A Layer-Anchoring Strategy for Enhancing Cross-Lingual Speech Emotion Recognition

Shreya G. Upadhyay\(^1\), Carlos Busso\(^2\), Chi-Chun Lee\(^1\)

\(^1\)National Tsing Hua University, Taiwan  \hspace{1em} \(^2\)University of Texas at Dallas, USA

shreya@gapp.nthu.edu.tw, busso@utdallas.edu, cclee@ee.nthu.edu.tw

Abstract

Cross-lingual speech emotion recognition (SER) is important for a wide range of everyday applications. While recent SER research relies heavily on large pretrained models for emotion training, existing studies often concentrate solely on the final transformer layer of these models. However, given the task-specific nature and hierarchical architecture of these models, each transformer layer encapsulates different levels of information. Leveraging this hierarchical structure, our study focuses on the information embedded across different layers. Through an examination of layer feature similarity across different languages, we propose a novel strategy called a layer-anchoring mechanism to facilitate emotion transfer in cross-lingual SER tasks. Our approach is evaluated using two distinct language affective corpora (MSP-Podcast and BIIC-Podcast), achieving a best UAR performance of 60.21% on the BIIC-podcast corpus. The analysis uncovers interesting insights into the behavior of popular pretrained models.

Index Terms: speech emotion recognition, large pretrained models, cross-lingual

1. Introduction

Recently, large pretrained models like Wav2Vec 2.0 [1], WavLM [2], Whisper [3], and Hubert [4] have gained popularity, offering versatile capabilities across diverse applications. Trained on extensive datasets, these models serve as potent resources for tasks beyond their original domains. A notable trend involves fine-tuning these pretrained models for specific tasks such as SER [5–8], phonetics [9, 10], speaker or language change detection [11], and speaker identification [12,13]. These models also serve as feature extractors, providing a rich source of abstract representations useful across different tasks.

These transformer-based models exhibit a hierarchical architecture and within this hierarchy, diverse levels of information are embedded in the transformer layers [14]. The layer information varies based on the task specificity. For example, in models trained for automatic speech recognition (ASR), initial layers capture fundamental acoustic details, while later layers encapsulate more complex lexical information. When employing these layer embeddings for different tasks, the ability to selectively choose or assign weights to layers that are more relevant can enhance the learning. Numerous studies not only use the final transformer layer but also utilize the information within other layers of the pretrained models [10, 15, 16]. They aim to utilize this valuable information in a weighted or averaged manner, aligning features more effectively with specific task requirements. For instance, in speaker verification tasks [15], various studies analyze different feature extraction methodologies developed upon pretrained models. Additionally, they explore regularization techniques and learning rate scheduling to stabilize the fine-tuning process, resulting in notable performance enhancements. Another research endeavor [16] emphasizes that speech representations derived from specific neural models like Transformers exhibit closer alignment with human perception, particularly regarding phonetic transcriptions. Moreover, English et al. [10] highlights the Transformer architectures’ capacity to effectively capture significant phonetic nuances.

Speech representations derived from specific neural models, such as different transformer layers in pretrained models exhibit enhanced efficiency in recognizing phoneme [17]. These pretrained model’s transformer architectures are adept at capturing substantial levels of phonetic information across various layers [10,17]. In the domain of cross-lingual SER, where phonetic alignment between diverse language corpora is advantageous [5], certain layers within large pretrained models beyond the final layer may hold greater importance [14]. These layers which directly encode acoustic cues and phonetic characteristics can form the fundamentals of tasks related to emotion recognition. Recognizing the potential efficiency gains in cross-lingual SER models, this paper introduces a novel approach known as Layer-anchoring. This method strategically aligns layers based on their similarity across the two language corpora. By prioritizing and aligning layers that exhibit greater commonality between the features of both corpora, the model can enhance its performance. This novel strategy acknowledges the task-specific nature of SER and leverages the hierarchical structure of pretrained models to optimize layer representation utilization in a nuanced and contextually relevant manner.

