EasyChair Preprint

№ 15443

Cross-Lingual Speech Emotion Recognition Using English and Mandarin on Thai Data

Kantapong Wonghirunruch

November 20, 2024

Cross-lingual Speech Emotion Recognition Using English and Mandarin on Thai Data

Kantapong Wonghirunruch*
Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, 10330, Thailand
(* corresponding author; email: Kantapong_wong@outlook.com)

Abstract

This study explores the efficacy of cross-lingual Speech Emotion Recognition (SER) using Thai as a target language with training sets in English and Mandarin. The study evaluates the adaptability of SER models across linguistic boundaries, emphasizing the challenges and potential of leveraging well-resourced languages to enhance emotion recognition capabilities in a language with fewer resources. Through a series of experiments, the research investigates three primary aspects: the performance of same-corpus training within Thai, cross-lingual model application from English and Mandarin to Thai, and the effectiveness of transfer learning techniques in improving model accuracy. The findings indicate that Mandarin facilitates more effective cross-lingual SER with Thai compared to English. However, despite the initial promise, models trained on Mandarin or English and applied to Thai did not outperform those trained directly on Thai in the same-corpus settings, suggesting limited benefits from cross-lingual training without sophisticated adaptation methods. Transfer learning emerged as a pivotal strategy, particularly when models pre-trained on large datasets in Mandarin were fine-tuned with Thai data, showing improved performance, and suggesting a scalable approach for deploying SER systems in multilingual contexts.

Keywords: speech emotion recognition, cross-lingual, Thai language, deep learning

1. Introduction

Speech Emotion Recognition (SER) is an essential component of intelligent human-computer interaction [1]. It aims to equip machines with the ability to interpret and respond to human emotions, as understanding the emotional state of a speaker is crucial for effective communication and social interaction. By analyzing the emotional content of speech, machines can gain insights into the speaker's intentions, attitudes, and beliefs, enabling them to respond appropriately and build affinity with the speaker.

The practical applications of SER span diverse domains such as healthcare, education, customer service, and entertainment. In mental health assessment, SER can detect early signs of depression, anxiety, or stress from speech signals offering a non-intrusive method for monitoring emotional well-being [2]. In customer service, it enhances the quality of interactions by identifying the customer's emotional state and tailoring responses accordingly [3]. Moreover, integrating SER in virtual assistants and other interactive systems allows for more natural and empathetic interactions, significantly enhancing user experience, engagement, and satisfaction.

Recent advancements in machine learning and signal processing have significantly improved the accuracy and robustness of SER systems [4], [5]. These systems extract meaningful features from speech signals and employ sophisticated algorithms to classify emotions. While successful in monolingual settings, where training and testing data are in the same language [4], [6], SER systems face challenges in cross-lingual scenarios due some unknown language-specific features that may not generalize well across different languages.

Cross-lingual SER refers to the process of training an emotion recognition model on speech data in one language and applying it to recognize emotions in speech data from another language. This approach holds significant promise for languages with limited annotated emotional speech datasets like Thai. However, it also presents unique challenges due to the linguistic and cultural differences

in emotional expression. Emotional expressions are influenced by cultural norms, intonation patterns, and phonetic characteristics, which can vary significantly between languages. Despite these challenges, cross-lingual learning has been explored in several languages, including French [7], Mandarin [8], Urdu [9], Thai [10], or even in a multilingual setting as seen in [11], its feasibility and effectiveness in Thai still remain largely unexplored.

This study addresses the challenge of limited availability of annotated emotional speech datasets in Thai by investigating cross-lingual SER using a unique dataset that includes equal parts of English and Mandarin speech samples. Each language variant consists of the same speech sentences with the same translation, spoken by the same number of actors, facilitating direct comparison under controlled conditions. This setup allows for a detailed examination of how linguistic variations impact emotion recognition in Thai speech.

Additionally, this research seeks to compare the relative effectiveness of English and Mandarin training sets in recognizing emotions in Thai speech. English, as a widely spoken and studied language, has a vast repository of annotated emotional speech data, making it a strong candidate for cross-lingual training. Similarly, Mandarin, with its tonal nature and rich prosodic features, offers a unique perspective for cross-lingual SER. By systematically comparing the outcomes of using English and Mandarin training sets on a Thai test set, this study endeavors to identify best practices and potential pitfalls in cross-lingual SER.

2. Research Methodology

2.1 Datasets

This study utilizes two primary datasets: the Thai Speech Emotion Dataset (THAI SER) [12] and the Emotion Speech Dataset (ESD) [13], which includes both English and Mandarin speech samples. Table 1 gives an overview of the emotion labels covered in each dataset. However, since the datasets contain differing emotion labels, the emotions were re-categorized into three broader groups: negative, positive, and neutral. This re-categorization ensures a more cohesive comparison across the datasets while preserving as much data as possible. Table 2 details the composition of each dataset after re-categorization, along with the emotions included in each emotion group.

Table 1: Composition of datasets prior to re-categorization

Emotion	THAI SER	ESD (English)	ESD (Mandarin)
neutral	4172	3500	3500
happiness	2974	3500	3500
sadness	1594	3500	3500
anger	1973	3500	3500
frustration	3469	-	-
surprise	-	3500	3500
Total	14182	17500	17500

Table 2: Composition of datasets after re-categorization, along with the emotions included in each emotion group

Emotion Group	THAI SER	ESD (English)	ESD (Mandarin)
negative	7036 (*anger*, *sadness*, *frustration*)	7000 (*anger*, *sadness*)	7000 (*anger*, *sadness*)
positive	2974 (*happiness*)	7000 (*happiness*, *surprise*)	7000 (*happiness*, *surprise*)
neutral	4172 (*neutral*)	3500 (*neutral*)	3500 (*neutral*)
Total	14182	17500	17500

2.1.1 Thai Speech Emotion Dataset (THAI SER)

The dataset consists of Thai speech samples expressing five emotions: *neutral*, *happiness*, *sadness*, *anger*, and *frustration*. It features 200 actors (87 males and 113 females), offering a total of 41 hours and 36 minutes of recordings across 27,854 utterances. The dataset is divided into 100 groups, with 80 professionally recorded in a studio (21,850 utterances) and 20 recorded via Zoom (6,004 utterances). To enhance data reliability, only utterances with a majority agreement value of 0.71 or higher were selected. This threshold ensures that the final dataset consists of reliably labeled emotional speech samples. As a result, 14,182 utterances were included in the final dataset.

