ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890262

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

Shreya G. Upadhyay[1], Ali N. Salman[2], Carlos Busso[2,3], Chi-Chun Lee[1]
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan.
[2]Department of Electrical and Computer Engineering, University of Texas at Dallas, USA.
[3]Language Technologies Institute, Carnegie Mellon University, USA.
shreya@gapp.nthu.edu.tw, ali.salman@utdallas.edu, busso@cmu.edu, cclee@ee.nthu.edu.tw

***Abstract*—Cross-corpus *speech emotion recognition* (SER) plays a vital role in numerous practical applications. Traditional approaches to cross-corpus emotion transfer often concentrate on adapting acoustic features to align with different corpora, domains, or labels. However, acoustic features are inherently variable and error-prone due to factors like speaker differences, domain shifts, and recording conditions. To address these challenges, this study adopts a novel contrastive approach by focusing on emotion-specific *articulatory gestures* as the core elements for analysis. By shifting the emphasis on the more stable and consistent articulatory gestures, we aim to enhance emotion transfer learning in SER tasks. Our research leverages the CREMA-D and MSP-IMPROV corpora as benchmarks and it reveals valuable insights into the commonality and reliability of these articulatory gestures. The findings highlight mouth articulatory gesture potential as a better constraint for improving emotion recognition across different settings or domains.**

***Index Terms*—speech emotion recognition, articulatory gestures, cross-corpus, transfer learning.**

I. Introduction

Speech emotion recognition (SER) systems are essential for improved user experiences across various applications, including automated call centers, education, entertainment, and medical fields [1]–[4]. In cross-corpus SER, aligning corpora from different domains and settings poses a significant challenge [5]. Existing research has introduced several techniques to address domain, label, and feature discrepancies, such as transfer learning, semi-supervised learning, and few-shot or zero-shot learning to improve model generalization [6]–[8]. Numerous approaches like optimizing distance metrics [9], adversarial training [10], GANs to generate synthetic data [11], and phonetic-based feature alignments [12] have also been explored. These SER methods mostly focused on handling acoustic feature mismatch between corpora, given their strong correlation with emotion and ease of recording.

In our previous work [13], we introduced a phoneme-anchoring approach to enhance cross-corpus alignment in SER. This method focused on identifying stable sub-units by leveraging shared vowel-phoneme emotion-specific acoustic spaces to match acoustic distributions across corpora. By establishing stable phoneme-based anchors, we hypothesized that similar phonemes would yield similar acoustic features. Unlike many previous approaches that attempted to directly match acoustic feature distributions, our stable phoneme-anchoring method led to improved SER performance in cross-corpus settings. However, a critical question remains: Are acoustic features the most stable anchors for cross-corpus alignment in SER tasks? While acoustic features play a key role in conveying emotion, they are also vulnerable to noise, microphone quality, and recording environment variations, which can undermine SER accuracy in cross-corpus scenarios. We believe that our previous idea can achieve further improvement by incorporating more stable anchoring units.

Acoustic signals and articulatory features are intrinsically linked, offering complementary insights into emotion recognition [14]–[16]. Significant research has explored mapping between these modalities, such as converting acoustic to articulatory and vice versa [15]. Emotions are closely tied to articulatory movements, particularly in facial expressions, where the mouth region is crucial due to its role in speech production [17]. Unlike broader facial features, mouth gesture are more stable because of their limited physical range, making them valuable for emotion recognition tasks [18], [19]. Focusing on the more stable aspect of speech production (such as *articulatory gesture*) can offer a promising alternative. In this study, we adopt the definition of *articulatory gestures* (AG) as the coordinated actions of speech organs (in this case, the mouth) that produce distinct phonemes. This study aims to improve cross-corpus alignment using stable AG properties, hypothesizing that stable mouth articulation patterns should result in similar acoustic characteristics.

