ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095250

# PHONETIC ANCHOR-BASED TRANSFER LEARNING TO FACILITATE UNSUPERVISED CROSS-LINGUAL SPEECH EMOTION RECOGNITION

*Shreya G. Upadhyay*[1], *Luz Martinez-Lucas*[2], *Bo-Hao Su*[1], *Wei-Cheng Lin*[2], *Woan-Shiuan Chien*[1], *Ya-Tse Wu*[1], *William Katz*[3], *Carlos Busso*[2], *Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Department of Electrical and Computer Engineering, University of Texas at Dallas, USA
[3]Department of Speech Language and Hearing, University of Texas at Dallas, USA

## ABSTRACT

Modeling cross-lingual *speech emotion recognition* (SER) has become more prevalent because of its diverse applications. Existing studies have mostly focused on technical approaches that adapt the feature, domain, or label across languages, without considering in detail the similarities between the languages. This study focuses on domain adaptation in cross-lingual scenarios using phonetic constraints. This work is framed in a twofold manner. First, we analyze emotion-specific phonetic commonality across languages by identifying common vowels that are useful for SER modeling. Second, we leverage these common vowels as an anchoring mechanism to facilitate cross-lingual SER. We consider American English and Taiwanese Mandarin as a case study to demonstrate the potential of our approach. This work uses two in-the-wild natural emotional speech corpora: MSP-Podcast (*American English*), and BIIC-Podcast (*Taiwanese Mandarin*). The proposed unsupervised cross-lingual SER model using these phonetical anchors outperforms the baselines with a 58.64% of *unweighted average recall* (UAR).

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

## 1. INTRODUCTION

Building Speech Emotion Recognition (SER) strategies to improve its generalization across different domains is a crucial step to enable diverse applications across fields, including healthcare, security, education, and entertainment [1]. A common formulation for cross-corpus SER models aims to mitigate mismatches between the source and target domains. These approaches include strategies to compensate for features, domains, or label mismatches using techniques such as transfer learning, semi-supervised learning, and few-shot learning [2, 3]. Other approaches include explicitly optimizing to decrease a distance metric between source and target features (e.g., Wasserstein [4]), utilizing adversarial training to prevent domain memorization [5], or introducing additional synthetic domain-specific data that is produced by a Generative Adversarial Network (GAN)-based model [6]. Although these models are useful, they tend to come purely from a computational angle. When dealing with SER tasks that require cross-lingual domain adaptation, it is expected that knowledge about the languages can offer new modeling opportunities.

One of the most important unsupervised cross-corpus settings is cross-lingual applications, where SER models are trained on one language and tested on another. Having a strong cross-lingual SER strategy can facilitate the development of SER for languages with less resources. Previous studies have predominantly treated cross-lingual scenarios as a language-agnostic problem [3,7–11], which limits language domain adaptation. Emotion perception and the acoustic feature space depend on the language [12]. Understanding the similarities between languages can lead to better cross-lingual SER domain adaptation strategies. Previous studies have shown that discriminative emotional information can be observed even at the phonetic-level [13]. Interestingly, some of these emotional patterns at phone-level generalize to other languages [14]. Simple phoneme-class dependent emotion classifiers [13] and fine-tuned deep models (e.g., Wav2Vec2) with emotion-dependent phoneme transcriptions [15] can effectively improve emotion recognition rates. The similarities across languages at the phone-level may not necessarily be preserved at higher syntactic units (e.g., word, phrase or sentence-level). Therefore, we explore anchoring our unsupervised cross-lingual SER model to specific phonetic commonalities across the target languages. As a case study, we focus on transferring from American English (intonation language) to Taiwanese Mandarin (tonal language).

This study proposes a cross-lingual approach that leverages phonetic similarities across languages to anchor our transfer learning strategy. We rely on two large-scale in-the-wild natural speech emotion corpora: the MSP-Podcast (American English) and BIIC-Podcast (Taiwanese Mandarin) corpora. Our study involves two parts: First, we analyze the emotion-specific commonality at the phonetic-level be-

This work was supported by the NSTC under Grants 110-2221-E-007-067-MY3 and 111-2634-F-002-023, and the NSF under Grant CNS-2016719.

tween American English and Taiwanese Mandarin. We rely on two perspectives: phonological references and by building phoneme-level SER. We observe that some vowels present emotion-specific commonality across these languages. Second, we devise an anchoring mechanism that leverages the phonetic commonalities across languages. Using a contrastive formulation, we demonstrate that the proposed anchoring mechanism over these target vowels can facilitate cross-lingual SER. Our proposed anchor-based domain adapted cross-lingual SER achieves 6.89% improvement in *unweighted average recall* (UAR) over the model with no domain adaptation.