Given that WavLM [2] has currently secured the top position on the SUPERB benchmark [18] (retrieved on March 10, 2024), our experimentation will focus on utilizing WavLM and analyzing its performance with our proposed layer-anchoring algorithm. To address cross-lingual scenarios, we also use the multilingual pretrained model, Whisper [3] a widely adopted in recent studies [19, 20], to compare its performance and insights with those of the monolingual model (WavLM). This study employs two distinct language corpora: the MSP-Podcast (American English) [21] and the BIIC-Podcast (Taiwanese Mandarin) [22] corpora. First, we analyze layer similarities within the pretrained model’s encoded features across both corpora. This analysis reveals layers exhibiting better commonality between the corpora than the final layer. Building upon this insight, we implement the layer anchoring mechanism (LAM) to develop a cross-lingual SER model. Our proposed model, referred to as a layer-anchoring mechanism with a group of layers (LAM-GL), outperforms alternative approaches achieving 60.21% unweighted average recall (UAR) with WavLM features and 59.65% with Whisper encoded features over the BIIC-podcast corpus.
both the utterance-level and phonetic-level. Here, we only use
and MSP-P corpora across four primary emotions, including
similarity between layer representations of the BIIC-P
prior study [5] has shown that cross-lingual contexts may reveal
and Whisper. We analyze the layer-similarity at both utterance and
comparison, we employ two off-the-shelf models: WavLM and
language corpora (the MSP-P and the BIIC-P). To explore this
ings, we aim to examine the degree of layer-similarity encoded
within these model’s layer representations across two distinct
language corpora (the MSP-P and the BIIC-P) across each emotional
category as presented in Figure 1. Upon examining plots depicted in Figure 1, we observe differing layer similarity behaviors between the WavLM and Whisper models’ feature embeddings. In WavLM, later layers, such as layer 11 show higher similarity, while the Whisper model’s initial layers (1 to 5) exhibit greater similarity. This trend is constant across different emotions. WavLM shares a similar phenomenon with wav2vec2.0 [14], suggesting that high-level features may be more general and potentially lead to higher similarity. However, this assumption contrasts with the Whisper model, possibly due to distinct training methodologies. These findings prompt further exploration into phonetic-level layer similarity to determine any distinct observations compared to utterance-level analysis.

**2.2. Layer Similarities**

Previous research using large pretrained models on different tasks reveals that each layer contributes different levels of information during task learning [10, 17]. Building on these findings, we aim to examine the degree of layer-similarity encoded within these model’s layer representations across two distinct language corpora (the MSP-P and the BIIC-P). To explore this comparison, we employ two off-the-shelf models: WavLM and Whisper. We analyze the layer-similarity at both utterance and phonetic levels. We include phonetic-level analysis because our prior study [5] has shown that cross-lingual contexts may reveal shared phonetic commonalities. Figure 1 visually illustrates the cosine similarity between layer representations of the BIIC-P and MSP-P corpora across four primary emotions, including both the utterance-level and phonetic-level. Here, we only use training samples, excluding the test samples.

**Utterance-Level Layer Similarity:** To assess the presence of similarities across layers in the two corpora for different emotions over the whole utterance, we extract all-layer feature representations from the considered pretrained models (WavLM and Whisper) for both the MSP-P and the BIIC-P corpora using entire utterances. Subsequently, we compute the layer-similarity using the cosine similarity metric between the layer representations of the MSP-P and the BIIC-P corpora across each emotional category as presented in Figure 1. From Figure 1, a clear pattern emerges in the similarity of layer features at the vowel level between the WavLM and Whisper models, aligning with the observations made at the utterance-level. Specifically, the WavLM model displays a trend of increasing layer similarity in its later layers for all vowel phones, contrasting starkly with the Whisper model, where greater similarity is observed in the initial layers. Investigating inter-vowel segment similarities within [u/a, r, a, i, a, u], Figure 1 reveals variations across layers for different emotions. Nonetheless, an overall high degree of similarity is observed between the corpora at the corresponding layers.

The above analyses highlight that due to task specificity and the hierarchical nature of models, in self-supervised learning models like WavLM, later layers encapsulate more abstract patterns and language-specific phonetic nuances as the model learns to predict future speech tokens. Conversely, Whisper be-
Table 1: Selected layer for WavLM and Whisper model.