Mouth articulation data is available through methods like electromagnetic articulography (EMA) [20] and real-time magnetic resonance imaging (MRI) [21], [22]. However, these methods are challenging to record and have limited corpora. Inspired by past research using marker information to analyze AG [18], [23], this work focuses on using mouth landmarks extracted from the visual modality as a representation of AG. This work introduces the concept of incorporating constraints on AG into transfer learning to improve emotion recognition accuracy across corpora. We evaluate our approach using two multimodal datasets, CREMA-D [24] and MSP-IMPROV [25]. Our proposed cross-modal anchoring idea, *articulatory gesture-anchored cross-corpus SER* (AG-CC), shows improved performance compared to the considered baseline.

II. Articulatory Gesture Analysis

A. Multi-Modal Affective Corpora

The **CREMA-D** [24] (CREMA) is a publicly available resource for emotion recognition research. It includes approximately 7.5 hours of recordings from 91 actors, each performing seven categorical emotions, and primary attributes across 12 scripts. With around 7,440 utterances averaging 3 to 4 seconds each, the corpus provides a rich source of multimodal emotional expressions through both audio and video recordings.

The **MSP-IMPROV** [25] (IMPROV) corpus includes approximately 8.5 hours of recordings, consisting of 8,438 prompted and spontaneous emotional sentences, with each utterance averaging about 4 seconds in length. This corpus is specifically designed for emotion recognition tasks and provides both audio and video recordings.

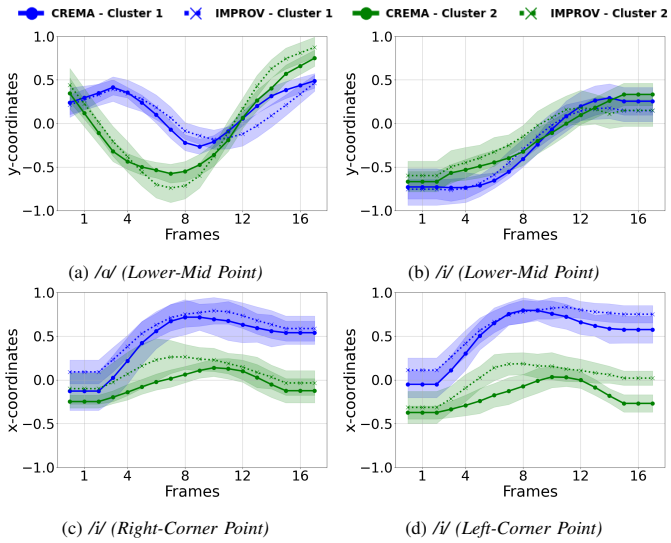
We select these corpora for their diverse emotional expressions and multimodal data. IMPROV provides naturalistic, conversational emotions with varied intensity, while CREMA offers more controlled, scripted emotional expressions that are often more intense. In this study, we only focus on four major emotions: *Neutral, Anger, Happiness*, and *Sadness*. The phoneme information for both the corpora is obtained using the Montreal Forced Aligner (MFA) [26].

B. AG Feature Extraction and Prepossessing

Our goal in this step is to extract robust features that enable the comparison of the mouth region across different subjects and corpora. To achieve this, we first use OpenFace [27] to detect the face bounding box and identify 68 2D landmark points that capture consistent facial features (e.g., eyes, chin, lips). We then rotate the landmarks to align the distance between the eyes parallel to the x-axis and normalize by the inter-pupil distance, reducing speaker-specific variations and ensuring consistent results from the same speaker across different sessions [28], [29]. This work examines mouth region AG by analyzing twelve key landmarks (48 to 59) that define the outer mouth shape. Building on insights from our previous research [13], we focus on six vowel phonemes: {ɑ, ə, ɛ, i, æ, u}. Phoneme-specific AG segments are extracted by segmenting frames based on phonetic boundaries.