2.1.2 Emotion Speech Dataset (ESD)

The ESD comprises 35,000 parallel utterances from 20 actors, including 10 native English speakers and 10 native Mandarin speakers, each contributing 17,500 utterances. The dataset spans five emotion classes: *neutral*, *happiness*, *anger*, *sadness*, and *surprise*, and contains over 29 hours of speech data recorded in a controlled acoustic environment. Its equal representation of English and Mandarin utterances, with identical translations, makes it ideal for multi-speaker and cross-lingual emotional voice conversion studies, applicable to SER tasks. This dataset does not calculate agreement values as it does not involve multiple independent raters. Consequently, all 35,000 utterances are used without any filtering based on agreement.

2.2 Feature Extraction

To implement the feature extraction method, we utilized the Python package *librosa* [14]. The process involved applying a 23 Mel-band LMFB to raw audio files with a sample rate of 22050 Hz (Figure 1), with a 93ms window (n_fft = 2048) and a 23ms shift (hop_length = 512), with no speed or acceleration coefficients. Each utterance was padded to a length of 154,350 frames (equivalent to 7 seconds of audio), using the "wrap pad" method. Figure 2 illustrates a spectrogram generated using STFT with the selected parameters, while Figure 3 presents a resulting spectrogram after computing the LMFB, ultimately producing an array with dimensions 302 x 23 for each data sample.

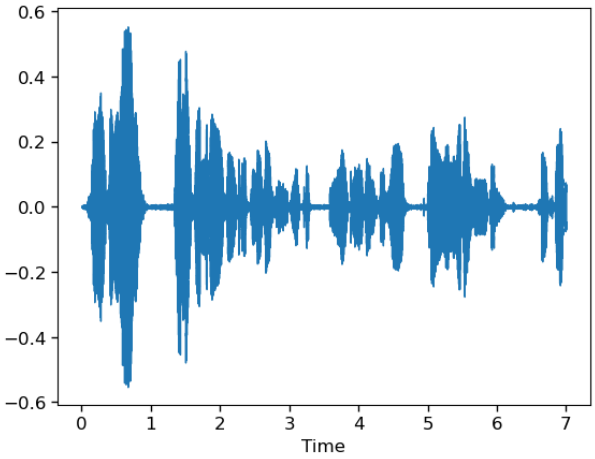


Figure 1: An example of an audio waveform, representing a neutral emotion from the THAI SER dataset

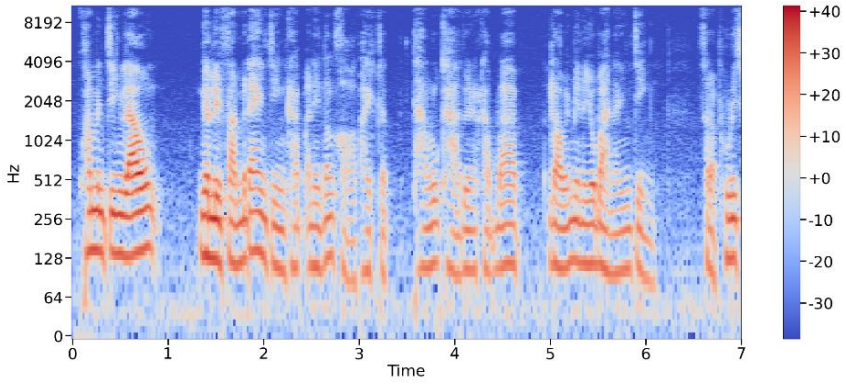


Figure 2: A spectrogram generated by applying STFT with n_fft = 2048 and hop_length = 512 to the audio file from Figure 1, resulting in an array with dimensions of 302 x 1025

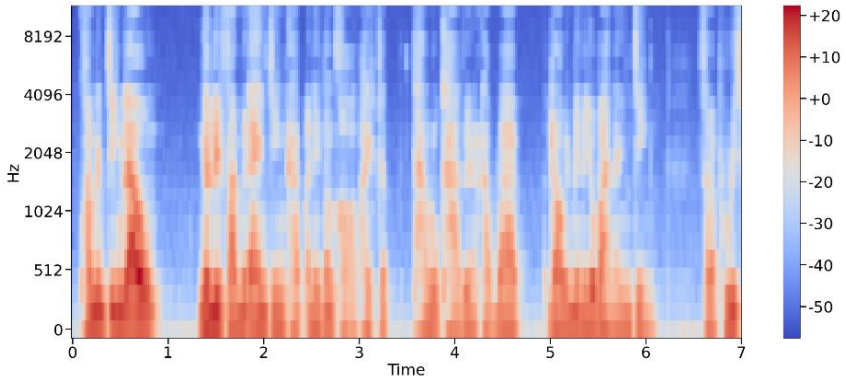


Figure 3: A spectrogram obtained after applying a 23-mel band LMFB to the output in Figure 2, reducing the array dimensions to 302 x 23, thus completing the preprocessing step

2.3 Model Evaluation

In multiclass classification, evaluating the performance of a model involves calculating metrics that provide insights into how well the model distinguishes between different classes. Recall, also known as sensitivity or true positive rate, is a key metric in this context. It measures the ability of a model to correctly identify all the relevant instances for each class. Unlike accuracy, which measures

the overall correctness of predictions, recall focuses on the effectiveness of identifying specific classes.

2.3.1 Recall in Multiclass Context

In multiclass settings, recall can be computed for each class individually. It answers the question: *For each class, how many of the actual instances were correctly identified by the model?* The calculation involves determining true positives (correct predictions for a class) and false negatives (instances of a class incorrectly predicted as another class). Equation (1) is for calculating recall for each class.

$$\text{Recall}_i = \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Negatives}_i} \quad (1)$$

Here, True Positives_i are the correctly predicted instances of class i, and False Positives_i are the instance of class *i*, that were missed by the model.

