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

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

\textsuperscript{1}Department of Electrical Engineering, National Tsing Hua University, Taiwan
\textsuperscript{2}Department of Electrical and Computer Engineering, University of Texas at Dallas, USA
\textsuperscript{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-
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, r, a, a/a, a, 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].

![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.](image)

Figure 1 shows the vowel space (F1 vs. F2 values) 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 /a/, 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 /a/. 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 /e/ 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-
Sadness shows the cross-lingual analysis (M→B) results 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 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 /a/ lead to better UAR, compared to the other vowels for both corpora. For Happiness, the SER models for /i/ and /a/ have similar performance for both corpora. For Anger, the SER models for /o/a/ and /a/ achieve similar good performance for both corpora. For Sadness, the SER models for /i/ and /a/ 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.

**4. ANCHOR-BASED CROSS-LINGUAL SER**

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 \( \mathbb{G} \) in Table 1) to facilitate the design of cross-lingual SER.

![Fig. 2: Emotion-specific average for common vowels in the F1-F2 space for sentences from the MSP-P and BIIC-P corpora.](image-url)
are the positive and negative feature representations 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,

$$ L_{ad} = \sum_{i \in I} [d(f(X^{s \text{ph}}_i), f(X^{s \text{ph}}_i)) - d(f(X^{t \text{ph}}_i), f(X^{t \text{ph}}_i)) + \alpha] $$

where \( d \) represents the Euclidean distance function, \( f(X^{s \text{ph}}_i) \) is the feature representation for the target domain, and \( f(X^{t \text{ph}}_i) \) and \( f(X^{s \text{ph}}_i) \) 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,

$$ L_{\text{total}} = L_{cc} + L_{ad} \times \lambda $$

where \( L_{cc} \) 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 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 indicates 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