## 2. CROSS-LINGUAL CORPUS

**The MSP-Podcast (MSP-P)** [16] corpus contains a total of 166 hours of emotional speech in English (v1.10). The speech samples are obtained from recordings available on audio-sharing websites. We used the *Montreal forced aligner* (MFA) [17] to extract the phones with their boundary alignment. The MFA outputs phones in ARPABET notation, so we convert the phones to the *International Phonetic Alphabet* (IPA) notation using the commonly used mapping found in the study of Rice [18].

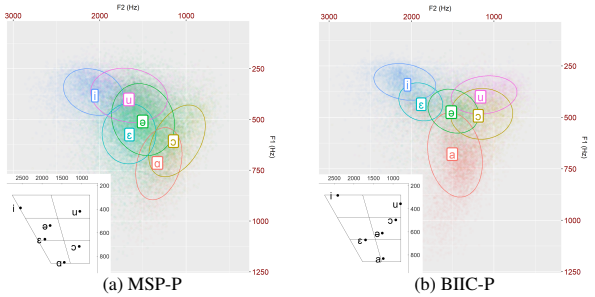
**The BIIC-Podcast (BIIC-P)** corpus is a new SER database that we are currently collecting. The speech samples come from *Taiwanese Mandarin* podcasts, collected with a similar protocol as the one used for the MSP-P corpus. The number of emotional annotations ranges from 3-6 per sample. The samples are annotated using eight primary emotional categories (*Neutral, Happiness, Anger, Sadness, Disgust, Contempt, Fear, Surprise*) and three emotional attributes (*Arousal, Valence, Dominance*). The corpus also includes transcriptions. In this work, we used around 20 hours of data. We first train a *Taiwanese Mandarin* forced aligner [19] using the *Formosa* database. Then, we convert the phones to IPA notation using the mapping in Liao et al. [19].

## 3. EMOTION-SPECIFIC COMMONALITY

We analyze the corpora from the perspectives of phonetic analyses and emotion-specific SER results to find emotion-specific commonalities in the set of "common ground" vowels. This work considers the following set of common vowels: {i, ɛ, ə, ɑ/a, ɔ, u}. We use around 12 hours of data from each corpora, matching the emotional distributions across them. Our analysis considers the emotional classes of happiness, anger, sadness, and neutrality.

### 3.1. Phonetic Analysis

The phonetic analysis aims to observe similarities across vowels obtained from the emotional recordings in *American English* and *Taiwanese Mandarin*. We analyze the vowel space using the first two formants (F1 and F2) estimated from the MSP-P and BIIC-P corpora using Praat [20].




(a) MSP-P (b) BIIC-P

**Fig. 1**: Vowel F1-F2 plots of common vowels for the MSP-P and BIIC-P corpora; lower-left corner shows the canonical placement of the vowels considered in the literature.

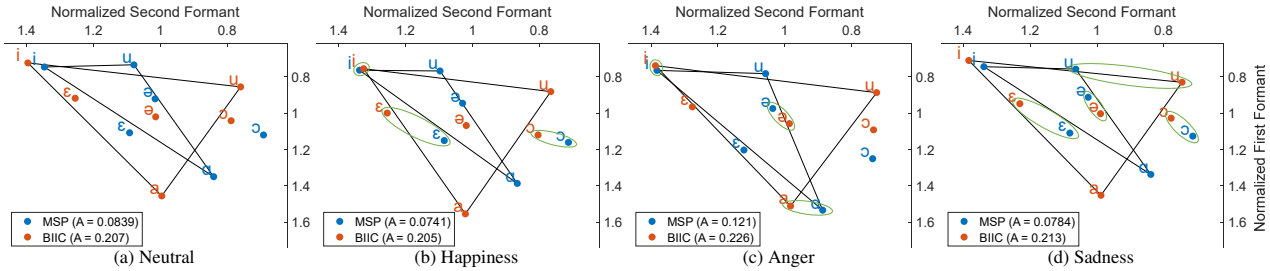
Figure 1 shows the vowel space (F1 *vs.* F2 plots) for the samples across emotions. The superimposed graphs on the lower left provide the location of the vowels known from previous literature for English and Mandarin [21–23]. Figure 1a shows that the English's vowel /u/ is high fronted (reported in [24]), which could be due to different gender ratios or dialects of the speakers. The placement for this vowel is different in Chinese. Overall, Figure 1 shows similar results for vowels in English and Chinese. In both languages, the plots show that this set of common vowels span most of the F1-F2 space, and their positions are consistent with what it is expected from the literature. From Figure 1, we see some visible vowel commonality over corpora, such as vowels /i/ and /ə/, which cover similarity regions in their respective languages.