In this study, recall is prioritized as the primary performance metric due to the challenges posed by imbalanced datasets. Emotional categories, such as negative emotions, are often over-represented in datasets, while others are under-represented. In such cases, precision alone might not reflect the model's true performance, as it could prioritize majority classes and overlook the minority ones.

Recall provides a more sensitive evaluation of how well the model identifies all instances of a particular emotion, regardless of their frequency, making it particularly suitable for SER tasks. In real-world applications like mental health monitoring or customer service, missing key emotional states, especially minority emotions like anxiety or frustration can lead to inadequate responses. High recall ensures that the system captures the full range of emotional expressions.

Furthermore, many studies in the SER field emphasize recall due to its importance in emotion detection tasks, especially when dealing with class-imbalanced datasets. For instance, research such as [5] and [7] with the use of unweighted average recall (UAR), and many more that utilized unweighted accuracy (UA) calculated from an average of recalls of each class. By aligning with these established practices, we ensure consistency with prior work and focus on improving the detection of all emotion categories, particularly those that are less frequent but equally critical.

Finally, while recall is our primary focus, the study includes additional metrics like precision and F1 score in Appendix A for a more comprehensive view of the model's performance across all metrics.

To evaluate the overall model performance across all classes, the next two metrics are used.

2.3.2 Micro-Averaged Recall or Weighted Accuracy (WA)

This method aggregates the contributions from all classes to compute recall. It sums up the true positives and false negatives across all classes before computing the recall, giving more weight to classes with more instances. Micro-average recall is often used when you want to assess the model's overall performance, considering the size of each class.

$$\text{Micro Recall} = \frac{\sum_{i=1}^{N} \text{True Positives}_i}{\sum_{i=1}^{N} (\text{True Positives}_i + \text{False Negatives}_i)} \quad (2)$$

Micro-Averaged Recall, often referred to as "weighted accuracy" (WA), "normal accuracy", or simply "accuracy" in other studies, is a metric that measures the proportion of correctly predicted samples across all classes. The term "weighted" in this context refers to the fact that this measure accounts for the number of instances in each class, providing a sense of overall performance without being affected by class imbalances.

2.3.3 Macro-Averaged Recall or Unweighted Accuracy (UA)

This method calculates the recall for each class independently and then averages these values. It treats all classes equally, regardless of their frequency in the dataset. This approach is particularly useful when class distribution is imbalanced, as it emphasizes the model's performance on all classes, including minority ones.

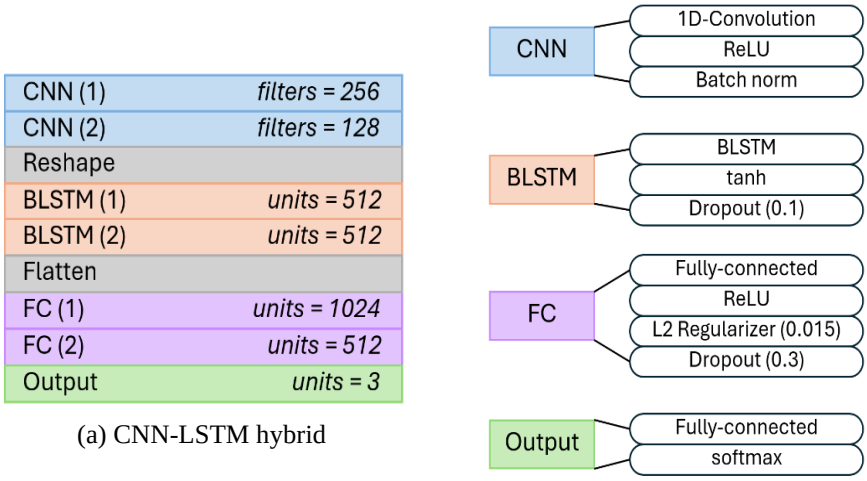
$$\text{Macro Recall} = \frac{1}{N}\sum_{i=1}^{N} \text{Recall}_i \quad (3)$$

Macro-Averaged Recall is often described as “unweighted accuracy (UA)” or “balanced accuracy”. It is termed "unweighted" because it does not weigh the contribution of each class by the number of instances in that class. Instead, it considers the performance across classes equally, which helps in giving a more “balanced view” of the model's performance. This metric is particularly valuable when the goal is to ensure that the model performs well across all classes, rather than excelling in classes with more samples while neglecting others. Since for imbalanced datasets UA is a more relevant characteristic, we rather concentrated our efforts on getting a high UA.

2.4 Model Architecture

The model implemented in subsequent experiments is the combination of 1-dimensional CNNs and BLSTMs. This integration leverages the strong feature extraction capabilities of CNNs at the initial stage, followed by the comprehensive sequential analysis of BLSTMs. Such a combination significantly enhances the model's proficiency in identifying subtle emotional nuances, which might be missed when each technique is applied independently. This architecture was selected for its widespread adoption and proven efficacy in the domain, as evidenced by its prevalence in scholarly literature and successful deployment in similar tasks.

Figure 4a outlines the model's architecture, presenting a sequence of modules along with their respective parameters. Figure 4b provides a more detailed exploration of each module described in Figure 4a. Note that all CNN layers are configured with a kernel size of 3 and a stride of 1. Notably, there is no max-pooling layer following these CNN layers, allowing for a continuous convolution process that preserves the spatial resolution throughout the network.



(a) CNN-LSTM hybrid

(b) Network modules

Figure 4: Model architecture used in the study. These diagrams detail the structure of the model

In the Tensorflow Keras implementation, each model was trained using the Adam optimizer with an initial learning rate of 0.0001. Sparse categorical crossentropy was chosen as the loss function. To optimize training, a ReduceLROnPlateau approach was applied, featuring a patience of 4 epochs and a multiplicative factor of 0.8, with monitoring focused on validation accuracy. The training process

extended over 40 epochs, using a batch size of 16. Model checkpoints were strategically implemented to save versions of the model that achieved the highest validation accuracy. To ensure robustness and reliability of the results, each training session was repeated 10 times. This repetition allowed for averaging the outcomes and recording evaluation metrics, including weighted accuracy (WA), unweighted accuracy (UA), and recalls for each class, along with their respective standard deviations (SD). This comprehensive approach ensured a detailed assessment of the model's performance across different runs.

