Analysis of Phonetic Level Similarities Across Languages in Emotional Speech

Pravin Mote[1,2], *Abinay Reddy Naini*[1,2], *Donita Robinson*[3], *Elizabeth Richerson*[3], *Carlos Busso*[1]

[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Electrical and Computer Engineering, The University of Texas at Dallas, USA
[3]Laboratory for Analytic Sciences, North Carolina State University, USA

{pmote,anaini}@andrew.cmu.edu, {drobins7,ericher}@ncsu.edu, busso@cmu.edu

Abstract

Unsupervised domain adaptation offers significant potential for cross-lingual *speech emotion recognition* (SER). Most relevant studies have addressed this problem as a domain mismatch without considering phonetical emotional differences across languages. Our study explores universal discrete speech units obtained with vector quantization of wavLM representations from emotional speech in English, Taiwanese Mandarin, and Russian. We estimate cluster-wise distributions of quantized wavLM frames to quantify phonetic commonalities and differences across languages, vowels, and emotions. Our findings indicate that certain emotion-specific phonemes exhibit cross-linguistic similarities. The distribution of vowels varies with emotional content. Certain vowels across languages show close distributional proximity, offering anchor points for cross-lingual domain adaptation. We also propose and validate a method to quantify phoneme distribution similarities across languages.
Index Terms: Speech Emotion Recognition, Discrete Speech Units, Cross-Lingual Emotion Analysis

1. Introduction

Important progress has been made in *speech emotion recognition* (SER) over the last years [1–4]. There is a growing need to develop a generalized SER system capable of handling various languages and cultural nuances [5]. This task is challenging due to the complex nature of emotion, cultural differences in the externalization of emotions, and phonetic variability across languages. As globalization advances, *human-computer interaction* (HCI) [6] is extending beyond language barriers. Cross-lingual SER performance, as extensively studied by Feraru et al. [7], declines sharply when the languages are dissimilar, particularly if they belong to different language families. Therefore, it is essential to address the variations in emotional expression that arise from different linguistic and cultural factors [8] to enhance the SER generalization capabilities across languages.

Prominent labeled datasets for SER are available in languages such as English [9, 10], Taiwanese Mandarin [11], and Russian [12], but most languages still lack reliable resources for SER. Languages with low resources will require effective transfer learning strategies. Some adaptation strategies have led to improvements, such as training models on multiple datasets to capture diverse data distributions [13, 14] or placing greater emphasis on selective crucial data points [15]. However, these supervised approaches often fall short due to the scarcity of labeled data and the high costs associated with data annotation. General unsupervised domain adaptation strategies are effective. For example, studies have used the ladder network strategy [16–19], which emphasizes reconstruction tasks as auxiliary tasks to reduce mismatches between source and target domains, or adversarial domain adaptation [20,21], which focuses on creating shared discriminative representations where source and target domains are indistinguishable. However, we hypothesize that cross-lingual emotion recognition strategies should consider phonetic and linguistic similarities between languages.

Cross-lingual SER faces challenges due to inconsistencies in prosody, phonetic structure, and syntactic patterns across languages when expressing emotions. The differences can also include mismatches in the distribution of emotions [22]. Several studies have explored linguistic properties. For instance, Pasad et al. [23] conducted a layer-wise investigation into semantic, syntactic, and morphological properties of *self-supervised learning* (SSL) models. Choi et al. [24] suggested that features extracted from various SSL models exhibit stronger phonetic alignment than semantic alignment, indicating that features representing similar phonemes are closer in the feature space compared to those representing synonymous words. These studies suggest that phonetic-level variations serve as valuable indicators of linguistic correlations, aiding in designing effective cross-lingual unsupervised adaptation strategies. For instance, Upadhyay et al. [25–27] leverage phonetic commonality as an anchor point in transfer learning to enhance cross-lingual SER. These studies demonstrates the need to understand and leverage strategies to compare the acoustic differences across languages.

