Phonetically-Anchored Domain Adaptation for Cross-Lingual Speech Emotion Recognition

Shreya G. Upadhyay[1], Luz Martinez-Lucas[2], *Student Member, IEEE,*
William Katz[3], Carlos Busso[2,4], *Fellow, IEEE*, Chi-Chun Lee[1], *Senior Member, IEEE*

Abstract—The prevalence of cross-lingual *speech emotion recognition* (SER) modeling has significantly increased due to its wide range of applications. Previous studies have primarily focused on technical strategies to adapt features, domains, and labels across languages, often overlooking the underlying universalities between the languages. In this study, we address the language adaptation challenge in cross-lingual scenarios by incorporating vowel-phonetic constraints. Our approach is structured in two main parts. Firstly, we investigate the vowel-phonetic commonalities associated with specific emotions across languages, particularly focusing on common vowels that prove to be valuable for SER modeling. Secondly, we utilize these identified common vowels as anchors to facilitate cross-lingual SER. To demonstrate the effectiveness of our approach, we conduct case studies using *American English*, *Taiwanese Mandarin*, and *Russian* using three naturalistic emotional speech corpora: the MSP-Podcast, BIIC-Podcast, and Dusha corpora. The proposed unsupervised cross-lingual SER model, leveraging this phonetic information, surpasses the performance of the baselines. This research provides insights into the importance of considering phonetic similarities across languages for effective language adaptation in cross-lingual SER scenarios.

Index Terms—speech emotion recognition, domain adaptation, cross-lingual, transfer learning.

1 Introduction

Effective generalized *speech emotion recognition* (SER) holds significant capability across a wide spectrum of applications, spanning various domains, such as the development of intelligent agents, social robots, voice assistants, and automated call center systems [1]–[3], with applications in healthcare, security, education, and entertainment [4]. Generalized SER demonstrates its effectiveness in diverse cross-context scenarios such as cross-lingual, cross-corpus, and cross-domain. Prior research has predominantly approached all cross-context challenges from a computational perspective, often considering cross-lingual scenarios as cross-corpus tasks by assuming language-agnostic contexts [5]–[10]. However, cross-lingual tasks diverge from cross-corpus cases, primarily because emotion perception and the acoustic feature space are language-dependent [11], suggesting that understanding the language-specific information provides valuable insights for cross-lingual adaptation strategies. The traditional approach to this problem has been from the computational standpoint by considering this as a data-matching issue and aiming to mitigate disparities between the source and target domains. This perspective might emerge due to constraints related to data availability and the inclination to repurpose established methodologies. To bridge these gaps, various techniques such as transfer learning, semi-supervised learning, and few-shot learning have been widely employed to reduce the differences in features, domains, or labels [8], [12]. Moreover, methods such as Wasserstein distance optimization [13], adversarial training [14], and the use of synthetic domain-specific data generated through Generative Adversarial Networks (GANs) [15] have been employed in the cross-context SER modeling. While these computational approaches within the cross-corpus framework have proven effective, they often lack the integration of linguistic knowledge when addressing tasks with different languages that require cross-lingual domain adaptation.

Rather than adopting a purely computational approach, our perspective includes a linguistic dimension, anchoring over universal attributes across different languages in emotional contexts. Speech serves as the carrier for both language and emotion. In cross-lingual scenarios, where the goal is to transfer emotions from one language to another, we contend that the anchor should be something universally recognized, specifically, a linguistic structure, and more precisely, phonemes. This choice is grounded in the fundamental concept that language is based on meaningful units of sound, and phonemes, comprised of vowels and consonants. Phonetically well-structured sounds contribute valuable linguistic cues to spoken information [16], [17]. Prior studies have shown that emotional information can be detected at the phonetic level [18], with specific emotional patterns at this level exhibiting applicability across various languages [19]. Moreover, vowel articulation plays a significant role in conveying emotional content, as emotions are often distinctly expressed through vowel sounds [20], [21]. It has also been observed that vowel sounds hold more significance in defining acoustic cues for emotions than consonants [16], [17].

By leveraging these insights, we can enhance the gen-

[1]*Department of Electrical Engineering, National Tsing Hua University, Taiwan.* [2]*Department of Electrical and Computer Engineering, University of Texas at Dallas, USA.* [3]*Department of Speech Language and Hearing, Department of Electrical and Computer Engineering, University of Texas at Dallas, USA.,* [4]*Language Technologies Institute, School of Computer Science, Carnegie Mellon University, USA.*
E-mail: shreya@gapp.nthu.edu.tw, {Luz.Martinez-Lucas, wkatz, busso}@utdallas.edu, cclee@ee.nthu.edu.tw

Manuscript received xxx xx, 2023; revised xxx xx, 2024.

eralization of SER for cross-lingual cases, a concept effectively demonstrated in our prior work [22]. In particular, our earlier study illuminated the phonetic commonalities that constitute the foundation of cross-lingual SER. Vowels, distinguished by their unique acoustic characteristics, served as anchors for transferring emotions between different languages, conveying emotional subtleties that transcend traditional method barriers. Our contrastive learning approach uses these vowel commonalities as constraints to mitigate variability between languages and enhance the efficacy of cross-lingual SER. This previous study based on phonetic anchoring showed that certain vowels exhibit commonalities between the two languages specific to emotions, revealing increased similarity between corpora after emotion modulations. Here, we select anchors by training the SER with vowel-specific segments to train the specific emotion recognition model and comparing them across corpora to identify vowels that provide better recognition under that emotion category. These common-performing vowels were used as anchor candidates, and implementing the anchoring mechanism with these vowels led to improved performance. This initial analysis provides insights and better cross-lingual SER performance, supporting our hypothesis. However, in our previous work, the anchor selection method required additional model training, which inherently demands a significant amount of data. When data is limited, it becomes difficult to fully trust the model's performance. Thus, while our previous work demonstrated some effectiveness, it also had limitations due to the need for additional data and the challenge of model reliability.

In our current research, we extend upon our prior work [22], introducing several novel aspects to our study. Firstly, our previous work successfully emphasized the importance of vowel commonalities in bridging linguistic disparities. In this study, we delve deeper into phonetic common features by expanding our analysis to encompass not only monophthong vowels but also diphthongs. This addition allows us to gain a more comprehensive understanding of vowel behavior in emotion recognition. Secondly, we refine our anchoring selection and mechanism more tied to the modeling aspect. Given the prevalent use of large pre-trained models (e.g., Wav2vec2.0) in SER for encoding emotional content, our phonetic analysis also centers around this specific emotional encoder to enhance its effectiveness. Lastly, we also introduce an additional cross-attention mechanism into the proposed architecture. This modification aims to provide a more holistic comprehension of phonetic commonalities and integrate them into the context of cross-lingual SER scenarios for better language adaptation. Overall, our study presents a novel attention-based phonetic-anchored domain adaptation technique for cross-lingual SER, leveraging the commonalities in vowels across languages to enhance our transfer learning strategy for SER.

Specifically, we conduct a case study involving two contrastive languages: *American English* (an intonation-based language) and *Taiwanese Mandarin* (a tonal language), utilizing two extensive in-the-wild natural speech emotion datasets, the MSP-Podcast (*American English*) and BIIC-Podcast (*Taiwanese Mandarin*) corpora. The research is divided into two primary segments: First, we investigate the emotion-specific commonalities at the phonetic level, analyzing phonological references (in terms of F1 *v/s*. F2 formants) and encoded feature representations for emotion-specific vowels (using Wav2vec2.0). By extensively examining speech data across diverse emotional contexts, the insights uncover the significant associations between particular vowel qualities and emotions, some of which are common across languages. These findings highlight the cross-linguistic convergence in linking phonemes with emotions, suggesting that, despite variations in emotional expression across languages, there exist underlying patterns that tie certain vowels to distinct emotional states, transcending linguistic boundaries. Second, this study introduces an anchoring mechanism to enhance cross-lingual SER by leveraging the phonetic commonalities identified earlier. The anchoring mechanism is designed to exploit the common behavior of vowel phonemes in the target languages and by employing a contrastive formulation, we demonstrate the effectiveness of this approach in improving cross-lingual SER performance. Our proposed attention-based group-anchored cross-lingual SER model (AGA-CL) achieves a 6.89% improvement in unweighted average recall (UAR) for a 4-category SER task compared to models that do not incorporate any domain adaptation technique (CL). Our proposed approach is further validated on crosslingual experiments for *Russian*, using the DUSHA corpus, demonstrating consistent improvements. This significant enhancement highlights the efficacy of our approach in leveraging phonetic commonalities across languages to enhance the performance of cross-lingual SER.

2 Related Work

This paper approaches the topic from a linguistic perspective by examining emotion-specific vowel commonalities in the context of unsupervised cross-lingual research. Related work in this paper is divided into two sections: the first focuses on the computational study of cross-lingual research, and the second on the emotional patterns observed on vowels.

2.1 Computational Study Of Cross-Lingual Data

Unsupervised cross-lingual SER offered a promising way to reduce the need for labeled data in the target domain, making SER models more adaptable across languages. Existing research explored various strategies [5], [8], [23]–[25]. Some popular techniques included addressing source-target domain disparities through transfer learning, semi-supervised learning, or few-shot learning [8], [12]. These methods adapted models to the target domain using source domain knowledge, reducing the need for labeled data. Another approach used synthetic domain-specific data, generated by methods like GANs or data augmentation [5], [13]–[15], to improve emotion recognition by simulating target domain samples. Additionally, methods using pseudo-labeling [25] treated cross-lingual SER as a multi-label classification problem by utilizing unsupervised techniques to overcome limited labeled data challenges in target domains. These strategies offered computational solutions for cross-lingual SER.

2.1.1 Addressing Disparities Of Source and Target

Addressing the disparities between the source and target domains in cross-lingual SER tasks was a common focus in the literature. The transfer learning approach was a widely used approach that leveraged knowledge learned from a source language to enhance performance in a target language [6], [7]. Another avenue for improving model generalizability was through the integration of multi-task learning, where additional information such as gender and naturalness was incorporated into the learning process [23]. Alternatively, few-shot learning had been proposed, specifically adapted to the target domain by enabling the model to learn emotions from source domain samples [8]. However, few-shot learning faced challenges in practical applications due to the reliance on selecting a suitable support set and the effectiveness of few-shot learning was highly dependent on the composition and quality of the support set, introducing complexities and limitations in real-world scenarios [26]. Additionally, self-supervised learning had been explored as a means to bridge the gap between the source and target domains by leveraging the inherent structure and patterns within the data [10].