2.5 Experiments

For all subsequent experiments, the test sets comprised approximately 2000 - 2200 samples from subsets of the THAI SER dataset. These samples were extracted semi-randomly each training from 100 predefined groups (80 professionally recorded and 20 recorded via Zoom as mentioned in *section 2.1.1*). The semi-random extraction was managed through an algorithm designed to select groups from both recording types (12 from professional and 3 from Zoom) to maintain a proportional representation of the original dataset. This ensures that the data samples preserve the original group ratios and exhibit consistent data distribution across the three emotion groups. Each resulting test set contains about 2000 - 2200 samples, mirroring the diversity and balance of the larger dataset.

After setting aside the test samples, the remaining data were adjusted to fit the amount requirements of each specific experiment and then further divided equally from each group into training and validation sets at an 84% to 16% ratio, respectively. This split was carefully calculated to ensure each training set was correctly sized for the various experimental setups. This systematic approach of dividing the dataset into training, validation, and test segments was consistently applied across all training instances, facilitating a rigorous and fair evaluation of model performance across different configurations.

2.5.1 Experiment 0: Same-corpus Training

The objective of Experiment 0 is to evaluate the performance of the model using a same-corpus SER approach on the THAI SER dataset, where both training and test data are sourced from the same corpus. This experiment serves to establish baseline metrics for comparison with cross-lingual results from later experiments.

Three distinct training sample sizes were selected to facilitate comparative analysis:

1) A full dataset with 10,000 samples used for training and approximately 2,000 for validation,
2) A medium dataset with 5,000 training samples and approximately 1,000 for validation,
3) A small dataset with 1,000 samples for training and about 200 reserved for validation.

For each sample size, the training was repeated 10 times to ensure robust results, with performance metrics averaged across the 10 instances. This yielded a total of 30 models (10 per sample size), with the corresponding results averaging into three sets of evaluation metrics. Standard deviations were also calculated to capture the variability across training runs, providing a comprehensive view of the model’s performance under different data constraints.

2.5.2 Experiment 1: Cross-lingual Training

The objective of Experiment 1 is to evaluate the models' capacity to generalize emotion recognition across languages by training on English and Mandarin datasets (from the ESD dataset) and testing on the Thai SER dataset without the use of transfer learning techniques.

For training, the full ESD English and ESD Mandarin datasets were used, each comprising 17,500 samples. These were split into 15,400 samples for training and 2,100 for validation, maintaining an 88% to 12% ratio.

Each training scenario was repeated 10 times, and the performance metrics were averaged across these iterations, resulting in 20 models in total, 10 trained on English and 10 on Mandarin. The evaluation metrics from these models were then averaged to create two distinct sets of results (one

for English and one for Mandarin), with standard deviations provided to illustrate the variability in performance across the multiple training runs.

2.5.3 Experiment 2: Cross-lingual Training with Transfer Learning

The objective of experiment 2 is to assess the effectiveness of transfer learning by adapting pre-trained models from English and Mandarin (from experiment 1) to the Thai linguistic context. The goal is to determine whether fine-tuning these models on Thai data can enhance cross-lingual emotion recognition performance.

Starting with the 20 models from Experiment 1 (10 models trained on English and 10 on Mandarin), each model underwent fine-tuning on the THAI SER dataset. During this process, the CNN layers responsible for extracting low-level features from the speech data were frozen, preserving their learned representations from the English and Mandarin datasets. The unfrozen layers were re-trained using varying sizes of Thai data, allowing the model to adapt to the Thai language's specific emotional patterns. This approach enables efficient adaptation while reducing the risk of overfitting, as the core feature extraction processes remain unchanged (Figure 5 illustrates this structure).

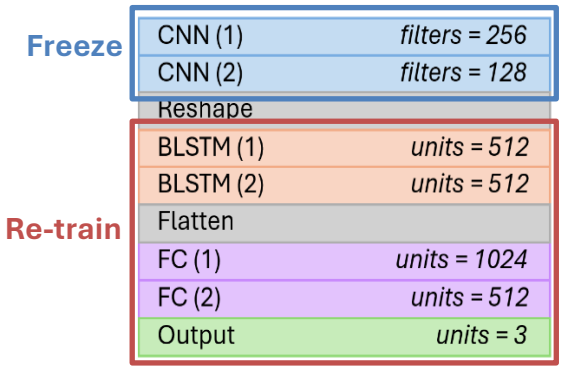


Figure 5: Fine-tuning process of models from Experiment 1. The two CNN layers (frozen in this step) retain the feature extraction capabilities learned from the English and Mandarin datasets, while the remaining layers are re-trained using Thai SER data of varying sizes to adapt the model to the Thai language

The three different transfer learning sizes of the Thai dataset are:

1) A full transfer dataset with 10,000 samples and approximately 2,000 for validation,
2) A medium transfer dataset with 5,000 samples and about 1,000 for validation,
3) A small transfer dataset with 1,000 samples and around 200 for validation.

This setup led to 60 models in total (20 initial models each fine-tuned on three different Thai dataset sizes). For each combination of language and dataset size, the 10 models were averaged, resulting in six distinct sets of evaluation metrics (three for English and three for Mandarin). Each set provides an average performance along with standard deviations to reflect the variability of the results across different training iterations.

The experimental design involving different sizes of the Thai SER dataset: 1,000, 5,000, and 10,000 samples, was specifically chosen to align with the availability of non-Thai samples, each numbering around 15,000. This setup not only facilitates straightforward comparative analysis due to round numbers but also tests the model's effectiveness across varying data volumes with a significant focus on scalability.

3. Research Results and Discussion

The primary objective of this study is to explore the potential of leveraging extensive annotated speech resources from well-resourced languages, specifically English and Mandarin, to enhance

speech emotion recognition (SER) capabilities in a linguistically limited dataset like Thai. This investigation is driven by the need to understand how effectively data from linguistically and culturally diverse languages can be applied to improve emotion recognition systems in languages with fewer resources.