Figure 2 shows a plot of the average F1 and F2 values with respect to four emotional classes: *Neutral*, *Happiness*, *Anger*, and *Sadness*. The data in Figure 2 are normalized using the Nearey normalization [25] to remove speaker differences due to individual vocal tract disparities and gender. Figure 2a shows that for *Neutral* speech, the closest distances across languages for corresponding vowels are /i/ and /ə/. In fact, the two vowels show this trend over all four emotions. These vowels are potential candidates for serving as anchors in our transfer learning strategy due to their similarity across both languages. In Figure 2, we circle in green vowels that are closer across languages for *Happiness*, *Anger*, and *Sadness* than for *Neutral* (e.g., the distance between /ɛ/ across languages is smaller for *Happiness* than for *Neutral*). These plots are useful to identify vowels that have similar responses across languages in the presence of emotions.

Phonetic examinations using these figures provides initial insights into phonological similarities across languages. These insights point toward the existence of some candidate emotion-specific vowel commonality (as seen in the vowel format plots) that can inspire cross-lingual SER strategies.

### 3.2. Emotion-Specific SER Analysis

We evaluate performance of vowel-based SER models. We use the 768 dimension wav2vec 2.0 feature vector [26], extracted after the phone-level segmentation [27]. These acoustic features are the input of a transformer encoder that gen-




(a) Neutral (b) Happiness (c) Anger (d) Sadness

**Fig. 2**: Emotion-specific average for common vowels in the F1-F2 space for sentences from the MSP-P and BIIC-P corpora.

erates a self-attention hidden embedding. These embeddings are used as encoded features for modeling emotion-specific SER. The model uses the Adam optimizer, with a stochastic gradient decent algorithm. The learning rate and decaying factor are set to 0.0001. The models are trained for a maximum of 100 epochs, with a batch size of 128 with early stopping. The cost function is the binary cross-entropy loss and UAR is used as the evaluation metric. Table 1 shows vowel-dependent SER models trained and tested with different corpus combinations. For example, M→B indicates a cross-lingual experiment training the model with the MSP-P and testing it with the BIIC-P.

***Within-Corpus Vowel Discriminability Analysis:*** By analyzing Table 1, in matched conditions (M→M and B→B), we can see that some vowels do carry similar discriminative capacity from a SER modeling perspective. The performances highlighted in bold in matched cases show that some vowels have similar UAR (within a 2-3% UAR range). In the case of *Neutral*, the SER models for /i/ and /ə/ lead to better UAR, compared to the other vowels for both corpora. For *Happiness*, the SER models for /i/, /ə/, and /ɑ,a/ have similar performance for both corpora. For *Anger*, the SER models for /ɑ,a/, and /ɛ/ achieve similar good performance for both corpora. For *Sadness*, the SER models for /ɛ/, /ɔ/, and /u/ show similar performances over corpora. Because our study aims to identify vowels that perform well in both corpora in vowel-based SER, this selection of vowels are good candidates as anchors.

***Cross-Lingual Vowel Discriminability Analysis:*** Table 1 also shows the cross-lingual analysis (M→B) SER performances for each emotion. Given the larger size of the MSP-P corpus, we only formulate cross-lingual experiments by training with MSP-P and testing with BIIC-P. The bold values in this section highlight the best vowel for that specific emotion in this cross-lingual setting. The results indicate that some vowel-specific SER models trained with the MSP-P corpus do not work well in recognizing emotions for the BIIC-P samples. For example, the SER model for /ɔ/ for *Sadness* shows low performance as compared to other emotions, even in the matched condition SER models for /ɔ/ have relatively good performance for both languages. The table depicts that even though some vowels have emotion-specific commonality over corpora, these vowel-dependent models are not effective without compensating for the corpus-wise variability.