2.1.2 Leveraging Adversarial Networks

In cross-lingual SER, another popular direction is to utilize adversarial-based methods, which leveraged domain adversarial training to learn representations that were invariant across different language corpora. For instance, a Generative Adversarial Network (GAN)-based method was proposed by Latif et al. [5] for unsupervised domain adaptation in multilingual SER, enabling the learning of language-invariant feature representations from source to target languages. However, GANs often faced challenges in training and were prone to convergence failures [27]. In contrast, Abdelwahab and Busso [24] proposed a Domain Adversarial Neural Network (DANN) approach, which aimed to generate a domain-invariant feature representation that minimized the discrepancy between source and target domain features . However, the effectiveness of domain adversarial training heavily relied on the distribution of the two databases, and adversarial attacks and instabilities could arise during training when the data points exhibited significant dissimilarities [28].

2.1.3 Employing Pseudo Labelling

Unsupervised cross-lingual SER has also been approached as a multi-label classification problem in the literature by drawing inspiration from ensemble learning [29]. This learning strategy combined decisions from multiple related emotion features to reduce the impact of individual emotion labels and enhance the robustness of the model. Furthermore, dynamic external memory was designed to store and update the source domain features effectively, ensuring that all samples from the source domain were utilized during training. The features from the source domain were then stored in the dynamic external memory. Pseudo multi-labels were assigned to the target domain data by calculating the similarities between the features in the dynamic external memory and the features in the target domain.

2.2 Affective Study Of Vowels Behavior

Numerous linguistic studies have underscored the behavior of vowel phonemes in conveying emotions and their contribution to the expression of distinct emotional states due to their unique characteristics [20], [21], [30]. There is substantial evidence supporting robust phoneme-emotion correspondences, which are likely universal [31]. However, some linguistic research has shown that these phoneme-emotion associations can be language-specific [32]. Vowels, due to their distinctive acoustic properties, are recognized as key elements in speech that effectively encode and transmit emotional nuances [16], [17], [33]. This perspective motivates the investigation of vowel behaviors to unlock the potential of more generalized and language-adapted emotion recognition systems.

2.2.1 Emotion-Specific Vowel Behavior

Several studies have highlighted the significance of vowels in expressing emotional content, with various investigations exploring the connection between specific vowels and emotions [17], [18], [34]. Insights from psychology suggest that emotions may not solely be conveyed through entire words but also through individual sounds or phonemes. This perspective leads to the intriguing observation that the prevalence of /i/ sounds in positive emotions and /o/ and /u/ sounds in negative emotions, observed across various languages, appears linked to the concurrent activation of specific facial muscles [34]. Studies investigating the relation between sound and linguistic meaning provide substantial evidence of an implicit connection between the articulatory and acoustic properties of vowels and emotions [35], [36], using analyses of vowel formant frequency and formant dispersion. Specifically, extremes in vowel acoustic features have emerged as significant predictors for discerning emotional tone [35]. Other studies [36] suggest that perceived emotion from specific phoneme combinations relies on their inherent acoustic features. An articulatory and acoustic analysis of Mandarin Chinese vowels [37] highlighted the emotional sensitivity of the /a/ vowel, with greater F1 values observed in emotional states such as happiness and *anger*, and smaller F1 values in *sadness* and *neutral* states. Additionally, emotional states of happiness and *anger* exhibited larger vocalic triangles compared to *neutral* and *sadness*, aligning with findings in English-language studies [38]. These findings suggest the presence of vowel-specific acoustic attributes influencing human emotions.

2.2.2 Vowel-Specific Emotion Encoding

Many studies have emphasized that vowels can serve as distinct markers closely associated with specific emotional states in language. Some investigations reveal a strong correlation between the positions of average F1/F2 values extracted at the vowel level and the speaker's emotional *arousal* and *valence* values [39], [40]. Notably, Shah et al. [40] suggest that incorporating articulatory information significantly enhances *valence*-based classification performance for both within-corpus and cross-corpus emotion recognition, particularly effective in distinguishing happiness from other emotional states. Approaches like employing phoneme-class-dependent emotion classifiers [18] and utilizing deep

TABLE 1: Statistics of the corpora used in this study.

	Overall	neu	hap	ang	sad
MSP-P	49018	25467	2773	15821	4957
BIIC-P	18980	7933	6000	2763	2284

models fine-tuned with emotion-dependent phoneme transcriptions [17] have shown effectiveness in enhancing emotion recognition. Several studies in emotion recognition have also adopted a vowel-centric approach, consistently supporting that among various speech units, the vocalic nucleus is most perceptually significant [17], [19], [33].

Conversely, there is a growing body of research exploring sound-meaning associations [41], and sound symbolism [42], which have been observed across multiple languages. This suggests that certain vowels may consistently convey emotional connotations across diverse languages. These emotionally significant vowels can function as connectors that facilitate cross-lingual emotion recognition. Recent investigations have aimed to bridge the gap between linguistic analyses of vowels and the practical applications of this knowledge in the context of cross-lingual SER. For instance, our previous work [22] has illuminated the potential of phonetic commonalities, specifically focusing on vowels, to serve as bridges between languages in emotion recognition.

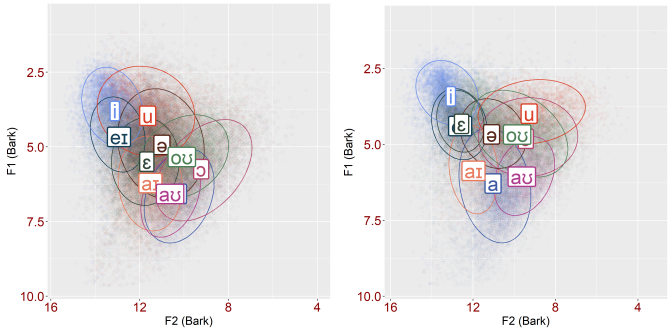
Emerging insights from these related studies underscore the advantages of integrating linguistic knowledge when dealing with cross-lingual SER. This approach broadens the perspective on cross-lingual SER, extending beyond the computational viewpoint and encompassing linguistic considerations. We facilitate the enhancement of cross-lingual SER systems in their adaptation to the target language by gaining an understanding of the language's phonetic characteristics specifically the *Vowels* that are specific to each language but share some commonalities, and implementing mechanisms to align these phonetic elements.

3 Corpora

This work considers the MSP-Podcast [43] and BIIC-Podcast [22] SER corpora to evaluate our proposed idea. The MSP-Podcast and BIIC-Podcast corpora have undergone meticulous annotation by human raters, ensuring the availability of accurate emotion labels as ground truth. These corpora serve as valuable and trustworthy resources for training and evaluating SER models, allowing researchers to assess the efficacy and applicability of their proposed methodologies. Table 1 provides an overview of the sample counts drawn from both corpora for this study.

3.1 MSP-Podcast Corpus

The MSP-Podcast (MSP-P) corpus [43] is used as a benchmark for our research, as it comprises a total of 166 hours of emotional speech in *American English* (v1.10). It is increasingly being used for research on SER due to its scale and availability of emotionally balanced dialogues from multiple speakers. These speech samples were collected from podcast recordings available on various audio-sharing websites. Each sample in the corpus is annotated by at least five different workers with primary emotions (*neutral, happiness, anger, sadness, disgust, contempt, fear*, and *surprise*), secondary emotions, and emotional attributes (*arousal, valence*, and *dominance*). To examine the phonetic components of the MSP-P corpus, we employed the Montreal Forced Aligner (MFA) [44]. The MFA is particularly valuable as it offers precise boundary alignments for phones, using the ARPABET notation. To ensure consistency across languages, we utilize the International Phonetic Alphabet (IPA) [45], a widely recognized standard in linguistic research, for phonemic symbols. This conversion process was executed following a well-established mapping by Rice [46]. By employing this mapping, we ensured that the phones extracted from the MSP-P corpus were represented in a phonetically accurate and standardized manner, facilitating our subsequent phonetic analysis and comparison across languages.



(a) MSP-Podcast (b) BIIC-Podcast

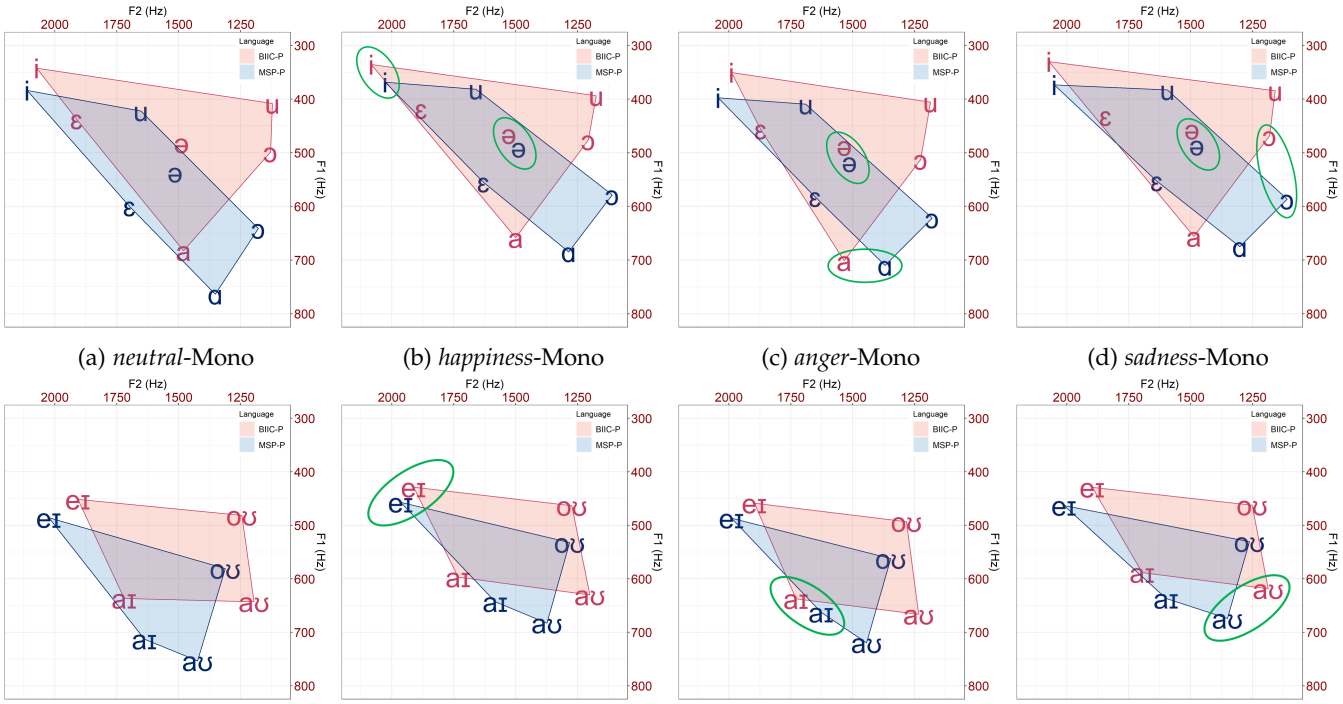
Fig. 1: Considered vowels in vowel space (F2 *v/s.* F1 formants) of MSP-P and BIIC-P corpora.