The research questions are designed to probe deep into the comparative effectiveness of utilizing resources from different language families, comparing a set of European languages represented by English with a set of Asian languages represented by Mandarin. This comparison seeks to identify which language resources are more compatible with the Thai language and more effective at improving SER accuracy in a cross-lingual context. Furthermore, the study examines the scalability of these approaches: if the experiments were scaled up significantly, which language would prove more advantageous for enhancing the Thai SER systems?

3.1 Results of Experiment 1: Cross-lingual Training

Experiment 1: Cross-lingual Training tests the effectiveness of models trained on English and Mandarin datasets when applied to recognizing emotions in Thai speech. Table 3 outlines the performance outcomes of models trained on English and Mandarin datasets when tested on the Thai SER dataset. The values represent averaged results of 10 models for micro recall (WA), macro recall (UA), and emotion-specific recalls, adjusted by their standard deviations (SD). This presentation underscores the challenges and successes of cross-lingual emotional recognition, particularly highlighting the models' effectiveness in transferring learned emotional cues from one language to another.

Table 3: Evaluation metrics from Experiment 1: Cross-lingual Training. Each cell displays the average performance ± SD on Thai test set, using English and Mandarin as a training set without any transfer learning method over 10 training iterations. Values in bold indicate the highest value for each metric across the two training sets

Training Set	WA	UA	negative recall	positive recall	neutral recall
English	**49.67** ± 1.26	41.13 ± 1.47	**81.00** ± 4.65	35.76 ± 7.29	6.65 ± 2.30
Mandarin	48.82 ± 1.90	**48.72** ± 2.58	46.73 ± 9.94	**42.93** ± 5.26	**56.49** ± 13.00

The performance of models trained on English and Mandarin, then tested on Thai, highlighted the complexity involved in cross-lingual emotion recognition. Models trained on Mandarin demonstrated relatively robust generalization across all emotion groups, including neutral emotions, which typically had fewer data points. This resulted in both Weighted Accuracy (WA) and Unweighted Accuracy (UA) reaching similar levels, despite the high variability in recall values for negative and neutral emotions. In contrast, models trained on English excelled in recognizing negative emotions but showed marked difficulties with others, particularly neutral emotions. This significantly lowered the UA, suggesting potential cultural or linguistic biases in how emotions are expressed and recognized. These biases highlight the impact of linguistic nuances, as all other variables such as recording tools, speech translation, data quantity, and preprocessing methods were consistent across the datasets. The only variable was the language of the data, pinpointing the challenges in training SER systems across languages with inherently different emotional expressivity.

3.2 Combined results of Experiment 2 and Experiment 0

Experiment 2: Cross-lingual training with transfer learning explores how models pre-trained on English and Mandarin can be adapted through transfer learning techniques to improve performance on Thai data, accompanied by Experiment 0: Same-corpus Training which serves as a baseline, examining the performance of SER models solely within the Thai language dataset.

Table 4: Combined evalation metrics from Experiment 2 and Experiment 0 for comparitive analysis. The values in each cell reflect the average performance ± SD on Thai test set over 10 training iterations. Values in bold indicate the highest value for each metric across all training/transfer learning sizes

Training/Transfer Learning Size	Initial Models	WA	UA	negative recall	positive recall	neutral recall
Full	English	69.27 ± 0.82	66.81 ± 0.97	74.68 ± 2.95	56.47 ± 5.52	69.29 ± 4.82
	Mandarin	71.21 ± 0.97	69.20 ± 0.96	75.25 ± 3.72	59.91 ± 3.38	**72.45** ± 3.66
	Thai	**72.28** ± 1.18	**70.17** ± 1.11	**77.56** ± 2.80	**62.66** ± 3.23	70.28 ± 2.40

(a) Full training/transfer learning size

Training/Transfer Learning Size	Initial Models	WA	UA	negative recall	positive recall	neutral recall
Medium	English	66.25 ± 1.36	63.49 ± 1.40	71.87 ± 5.54	50.91 ± 4.24	67.69 ± 5.25
	Mandarin	66.75 ± 1.16	**64.58** ± 1.59	71.11 ± 2.51	**54.41** ± 4.93	**68.23** ± 5.38
	Thai	**67.15** ± 1.19	63.92 ± 2.01	**74.59** ± 2.76	51.20 ± 7.07	65.98 ± 3.77

(b) Medium training/transfer learning size

Training/Transfer Learning Size	Initial Models	WA	UA	negative recall	positive recall	neutral recall
Small	English	58.72 ± 1.61	54.17 ± 3.18	**68.32** ± 7.24	34.92 ± 9.34	59.28 ± 9.95
	Mandarin	**59.75** ± 2.41	**56.81** ± 2.55	65.08 ± 4.25	41.65 ± 4.88	**63.70** ± 3.69
	Thai	57.13 ± 1.84	54.20 ± 1.52	63.60 ± 5.09	**41.75** ± 4.86	57.25 ± 5.77

(c) Small training/transfer learning size

To facilitate the comparison between the two experiments, the results from both were tabulated based on the training and transfer learning sizes of the THAI SER dataset: a full dataset consisting of 10,000 samples, a medium dataset with 5,000 samples, and a small dataset comprising 1,000 samples. Table 4 presents the micro recall (WA), macro recall (UA), and recalls for negative, positive, and neutral emotion groups across various training and transfer learning sizes. The values in each cell reflect the average performance on Thai test set over 10 training iterations, with standard deviations indicating the variability of results (See Appendix B for individual experiment results).

Tables 4a and 4b illustrate that for both full and medium transfer learning sizes, models trained on Mandarin exhibited superior generalization across all emotional categories compared to their English counterparts, achieving higher accuracy metrics. However, it is notable that even these enhanced metrics did not surpass those achieved by the models trained solely on the Thai SER dataset within the same-corpus context. While the differences in performance metrics were not dramatically significant, they consistently demonstrated the robustness of the Mandarin models in these settings.

On the other hand, Table 4c reveals a particularly compelling outcome for the small transfer learning size. Here, Mandarin models not only outperformed the English models but also exceeded the performance of the models trained directly on the Thai SER dataset, moreover, displaying notably lower variability. This outcome highlights the potential effectiveness of Mandarin as a base for cross-lingual adaptation in SER applications, especially in scenarios involving limited training data.