<table>
<thead>
<tr>
<th></th>
<th>WavLM</th>
<th>Whisper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group-Layers (GL)</td>
<td>[8, 9, 11]</td>
<td>[1, 3, 5]</td>
</tr>
<tr>
<td>Best-Layer (BL)</td>
<td>[11]</td>
<td>[2]</td>
</tr>
<tr>
<td>Worst-Layers (WL)</td>
<td>[5, 6, 7]</td>
<td>[7, 10, 11]</td>
</tr>
<tr>
<td>Random-Layers (RL1)</td>
<td>[2, 6, 9]</td>
<td>[2, 6, 9]</td>
</tr>
<tr>
<td>Random-Layers (RL2)</td>
<td>[1, 5, 12]</td>
<td>[1, 5, 12]</td>
</tr>
<tr>
<td>Random-Layers (RL3)</td>
<td>[3, 7, 11]</td>
<td>[3, 7, 11]</td>
</tr>
</tbody>
</table>

As per our hypothesis, aligning layer features exhibiting high layer-similarities across different language corpora and imposing constraints on those layers can enhance the effectiveness of emotion transfer in CL-SER tasks. To anchor on more similar layers, we select specific sets of layers based on the findings of our previous experience, we have observed that a group of anchors tends to outperform individual ones. Therefore, the table includes group-layers (GL), representing clusters of highly similar layers under different settings. Drawing from our previous work, we integrate LAM over the entire utterance.

2.3. Unified Layer Selection

As per our hypothesis, aligning layer features exhibiting high layer-similarities across different language corpora and imposing constraints on those layers can enhance the effectiveness of emotion transfer in CL-SER tasks. To anchor on more similar layers, we select specific sets of layers based on the findings of our previous experience, we have observed that a group of anchors tends to outperform individual ones. Therefore, the table includes group-layers (GL), representing clusters of highly similar layers. Additionally, we identify the best-layer (BL) and worst-layers (WL), along with three sets of random-layers (RL1, RL2, RL3). Except for BL, we select the top three layers for all cases concerning training CL-SER in the subsequent section.

3. Layer Anchored Cross-Lingual SER

In all our experiments, we consider the MSP-P (source) and BIIC-P (target) corpora as benchmarks to test our idea. The WavLM and Whisper embedding feature vectors are used as the pretrained layer representations. For the CL-SER architecture, we use the transformer with 4-fully connected layer architecture. Following the same model presented in our previous work [5] with an attention-weighted layer feature pooling concept as our backbone SER architecture. We employ the Adam optimizer with a learning rate of 0.0001 and a decay factor of 0.001, and back-propagation is done with the cross-entropy loss function. The network undergoes a maximum of 70 epochs and a batch size of 64 with early stopping. To evaluate model performances, we use the UAR metric. Since both utterance-level and phonetic-level similarities yield similar insights from Section 2, we integrate LAM over the entire utterance.

Our investigations detailed in Section 2 offer initial insights suggesting that specific layers may exhibit more similarity and can enhance emotion modulation across both corpora. Motivated by these findings, we devise a layer anchoring mechanism aim at incorporating the layer-anchoring constraint in the CL-SER modeling (Figure 2). Our proposed unsupervised CL-SER comprises two branches: (1) a conventional emotion classification branch tasked with classifying emotions, and (2) a layer anchoring mechanism (LAM) branch that identifies the layers in the transformer that increase the similarities across languages at the phonetic level. Equation 1 outlines the LAM loss.

$$L_{\text{avg}} = \sum_{i=1}^{12} \alpha_i L^{(i)}_{\text{src}}$$  

(2)

where $L_{\text{avg}}$ denotes the attentional weighted average of the feature representations from all layers, and $\alpha_i$ denotes the attention weight assigned to the layer $i$. Here $\sum_{i=1}^{12} \alpha_i = 1$ and $\alpha_i \geq 0$ for all $i$. The weight $\alpha_i$ can be computed using Equation 3.

$$\alpha_i = \frac{e^{\gamma_1}}{\sum_{j=1}^{12} e^{\gamma_j}}$$  

(3)

Where $\gamma_1$ represents the attention score for layer $i$. The attention scores are the learnable parameter.

The overall loss for the first branch is the sum of the cross-entropy loss for the classic SER task and the CORAL loss for the layer anchoring mechanism. The complete loss is calculated using Equation 4.

$$L_{\text{total}} = L_{\text{SER}} + \gamma \ast L_{\text{CORAL}}$$  

(4)

where $L_{\text{SER}}$ and $L_{\text{CORAL}}$ are the losses for the emotion classification and domain adaptation tasks. $\gamma$ is the regularization parameter which is set to a constant value ($\gamma = 0.5$).