3.2 BIIC-Podcast Corpus

We use the BIIC-Podcast (BIIC-P) corpus [22] for evaluating our idea on cross-lingual SER tasks. It consists of speech samples extracted from *Taiwanese Mandarin* language. Each sample in the BIIC-P corpus is annotated with emotional labels, with each sample having between 3 to 6 annotations available. The emotional annotations cover eight primary emotional categories, namely *neutral, happiness, anger, sadness, disgust, contempt, fear*, and *surprise*. Additionally, the dataset includes 7-point Likert scale labels for *arousal, valence*, and *dominance*. Manual transcriptions are available for all samples in the corpus. For this research, we utilized approximately 137 hours of data from the BIIC-P corpus (v.1.0). To facilitate further phonetic analysis and phonetic-anchored SER modeling, we first train a *Taiwanese Mandarin* forced aligner using the *Formosa* [47] corpus. This aligner allows us to align the speech samples to their corresponding phonetic segments. The phonetic segments are then converted to IPA phonemic notation using the mapping provided by Liao et al. [47].

4 Phonetic Commonality Analyses

This section explores the estimation of vowel commonalities in both the MSP-P and BIIC-P corpora. Based on the language phonetic universality fact, understanding the common vowels (monophthongs and diphthongs) across



(a) *neutral*-Mono (b) *happiness*-Mono (c) *anger*-Mono (d) *sadness*-Mono

(e) *neutral*-Diph (f) *happiness*-Diph (g) *anger*-Diph (h) *sadness*-Diph

Fig. 2: Visualization of monophthongs and diphthongs vowel polygons representing the vowel space for the emotions *neutral*, *happiness*, *anger*, and *sadness* across the MSP-P and BIIC-P corpora.

emotions and languages can greatly enhance the effectiveness of SER systems in accurately detecting and classifying diverse emotions across languages. To explore these common vowels, we conduct a comprehensive analysis utilizing two distinct corpora (MSP-Podcast and BIIC-Podcast) from different perspectives. Firstly, we explore the conventional formant-based vowel spaces to find potential commonalities across the corpora. Subsequently, we also examine vowel correlations within the context of Wav2vec2.0 [48] which plays a crucial role in our SER modeling effort.

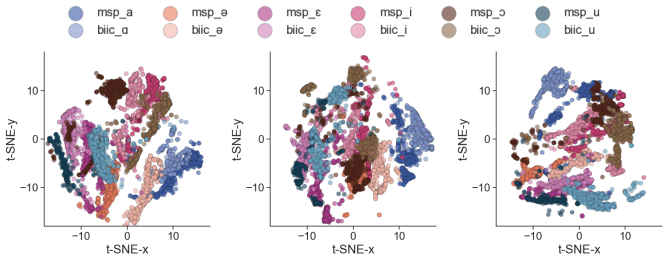
In this analysis, we utilize a limited amount of labeled data from the validation sets of both corpora, specifically using 500 labeled samples for each emotion category from both the source and target datasets. This allows us to analyze the behavior and select vowels as anchors based on their commonalities across the corpora. Importantly, no emotional labels from the target set are used during training, which is why our method is referred to as unsupervised cross-lingual transfer learning.

4.1 Formant-Based Phonetic Analyses

We conduct a vowel space analysis of both corpora by using their formants. This analysis is carried out both holistically and on an emotion-specific basis to identify common vowels between the MSP-P and BIIC-P corpora. In our prior research [22], we concentrated solely on monophthong vowels. In this study, we broaden our perspective by incorporating diphthongs to facilitate a more comprehensive utilization and understanding of vowel phonemes.

Vowel Space Commonalities: The vowel space, depicted in Fig. 1, provides a visualization of the F1 *vs.* F2 plots for the speech samples across various emotions. Upon examining Fig. 1, it is evident that the *English* vowel /u/ appears to be high-fronted, which aligns with findings reported in [22], [49]. This discrepancy in vowel placement could be attributed to factors such as variations in gender ratios or dialects among the speakers. Notably, the placement of this vowel differs in BIIC-P compared to MSP-P. Overall, Figure 1 demonstrates similar vowel characteristics in *English* and *Mandarin*. The plots indicate that the set of common vowels chosen for analysis spans a substantial portion of the F1-F2 space, and their positions align with the expected placement discussed in existing literature [50]–[52]. By examining Figure 1, we can observe certain visible commonalities in vowel distribution across corpora. For instance, vowels /i/ and /ə/ exhibit similarity regions in their respective languages. This observation suggests the presence of vowel commonality between *English* and *Taiwanese Mandarin*.

Specific-Emotion Vowel Space Commonalities: Fig. 2 presents a comprehensive visualization of the average F1 and F2 values for four emotional classes: *neutral*, *happiness*, *anger*, and *sadness*. The data depicted in Fig. 2 have been subjected to Nearey normalization [53] to eliminate speaker-related variations arising from differences in vocal tract characteristics and gender. Specifically, Fig. 2 highlights the vowel distances observed in *neutral* speech, revealing that the closest distances across languages for corresponding vowels occur between /i/ and /ə/. This consistent trend holds true across all four emotional categories. These emotion-specific vowels exhibit commonalities across both languages, making them potential candidates for serving as anchors in our transfer learning strategy.



(a) *happiness* (b) *anger* (c) *sadness*

Fig. 3: t-SNE visualization illustrating the Wav2vec2.0 feature representations of vowels across both the MSP-P and BIIC-P corpora for the emotions being studied.

In Figure 2, we encircle in green those vowels that exhibit closer distances across languages for *happiness*, *anger*, and *sadness* compared to *neutral* (e.g., the distance between /ɛ/ across languages is smaller for *happiness* than for *neutral*). These highlighted regions provide valuable insights into vowels that consistently display similar responses across languages when emotions are present.

This formant-based phonetic analysis uncovers the significance of identifying vowel commonalities as key anchor points for our transfer learning approach in cross-lingual SER. The visualizations shown in the plots help us to identify vowels that consistently demonstrate similar behavior across languages. This knowledge contributes to more precise and reliable emotion recognition in diverse linguistic contexts. Building upon these observations, our next objective is to determine the vowel anchors that are optimal for expressing specific emotions and are common between the two languages.

4.2 Wav2vec2.0 Phonetic Commonality

Numerous investigations have explored the connection between Wav2vec2.0, a widely-used self-supervised speech model, and phonemes, which serve as the fundamental building blocks of speech. These studies offer valuable insights into the model's learning process and its representation of phonetic information, providing a deeper understanding of both end-to-end speech recognition models and self-supervised speech models. Notably, previous research has demonstrated a strong correlation between the learned representations of Wav2vec2.0 and phonemes, suggesting that the model's encoder can acquire phonetic information in an unsupervised manner [54], [55]. Furthermore, some studies have shown that pre-trained Wav2vec2.0 embeddings contain valuable phonetic information related to the manner and place of articulation [56], [57]. This finding is significant as it implies that these embeddings can be effectively utilized in various downstream tasks, including automatic speech recognition, speaker identification, and SER. Consequently, the ability of end-to-end models like Wav2vec2.0 to recognize speech without explicit phonetic supervision sets the foundation for the development of more efficient and accurate systems for SER.

To understand the degree of similarity in the Wav2vec2.0 features between MSP-P and BIIC-P samples in the context of vowel phonemes, we conducted an analysis by examining their vowel-centered feature closeness. To gain insights into the emotional samples present in the utilized corpora, we fine-tune the Wav2Vec2.0 model on both the BIIC-P and MSP-P datasets together for the emotion recognition downstream task. This approach ensures that both datasets are represented in the same feature space, facilitating a more accurate comparison. We then utilize the final layer of the fine-tuned Wav2Vec2.0 model to extract features. To estimate the phoneme-centered features, we segment these features by taking a 120ms window on either side of the phoneme's center time frame, considering this as the phoneme-specific segment features, following the chunk-based method outlined in [58]. These features are subsequently used to generate t-SNE plots for visualization.

Fig. 3 presents the t-SNE visualization for monophthongs vowels, depicting the Wav2vec2.0 feature representations of vowels across both the MSP-P and BIIC-P corpora for the emotions under investigation. These plots illustrate the presence of vowel clusters that demonstrate closer proximity across these two distinct language corpora. For instance, when we analyze the corpus vowel sets containing /i/ (msp_i and biic_i), we notice some level of closeness in all the plots. Similarly, we can observe varying levels of closeness, whether high or low, between all corpus vowel phoneme clusters presented in Fig. 3. Furthermore, they reveal that certain vowels exhibit different levels of similarity in particular emotions. For example, in the context of *happiness*, we observe that the vowel sets of /i/ (msp_i and biic_i), as well as /ɑ/a/ (msp_ɑ and biic_a) cluster closely together. Similarly, in the case of *anger*, the vowel sets of /ɑ/a/ (msp_ɑ and biic_a) and /ɔ/ (msp_ɔ and biic_ɔ), appear to be more closely related. These patterns of vowel cluster closeness can be observed across all emotions, although they are somewhat less pronounced in the case of *sadness*. In this paper's representation, we only show the feature closeness for monophthong vowels but we also observe similar insights for diphthong vowels.