3.3 Results Discussion

In Experiment 1, the exploration of cross-lingual SER capabilities without the aid of transfer learning or domain adaptation techniques and ensuring that the two datasets used are consistent with one another as possible, the study revealed a notably better compatibility between Mandarin and Thai compared to English. This outcome was evidenced by Mandarin's superior generalization across all emotional categories within the Thai SER context. The findings suggest that Mandarin, possibly due to linguistic and phonetic similarities or cultural proximities with Thai, facilitates a more effective cross-lingual adaptation in SER applications.

Despite the initial results of Experiment 1 showing moderate performance, with both Weighted Accuracy (WA) and Unweighted Accuracy (UA) hovering just below 50%, this underscored the need for implementing a form of transfer learning. In Experiment 2, a basic transfer learning approach was utilized, where the convolutional neural network (CNN) modules of the initial models were frozen to prevent changes in their parameters, allowing for the retraining of the remainder of the model using subsets of the Thai SER dataset. This approach was tested using varying sizes of the Thai dataset, comprising full (10,000 samples), medium (5,000 samples), and small (1,000 samples) datasets to facilitate a detailed comparison.

In Experiment 2, we conclude that even with basic transfer learning approach can significantly enhance the performance of both English and Mandarin models when applied to the Thai dataset, with more pronounced improvements observed as the size of the Thai data used for fine-tuning increased. This success demonstrates the efficacy of transfer learning in bridging linguistic gaps in SER applications, suggesting that adapting models across languages can effectively enhance emotion recognition capabilities. Notably, the models initially trained with Mandarin data consistently yielded higher WA and UA across all sizes of Thai data used for fine-tuning, compared to their English counterparts. Although these differences were not significant, they were consistent, suggesting a slightly better generalization capability of the Mandarin models to the Thai language context.

4. Conclusions and Recommendations

The experiments demonstrate that Mandarin exhibits more consistent compatibility with Thai in the context of Speech Emotion Recognition (SER) than English. This compatibility can significantly enhance SER performance for Thai, a language with limited resources, by leveraging the more widely available Mandarin dataset. Given the shared tonal characteristics and cultural similarities between Mandarin and Thai, the use of Mandarin data helps bridge the gap in Thai's limited dataset. However, despite this advantage, the performance gap between models trained on Mandarin and those trained solely on Thai datasets raises a critical question: *is cross-lingual training from Mandarin necessary when similar results can be achieved with native Thai data?* Nevertheless, these experiments were not just about matching performance but also exploring scalability. The success of Mandarin models, particularly with smaller subsets of Thai data, suggests the viability of scaling up cross-lingual systems. For instance, using 150,000 Mandarin samples (instead of 15,000 samples used in the experiments) to potentially enhance performance on all 10,000 Thai samples (instead of 1,000 samples used in the small size instance) could eliminate the need for extensive new data collection, which is a significant challenge in Thai SER development.

This scalable approach, backed by the results from the smaller dataset configurations, along with more advanced transfer learning or domain adaptation techniques, and a more refined feature extraction method proposes a strategic use of extensive Mandarin resources to bolster SER capabilities for Thai, suggesting a path forward where large-scale linguistic resources can be effectively leveraged to improve SER systems without the burdensome requirement for new data annotation in under-resourced languages. This strategic use of existing datasets could significantly

impact the development of SER applications, making them more accessible and effective in multilingual contexts.

Lastly, while this study provides valuable insights into cross-lingual SER, particularly the effectiveness of transfer learning techniques, it acknowledges that the models trained on Mandarin or English did not outperform those trained directly on Thai datasets. This highlights the limitations of using basic fine-tuning without exploring more advanced adaptation techniques. Future work could benefit from investigating methods such as domain adversarial training (DAT), which helps reduce discrepancies between source and target domains, or data augmentation strategies that increase the diversity and robustness of training data. Both techniques have been successfully used in similar tasks, such as in [4] and [6], respectively, and may provide improved generalization performance across languages.

5. Acknowledgements

I would like to express my deepest gratitude to my advisor, Prof. Seksan Kiatsupaibul, Ph.D., for his support, guidance, and valuable insights throughout this research. I also extend my appreciation to my family and friends for their constant encouragement. Finally, I am grateful to the Department of Statistics, Chulalongkorn Business School, for providing the resources and environment needed to complete this thesis. Thank you to all who made this journey possible.

6. References

[1] Ayadi ME, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 2011; 44(3): 572-587.

[2] Hongbo W, Yu L, Xiaoxiao Z, Xuyan T. Depression Speech Recognition With a Three-Dimensional Convolutional Network. Frontiers in Human Neuroscience. 2021; 15.

[3] Xutong L, Rongheng L. Speech Emotion Recognition for Power Customer Service. In 2021 7th International Conference on Computer and Communications (ICCC); 2021. p. 514-518.

[4] Milner R, Jalal MA, Ng RWM, Hain T. A Cross-Corpus Study on Speech Emotion Recognition. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); 2019. p. 304-311.

[5] Cai X, Wu Z, Zhong K, Su B, Dai D, Meng H. Unsupervised Cross-Lingual Speech Emotion Recognition Using Domain Adversarial Neural Network. In 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP); 2021. p. 1-5.

[6] Braunschweiler N, Doddipatla R, Keizer S, Stoyanchev S. A Study on Cross-Corpus Speech Emotion Recognition and Data Augmentation. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); 2021. p. 24-30.

[7] Neumann M, Thang Vu Ng. CRoss-lingual and Multilingual Speech Emotion Recognition on English and French. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018. p. 5769-5773.

[8] Xiao Z, Wu D, Zhang X, Tao Z. Speech emotion recognition cross language families: Mandarin vs. western languages. In 2016 International Conference on Progress in Informatics and Computing (PIC); 2016. p. 253-257.

[9] Latif S, Qayyum A, Usman M, Qadir J. Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages. In 2018 International Conference on Frontiers of Information Technology (FIT); 2018. p. 88-93.

[10] Mekruksavanich S, Jitpattanakul A, Hnoohom N. Negative Emotion Recognition using Deep Learning for Thai Language. In 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON); 2020. p. 71-74.