**Table 1**: Emotion-specific SER performance over common vowels in both corpora; the *Exp* column shows the vowels as a phonetic constraint used in *GA-CL* as 𝔾, *BA-CL* as 𝔹, and *WA-CL* as 𝕎. GA = group anchored, BA = best anchored, and WA = worst anchored.

| | | Neutral | | Happy | | Angry | | Sad | |
|---|---|---|---|---|---|---|---|---|---|
| | | UAR | Exp | UAR | Exp | UAR | Exp | UAR | Exp |
| /i/ | M→M | **75.78** | 𝔾 | **76.47** | 𝔾𝔹 | 73.36 | | 65.30 | |
| | B→B | **77.62** | | **75.04** | | 72.53 | | 67.87 | |
| | M→B | 60.8 | | 60.28 | | 60.19 | | **59.96** | |
| /ɛ/ | M→M | 69.45 | | 73.90 | | **75.78** | 𝔾 | **67.34** | 𝔾 |
| | B→B | 75.66 | | 68.24 | | **75.22** | | **70.19** | |
| | M→B | 58.34 | | 60.10 | | 55.53 | | 51.76 | |
| /ə/ | M→M | **76.34** | 𝔾,𝔹 | **75.78** | 𝔾 | 73.65 | | 64.35 | 𝕎 |
| | B→B | **77.15** | | **75.50** | | 72.52 | | 65.19 | |
| | M→B | 61.55 | | **63.23** | | **63.89** | | 50.40 | |
| /ɑ,a/ | M→M | 69.36 | 𝕎 | **75.61** | 𝔾 | **76.56** | 𝔾,𝔹 | 67.45 | |
| | B→B | 75.31 | | **74.31** | | **75.14** | | 68.34 | |
| | M→B | **61.93** | | 61.41 | | 61.45 | | 53.02 | |
| /ɔ/ | M→M | 74.38 | | 72.53 | 𝕎 | 70.89 | | **68.76** | 𝔾,𝔹 |
| | B→B | 76.19 | | 70.99 | | 74.62 | | **70.82** | |
| | M→B | 58.93 | | 57.82 | | 59.20 | | 58.48 | |
| /u/ | M→M | 76.45 | | 77.01 | | 70.35 | 𝕎 | **66.89** | 𝔾 |
| | B→B | 73.36 | | 71.23 | | 72.29 | | **69.28** | |
| | M→B | 51.04 | | 52.69 | | 53.02 | | 52.24 | |

From both phonetic analysis and within- and cross- corpus vowel-specific emotion recognition experiments, there seems to be phonological similarity over some of these common vowels, similar emotion discrimination ability, and difficulty in directly transferring learned emotion discrimination for certain vowels. With these insights and observations, we select some commonly behaving vowels (marked with 𝔾 in Table 1) to facilitate the design of cross-lingual SER.

## 4. ANCHOR-BASED CROSS-LINGUAL SER

Our analyses in Section 3 provide initial evidences that certain vowels can be phonetically-similar after emotion modulation across the two languages. Inspired by these findings, we design an anchoring mechanism to integrate the phonetic constraint in cross-lingual modeling (Fig. 3). Our proposed unsupervised cross-lingual SER contains two branches: (1) the conventional emotion classification branch for classifying emotions, and (2) the vowel domain adaptation branch that integrates the phonetic constraint. Equation 1, gives the classification loss,

$$L_{ec} = \mathbb{E}_{X_S, y_S}[\|CE(T(X_S), y_S)\|] \quad (1)$$

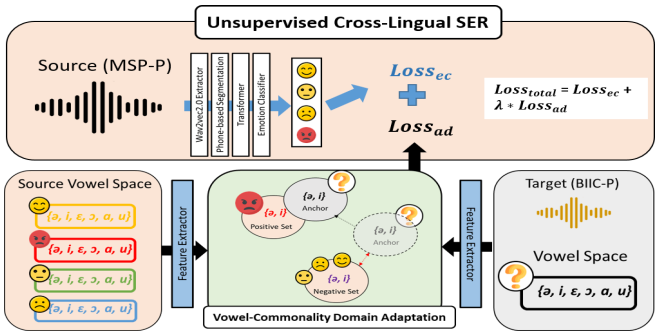



**Fig. 3**: Proposed contrastive learning approach using emotion-specific commonality-based anchoring mechanism for cross-lingual SER.

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.

The vowel domain adaptation branch performs the anchoring mechanism on the two corpora by imposing phonetic knowledge as a constraint to leverage the similarity between the two languages for certain phones, which leads to better regularization. Our formulation rely on the triplet loss function. Specifically, the segments from emotion-specific vowels in the target domain act as $Anchors$. The $Positives$ samples are the vowel segments from the source domain for the same set of vowels to transfer specific-emotion knowledge. The $Negatives$ samples are the vowel segments from the source domain for the same vowel set, but with different emotions.