5 Anchor Candidates Selection

There are distinct ways to select vowel anchor candidates from a set of common vowels across two languages. On the one hand, a straightforward approach is to consider basic vowel-space formant commonalities as selection criteria. However, this may not be the most efficient method given that modern SER models often use more complex features, not just basic ones, sometimes encoded from advanced models like Wav2vec2.0. As demonstrated in our previous work [22], we adopted a feature-centric strategy that engaged in vowel-centered emotion-specific modeling. This approach allowed us to identify vowels that prominently contribute to the expression of specific emotions while exhibiting consistent behavior across both languages. Although this approach has its limitations, such as potential accuracy issues with the model and the need for sufficient labeled training data, in our current study, we opt for a more direct approach. We utilize vowel-centered feature distances and similarities to pinpoint vowels displaying emotion-specific behavior across languages. Among the various strategies for anchor selection, we favor this approach due to its simplicity

TABLE 2: Similarity and distance metrics across common vowels in both corpora, specific to each emotion and measured in terms of cosine similarity (CS) and Euclidean distances (ED) and the combined score (Com).

		Mono						Diph			
		/i/	/ɛ/	/ə/	/ɑ,a/	/ɔ/	/u/	/aɪ/	/eɪ/	/aʊ/	/oʊ/
neu	CS	0.96	0.87	0.85	0.97	0.93	0.81	0.81	0.85	0.86	0.87
	ED	2.01	3.21	3.68	2.45	2.12	3.91	4.01	3.98	3.91	3.28
	Com	**0.81**	0.67	0.62	0.78	0.79	0.58	0.57	0.59	0.60	**0.66**
hap	CS	0.96	0.89	0.92	0.96	0.86	0.79	0.92	0.91	0.88	0.83
	ED	2.21	3.04	2.24	2.26	3.33	3.47	3.24	3.16	4.21	4.66
	Com	**0.80**	0.69	0.77	0.79	0.65	0.61	**0.69**	0.68	0.59	0.53
ang	CS	0.86	0.92	0.93	0.94	0.81	0.79	0.89	0.94	0.92	0.81
	ED	3.97	2.96	3.02	2.78	3.58	3.73	4.02	3.04	3.15	3.49
	Com	0.60	0.71	0.71	**0.74**	0.61	0.58	0.61	**0.72**	0.70	0.61
sad	CS	0.91	0.91	0.84	0.91	0.93	0.82	0.84	0.83	0.89	0.88
	ED	3.64	3.02	3.77	3.25	2.81	3.67	5.36	5.62	4.19	4.98
	Com	0.65	0.70	0.61	0.68	**0.73**	0.60	0.47	0.45	**0.60**	0.53

in estimation, not requiring extensive memory or a large number of labeled samples.

To choose suitable vowels for the proposed anchoring mechanism of our study, we use Wav2vec2.0 features for training. We compute multidimensional distances based on vowel-centered feature representations from both corpora. This estimation of multidimensional distances of common ground vowels is performed using the Cosine Similarity (CS) and Euclidean Distances (ED), allowing us to capture the closeness in terms of similarity and distance metrics across different emotions. Table 2 presents the results of the emotion-specific multidimensional CS and ED for common vowels in both corpora. In the process of selecting anchor candidates based on the metrics (CS and ED), our objective is to achieve a low ED and high CS to identify closer vowels. To make both metrics comparable, we start by normalizing the ED values within the range [0, 1], referred to as ED_{norm}. Subsequently, we apply an inverse transformation to the normalized ED values (ED_{inv}) using Equation 1 and compute an average score using CS and the transformed ED scores. This average score is referred to as the combined score (*Com*), as shown in Equation 2. The candidates chosen for the anchoring strategy, as shown in Table 3, are selected based on these combined scores.

$$ED_{\text{inv}} = 1 - ED_{\text{norm}} \tag{1}$$

$$Com = \frac{CS + ED_{\text{inv}}}{2} \tag{2}$$

There can be several possible approaches for anchor selections. One option is to choose the best anchor, which exhibits more feature similarity compared to other vowels. However, based on our previous work [22] discussions, it is apparent that opting for a group of top anchors can yield better SER performance. This approach is more rational than selecting a single vowel as an anchor because, from the perspective of speech production, emotional speech cannot be exclusively associated with a single vowel. For our subsequent experiments, we have chosen both the best anchors (BA) and group anchors (GA). These vowel anchors are selected based on the ranked combined scores for both monophthongs and diphthongs for each emotion. For BA, we pick the top-ranked vowels, and for GA, we select the top 50% of vowels present from both monophthongs and diphthongs, separately. To comprehensively evaluate the performance of our method, we have incorporated the worst anchors (WA) in our experiment. These WA represent the lowest-ranked vowels. Regarding the WA, we observe that the vowel /u/ consistently ranked at the bottom for all emotion candidate selections. To have a better comparison, we also included the second lowest-ranked vowels in our analysis. Table 3 shows the anchor candidates included in the BA, WA, and GA sets which, respectively, represent vowels with the highest, lowest, and group with the utmost closeness in terms of vowel feature similarity. For example, in *happiness*, /i/ is selected as BA, /ɔ/ and /u/ as the WA, and /i/, /ə/, and /ɑ/a/ are considered as GA for the monophthongs (Mono) category. Similarly, in the diphthongs (Diph) category, /aɪ/ is selected for BA, /oʊ/ is selected for WA, and /eɪ/ and /aɪ/ are selected for GA.

TABLE 3: Anchor vowel candidates selected for the proposed anchoring mechanism based on their higher cosine similarity (CS) and smaller Euclidean distances (ED) among common vowels in both corpora. The anchor candidates include Best Anchors (BA), Worst Anchors (WA), and Group Anchors (GA), which represent the vowels with the best, worst, and group-wise closeness in terms of vowel feature similarity.

	Mono			Diph		
	BA	WA	GA	BA	WA	GA
neu	i	ə, u	i, ɑ/a, ɔ	oʊ	aɪ	oʊ, aʊ
hap	i	ɔ, u	i, ə, ɑ/a	aɪ	oʊ	eɪ, aɪ
ang	ɑ/a	ɔ, u	ɑ/a, ə, ɛ	eɪ	oʊ	eɪ, aʊ
sad	ɔ	ə, u	ɛ, ɔ, ɑ/a	aʊ	eɪ	aʊ, oʊ

Once the Anchor candidates have been selected, the subsequent step is to implement the vowel phonemes-based anchoring mechanism within the cross-lingual SER modeling framework.

6 Cross-Lingual SER Modelling

This study evaluates the performance of SER models that leverage vowel phoneme commonalities across languages. In Section 4.1, our analyses yield preliminary evidence suggesting that specific vowels exhibit phonetic commonalities following emotional modulation in both *American English* (MSP-P) and *Taiwanese Mandarin* (BIIC-P) languages. Motivated by these findings, we propose an anchoring mechanism to incorporate the phonetic constraint into our cross-lingual modeling approach, as illustrated in Fig. 4. Our cross-lingual SER framework consists of two branches: (1) the conventional emotion classification branch, which focuses on accurately classifying emotions, and (2) the phonetically-anchored domain adaptation branch, which integrates the phonetic anchor-based constraint in learning.

6.1 Emotion Classification Branch

We utilize a 768-dimensional Wav2Vec2.0 feature vector [48], derived from phoneme-level segmentation [58], as detailed in section 4.2. Subsequently, these acoustic features are fed

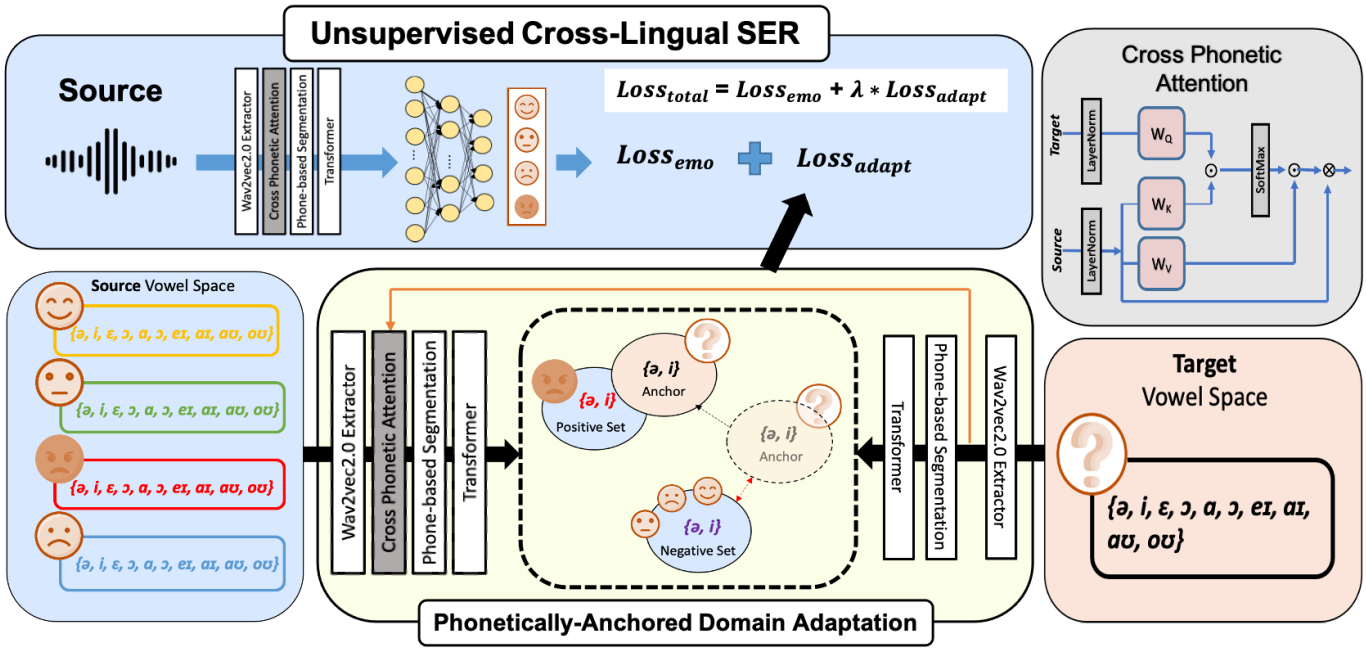


Fig. 4: Proposed phonetic-based unsupervised SER architecture.

into a transformer-based encoder network to produce self-attention hidden embeddings, serving as a conventional SER network in our proposed approach. Before performing phoneme-centric segmentation and passing the Wav2Vec2.0 features to our transformer-based network, we introduce a novel element in the model architecture: cross-phonetic attention. This addition enables information transfer from the target corpus to the source corpus before any processing of the Wav2vec2.0 features, enhancing the model's capabilities. We take it as a crucial step, particularly in the initial stage where the features primarily have phoneme-related information as it is extracted from the pre-trained ASR model (e.g., Wav2vec2.0). By applying cross-phonetic attention, we aim to enhance the alignment and transfer of phonetic knowledge from the target corpus to the source corpus. This similarity enables the model to benefit from the specific phonetic characteristics and nuances present in the target language, leading to improved performance in cross-lingual SER. Later these features are utilized in the domain adaptation branch. Here, Wav2Vec2.0 is used solely as a feature extractor without fine-tuning, while all other model parameters, including those in the transformer-based network, are updated during SER training. The classification loss is defined by Equation 3, and it serves as a means to optimize the model during training.

$$L_{emo} = \mathbb{E}_{X_S, y_S}[CE(T(X_S), y_S)] \tag{3}$$

where CE is the cross-entropy function, T is the transformer function, X_S is the source features, and y_S is the emotional labels.