[11] Zehra W, Javed AR, Jalil Z, Gadekallu T, Kahn H. Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex & Intelligent Systems. 2021; 7.
[12] STEC- depa Thailand Artificial Intelligence Research Institute & Advanced Info Service. airesearch. [Online]; 2021. Available from: https://airesearch.in.th/releases/speech-emotion-dataset/.
[13] Zhou K, Sisman B, Liu R, Li H. Seen and Unseen emotional style transfer for voice conversion with a new emotional speech dataset. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021.
[14] McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, et al. librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th Python in Science Conference; 2015. p. 18-24.

7. Appendix A

A more comprehensive view of the model's performance in each experiment across all metrics, including recalls, precisions, and F1 scores of each class, starting with equations to compute each of the metric.

$$\text{Precision}_i = \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Positives}_i} \quad (4)$$

$$\text{Macro Precision} = \frac{1}{N}\sum_{i=1}^{N} \text{Precision}_i \quad (5)$$

$$\text{F1} - \text{Score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} = \frac{2 \times \text{True Positives}_i}{2 \times \text{True Positives}_i + \text{False Positives}_i + \text{False Negatives}_i} \quad (6)$$

$$\text{Macro F1} - \text{Score} = \frac{1}{N}\sum_{i=1}^{N} \text{F1} - \text{Score}_i \quad (7)$$

In single-label multi-class classification scenario such as in this study, the values for micro recall, micro precision, and micro F1 score are equal to one another, thus sharing a single value for each experiment.

Table 5: Evaluation metrics from Experiment 1: Cross-lingual Training with all the metrics. The values in each cell represent the average performance ± SD on Thai test set over 10 training iterations

Training Set	micro recall (WA)	macro recall (UA)	negative recall	positive recall	neutral recall
English	49.67 ± 1.26	41.13 ± 1.47	81.00 ± 4.65	35.76 ± 7.29	6.65 ± 2.30
Mandarin	48.82 ± 1.90	48.72 ± 2.58	46.73 ± 9.94	42.93 ± 5.26	56.49 13.00

(a) Recall related metrics from Experiment 1

Training Set	micro precision	macro precision	negative precision	positive precision	neutral precision
English	49.67 ± 1.26	47.57 ± 2.54	50.99 ± 0.74	45.34 ± 4.31	46.39 ± 6.67
Mandarin	48.82 ± 1.90	47.66 ± 2.13	57.99 ± 3.75	36.89 ± 5.94	48.11 ± 2.73

(b) Precision related metrics from Experiment 1

Training Set	micro F1 score	macro F1 score	negative F1 score	positive F1 score	neutral F1 score
English	49.67 ± 1.26	37.79 ± 1.79	62.53 ± 1.60	39.38 ± 3.96	11.46 ± 3.59
Mandarin	48.82 ± 1.90	47.14 ± 1.76	50.97 ± 5.78	39.33 ± 3.81	51.12 ± 5.41

(c) F1 score related metrics from Experiment 1

Notably, in Experiment 1, the difference between recall and precision is especially evident in cross-lingual training with English data. In Table 5a, while the recall for negative emotions reached 81%, the recall for neutral emotions was only 6.65%, showing poor generalization across emotion groups. In contrast, in Table 5b, precision values were more balanced, indicating that recall is better suited for identifying how well the model captures each emotional category, especially in imbalanced scenarios like SER. This justifies the use of recall as a primary metric, as it ensures the system recognizes emotions across all classes, not just the majority ones.

Table 6: Combined evalation metrics from Experiment 2 and Experiment 0 in all metrics for full training/transfer learning size. The values in each cell reflect the average performance ± SD on Thai test set over 10 training iterations.

Training/Transfer Learning Size	Initial Models	micro recall (WA)	macro recall (UA)	negative recall	positive recall	neutral recall
Full	English	69.27 ± 0.82	66.81 ± 0.97	74.68 ± 2.95	56.47 ± 5.52	69.29 ± 4.82
	Mandarin	71.21 ± 0.97	69.20 ± 0.96	75.25 ± 3.72	59.91 ± 3.38	72.45 ± 3.66
	Thai	72.28 ± 1.18	70.17 ± 1.11	77.56 ± 2.80	62.66 ± 3.23	70.28 ± 2.40

(a) Recall related metrics for full training/transfer learning size

Training/Transfer Learning Size	Initial Models	micro precision	macro precision	negative precision	positive precision	neutral precision
Full	English	69.27 ± 0.82	68.16 ± 0.76	72.23 ± 1.40	63.56 ± 3.83	68.69 ± 3.91
	Mandarin	71.21 ± 0.97	70.18 ± 1.39	73.81 ± 1.62	65.69 ± 3.16	71.04 ± 3.65
	Thai	72.28 ± 1.18	70.99 ± 1.40	75.73 ± 1.31	67.17 ± 3.20	70.08 ± 2.21

(b) Precision related metrics for full training/transfer learning size

Training/Transfer Learning Size	Initial Models	micro F1 score	macro F1 score	negative F1 score	positive F1 score	neutral F1 score
Full	English	69.27 ± 0.82	67.19 ± 0.77	73.38 ± 1.30	59.46 ± 1.94	68.72 ± 0.97
	Mandarin	71.21 ± 0.97	69.52 ± 0.99	74.43 ± 1.30	62.55 ± 1.69	71.57 ± 0.95
	Thai	72.28 ± 1.18	70.48 ± 1.12	76.60 ± 1.35	64.71 ± 1.47	70.14 ± 1.60

(c) F1 score related metrics for full training/transfer learning size

Table 7: Combined evalation metrics from Experiment 2 and Experiment 0 in all metrics for medium training/transfer learning size. The values in each cell reflect the average performance ± SD on Thai test set over 10 training iterations.