4. Experiment Results and Analyses

4.1. Performance Comparison

Table 2 shows the performance table, which includes results of our proposed idea LAM-GL, the baseline models, and the ablation results with WavLM and Whisper model’s encoded layer representations.
We assess three baseline methods: ensemble learning [24], which combines predictions from diverse models to enhance recognition accuracy in cross-lingual scenarios (Ensemble); few-shot learning [25], adapting models to target domains with limited labeled data (Few-shot); and our previously proposed Phonetic-constraint-based anchoring (PA) method [5], used for learning in a common phonetic space for SER. Compared with the baselines in Table 2 for the MSP-P→BIIC-P task, our layer anchoring approach (LAM-GL) yields superior performance. Specifically, compared to models using only the last layer (the 12th layer), LAM-GL achieves a UAR of 60.21%, surpassing Ensemble [24] at 52.18%, Few-shot [25] at 53.62%, and PC [5] at 58.14%. This enhanced performance is also evident with Whisper features. Additionally, drawing on prior work employing layer information for tasks like speaker identification and phone recognition [15], we integrate these methods with PC, which outperforms Ensemble and Few-shot, denoted as PC-Avg [5, 15] and PC-Atn [5, 15]. Comparing these models with LAM-GL reveals further performance improvements, with WavLM features achieving 2.15% and 1.38% UAR for PC-Avg and PC-Atn models, respectively. The same patterns are observed with Whisper features, suggesting that our LAM-GL model, which aligns more similar layers across various corpus features, improves utility for CL-SER.

To validate our LAM-GA method, we extended our investigation beyond the selected layers, exploring whether aligning any random layer of the two corpora or all layers is acceptable or if precise selection is necessary. We train LAM-GL models with diverse configurations: utilizing all layers (LAM-GL), the best layer (LAM-BL), the worst layer (LAM-WL), and random selections (LAM-RL1, LAM-RL2, LAM-RL3). Table 1 presents the selected layers for these analyses. The MSP-P→BIIC-P task results from Table 2 indicate that anchoring all layer models (LAM-AL) does not enhance performance over LAM-GL, yielding 58.54% for WavLM and 57.97% for Whisper UAR. Similarly, using the best layer (LAM-BL), which achieves the UAR of 59.16% for WavLM and 58.11% for Whisper, is not better than the LAM-GL strategy. Furthermore, the performance of the worst layer (LAM-WL) suggests that aligning dissimilar layers can adversely affect model performance, with WavLM features scoring 58.01% and Whisper features 57.72%. Despite generating three random sets for both WavLM and Whisper model features, results obtained for LAM-RL1, LAM-RL2, and LAM-RL3 did not outperform our LAM-GA model. For instance, using WavLM features, we achieved 58.94%, 58.55%, and 57.23% UAR with LAM-RL1, LAM-RL2, and LAM-RL3, respectively. This underscores the significance of precise layer selection for the LAM.

As a validation of our concept, we incorporate cross-lingual SER assessments utilizing the BIIC-P corpus as the source and the MSP-P corpus as the target. The results are presented in Table 2. In the BIIC-P→MSP-P task, our proposed model LAM-GL demonstrates superior performance compared to other models listed in Table 2, achieving 56.68% with WavLM features and 56.37% with Whisper features. The overall analysis of Table 2 for the BIIC-P→MSP-P task confirms a similar trend to the one observed in the MSP-P→BIIC-P task.

### 4.2. CL-SER With Different Layer Selection Strategy

In this investigation, we explore the performance of LAM-GL compared to LAM across various phoneme groups (vowel-based (Vowl), consonant-based (Cons)), as well as an utterance-based approach (Uttr) over specific emotion detection task. We segment utterances based on different phoneme groups (Vowl, Cons), selecting layers to train our LAM-GA model. Results in Table 3 for the MSP-P→BIIC-P task across four primary emotions reveal that while the selected layers remain relatively consistent across the Uttr, Vowl, and Cons strategies, notable performance differences emerge across different emotions. Particularly, the Vowl approach demonstrates superior performance for emotions such as Happiness and Anger, achieving 74.21% UAR and 75.55% UAR, respectively, with WavLM features. A similar trend is observed with the Whisper model. This significant finding suggests that vowels exhibit a higher level of commonality over the two different language corpora features, potentially facilitating more efficient emotion transfer compared to considering the entire utterance.