6.2 Vowel-Based Anchoring Mechanism

The phonetic-anchored domain adaptation branch incorporates an anchoring mechanism to establish a connection between the two corpora by leveraging phonetic knowledge as a constraint. This constraint aims to exploit the similarity between the two languages for certain phonetic units, resulting in improved regularization. Our approach utilizes a triplet loss function for this purpose. Specifically, the vowel segments corresponding to specific emotions in the target domain are considered as $anchors$. The vowel segments from the source domain for the same set of vowels, but with the intent of transferring emotion-specific knowledge, are treated as $positives$. Conversely, the vowel segments from the source domain for the same set of vowels but with different emotions serve as $negatives$.

Using these $anchors$, $positives$, and $negatives$ samples, we calculate the triplet loss to match the source and target domain to integrate the vowel similarity as a constraint in cross-lingual SER learning. This adaptation loss is calculated using Equation 4,

$$\mathrm{L}_{adapt} = \sum_i^N [d(f(X_i^{t_{Gph}}), f(X_i^{s_{pGph}})) - d(f(X_i^{t_{Gph}}), f(X_i^{s_{nGph}})) + \alpha] \tag{4}$$

where d represents the Euclidean distance function, $f(X_i^{t_{ph}})$ is the feature representation for the target domain, and $f(X_i^{s_{pph}}))$ and $f(X_i^{s_{nph}})$ are the positive and negative feature representations of the source domain for the same vowel set, respectively. α represents the margin. The complete loss is calculated using Equation 5,

$$L_{total} = L_{emo} + \lambda * L_{adapt} \tag{5}$$

where L_{emo} and L_{adapt} are the losses for the emotion classification and domain adaptation tasks. Here, the parameter λ is set to the constant value of 0.5.

7 Experiment Results and Analyses

7.1 Experimental Settings

7.1.1 Parameters

To optimize the model, we utilize the Adam optimizer combined with a stochastic gradient descent algorithm. The

TABLE 4: The table presents the performance results (in terms of UAR, averaged over 10 runs) of the considered baseline models for both 4-category and binary emotion SER. It includes the statistical test over the baseline models and the proposed AGA-CL ("Our") model performances, denoted by asterisks (* for p $<$ 0.1, ** for p $<$ 0.05, *** for p $<$ 0.01).

		4-CAT	neu	hap	ang	sad
M→B	GAN [59]	56.91**	**67.65**	69.20**	66.75***	61.23***
	NMF [7]	55.68**	65.21	67.13**	65.34***	65.19***
	Ensemble [6]	53.90**	64.10*	67.00***	61.05***	65.77***
	Few-shot [8]	54.32***	65.26	58.92**	68.13***	67.26***
	Our	**58.14**	66.88	**71.08**	**74.63**	**69.29**
B→M	GAN [59]	53.33**	**62.31**	59.64***	61.22**	55.60**
	NMF [7]	51.15**	61.18	56.36**	56.77**	53.12***
	Ensemble [6]	52.18**	62.29	54.67***	54.90***	53.00***
	Few-shot [8]	53.62**	60.34	58.29***	58.61**	57.71***
	Our	**55.49**	61.13	**60.98**	**61.34**	**60.05**

learning rate is set to 0.0001, and a decaying factor of 0.001 is applied to enhance the training process. The models undergo training for a maximum of 70 epochs, utilizing a batch size of 64. Early stopping is implemented to prevent overfitting during training. The SER models are trained for both multi-class classification (using 4 primary emotion categories) and binary classification (for emotion-specific modeling). The cross-entropy loss function and the triplet loss function for anchoring mechanism implementation are employed as the cost function and the evaluation metric used is the unweighted average recall (UAR). To evaluate our proposed idea, we partition the given corpora into predefined train-validate-test splits for training, validating, and testing.

7.1.2 Experiment Models

Baseline Experiments: In our research, we evaluate the performance of our proposed approach by comparing it to several existing methods used as baselines. The first baseline method is *GAN*, proposed by Su et al. [59]. It utilizes adversarial learning and a multi-source GAN framework to transfer emotion-related information across different corpora. Another method based on Non-negative Matrix Factorization (NMF), is employed for transfer subspace learning in SER, as described by Luo et al. [7]. Additionally, we consider ensemble learning as presented in the work of Zehra et al. [6], which we refer to as *Ensemble*. It combines predictions from multiple models trained on diverse corpora to enhance recognition accuracy and robustness in cross-lingual scenarios. Moreover, we include the few-shot learning approach proposed by Ahn et al. [8] that addresses the limited availability of labeled data in the target corpora by leveraging knowledge from the source corpora and adapting the model to the target domain. We refer to this method as *Few-shot*. By comparing the performance of our proposed method against these baselines, we can thoroughly assess its effectiveness and potential for improving cross-corpus SER.

Ablation Experiments: To gain a comprehensive understanding of the proposed approach, the *AGA-CL* (Attention-based Group-Vowel-Anchored Cross-Lingual SER) model and its variants, we conduct a series of ablation experiments to analyze the effectiveness of different components. These experiments include the Group-Vowel-Anchored (*GA-CL*) approach, which utilizes a group of vowels (GA) that exhibits better behavior across the corpora, the Best-Vowel-Anchored (*BA-CL*) approach, which selects the most well-behaving vowel (BA) across the corpora, and the Worst-Vowel-Anchored (*WA-CL*) approach, which examines the performance using the least well-behaving vowel (BA) across the corpora. These ablation methods are applied to both monophthongs (Mono), diphthongs (Diph), and a combination of both (Both) cases. Through these experiments, we aim to assess the impact of different vowel selection strategies on the overall performance of the *AGA-CL* model in cross-lingual SER tasks.

7.2 SER Performance Comparison

7.2.1 Baseline Comparisons

Table 4 illustrates the results obtained from the SER models that are trained and tested using different combinations of corpora, considering different baseline models mentioned in Section 7.1.2. The table provides a comprehensive overview of the model performance (average over 10 runs) for each experimental setup. For instance, the M→B experiment represents the cross-lingual scenario where the model is trained on the MSP-P and tested on the BIIC-P, while B→M indicates the reverse setup with the BIIC-P as the source and MSP-P as the target. Upon analyzing the table, we can observe that our proposed model, *AGA-CL*, exhibits significant improvements compared to the baselines for both the 4-category (*4-CAT*) and binary (*neu*, *hap*, *ang*, *sad*) SER models. For *4-CAT*, *AGA-CL* achieves a higher UAR with an increment of 1.23% and 2.16% on the M→B and B→M tasks, respectively, outperforming the *GAN* model. Moreover, *AGA-CL* surpasses the *NMF* and *Ensemble* approaches, achieving improved UARs of 2.46% and 4.24% on the M→B task and 4.34% and 3.31% for the B→M task, respectively. Additionally, when compared to the *Few-shot* technique, *AGA-CL* achieves a UAR improvement of 3.82% and 1.87% on the M→B and B→M tasks, respectively. When comparing binary tasks, our method outperforms baselines. For instance, in the M→B task, it achieves superior results with 71.08%, 74.63%, and 69.29% for *hap*, *ang*, and *sad*. In the B→M task, it attains 60.98%, 61.34%, and 60.05% for the same emotions. For *neu*, although our model's performance is not the highest, it remains competitively strong, reaching 66.88% and 61.13% for the M→B and B→M tasks, respectively.

We conduct statistical tests to evaluate the significance of performance differences between our proposed model and the baseline models. Paired t-tests are computed for this purpose. Table 4 presents the results of statistical significance tests (*) denoting p-values (p $<$ 0.1, p $<$ 0.05, p $<$ 0.01) between baseline models and our proposed AGA-CL ("Our") model. Upon reviewing this table, we can say that the performance differences between most baseline models and our model are statistically significant. These results highlight the significant advancements and superior performance of our proposed approach compared to state-of-the-art (SOTA) techniques, confirming the effectiveness and potential of our *AGA-CL* model in cross-lingual SER tasks.

TABLE 5: The SER performances (in terms of UAR, averaged over 10 runs) of the proposed model and the other experimented models for both 4-category and binary emotion SER.