C. Articulatory-Gesture Clustering

Mouth gestures are continuous and dynamic, making hard segmentation difficult without clear boundaries. Clustering provides a more effective approach for capturing distinct AG patterns. First, we segment the long landmark sequences into smaller, contextually relevant segments using phonetic boundaries. We then apply time-series *k*-means clustering [30] with Soft-DTW (Dynamic Time Warping) as the distance metric, grouping mouth shapes with similar articulatory patterns despite timing variations. Validation samples from all four emotion categories across both corpora are used to train the AG cluster model. Testing different *k* values (5 to 30), optimal cluster number is found to be 10 using the elbow method.



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

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

Fig. 1: Clustered AG patterns for /ɑ/ and /i/ from both corpora come from two AG clusters; shows the mean pattern at each frame, with standard deviation indicated over 50 samples for each vowel.

To evaluate the clustering model, we analyze how AG pattern variations for the same vowel are captured across different clusters. Fig. 1 presents examples using the vowels /ɑ/ and /i/. For this analysis, we focus on key points: the corner-left (48th landmark) and corner-right (54th landmark) for x-coordinate analysis and the lower mid-mouth point (57th landmark) for y-coordinate analysis. Fig. 1a and Fig. 1b show y-coordinate curves which is aligned with expected mouth movements. For /ɑ/, the downward AG curve in Fig. 1a reflects the lower mid-point moving down, while the upward trend in Fig. 1b for /i/ shows the lower mid-point rising, both consistent with expected articulation. For the x-coordinates, examining the corner points in Fig. 1c and Fig. 1d aligns with expected behavior, as the x-coordinates increase during the pronunciation of /i/ in their respective direction. The plots shown in Fig. 1 reveal distinct AG patterns across two clusters but show similar AG patterns from different corpora within the same cluster. This behavior is consistent across all the plots.

1) Cross-Corpus AG Cluster Overlap Analysis: In this analysis, we aim to identify overlap in AG clusters between two corpora by analyzing vowel-specific AG samples. We process these samples through an AG cluster model to determine their cluster IDs and then calculate the percentage of samples that are similarly clustered across the corpus using Equation 1. For example, if Corpus 1 has clusters 1 and 2, and Corpus 2 includes clusters 1, 2, and 3, we focus on the common clusters for similarity evaluation. To account for potential discrepancies and ensure accuracy, we use a 25% threshold to exclude very less common clusters, averaging similarity measures only for clusters exceeding this threshold. Table I shows the cross-corpus AG cluster Overlap analyses results.

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

TABLE I
PERCENTAGE OVERLAP OF AG CLUSTERS ACROSS CORPORA FOR SIX VOWELS, INCLUDING AVERAGE OVERLAP FOR EACH EMOTION-SPECIFIC CLUSTER.

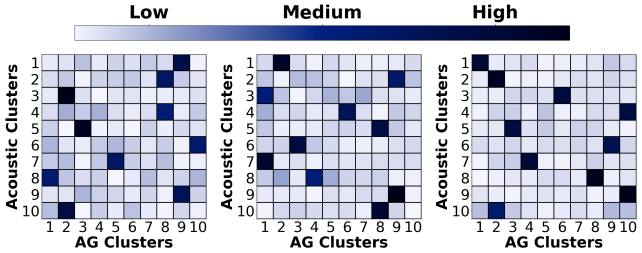
AG Cluster	/ɑ/	/ə/	/ɛ/	/i/	/æ/	/ɯ/
Cluster_1	75.3	2.1	60.6	71.1	78.9	0
Cluster_2	70.4	79.4	3.6	79.9	75.3	75.5
Cluster_3	61.2	64.5	77.1	0	76.7	10.8
Cluster_4	0	0	78.2	69.3	0	76.2
Cluster_5	0	0	0	5.7	15.5	0
Cluster_6	0	66.8	0	0	12.3	0
Cluster_7	68.4	0	18.4	2.5	0	74.5
Cluster_8	7.1	0	0	0	0	0
Cluster_9	0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

where, $K_{c,t}$ represents the number of common clusters between the two corpora after thresholding. $N_{1,i}$ and $N_{2,i}$ denote the number of samples in cluster i for the source and target, respectively, with N_1 and N_2 being the total number of samples in the source and target, respectively.