This study proposes a methodology to compare emotional differences at the phoneme level across languages. Our strategy calculates acoustic distributions by clustering the wavLM embedding space generated from a merged set of databases in English, Taiwanese Mandarin, and Russian. Then, we compare the vowel- and emotion-level distributions, analyzing the variation of phonemes across emotions and languages. We employed the *Kullback–Leibler* (KL) divergence to quantify differences in phonetic distributions across emotions and languages. This strategy is ideal for investigating the similarities between languages in the expression of emotions. Our findings demonstrate that certain vowels exhibit common distributions across languages. Furthermore, the cross-lingual distributional gap of vowels varies depending on the emotion expressed. Additionally, we utilized the Kendall-Tau rank correlation coefficient to demonstrate the consistency of this quantification strategy. This study opens research opportunities to design cross-lingual strategies for SER that are grounded in the phonetical content of the target languages.

2. Cross-lingual Datasets

We conducted experiments on three datasets with a focus on English, Taiwanese Mandarin, and Russian. The MSP-Podcast corpus (v1.11) [9] is an emotional dataset of natural speech in the wild derived from American English podcasts. The database

Table 1: *Set used to create the discrete speech units using balanced contributions of the three speech emotional databases.*

Dataset	Sentences	Frames
MSP-Podcast	10,000	945,383
BIIC-Podcast	10,000	923,265
Dusha	10,000	813,796
Total	30,000	2,682,444

includes 238 hours of emotional speech. The BIIC-Podcast corpus (v1.0) [11] is another in-the-wild emotional dataset sourced from Taiwanese Mandarin podcasts. The corpus consists of 157 hours of emotional speech. The third database is the Dusha corpus [28]. We used the crowd subset of the Dusha dataset, which comprises 255 hours of emotional Russian speech recordings. Unlike the MSP-Podcast and BIIC-Podcast datasets, the Dusha corpus features acted emotions. These three datasets differ significantly in the linguistic structures, including phonetics and syntax, and the cultural nuances of emotional expression, making them an ideal choice for cross-lingual analysis. Since the Dusha dataset provides categorical emotion labels, experiments are conducted using four emotion classes: Happiness, Sadness, Anger, and Neutral state.

Our analysis requires phonetic information. We employed the *Montreal Forced Aligner* (MFA) [29] to generate phoneme boundary alignments for the MSP-Podcast and Dusha datasets. MFA provides phonetic transcriptions in ARPABET notation, which we translated into the *international phonetic alphabet* (IPA) notation using mappings from Lloyd [30], for the MSP-Podcast corpus, and Trofimov and Jones [31], for the Dusha corpus. For the BIIC-Podcast, we utilized a forced aligner trained on the Formosa database [32]. We convert the phonetic transcriptions to IPA notation using the mapping in Liao et al. [32].

3. Methodology

This section outlines our methodology for creating discrete speech units to analyze phonetic-level differences in the expression of emotion across languages. It describes the comparative approach employed to assess the vowel distribution across languages and evaluate their differences. In this analysis, we focus on the following subset of vowels represented in IPA notation for comparison: [a, e, i, o, u, ə]. For the emotion-wise analysis, we focused on four specific emotion categories: happiness, sadness, anger, and neutral state.

3.1. Discrete Speech Units using Vector Quantization

The first step in our approach is to transform continuous speech SSL frames into discrete speech units. The goal is to establish a common, unified framework that integrates all the acoustic and emotional variability observed across the target languages. Then, we will estimate distributions to quantify the similarities and differences across phonetic units, emotions and languages.

We rely on wavLM Large SSL [33], although we found similar observations when the analysis was implemented with Whisper [34]. We utilized a 1,024-dimensional wavLM representation [33], sourced from the HuggingFace library, and fine-tuned the wavLM Large model specifically for SER as the downstream task using the MSP-Podcast dataset [9]. The WavLM frames are estimated at 50 frames per second (i.e., 20ms per frame) for audio sampled at 16 kHz. To obtain consistency across emotions, we sub-sample all three datasets to match inter-category distributions by selecting an equal number of sentences per emotion. Table 1 reports the contribution of sentences and frames used from each dataset. The unified space comprises approximately 2.6 million frames.

