_i, \mathbf{z}_n) + \alpha]$$



Overall Loss

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{ER}} + \gamma * \mathcal{L}_{\text{AG}}$$





Experimental Results

Baseline Models

- 1) Phonetic-Anchoring (PA) [1]
  - Based on the common phoneme as an anchoring unit
- 2) Layer-Anchoring (LA) [2]
  - Based on the common pretrained model's layers as an anchoring unit

[1] S. G. Upadhyay, L. Martinez-Lucas, B.-H. Su, W.-C. Lin, W.-S. Chien, Y.-T. Wu, W. Katz, C. Busso, and C.-C. 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.
[2] S. G. Upadhyay, C. Busso, and C.-C. Lee, "A layer-anchoring strategy for enhancing cross-lingual speech emotion recognition," arXiv preprint arXiv:2407.04966, 2024.

Performance Comparison

C→I: CREMA-D to MSP-IMPROV
I→C: MSP-IMPROV to CREMA-D

		4-Category	Neutral	Anger	Happy	Sad
Upper Bound	C→C	66.36	89.44	88.27	85.35	79.51
	I→I	62.10	87.84	85.33	83.68	75.05
PA [1]	C→I	55.33	75.35	73.33	74.82	66.98
	I→C	53.18	75.46	72.14	73.64	63.35
LA [2]	C→I	56.04	78.24	75.49	73.50	66.73
	I→C	52.95	**77.67**	76.52	74.52	**67.73**
AG	C→I	**57.37**	**78.51**	**75.03**	**76.74**	**68.10**
	I→C	**53.83**	77.46	**77.72**	**75.30**	65.85

PA: Phoneme-Anchored Model
LA: Layer-Anchored Model
AG: Articulatory-Gesture Anchored Model

- For 4-Category SER: AG Outperforms PA and The LA
- For Specific-Emotions: Most AG's are outperforming

[1] S. G. Upadhyay, L. Martinez-Lucas, B.-H. Su, W.-C. Lin, W.-S. Chien, Y.-T. Wu, W. Katz, C. Busso, and C.-C. 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.
[2] S. G. Upadhyay, C. Busso, and C.-C. Lee, "A layer-anchoring strategy for enhancing cross-lingual speech emotion recognition," arXiv preprint arXiv:2407.04966, 2024.

Ablation: Hard-AG comparison

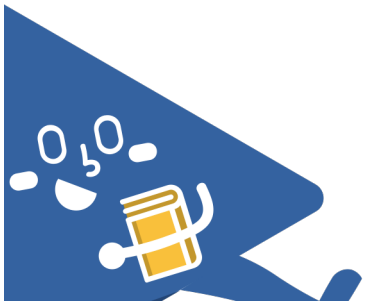
Hard-AG: Hard Segmentation
(NO Clustering)

C→I: CREMA-D TO MSP-IMPROV
I→C: MSP-IMPROV TO CREMA-D

		4-CAT	Neutral	Anger	Happy	Sad
AG	C→I	**57.37**	**78.51**	**75.03**	**76.74**	**68.10**
	I→C	**53.83**	77.46	**77.72**	**75.30**	65.85
Hard-AG	C→I	54.87	72.35	74.46	71.90	69.41
	I→C	52.90	70.68	71.53	70.24	63.24

Articulatory gesture anchoring (AG) Outperforms Hard-AG

Conclusion



Conclusion and Future Work

- AG anchors provide a robust foundation for emotion transfer
- Analysis shows that there is a correlation between AG gestures and their acoustic counterparts
- Future work:
  - Improve Speaker Variability Handling
    - Expand emotional categories for a more inclusive SER model
  - Enhance Generalization
    - Integrate anchoring with advanced domain adaptation techniques

Thanks!

Carnegie Mellon University MSP Multimodal Speech Processing Laboratory

biic National Tsing Hua University

NSF

NSTC 國科會