Table I shows clustering results in two sections: cluster-specific overlap for each vowel and average overlap across all clusters for emotion-specific analysis. Higher values indicate greater AG cluster similarity across corpora for the given vowels. From Table I, we observe substantial overlap in certain clusters, suggesting there exist shared articulatory patterns, which is promising for cross-corpus analyses. For instance, the vowel /ɑ/ is clustered into five groups (1, 2, 3, 7, 8), with most showing high overlap, indicating strong consistency in gestures for this vowel. We can observe this pattern in other vowels as well.

Table I reveals varying levels of emotion-specific overlap between corpora for different vowel phonemes. For example, the vowel /ɑ/ has a high similarity score of 80.1% for *Happiness*, indicating a significant overlap in AG patterns across corpora in this emotional context. Likewise, /i/ shows higher similarity scores for *Happiness* (78.2%) and *Anger* (75.6%), suggesting consistent articulatory patterns for this vowel across datasets in these emotional states. These results are consistent with previous studies [13], [18].

2) AG-Acoustic Features Association Analysis: To assess how well samples grouped by AG clusters align with their corresponding acoustic features, we model an acoustic cluster system similar to the AG clustering approach described in Section II-C, using time-series K-means with 10 clusters for consistency and comparability. This gives us two clustering models: one based on AG and the other on acoustic features. First, we cluster the samples using the AG cluster model. Then, for each AG cluster, we extract 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 between AG-based and acoustic clustering.



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

Fig. 2: Visualization of association between AG cluster and acoustic cluster across different emotions.

We visualize the relationship between AG clusters and acoustic clusters using a heatmap, with rows representing AG clusters and columns representing acoustic clusters. Each cell shows how samples from an AG cluster are distributed across acoustic clusters. Given that the AG and acoustic models are trained separately, a diagonal pattern is not expected. However, if samples from an AG cluster mostly align with fewer acoustic clusters, it indicates a stronger association between the AG and acoustic features. Fig. 2 illustrates that acoustic embeddings often cluster in less number of clusters within each AG cluster, reflecting that there is a correlation between AG gestures and their acoustic counterparts. However, no consistent pattern is observed across acoustic clusters. For example, AG cluster_1 aligns mostly with acoustic cluster_9 in *Happiness* but with clusters_2 in *Anger*. This suggests a one-to-one mapping from AG clusters to acoustic clusters, where each type of mouth articulation corresponds to a specific set of acoustic features. In contrast, the mapping from acoustic clusters to AG clusters is more one-to-many, indicating that a single type of acoustic feature can arise from various mouth articulations.

III. AG-ANCHORED CROSS-CORPUS SER

To improve the cross-corpus emotion transfer task, we propose an AG clusters-based cross-modal anchoring method for the 4-category SER. Here, we implement a constraint on the common AG clusters over both corpora to align their acoustic space features, called AG-anchored loss ($\mathcal{L}_{\text{AG}}$) shown in Equation 2. For each target sample, we form triplets using: (1) Anchor: the acoustic embedding of the target sample from the common AG clusters, (2) Positive: an embedding from the same AG cluster and emotion category but from the source dataset, and (3) Negative: an embedding from a different AG cluster within the same emotion category from the source corpus. We adjust the distance between the anchor and negative samples using a weight factor $w_{i,n}$, based on cluster centroid distances, and compute the soft triplet loss accordingly as shown in Equation 2.

$$\mathcal{L}_{\text{AG}} = \sum_{i} [d(\mathbf{z}_i, \mathbf{z}_p) - w_{i,n} \cdot d(\mathbf{z}_i, \mathbf{z}_n) + \alpha] \tag{2}$$

where $\mathbf{z}_i$, $\mathbf{z}_p$, and $\mathbf{z}_n$ represent the anchor, positive, and negative sets, respectively. $w_{i,n}$ is soft weight, estimate using