We obtain discrete speech units using the *k*-means clustering algorithm, which is applied to the unified feature space to partition the acoustic space into a predetermined number of bins (implemented using 1,024). We fix the centroids and cluster boundaries obtained from vector quantization, using these discrete speech units consistently in the rest of the analysis.

3.2. Quantifying Acoustic Similarity

Using the phonetic alignment boundaries (Sec. 2), frames containing specific vowels are projected onto this clustered unified space. The distribution of each vowel is determined by counting the number of frames within each cluster. By using a common unified space for all distributions, meaningful comparisons can be made across various phonetic units, languages, and emotions.

We evaluate the similarity between two distributions using the *Kullback-Leibler* (KL) divergence, quantifying the statistical distance between two distributions. Lower KL divergence values indicate greater similarity between the distributions. Given that the KL divergence is asymmetric, we compute it in both directions, and the results are subsequently averaged:

$$d(A,B) = \frac{KL(A,B) + KL(B,A)}{2} \tag{1}$$

where A and B are two distributions.

To validate the KL-divergence values obtained from our initial bin configuration, we replicated the experiment using different numbers of bins. Since direct comparisons of KL-divergence values across different bin sizes may not be meaningful due to scaling differences, we focus on comparing the relative rankings across different vowels and emotions between two specific settings. For this purpose, we assess the consistency of the KL-divergence values across these bin configurations by employing the *Kendall-Tau* (KT) rank correlation coefficient, which measures the similarity in the ordering of two lists. The KT coefficient ranges from -1 to 1, where -1 indicates a perfect inverse correlation, 0 denotes no correlation, and 1 signifies a perfect correlation in the order of the ranks. If the relative trends are preserved with two different settings (i.e., similar ranking across either vowels or emotions), we can conclude that our strategy to estimate similarity in the acoustic space is consistent.

4. Emotion-wise Phonetic Analysis

The proposed methodology allows for flexible projection of any subset of dataset to measure the distribution to assess acoustic similarity. This section presents a comparative analysis of these distributions, examining languages, vowels, and emotions.

4.1. Acoustic Analysis for wavLM Discrete Speech Units

Figure 1 illustrates the distribution of the wavLM-based discrete speech units for each language using 2,000 clusters. In this figure, the distribution of the entire MSP-Podcast dataset serves as the reference to facilitate comparisons. The 2,000 bins of the wavLM space are sorted in descending order based on the reference frame counts within those clusters. The distribution of each language is then plotted while maintaining this sorted order to visualize the differences for Taiwanese Mandarin and

Table 2: *KL-divergence values for comparisons between the MSP-Podcast, BIIC-Podcast, and Dusha datasets for 2,000 bins in the wavLM-based discrete speech unit space. The comparisons consider the corresponding distributions for given phonetical and emotional classes across two languages. The column 'all-emo' and row 'all-vowels' gives the results across all emotions and vowels, respectively.*