		M→B							B→M						
Models		4-CAT	neu	hap	ang	sad	aro	val	4-CAT	neu	hap	ang	sad	aro	val
M→M		61.89	78.56	82.98	77.91	71.64	64.54	48.22	57.33	68.13	71.07	69.83	67.11	61.33	45.85
CL		52.01	65.26	63.89	63.40	59.33	52.81	36.28	49.15	56.21	53.76	55.32	54.20	50.47	35.43
Mono	BA-CL	55.69	65.31	66.67	62.22	57.45	54.55	39.59	53.87	55.43	52.65	56.17	51.02	53.63	37.28
	WA-CL	51.06	58.42	55.38	59.12	53.54	52.75	36.35	49.45	50.33	46.83	54.21	49.91	50.15	35.27
	GA-CL [22]	57.33	67.13	70.67	69.83	67.11	55.36	40.74	53.01	59.22	60.15	59.67	59.82	54.29	38.94
	AGA-CL	57.91	68.56	70.98	69.36	65.25	56.39	41.74	54.58	58.6	52.42	60.91	58.95	56.41	39.97
Diph	BA-CL	52.01	63.72	60.83	61.29	58.55	53.32	40.86	50.25	55.78	54.65	55.47	50.31	53.82	37.74
	WA-CL	45.60	58.15	51.57	57.35	55.92	50.98	37.42	49.12	53.56	53.77	52.46	48.22	50.63	35.38
	GA-CL	59.93	60.21	61.33	59.97	55.23	54.39	40.56	51.54	58.49	57.09	56.13	52.5	55.79	37.64
	AGA-CL	55.10	59.62	62.08	60.33	56.71	55.33	41.96	52.09	60.16	57.39	57.00	53.36	56.73	37.85
Both	BA-CL	54.26	64.66	65.04	68.97	64.45	55.33	40.74	53.31	57.03	55.32	56.68	55.28	54.35	38.34
	WA-CL	49.61	59.84	59.55	60.62	58.13	50.47	37.93	49.53	51.24	48.85	50.34	48.56	49.21	35.42
	GA-CL	57.49	67.75	70.63	70.27	68.49	56.42	41.24	54.12	59.75	60.56	60.82	59.23	56.86	38.35
	AGA-CL	**58.14**	**66.88**	**71.08**	**74.63**	**69.29**	**57.25**	**42.33**	**55.49**	**61.13**	**60.98**	**61.34**	**60.05**	**57.47**	**40.31**

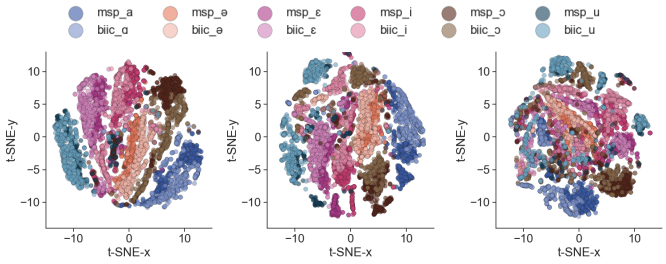
7.2.2 Proposed Architecture Evaluations

Table 5 provides a comprehensive overview of the performance of various variants of cross-lingual SER models, allowing for a thorough understanding of the proposed approach. The table includes a model without domain adaptation (*CL*). In contrast, *AGA-CL* approach, emerges as the top-performing model, achieving UARs of 58.14% and 55.49% on the M→B and B→M tasks, respectively. Among the unsupervised cross-lingual SER architecture considered in this work, the *AGA-CL* approach stands out, showcasing superior performance compared to the *CL* method with 6.23% and 6.34% for M→B and B→M tasks, respectively. Notably, the non-attention-based GA-CL model also exhibits absolute UAR gains of 5.48% and 4.67% on the M→B and B→M tasks, respectively. We also present the emotion-specific binary classifier results in Table 5, we can observe that all emotions categories *AGA-CL* has better results as compared to the *CL* model with UAR of 1.62%, 7.19% 11.23%, and 9.96% in the M→B task and UAR of 4.92%, 7.22% 6.02%, and 5.85% in the B→M task for *neutral*, *happiness*, *anger* and *sadness*, respectively.

Furthermore, we explore the phonetic anchoring approach using the best-vowel-anchored (*BA-CL*) and worst-vowel-anchored (*WA-CL*) models, which select a single vowel as the anchor. Table 5 specifies the performance for each model. The performances show that, for a 4-category SER, the *BA-CL* method has improvements over the *WA-CL* methods with 4.65% for the M→B task and 3.78% for the B→M task. The binary emotion SER results also reveal that the *BA-CL* model outperforms the *WA-CL* approaches, achieving improved UARs of 4.82%, 5.49%, 8.38%, and 6.32% on the M→B task and 5.89%, 6.47%, 6.34%, and 6.72% on the B→M task for *neutral*, *happiness*, *anger* and *sadness*, respectively. However, it is important to note that using a single vowel as the anchor (*BA-CL*) does not yield performance on par with the set of vowels utilized in our proposed model (*GA-CL* and *AGA-CL*). Also, when comparing our previous architecture *GA-CL* (proposed in our previous work [22]) with the current attention-based architecture *AGA-CL*, it is evident that *AGA-CL* demonstrates superior anchor utilization by consistently outperforming *GA-CL* in nearly all scenarios. This observation reinforces the idea that the incorporation of cross-phonetic attention significantly complements our mechanism. These findings validate the effectiveness of transfer learning based on the selected common phonetic anchors, which effectively incorporate essential information and facilitate language adaptation in cross-lingual SER tasks.

We extend our experiments to include *arousal* and *valence* dimensions. We categorize *arousal* and *valence* into three levels by dividing the 7-point Likert scale into low (1-2), medium (3-5), and high (6-7) categories. This approach captures calm/negative emotions, *neutral* states, and intense/positive emotions, ensuring a more interpretable structure for emotion classification while preserving the emotional range in the SER tasks. Table 5 presents the performance for *arousal* and *valence*. From Table 5, we observe a similar pattern in results for *arousal* and *valence* as seen with the emotional category results. For *arousal*, the *AGA-CL* model achieves the best performance (*Both*), with 57.25% for the M→B and 57.47% for B→M tasks. As for valence, which is more challenging to recognize, the *AGA-CL* model does not perform significantly better, but still delivers competitive results, achieving 42.33% in the M→B and 40.31% in the B→M tasks. The results demonstrate that our approach is effective not only for the emotional category task but also for other emotional attributes.

Upon observing the performance of our models in the only-*Mono* and only-*Diph* scenarios, as presented in Table 5, we notice that the models perform better in the only-*Mono* cases compared to the only-*Diph* cases. For instance, the *AGA-CL* models for *Mono* demonstrate an improvement of 2.81% in a 4-category SER task and 8.94% improvement in UAR for *neutral*, 8.90% for *happiness*, 9.03% for *anger*, and 8.54% for *sadness* in the binary SER models for the M→B train-test scenario. These observations are consistent with the results obtained in the B→M scenario as well. It indicates that the *Mono* case, being single phone vowels, exhibits more effective emotional transfers compared to the *Diph*, which involve a combination of two vowel sounds. Here, a model with "Both" *Mono* and *Diph* vowels gives better performance since it allows the model to capture a broader range of acoustic characteristics associated with different emotions.



(a) *happiness* (b) *anger* (c) *sadness*

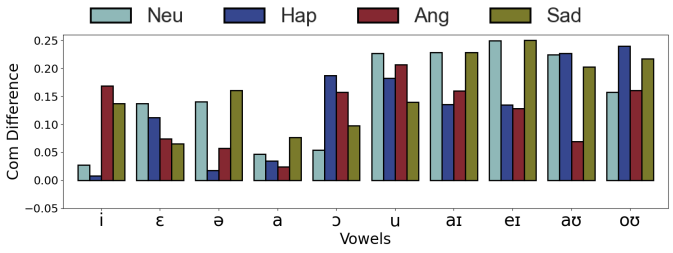
Fig. 5: t-SNE visualization depicting the feature representations of vowels coming from both corpora test sets from the proposed models (*AGA-CL*: with the anchoring mechanism) for M→B scenario for the considered emotions.

TABLE 6: Multidimensional distances, measured in *Combined score* (Com), within source and target vowel feature representations after the implementation of our proposed vowel anchoring mechanism.

		Mono						Diph			
		/i/	/ɛ/	/ə/	/ɑ,a/	/ɔ/	/u/	/aɪ/	/eɪ/	/aʊ/	/oʊ/
M→B	neu	**0.84**	0.81	0.76	0.83	0.83	0.81	0.80	**0.84**	0.83	0.82
	hap	0.80	0.80	0.79	**0.84**	0.81	0.79	0.83	0.83	**0.86**	0.81
	ang	0.77	0.78	**0.79**	0.76	0.76	**0.79**	0.77	**0.85**	0.77	0.78
	sad	0.79	0.77	0.77	0.76	0.83	0.74	0.80	0.84	0.83	0.81
B→M	neu	0.76	0.79	0.81	0.81	**0.84**	0.71	0.83	0.81	**0.85**	0.80
	hap	**0.85**	0.75	0.76	0.79	0.82	0.80	0.79	**0.83**	0.81	0.76
	ang	0.73	0.78	0.72	**0.80**	0.79	0.74	0.75	0.77	**0.82**	0.73
	sad	0.75	0.77	0.74	0.77	**0.78**	0.77	0.76	0.71	0.75	**0.79**

7.3 Before/After Anchoring Feature Space Analyses

In this study, we have chosen vowels as candidates for the anchoring mechanism based on their distances, enabling us to analyze vowel features and their changes in distances before and after the anchoring mechanism. Fig. 5 illustrates the t-SNE plots of vowel features extracted from specific-emotion binary SER AGA-CL-based models, focusing on *happiness, anger,* and *sadness* similar way as the one shown in Figure 3. The figure combines the vowel feature representations from the MSP-P and BIIC-P test sets within the same t-SNE plot corresponding to the M→B task. The overall analysis over these plots shows an enhanced closeness or overlap of vowel features between the two corpora for specific emotions. This overlapping tendency is particularly prominent in several vowels, indicating a better level of consistency among similar vowels in the MSP-P and BIIC-P corpora. Similar insights are observed in the B→M task as well. When we compare these findings with the t-SNE plots presented in Section 4.2, we can observe that, although initially only a few vowels are used as anchors, not only these anchor vowels become closer, but other vowels also exhibit some level of closeness. For instance, in the context of the *happiness*, it is not only /i/, /ə/, and /ɑ,a/ that display improved overlap, but also /ɛ/ and /u/. This trend of similar closeness among most vowel features is also observed in *neutral*, *anger*, and *sadness*. These findings highlight the effectiveness of the anchoring mechanism in promoting greater similarity and alignment of vowel features across languages.



(a) M→B

(b) B→M

Fig. 6: Variations in vowel feature proximity *before* and *after* applying the anchoring mechanism for both M→B and B→M scenarios, considering combined score (Com).