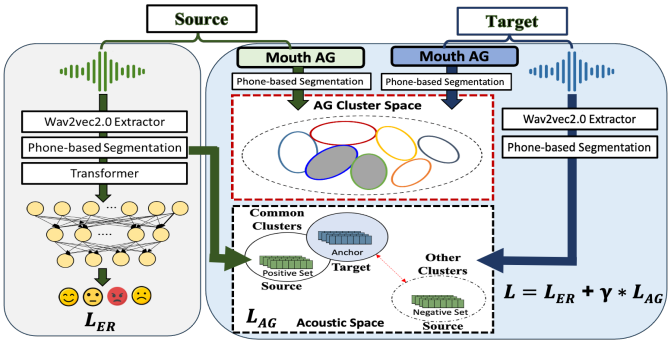


Fig. 3: Proposed mouth articulation-based anchoring architecture for cross-corpus SER.

the Equation 3. α is a margin parameter that enforces a minimum separation between the positive and negative pairs. Here the value of α is set to a constant value of 0.3.

$$w_{i,n} = exp(-\beta \cdot d(\mathcal{C}_{k_i}, \mathcal{C}_{k_n})) \tag{3}$$

where $d(\mathcal{C}_{k_i}, \mathcal{C}_{k_n})$ is the distance between the centroids of the clusters. β is a scaling parameter that controls the influence of cluster distances. Here the β value is set to 0.2.

This loss function, detailed in Equation 2, ensures that acoustic embeddings from the common AG cluster are closely aligned, while embeddings from different clusters are separated. By incorporating this loss into cross-corpus SER model training, we enhance the alignment of acoustic embeddings based on AG, improving cross-corpus SER. The total loss for each training batch combines the SER loss with the AG-anchored loss, as shown in Equation 4.

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{ER}} + \gamma * \mathcal{L}_{\text{AG}} \tag{4}$$

where $\mathcal{L}_{\text{ER}}$ is the conventional cross-entropy loss for 4-category emotion recognition and the $\mathcal{L}_{\text{AG}}$ is the soft weighted AG-anchored loss.

IV. Experiment Results

In our experiments, we benchmark using the CREMA and IMPROV corpora. We utilize Wav2vec2.0 [31] embeddings as pretrained features and apply a transformer with a 4-layer fully connected architecture, similar to our previous work [13], back-propagated using the loss function in Equation 4, with soft-weighted AG anchoring. The model is optimized with Adam, using a learning rate of 0.0001 and a decay factor of 0.001, and trained for up to 70 epochs with a batch size of 64 with early stopping. The performance is evaluated using the Unweighted Average Recall (UAR) metric.

In our study, we evaluate the effectiveness of the AG-CC method by comparing it with two baseline models: phoneme-anchored (PA-CC) [13] and layer-anchored (LA-CC) [32]. The PA-CC model leverages vowel phonemes as references to align emotional acoustic features in cross-corpora tasks, while the LA-CC approach aligns model layers to maintain consistent emotional patterns across corpora. In contrast, our AG-CC approach aligns these acoustic features through AG clusters to enhance cross-corpus SER.

TABLE II
SER performance (UAR) for baseline and proposed models: C→I (CREMA to IMPROV) and I→C (IMPROV to CREMA).

		4-CAT	Neu	Ang	Hap	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-CC [13]	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-CC [32]	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-CC	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

The cross-corpus SER performance of all considered models is presented in Table II. As evident from the results, our AG-CC approach outperforms all other models in both the 4-category (4-CAT) and binary SER tasks. Specifically, in the 4-category C→I task, where CREMA is the source and IMPROV is the target, AG-CC achieves a notable performance improvement of 2.02% over PA-CC and 1.33% over LA-CC. For binary SER in the C→I tasks, AG-CC delivers strong results across all emotion categories, for instance, *Anger* achieves 75.03% and *Happiness* reaches 76.74%. As a sanity check for our model, we also test for the I→C settings, where IMPROV is the source and CREMA is the target. Similar performance patterns are observed, with AG-CC surpassing PA-CC by 0.65% and LA-CC by 0.88%. Upper bound results for C→C and I→I SER tasks are also shown in Table II.