Phonemes	MSP-Podcast vs BIIC-Podcast					BIIC-Podcast vs Dusha					Dusha vs MSP-Podcast				
	Happiness	Neutral	Sadness	Anger	all-emo	Happiness	Neutral	Sadness	Anger	all-emo	Happiness	Neutral	Sadness	Anger	all-emo
a	0.49	0.51	0.64	1.11	0.48	0.66	0.90	1.02	1.47	0.62	0.60	0.95	0.89	1.05	0.45
o	0.71	0.78	0.87	1.31	0.63	1.29	1.55	1.59	1.88	0.91	1.25	1.48	1.62	1.81	0.77
u	3.41	3.77	3.58	2.64	2.04	1.52	1.73	1.89	1.91	0.91	1.55	2.19	1.91	1.71	0.91
i	0.57	0.59	0.61	0.98	0.42	0.60	0.71	0.84	1.09	0.46	0.83	1.12	1.00	1.06	0.63
e	0.77	0.89	0.91	1.43	0.64	1.29	1.54	1.85	2.29	0.93	1.82	2.28	2.65	2.52	1.42
ə	0.42	0.42	0.49	0.84	0.41	0.78	0.84	1.02	1.40	0.57	0.69	0.78	0.84	0.85	0.40
all-vowels	0.22	0.21	0.30	0.57	1.44	0.33	0.39	0.55	0.72	4.01	0.23	0.34	0.39	0.39	2.44

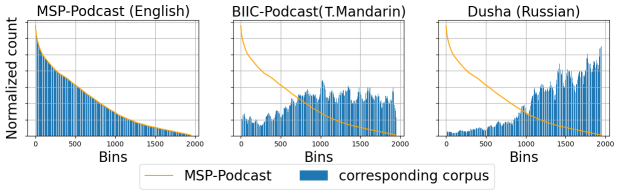

Figure 1: *Language-wise distribution of wavLM discrete speech unit frame assignments in comparison to the distribution assignments for the MSP-Podcast corpus.*

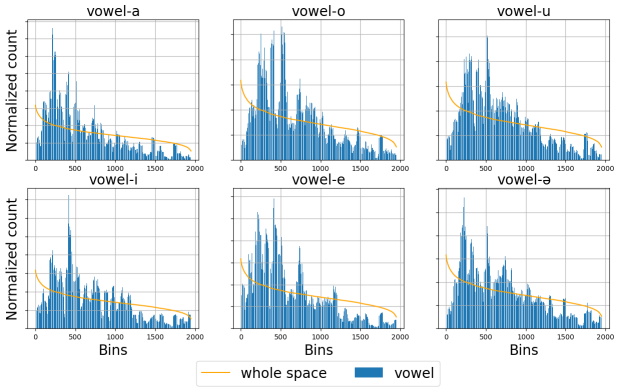

Figure 2: *Vowel-wise distribution across the entire wavLM discrete speech unit space. The three databases are combined for this graph. The reference is the assignments for all frames.*

Russian. Since the MSP-Podcast corpus serves as the reference, the left plot in Figure 1 shows that the distribution of the MSP-Podcast corpus precisely fits the orange line, which represents the reference distribution. The remaining two plots in Figure 1 illustrate that the distributions of the acoustic space for the BIIC-Podcast and Dusha corpora are markedly different from the acoustic space in the MSP-Podcast corpus. Clusters densely populated in the MSP-Podcast corpus frames contain relatively fewer frames from the other two datasets, indicating significant linguistic differences among the three languages.

Figure 2 presents the vowel-wise distribution relative to the entire wavLM feature space (orange line). The distributions of both vowels and the complete wavLM space are derived by merging all three datasets. The bins with the highest density vary for each vowel, indicating that vowels are concentrated around different areas within the feature space. However, these vowel clusters are not distinctly separated, suggesting that the

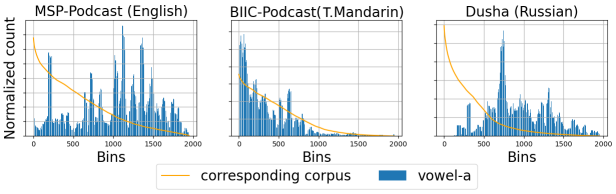

Figure 3: *Within corpus analysis of wavLM discrete speech unit distributions for vowel* a*. The reference (orange line) is the distribution for all frames from the corresponding corpus.*

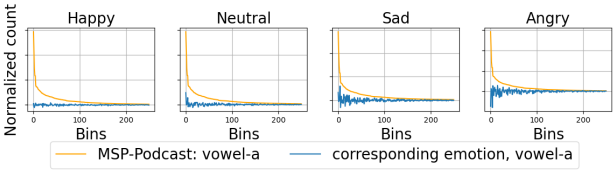

Figure 4: *Emotional variations in the wavLM discrete speech unit distributions for vowel* a *in the MSP-Podcast corpus. The reference (orange line) is the distribution for vowel* a *across emotions. Differently from Figures 1-3, the blue line gives the bin-wise differences in the distribution of vowel* a *for each specific emotion to capture the subtle differences.*

vowel centers are dispersed across languages.