In addition to the visual insights provided by the t-SNE plots in Fig. 5, which includes multi-dimensional vowel-centered features, we conduct a comprehensive analysis of feature closeness and distinctness similar to the Table 2. Specifically, we compute multi-dimensional similarity for all pairs of vowel features extracted from the MSP-P and BIIC-P test samples. For example, we examined the similarity in the vowel /i/ feature from both the MSP-P and BIIC-P corpora. These estimated commonalities in terms of combined score (Com) as elaborated in Section 4, are detailed in Table 6 for both corpus test sets and across the M→B and B→M train-test scenarios. As shown in Table 6, we observe that several vowels exhibit higher Com, suggesting greater similarity between these vowels. For instance, in the case of *anger*, vowels such as /ə/ and /ɔ/ demonstrate significant closeness, with Com with 0.79 each, respectively. However, it is worth observing that other vowels also show closer relations, indicating a trend of enhanced feature similarity across a broader set of vowels.

Upon careful observation from Tables 6 and 2, the analysis reveals that the Com values in Table 6 indicate a trend where some anchored phonemes either shift slightly or remain unchanged after applying constraints. For instance, in the M→B task for *hap*, the unanchored phoneme /ɛ/ aligns more closely across languages, reflected in the increased Com value of 0.80, while the anchored phoneme /i/ shows no significant change. This observation may result from the inherent characteristics of the phonetic anchoring process. Anchored phonemes are selected for their perceived commonality across languages, which may not perfectly correspond to the phonetic distances computed by the model. Thus, while unanchored phonemes can adjust and align more freely, anchored phonemes may exhibit variations due to the constraints of the anchoring process. These findings emphasize the complexity of cross-lingual phonetic alignment in SER tasks, highlighting the importance of careful interpretation when evaluating phonetic

patterns across languages.

We further cross-reference this level of commonality with the information shown in Table 2 discussed in Section 4.2 to understand the evolution of vowel closeness before and after the implementation of our concept. Fig. 6 shows a comparative analysis with Table 2 and Table 6 across all analyzed emotions and vowels and illustrates the changes in the proximity of vowel features *before* and *after* the implementation of our concept. These plots reveal a reduction in distances and an enhancement in similarity for the majority of vowels during the emotion transfer by our proposed cross-lingual SER framework. This result indicates that the proposed anchoring approach successfully aligns and enhances the similarity of vowel features across languages.

7.4 Extended Analyses

To better understand the adaptability and performance consistency of our approach across different models, we conduct an extended analysis by replacing Wav2Vec2.0 with the HuBERT [60] model. Here, we aim to find out whether a different pretrained architecture has influence on cross-lingual SER performance or not. Table 7 presents the results of the AGA-CL model using HuBERT embeddings. The results shown in Table 7 demonstrate that HuBERT performs competitively with Wav2Vec2.0, achieving 57.98% of UAR for the M→B task and 55.53% of UAR for the B→M task in the 4-CAT SER. For binary SER models, HuBERT shows competitive performance compared to Wav2Vec2.0, especially for emotions such as *neutral* and *anger* in the B→M task, achieving UAR scores of 63.46% and 62.13% for *neu* and *ang* respectively. However, as seen in Table 7, Wav2Vec2.0 generally outperforms HuBERT in overall evaluation.

To evaluate our idea in a more linguistically diverse context, we include *Russian*, one of the most widely spoken language alongside *American English* and *Taiwanese Mandarin*. *Russian* offers unique linguistic and phonetic features, such as vowel inventory, stress patterns, and prosody. Including *Russian* helps address languages with distinct traits, thereby validating the generalizability of our approach. We use the Dusha [61] (a *Russian*-language speech emotion corpus) dataset to evaluate and explore our proposed idea's generalization capability. The Dusha dataset is one of the largest open bi-modal collections for SER, including around 350 hours and over 300,000 *Russian* language audio recordings and transcripts. It features balanced acted recordings and an unbalanced real-life subset of podcasts, annotated through crowdsourcing. For this work, we only choose the 2000 samples for the analyses, using it as test set to evaluate the model.

Table 8 presents the results for the M→D task (MSP-P as the source and Dusha as the target) and the B→D task (BIIC-P as the source and Dusha as the target). Additionally, the Table 8 also includes Dusha's performance under direct cross-test (CL) as the baseline, which does not involve any transfer learning. The results shown in Table 8 indicate an improved UAR for *AGA-CL* compared to *CL*, with increases of 8.87% and 7.20% for the 4-CAT performance in the M→D and B→D tasks, respectively. For binary SER tasks

TABLE 7: The table presents the performance results (in terms of UAR) of the AGA-CL model with Wav2vec2.0 and HuBERT feature extractors.

		4-CAT	neu	hap	ang	sad
M→B	HuBERT	57.98	**67.43**	70.37	73.91	68.03
	Wav2vec2.0	**58.14**	66.88	**71.08**	**74.63**	**69.29**
B→M	HuBERT	55.33	**63.46**	59.35	**62.13**	59.68
	Wav2vec2.0	**55.49**	61.13	**60.98**	61.34	**60.05**

TABLE 8: The table shows the performance (UAR) of the proposed AGA-CL models using the Dusha (D) dataset as the target corpus, with MSP-P (M) and BIIC-P (B) as source corpora. Also includes the performance of Dusha under direct cross-test (CL), without any transfer learning applied.

		4-CAT	neu	hap	ang	sad
M→D	CL	42.04	53.73	53.45	55.20	50.67
	AGA-CL	**50.91**	**63.02**	**62.39**	**64.56**	**58.34**
B→D	CL	41.74	**54.06**	52.45	53.38	49.21
	AGA-CL	48.94	61.48	60.67	61.01	55.53

performances shown in Table 8, the *AGA-CL* outperforms *CL* across all emotion categories under both M→D and B→D tasks. From Table 8, we can observe that these cross-lingual tasks yield better performance with M→D (achieving 50.91% UAR) over B→D (achieving 48.94% UAR) for the 4-CAT SER task. This observation from Table 8 suggests that *English* (MSP-P) serves as a better source for *Russian* (Dusha) compared to *Taiwanese Mandarin* (BIIC-P). This performance difference can be attributed to the fact that both *Russian* and *English* are Indo-European languages and share some phonetic and syntactic similarities. These results support our hypothesis that linguistic and phonetic similarities between languages should be considered, rather than relying solely on general phonetic-agnostic domain adaptation strategies.

8 Discussion and Conclusion

This research introduces a novel approach to unsupervised cross-lingual SER utilizing the concept of leveraging language phonetic universality through the use of a phonetic anchoring mechanism. The idea is based on initial evidence suggesting that certain vowels exhibit emotion-specific commonality across different languages. By analyzing the commonality of vowels including monophthongs and diphthongs, in their multidimensional feature representations, the study identifies specific vowels that can serve as phonetic anchors to control the variability between languages, thereby enhancing unsupervised cross-lingual SER learning. The proposed model, named *AGA-CL*, leverages attention-based phonetically-anchored language domain adaptation and achieves superior performance compared to baseline models. With a UAR of 58.14% and 55.59% for M→B and B→M SER tasks, the *AGA-CL* outperforms the considered baseline models. Results highlight the effectiveness of the anchoring mechanism in enhancing cross-lingual SER.

It is important to recognize the limitations of our work. In this study, we explore the linguistic diversity across three widely used languages: *American English*, *Taiwanese*

Mandarin, and *Russian*. However, despite this diversity, our study specifically focuses on examining vowels, which represent just one aspect of these complex language systems. Other factors, such as consonantal variation, prosodic nuances, phonemic symbols, speaking styles, and articulatory gestures, are equally critical in understanding the full scope of cross-lingual variability and commonality. Among these, articulatory gestures are particularly noteworthy as they capture the physical movements involved in speech production, offering a shared physiological basis across languages. While phonetic and acoustic properties can be more variable, examining the finite set of human articulators may reveal concrete, universal insights into language commonalities. Addressing these additional aspects will likely be necessary to fully capture the intricate nature of these language groups. In spite of these limitations, the results of this study provide empirical evidence that this approach can be quite effective in other languages.

Further future work encompasses several technical directions. Firstly, we intend to enhance the generalization of our proposed approach by integrating this innovative phonetic knowledge-driven anchoring mechanism with advanced domain adaptation techniques. This combination holds the potential for achieving superior performance and wider applicability. Secondly, our study aims to expand the scope of analysis beyond vowels and incorporate common consonants and articulatory gestures, thus enhancing the cross-lingual SER performances by leveraging a broader set of phonetic constraints. Thirdly, considering the widespread adoption of large pre-trained models, we plan to conduct in-depth analyses that involve other large pre-trained encoders to evaluate their performance with the proposed architecture.

ACKNOWLEDGEMENTS

This work was supported by the NSTC under Grants 110-2634-F-002-050 and 110-2221-E-007-067-MY3, and the NSF under Grant CNS-2016719.

REFERENCES

[1] Jaime Cesar Acosta, "Using emotion to gain rapport in a spoken dialog system," 2009.

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

[3] Laurence Devillers, Christophe Vaudable, and Clément 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.

[4] Chi-Chun Lee, Kusha Sridhar, Jeng-Lin Li, Wei-Cheng Lin, Bo-Hao Su, and Carlos Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, 2021.

[5] Siddique Latif, Junaid Qadir, and Muhammad Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *2019 8th international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2019, pp. 732–737.

[6] 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.

[7] Hui Luo and Jiqing Han, "Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2047–2060, 2020.

[8] Youngdo Ahn, Sung Joo Lee, and Jong Won Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[9] Michael Neumann et al., "Cross-lingual and multilingual speech emotion recognition on English and French," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5769–5773.

[10] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Bjorn Wolfgang Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[11] Kristen A Lindquist, Lisa Feldman Barrett, Eliza Bliss-Moreau, and James A Russell, "Language and the perception of emotion.," *Emotion*, vol. 6, no. 1, pp. 125, 2006.

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

[13] 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.

[14] 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.

[15] Bo-Hao Su and Chi-Chun 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.

[16] Md Asif Jalal, Rosanna Milner, and Thomas Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition," in *Proceedings of Interspeech 2020*. International Speech Communication Association (ISCA), 2020, pp. 4113–4117.

[17] Jiahong Yuan, Xingyu Cai, Renjie Zheng, Liang Huang, and Kenneth Church, "The role of phonetic units in speech emotion recognition," *arXiv preprint arXiv:2108.01132*, 2021.

[18] 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.

[19] Christine SP Yu, Michael K McBeath, and Arthur M Glenberg, "Phonemes convey embodied emotion," in *Handbook of Embodied Psychology*, pp. 221–243. Springer, 2021.