### 5. Conclusion

This study introduces a novel approach that aims at reducing layer feature disparities between different language corpora through a layer anchoring strategy. By capitalizing on pretrained models and aligning similar layer features from the source language to those of the target language, we illustrate the efficacy of our method in harmonizing phonetic characteristics while mitigating discrepancies. Our experimentation and evaluation reveal that our layer-anchoring strategy (LAM-GA) achieves the best UAR of 60.21% by effectively facilitating emotion transfer in cross-lingual SER. Additionally, we uncover intriguing insights indicating that the selection of layers is not uniform across all pretrained models but varies depending on the task and the model’s training methodology. Our future work will delve deeper into the observed differences in specific emotion recognition using the LAM-GA method under various strategies, even when the layer disparities are minimal. Also, we will explore enhanced algorithms to accommodate multiple languages in the SER training.

### Table 2: Specific emotion recognition (in % UAR) with different layer selection strategies.

<table>
<thead>
<tr>
<th>Model</th>
<th>Layers</th>
<th>Neu</th>
<th>Hop</th>
<th>Ang</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>WavLM</td>
<td>[8, 9]</td>
<td>75.13</td>
<td>72.88</td>
<td>74.33</td>
<td>69.57</td>
</tr>
<tr>
<td>Whisper</td>
<td>[1, 2]</td>
<td>75.70</td>
<td>72.63</td>
<td>73.80</td>
<td>69.81</td>
</tr>
<tr>
<td>WavLM</td>
<td>[8, 9]</td>
<td>75.98</td>
<td>74.21</td>
<td>75.55</td>
<td>69.92</td>
</tr>
<tr>
<td>Whisper</td>
<td>[1, 3, 5]</td>
<td>74.83</td>
<td>75.02</td>
<td>74.19</td>
<td>70.46</td>
</tr>
<tr>
<td>WavLM</td>
<td>[9, 10, 11]</td>
<td>74.19</td>
<td>73.73</td>
<td>74.92</td>
<td>68.63</td>
</tr>
<tr>
<td>Whisper</td>
<td>[3, 5, 6]</td>
<td>73.34</td>
<td>73.97</td>
<td>73.30</td>
<td>69.95</td>
</tr>
</tbody>
</table>

### Table 3: The proposed model performance (in UAR) for each SER task with considered baselines. It includes the statistical test over the baselines and the proposed LAM-GL model performances, denoted by asterisks (* for p < 0.1, ** for p < 0.05.

<table>
<thead>
<tr>
<th>Top Layer</th>
<th>CC</th>
<th>WavLM Whisper</th>
<th>WavLM Whisper</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIIC-P</td>
<td>52.01%</td>
<td>51.87%</td>
<td>48.39%</td>
</tr>
<tr>
<td>MSP-P</td>
<td>52.18%</td>
<td>52.03%</td>
<td>51.75%</td>
</tr>
<tr>
<td>Few-shot</td>
<td>53.62%</td>
<td>52.74%</td>
<td>50.59%</td>
</tr>
<tr>
<td>Ensemble</td>
<td>58.14%</td>
<td>57.83%</td>
<td>55.35%</td>
</tr>
<tr>
<td>PC</td>
<td>58.06%</td>
<td>58.01%</td>
<td>55.24%</td>
</tr>
<tr>
<td>LAM-GL</td>
<td>60.21%</td>
<td>59.65%</td>
<td>56.68%</td>
</tr>
</tbody>
</table>

We introduce a novel approach that aims at reducing layer feature disparities between different language corpora through a layer anchoring strategy. By capitalizing on pretrained models and aligning similar layer features from the source language to those of the target language, we illustrate the efficacy of our method in harmonizing phonetic characteristics while mitigating discrepancies. Our experimentation and evaluation reveal that our layer-anchoring strategy (LAM-GA) achieves the best UAR of 60.21% by effectively facilitating emotion transfer in cross-lingual SER. Additionally, we uncover intriguing insights indicating that the selection of layers is not uniform across all pretrained models but varies depending on the task and the model’s training methodology. Our future work will delve deeper into the observed differences in specific emotion recognition using the LAM-GA method under various strategies, even when the layer disparities are minimal. Also, we will explore enhanced algorithms to accommodate multiple languages in the SER training.
6. 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.

7. References