As an example, Figure 3 illustrates the distribution of the vowel *a* across the three languages. The presence of highly dense bins that are distantly located indicates that vowel-a is distinctly clustered and separated across languages. We observe similar findings for other vowels. This observed disparity in distributions is pivotal for developing adaptation strategies aimed at bridging inter-lingual differences.

Figure 4 shows the variation in the distribution of vowel *a* within the MSP-Podcast for each emotion, compared to the overall distribution of vowel *a* across emotions (orange line). The orange line converges to zero within 250 bins, indicating that vowel *a* is tightly clustered in the wavLM feature space. The blue line represents the bin-wise differences in the distribution of vowel *a* for each specific emotion. Emotion-specific instances of vowel *a* are clustered in the proximity to each other. There are some changes, but they are small.

4.2. Distances across Vowels, Emotions and Languages

As outlined in Section 3.2, we employ the KL divergence to quantify the similarity between distributions. Table 2 reports the KL divergence values across pairs of languages (MSP-Podcast,

Table 3: *Consistency trend analysis for the MSP-Podcast and BIIC-Podcast comparisons using 500, 2k, and 4K bins (English versus Taiwanese Mandarin). Left table: KT-coefficients for vowel trends. Right table: KT-coefficients for emotional trends.*

	500-2k	2k-4k	4k-500
H	0.81	0.90	0.90
N	0.90	1.00	0.90
S	0.90	0.90	0.81
A	1.00	0.90	0.90
all	0.81	1.00	0.81
avg	0.88	0.94	0.86

	500-2k	2k-4k	4k-500
a	0.40	1.00	0.40
o	0.40	1.00	0.40
u	0.80	0.80	0.60
i	1.00	0.80	0.80
e	0.40	1.00	0.40
ə	0.40	0.80	0.60
all	1.00	1.00	1.00
avg	0.63	0.91	0.60

Table 4: *Consistency trend analysis for the BIIC-Podcast and Dusha comparisons using 500, 2k, and 4K bins (Taiwanese Mandarin versus Russian). Left table: KT-coefficients for vowel trends. Right table: KT-coefficients for emotional trends.*

	500-2k	2k-4k	4k-500
H	0.90	0.90	0.81
N	1.00	1.00	1.00
S	0.90	1.00	0.90
A	1.00	0.90	0.90
all	0.81	1.00	0.81
avg	0.92	0.96	0.88

	500-2k	2k-4k	4k-500
a	0.80	1.00	0.80
o	1.00	1.00	1.00
u	0.80	0.60	0.80
i	1.00	1.00	1.00
e	0.40	1.00	0.40
ə	0.80	1.00	0.80
all	1.00	1.00	1.00
avg	0.83	0.94	0.83

Table 5: *Consistency trend analysis for the Dusha and MSP-Podcast comparisons using 500, 2k, and 4K bins (Russian versus English). Left table: KT-coefficients for vowel trends. Right table: KT-coefficients for emotional trends.*

	500-2k	2k-4k	4k-500
H	0.71	1.00	0.71
N	0.90	1.00	0.90
S	0.90	1.00	0.90
A	0.90	1.00	0.90
all	0.81	0.90	0.90
avg	0.84	0.98	0.86

	500-2k	2k-4k	4k-500
a	0.80	0.80	0.60
o	1.00	1.00	1.00
u	1.00	0.80	0.80
i	0.80	1.00	0.80
e	0.80	0.60	0.80
ə	0.80	1.00	0.80
all	1.00	1.00	1.00
avg	0.89	0.89	0.83

BIIC-Podcast, and Dusha), categorized by vowel and emotion. The last column, labeled 'all-emo', corresponds to the KL divergence values for each vowel, aggregated across all emotions. Similarly, the last row, labeled 'all-vowels', provides the KL divergence values for each emotion, aggregated across all vowels.