[20] Anita Körner and Ralf Rummer, "Articulation contributes to valence sound symbolism.," *Journal of Experimental Psychology: General*, vol. 151, no. 5, pp. 1107, 2022.

[21] Anita Körner and Ralf Rummer, "What is preferred in the in–out effect: articulation locations or articulation movement direction?," *Cognition and Emotion*, vol. 36, no. 2, pp. 230–239, 2022.

[22] 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.

[23] Jaebok Kim, Gwenn Englebienne, Khiet P Truong, and Vanessa Evers, "Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning," *arXiv preprint arXiv:1708.03920*, 2017.

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

[25] Jin Li, Nan Yan, and Lan Wang, "Unsupervised cross-lingual speech emotion recognition using pseudo multilabel," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 366–373.

[26] Mateusz Ochal, Jose Vazquez, Yvan Petillot, and Sen Wang, "A comparison of few-shot learning methods for underwater optical

and sonar image classification," in *Global Oceans 2020: Singapore–US Gulf Coast*. IEEE, 2020, pp. 1–10.

[27] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[28] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh, "The limitations of adversarial training and the blind-spot attack," *arXiv preprint arXiv:1901.04684*, 2019.

[29] David Opitz and Richard Maclin, "Popular ensemble methods: An empirical study," *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.

[30] Daniel N. McIntosh RB Zajonc Peter S. Vig Stephen W. Emerick, "Facial movement, breathing, temperature, and affect: Implications of the vascular theory of emotional efference," *Cognition & Emotion*, vol. 11, no. 2, pp. 171–196, 1997.

[31] Asifa Majid, "Current emotion research in the language sciences," *Emotion Review*, vol. 4, no. 4, pp. 432–443, 2012.

[32] Noriko Iwasaki, David P Vinson, and Gabriella Vigliocco, "What do english speakers know about gera-gera and yota-yota?: A cross-linguistic investigation of mimetic words for laughing and walking," *Japanese-language education around the globe*, vol. 17, pp. 53–78, 2007.

[33] Bogdan Vlasenko, David Philippou-Hübner, Dmytro Prylipko, Ronald Böck, Ingo Siegert, and Andreas Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal emotions," in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–6.

[34] Ralf Rummer, Judith Schweppe, René Schlegelmilch, and Martine Grice, "Mood is linked to vowel type: The role of articulatory movements.," *Emotion*, vol. 14, no. 2, pp. 246, 2014.

[35] Jan Auracher, Winfried Menninghaus, and Mathias Scharinger, "Sound predicts meaning: Cross-modal associations between formant frequency and emotional tone in stanzas," *Cognitive Science*, vol. 44, no. 10, pp. e12906, 2020.

[36] Blake Myers-Schulz, Maia Pujara, Richard C Wolf, and Michael Koenigs, "Inherent emotional quality of human speech sounds," *Cognition & emotion*, vol. 27, no. 6, pp. 1105–1113, 2013.

[37] Aijun Li, Qiang Fang, Fang Hu, Lu Zheng, Hong Wang, and Jianwu Dang, "Acoustic and articulatory analysis on mandarin chinese vowels in emotional speech," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 38–43.

[38] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan, "Emotion recognition based on phoneme classes," in *Eighth international conference on spoken language processing*, 2004.

[39] Fabien Ringeval and Mohamed Chetouani, "Exploiting a vowel based approach for acted emotion recognition," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction: COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007. Revised Papers*. Springer, 2008, pp. 243–254.

[40] Mohit Shah, Ming Tu, Visar Berisha, Chaitali Chakrabarti, and Andreas Spanias, "Articulation constrained learning with application to speech emotion recognition," *EURASIP journal on audio, speech, and music processing*, vol. 2019, pp. 1–17, 2019.

[41] Damián E Blasi, Søren Wichmann, Harald Hammarström, Peter F Stadler, and Morten H Christiansen, "Sound–meaning association biases evidenced across thousands of languages," *Proceedings of the National Academy of Sciences*, vol. 113, no. 39, pp. 10818–10823, 2016.

[42] Anita Körner and Ralf Rummer, "Valence sound symbolism across language families: a comparison between japanese and german," *Language and Cognition*, vol. 15, no. 2, pp. 337–354, 2023.

[43] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[44] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 498–502.

[45] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999.

[46] Lloyd Rice, "Hardware & software for speech synthesis," *Dr. Dobb's Journal of Computer Calisthenics & Orthodontia*, vol. 1, no. 4, pp. 6–8, April 1976.

[47] Yuan-Fu Liao, Wu-Hua Hsu, Yu-Chen Lin, Yung-Hsiang Shawn Chang, Matús Pleva, Jozef Juhar, and Guang-Feng Deng, "Formosa speech recognition challenge 2018: data, plan and baselines," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 270–274.

[48] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[49] Cynthia G Clopper, David B Pisoni, and Kenneth De Jong, "Acoustic characteristics of the vowel systems of six regional varieties of American English," *The Journal of the Acoustical society of America*, vol. 118, no. 3, pp. 1661–1676, 2005.

[50] Gordon E Peterson and Harold L Barney, "Control methods used in a study of the vowels," *The Journal of the acoustical society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[51] Hanjun Liu and Manwa L Ng, "Formant characteristics of vowels produced by Mandarin esophageal speakers," *Journal of voice*, vol. 23, no. 2, pp. 255–260, 2009.

[52] James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[53] T.M. Nearey, *Phonetic Feature Systems for Vowels*, Indiana University (Bloomington). Linguistics Club. (Bd 224). Indiana University Linguistics Club, 1978.

[54] Teena tom Dieck, Paula-Andrea Pérez-Toro, Tomás Arias-Vergara, Elmar Nöth, and Philipp Klumpp, "Wav2vec behind the scenes: How end2end models learn phonetics," *Proc. Interspeech 2022*, pp. 5130–5134, 2022.

[55] Kwanghee Choi and Eun Jung Yeo, "Opening the black box of wav2vec feature encoder," *arXiv preprint arXiv:2210.15386*, 2022.

[56] Patrick Cormac English, John Kelleher, and Julie Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2022, pp. 83–91.

[57] Ankita Pasad, Bowen Shi, and Karen Livescu, "Comparative layer-wise analysis of self-supervised speech models," *arXiv preprint arXiv:2211.03929*, 2022.

[58] Wei-Cheng Lin and Carlos Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, 2021.

[59] Bo-Hao Su and Chi-Chun Lee, "Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-gan," *IEEE Transactions on Affective Computing*, 2022.

[60] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[61] Vladimir Kondratenko, Artem Sokolov, Nikolay Karpov, Oleg Kutuzov, Nikita Savushkin, and Fyodor Minkin, "Large raw emotional dataset with aggregation mechanism," *arXiv preprint arXiv:2212.12266*, 2022.

Shreya G. Upadhyay (S'23) is currently pursuing a PhD degree in Electrical Engineering at National Tsing Hua University (NTHU), Taiwan. She obtained her BE degree in computer engineering from Mumbai University, India in 2013, and her MTech degree in computer engineering from K. J. Somaiya College of Engineering, India in 2018. Her research interests include behavioral speech signal processing, speech emotion recognition, automatic speech recognition, and acoustic sound event detection. She is a student member of IEEE, ISCA, EURASIP, AAAC, and SPS.

Luz Martinez-Lucas (S'21) is a PhD Student in the Electrical and Computer Engineering Department at the University of Texas at Dallas (UTD). She did her Bachelor's in Electrical Engineering at UTD. Her research interests include affective computing, speech technology, and machine learning. She is a student member of IEEE, AAAC, and SIAM.

William Katz (SM '23) is a Professor Emeritus at the University of Texas at Dallas, (UTD), Department of Speech Language Pathology and Audiology. He received his B.A. in Biology from the University of California at Santa Cruz (1977) and his M.A. and Ph.D. degrees in Linguistics from Brown University, in 1985 and 1987, respectively. His research interests include speech and language in healthy and disordered individuals and the role of visual feedback in speech motor learning. He was a Fulbright Fellow at the University of Konstanz, Germany (1987-1988) and a visiting scholar at the University of Goettingen, Germany (1987-1988). He completed NIH-funded post-doctoral research in child language at the University of California San Diego (1986-1990), then came to UTD (1990), where he established the UTD Speech Production Laboratory. Over a 30 year period, this group has contributed numerous studies of prosodic processing as well as new methods for treating speech disorders after brain damage. He was a UTD Callier Research Scholar (2006, 2007) and is a Fellow of the Academy of Aphasia. He is a member of ISCA and a senior member of the American Speech and Hearing Association and the Acoustical Society of America.

Carlos Busso (S'02-M'09-SM'13-F'23) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. Carlos Busso is a Professor at Language Technologies Institute, Carnegie Mellon University, where he is also the director of the Multimodal Speech Processing (MSP) Laboratory. His research interest is in human-centered multimodal machine intelligence and application, focusing on the broad areas of speech processing, affective computing, and machine learning methods for multimodal processing. He has worked on speech emotion recognition, multimodal behavior modeling for socially interactive agents, in-vehicle active safety systems, and robust multimodal speech processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. His students received the third prize IEEE ITSS Best Dissertation Award (N. Li) in 2015, and the AAAC Student Dissertation Award (W.-C. Lin) in 2024. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). In 2023, he received the Distinguished Alumni Award in the Mid-Career/Academia category by the Signal and Image Processing Institute (SIPI) at the University of Southern California. He received the 2023 ACM ICMI Community Service Award. He is currently an associate editor of the IEEE Transactions on Affective Computing. He is a member of AAAC and a senior member of ACM. He is an IEEE Fellow and an ISCA Fellow.

Chi-Chun Lee (M'13, SM'20) is a Professor at the Department of Electrical Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degrees both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective computing, health analytics, and behavioral signal processing. He is an associate editor for the IEEE Transaction on Affective Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), the Journal of Computer Speech and Language (2021-), the APSIPA Transactions on Signal and Information Processing, and a TPC member for APSIPA IVM and MLDA committee. He serves as the general chair for ASRU 2023, an area chair for Interspeech 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, late breaking result chair for ACM ICMI 2023, sponsorship and special session chair for ISCSLP 2018, 2020. He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the NTHU Industry Collaboration Excellence Award (2021), and the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to 1st place in the Emotion Challenge in Interspeech 2009, and with his students won 1st place in the Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a co-author on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is also an ACM and ISCA member.