To further validate AG-CC, we compare it against the hard-segmented AG anchoring method (Hard-AG), as outline in Table II. In this comparison, Hard-AG refers to the use of fixed AG phonetic segments as anchors, consistent with our previous work [13] where we anchored on acoustic hard segments. The results, with 54.87% for C→I and 52.90% for I→C in the 4-CAT task, indicate that AG-CC's continuous clustering approach significantly outperforms the Hard-AG method. This is likely due to the continuous nature of AG, which lacks distinct boundaries, making clustering a better fit for capturing their continuous patterns.

V. Conclusion

We introduce the mouth articulation-based anchoring (AG-CC) approach to improve cross-corpus SER by aligning acoustic features across corpora through stable *articulatory gesture* (AG). By focusing on AG, which is more stable than acoustic features, we aim to enhance the generalization of SER systems across different domain corpora. The AG-CC method leverages stable AG anchors for cross-modal alignment, offering a robust foundation for emotion transfer. Our model AG-CC shows the better UAR with 57.37% for the 4-category cross-corpus SER task. Future work will concentrate on applying this concept in cross-lingual settings, improving AG-clustering strategies to better handle speaker variability, extending AG-CC evaluation to additional emotional categories, and optimizing performance across diverse acoustic environments.

REFERENCES

[1] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[2] J. C. Acosta, "Using emotion to gain rapport in a spoken dialog system," 2009.

[3] A. Tawari and M. Trivedi, "Speech based emotion classification framework for driver assistance system," in *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 2010, pp. 174–178.

[4] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[5] S. Zhang, R. Liu, X. Tao, and X. Zhao, "Deep cross-corpus speech emotion recognition: Recent advances and perspectives," *Frontiers in neurorobotics*, vol. 15, p. 784514, 2021.

[6] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2697–2709, 2020.

[7] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[8] X. Xu, J. Deng, Z. Zhang, Z. Yang, and B. W. Schuller, "Zero-shot speech emotion recognition using generative learning with reconstructed prototypes," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] J. Gideon, M. McInnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, October-December 2021.

[10] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[11] B.-H. Su and C.-C. Lee, "A conditional cycle emotion gan for cross corpus speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 351–357.

[12] S. G. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. N. Salman, C. Busso, and C.-C. Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *2023 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023.

[13] 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.

[14] 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.

[15] 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.

[16] Z. Zhang, M. Huang, and Z. Xiao, "A study of correlation between physiological process of articulation and emotions on mandarin chinese," *Speech Communication*, vol. 147, pp. 82–92, 2023.

[17] Y. Kim and E. Mower Provost, "Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 27–36.

[18] N. Sadoughi and C. Busso, "Expressive speech-driven lip movements with multitask learning," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 409–415.

[19] M. Shah, M. Tu, V. Berisha, C. Chakrabarti, and A. Spanias, "Articulation constrained learning with application to speech emotion recognition," *EURASIP journal on audio, speech, and music processing*, vol. 2019, pp. 1–17, 2019.

[20] D. Erickson, C. Zhu, S. Kawahara, and A. Suemitsu, "Articulation, acoustics and perception of mandarin chinese emotional speech," *Open Linguistics*, vol. 2, no. 1, 2016.

[21] J. Kim, A. Toutios, S. Lee, and S. S. Narayanan, "Vocal tract shaping of emotional speech," *Computer speech & language*, vol. 64, p. 101100, 2020.

[22] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[23] J. Wang, J. R. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," 2013.

[24] 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.

[25] 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 Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[26] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[27] 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.

[28] 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.

[29] 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.

[30] X. Huang, Y. Ye, L. Xiong, R. Y. Lau, N. Jiang, and S. Wang, "Time series k-means: A new k-means type smooth subspace clustering for time series data," *Information Sciences*, vol. 367, pp. 1–13, 2016.

[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[32] 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.