The KL divergence values between the MSP-Podcast and BIIC-Podcast corpora (English versus Taiwanese Mandarin), presented in Table 2, indicate that vowels *i* and ə exhibit the highest similarity across the 'all-emo' column. This observation aligns with the findings of Upadhyay et al. [25], where phonemes were analyzed within a vowel space defined by the first two formants. Moreover, vowel *i* shows the closest similarity between the BIIC-Podcast and Dusha databases (Taiwanese Mandarin versus Russian), while vowels *a* and ə are the most similar between the Dusha and MSP-Podcast corpora (Russian versus English). These variations highlight inter-lingual differences, which can be leveraged to learn discriminative features for cross-lingual adaptation.

Phonetic similarity patterns vary across emotions for different language pairs. In the MSP-Podcast and BIIC-Podcast pair (English versus Taiwanese Mandarin), happiness and neutral state exhibit the highest similarity for 'all-vowels', particularly *a* and ə, while sadness and anger show the greatest similarity for vowels *i* and ə. For the BIIC-Podcast and Dusha pair (Taiwanese Mandarin versus Russian), vowels *a* and *i* are more similar for happiness, whereas *i* and ə show higher similarity in anger and neutral state. For sadness, *i* is the most similar vowel. In the Dusha and MSP-Podcast pair (Russian versus English), vowel *a* exhibits the highest similarity for happiness, while ə remains the most similar vowel for sadness, anger and neutral.

Overall, the MSP-Podcast and BIIC-Podcast (English versus Taiwanese Mandarin) pair exhibits greater similarity compared to the other two language pairs, with ə being the most similar vowel and *u* the most dissimilar vowel. The phoneme distribution for neutral state shows the closest alignment. In contrast, the BIIC-Podcast and Dusha (Taiwanese Mandarin versus Russian) pair reveals the highest degree of dissimilarity, with *i* being the most similar vowel and *e* the most dissimilar vowel. The phoneme distribution for happiness is the most aligned vowel for these languages.

4.3. Consistency across Different Number of Clusters

To examine the consistency in the trends of the KL divergence values obtained using 2,000 bins, we replicated the experiment with 500 and 4,000 bins. We aim to assess how sensitive the results are to the number of discrete speech units. We evaluate the consistency of these values across different bin configurations using the KT rank correlation coefficient.

We report the KT-rank correlation for the results of the MSP-Podcast and BIIC-Podcast corpora in Table 3, BIIC-Podcast and Dusha corpora in Table 4, and Dusha and MSP-Podcast corpora in Table 5. Across all comparisons, the values demonstrate high KT-coefficient agreement, confirming the consistency of clustering and KL divergence values across different bin configurations. The Kendall-Tau rank correlation coefficients in Tables 3-5 collectively validate the robustness of the clustering and the reliability of the results obtained with the KL divergence values.

5. Conclusions

This study conducted a detailed phonetic-level analysis to estimate cross-lingual differences by examining the language-wise distributions of various datasets and emotion-specific vowels within a unified SSL feature space. The KL-divergence values, used to compare vowel distributions across different emotions, highlighted both the most similar and most distinct vowels between the three languages studied. Such analysis forms a foundational basis for developing unsupervised adaptation strategies to bridge cross-lingual variations.

The relevance of this analysis is supported by studies such as Upadhyay et al. [25] showing that transfer learning leveraging contrastive learning strategies based on phonetic proximity can improve cross-corpus SER performance. Additionally, research on speech units using discretized SSL features, such as the one presented in Polyak et al. [35], is closely aligned with the analysis conducted in this paper, offering potential for integration to enhance cross-lingual adaptation methods further.