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 2,

$$\mathrm{L}_{ad} = \sum_i^N [d(f({X_i}^{t_{ph}}), f({X_i}^{s_{p_{ph}}})) - d(f({X_i}^{t_{ph}}), f({X_i}^{s_{n_{ph}}})) + \alpha] \quad (2)$$

where $d$ represents the Euclidean distance function, $f({X_i}^{t_{ph}})$ is the feature representation for the target domain, and $f({X_i}^{s_{p_{ph}}})$ and $f({X_i}^{s_{n_{ph}}})$ are the positive and negative feature representations of the source domain for the same vowel set, respectively. $\alpha$ represents the margin. The complete loss is calculated using Equation 3,

$$L_{total} = L_{ec} + L_{ad} * \lambda \quad (3)$$

where $L_{ec}$ and $L_{ad}$ are the losses for the emotion classification and domain adaptation tasks. $\lambda$ is the regularization parameter.

### 4.1. Experiment Results

All the feature extraction and experimental settings are the same as the ones presented in Section 3. The systems are trained and back-propagated with Equation 3. The SER formulation is a binary emotion detection task, and a four-class emotion classification task in an unsupervised cross-lingual setting. The models are trained and evaluated on a fixed train, validation, and test sets. The BIIC-P test data used in this section is approximately 6 hours, which is not included in the analysis presented in Section 3.

**Table 2**: Cross-lingual SER performance (in UAR) with proposed group-vowel-anchored (*GA-CL*), feature-matching (FM-CL), and some ablation results with best-vowel-anchored (*BA-CL*) and worst-vowel-anchored (*WA-CL*).

| Models | 4-Category | Neu | Hap | Ang | Sad |
|---|---|---|---|---|---|
| CL | 51.75 | 65.61 | 62.77 | 64.47 | 58.53 |
| FM-CL | 56.92 | 70.40 | 67.32 | 69.83 | 65.59 |
| *GA-CL* | **58.64** | **72.83** | **69.69** | **70.15** | **68.17** |
| *BA-CL* | **55.33** | 70.23 | **68.74** | **67.83** | 63.91 |
| *WA-CL* | 55.21 | **70.43** | 61.45 | 66.26 | **64.62** |

Table 2 shows the performances (in UAR) for the cross-lingual SER models. In this initial work, we consider as baselines for performance comparison models with no domain adaptation (CL), and with simple feature matching (FM-CL), where $Anchors$ segments in the target domain are randomly selected. Our proposed model group-vowel-anchored (GA-CL) for unsupervised cross-lingual SER outperforms the *CL* and *FM-CL* models with absolute UAR gains of 6.89% and 2.72%, respectively. We also experimented our phonetic anchoring idea with the *most* and *least* commonly behaving vowel over corpora i.e., best-vowel-anchored (BA-CL) and worst-vowel-anchored (WA-CL). Table 1 indicate the single vowel selected for these two models. Table 2 shows that the *BA-CL* model has better performance than the *WA-CL* model for *Happiness* and *Anger*. Using a single vowel (BA-CL) as $Anchor$ is not as good as selecting the set of vowels used for our proposed model (GA-CL). These results confirm that transfer learning based on the selected common phonetic anchors can integrate important information that facilitates language adaptation in cross-lingual SER.

## 5. CONCLUSION

This paper proposed a phonetic anchoring mechanism for unsupervised cross-lingual settings based on initial evidence of emotion-specific commonality of vowels from two different languages. An emotion-specific commonality analysis indicated that some vowels are more similar between corpora after emotion modulations. Our contrastive learning approach used these vowels as phonetic constraints to control the variability between two languages, enhancing the learning for unsupervised cross-lingual SER. The proposed model *GA-CL* based on phonetic anchor-based transfer learning (58.64%) outperforms the *FM-CL* (56.92%) and *CL* (51.75%) baselines models. In future work, we plan to merge this novel phonetic knowledge-driven anchoring mechanism with recent SOTA approaches on domain adaptation for better generalization. In addition, we plan to include common ground consonants (particularly fricatives, affricates, and approximants) to improve cross-lingual SER performance.

## 6. REFERENCES


[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Wisha Zehra, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu, "Cross corpus multilingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, 2021.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] Paul Boersma and David Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341–345, January 2001.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.