The consistency of the KL-divergence values across vowels and emotions was further validated by the KT-rank correlation coefficients, underscoring the robustness of the findings. Future work will explore other SSL features, such as Whisper, which may uncover deeper semantic correlations across languages. We will also explore unsupervised or semi-supervised strategies that leverage the findings observed in this study.

6. Acknowledgments

This work was funded by NSF under grant CNS-2016719.

7. References

[1] J. Wagner *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 09, pp. 10 745–10 759, sep 2023.

[2] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada: IEEE, April 2018, pp. 5084–5088.

[3] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.

[4] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215–1227, April-June 2023.

[5] L. C. Matthews and B. Thakkar, "The impact of globalization on cross-cultural communication," *Globalization-education and management agendas*, pp. 325–340, 2012.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[7] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 125–131.

[8] K. Scherer, R. Banse, and H. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, p. 76, January 2001.

[9] R. Lotfian and C. 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, October-December 2019.

[10] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 1823–1827.

[11] S. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. Salman, C. Busso, and C.-C. Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023, pp. 1–8.

[12] l. Kondratenko, N. Karpov, A. Sokolov, N. Savushkin, O. Kutuzov, and F. Minkin, "Hybrid dataset for speech emotion recognition in Russian language," in *ISCA Interspeech 2023*, Dublin, Ireland, August 2023, pp. 2958–1796.

[13] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," in *Speaker Classification II*, ser. Lecture Notes in Computer Science, C. Müller, Ed. Berlin, Germany: Springer-Verlag Berlin Heidelberg, August 2007, vol. 4441, pp. 43–56.

[14] S. Amiriparian, F. Packań, M. Gerczuk, and B. Schuller, "Ex-hubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets," *arXiv preprint arXiv:2406.10275*, 2024.

[15] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, July 2013.

[16] 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, September 2020.

[17] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and J. Yi, "Speech emotion recognition using semi-supervised learning with ladder networks," in *Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018)*, Beijing, China, 2018, pp. 1–5.

[18] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[19] A. Reddy Naini *et al.*, "Unsupervised domain adaptation for preference learning based speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes island, Greece, June 2023, pp. 1–5.

[20] 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.

[21] J. Gideon *et al.*, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, October-December 2021.

[22] A. Reddy Naini *et al.*, "Domain-specific adaptation in speech emotion recognition using emotional distribution alignment," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India, April 2025.

[23] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] K. Choi, A. Pasad, T. Nakamura, S. Fukayama, K. Livescu, and S. Watanabe, "Self-supervised speech representations are more phonetic than semantic," *arXiv preprint arXiv:2406.08619*, 2024.

[25] S. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[26] S. Upadhyay, C. Busso, and C.-C. Lee, "A layer-anchoring strategy for enhancing cross-lingual speech emotion recognition," in *Interspeech 2024*, Kos Island, Greece, Sept. 2024, pp. 4693–4697.

[27] S. Upadhyay, A. Salman, C. Busso, and C.-C. Lee, "Mouth articulation-based anchoring for improved cross-corpus speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*, Hyderabad, India, April 2025, pp. 1–5.

[28] V. Kondratenko *et al.*, "Large raw emotional dataset with aggregation mechanism," *arXiv preprint arXiv:2212.12266*, 2022.

[29] 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.

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

[31] M. Trofimov and D. Jones, *The pronunciation of Russian*. The University Press, 1923.

[32] Y. Liao *et al.*, "Formosa speech recognition challenge 2018: data, plan and baselines," in *2018 11th International Symposium on Chinese Spoken Language Processing*, 2018, pp. 270–274.

[33] S. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, October 2022.

[34] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 23–29 Jul 2023, pp. 28 492–28 518.

[35] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.