Training/Transfer Learning Size	Initial Models	micro recall (WA)	macro recall (UA)	negative recall	positive recall	neutral recall
Medium	English	66.25 ± 1.36	63.49 ± 1.40	71.87 ± 5.54	50.91 ± 4.24	67.69 ± 5.25
	Mandarin	66.75 ± 1.16	64.58 ± 1.59	71.11 ± 2.51	54.41 ± 4.93	68.23 ± 5.38
	Thai	67.15 ± 1.19	63.92 ± 2.01	74.59 ± 2.76	51.20 ± 7.07	65.98 ± 3.77

(a) Recall related metrics for meduim training/transfer learning size

Training/Transfer Learning Size	Initial Models	micro precision	macro precision	negative precision	positive precision	neutral precision
Medium	English	66.25 ± 1.36	64.85 ± 1.94	69.44 ± 1.88	58.53 ± 3.33	66.58 ± 4.83
	Mandarin	66.75 ± 1.16	65.31 ± 1.00	70.39 ± 2.47	59.31 ± 3.43	66.23 ± 3.26
	Thai	67.15 ± 1.19	65.66 ± 1.39	70.14 ± 2.30	60.18 ± 3.52	66.66 ± 2.94

(b) Precision related metrics for meduim training/transfer learning size

Training/Transfer Learning Size	Initial Models	micro F1 score	macro F1 score	negative F1 score	positive F1 score	neutral F1 score
Medium	English	66.25 ± 1.36	63.84 ± 1.30	70.46 ± 1.87	54.24 ± 2.01	66.82 ± 2.25
	Mandarin	66.75 ± 1.16	64.72 ± 1.23	70.68 ± 1.23	56.48 ± 2.07	67.00 ± 2.40
	Thai	67.15 ± 1.19	64.45 ± 1.65	72.22 ± 0.97	54.91 ± 3.66	66.22 ± 2.17

(c) F1 score related metrics for meduim training/transfer learning size

Table 8: Combined evalation metrics from Experiment 2 and Experiment 0 in all metrics for small training/transfer learning size. The values in each cell reflect the average performance ± SD on Thai test set over 10 training iterations.

Training/Transfer Learning Size	Initial Models	micro recall (WA)	macro recall (UA)	negative recall	positive recall	neutral recall
Small	English	58.72 ± 1.61	54.17 ± 3.18	68.32 ± 7.24	34.92 ± 9.34	59.28 ± 9.95
	Mandarin	59.75 ± 2.41	56.81 ± 2.55	65.08 ± 4.25	41.65 ± 4.88	63.70 ± 3.69
	Thai	57.13 ± 1.84	54.20 ± 1.52	63.60 ± 5.09	41.75 ± 4.86	57.25 ± 5.77

(a) Recall related metrics for small training/transfer learning size

Training/Transfer Learning Size	Initial Models	micro precision	macro precision	negative precision	positive precision	neutral precision
Small	English	58.72 ± 1.61	57.32 ± 1.60	60.71 ± 2.89	51.65 ± 2.82	59.61 ± 3.76
	Mandarin	59.75 ± 2.41	57.97 ± 2.78	62.77 ± 2.45	51.52 ± 5.10	59.63 ± 2.92
	Thai	57.13 ± 1.84	55.39 ± 2.13	60.52 ± 1.88	49.07 ± 4.90	56.58 ± 2.00

(b) Precision related metrics for small training/transfer learning size

Training/Transfer Learning Size	Initial Models	micro F1 score	macro F1 score	negative F1 score	positive F1 score	neutral F1 score
Small	English	58.72 ± 1.61	54.54 ± 2.77	63.96 ± 2.58	40.94 ± 7.05	58.72 ± 4.88
	Mandarin	59.75 ± 2.41	57.08 ± 2.59	63.84 ± 2.73	45.89 ± 4.03	61.51 ± 2.41
	Thai	57.13 ± 1.84	54.51 ± 1.64	61.89 ± 2.52	44.92 ± 4.03	56.72 ± 2.71

(c) F1 score related metrics for small training/transfer learning size

In both Experiments 0 and 2, the results for precision and F1 score largely reflect the same patterns observed with recall. Specifically, in cases where recall values are high or low for certain emotion groups or dataset sizes, the corresponding precision and F1 score metrics follow the same trends. This suggests that unlike in Experiment 1 where recall and precision showed contrasting behaviors, in Experiments 0 and 2, all three metrics—recall, precision, and F1 score—are closely aligned, indicating consistent model performance across these different evaluation measures.

Appendix B

Table 9 and Table 10 show individual results from Experiment 0 and Experiment 2, respectively, as opposed to the combined result seen in Table 4.

Table 9: Evaluation metrics from Experiment 0: Same-corpus Training. The values in each cell represent the average performance ± SD on Thai test set over 10 training iterations

Training Set	Training size	WA	UA	negative recall	positive recall	neutral recall
Thai	Full	72.28 ± 1.18	70.17 ± 1.11	77.56 ± 2.80	62.66 ± 3.23	70.28 ± 2.40
	Medium	67.15 ± 1.19	63.92 ± 2.01	74.59 ± 2.76	51.20 ± 7.07	65.98 ± 3.77
	Small	57.13 ± 1.84	54.20 ± 1.52	63.60 ± 5.09	41.75 ± 4.86	57.25 ± 5.77

Table 10: Evaluation metrics from Experiment 2: Cross-lingual with Transfer Learning. These tables present the average performances ± SD on Thai test set over 10 training iterations

Initial Model	Transfer Learning size	WA	UA	negative recall	negative recall	neutral recall
English	Full	69.27 ± 0.82	66.81 ± 0.97	74.68 ± 2.95	56.47 ± 5.52	69.29 ± 4.82
	Medium	66.25 ± 1.36	63.49 ± 1.40	71.87 ± 5.54	50.91 ± 4.24	67.69 ± 5.25
	Small	58.72 ± 1.61	54.17 ± 3.18	68.32 ± 7.24	34.92 ± 9.34	59.28 ± 9.95

(a) Evaluation metrics from Experiment 2 with initial models trained on English data

Initial Model	Transfer Learning size	WA	UA	negative recall	positive recall	neutral recall
Mandarin	Full	71.21 ± 0.97	69.20 ± 0.96	75.25 ± 3.72	59.91 ± 3.38	72.45 ± 3.66
	Medium	66.75 ± 1.16	64.58 ± 1.59	71.11 ± 2.51	54.41 ± 4.93	68.23 ± 5.38
	Small	59.75 ± 2.41	56.81 ± 2.55	65.08 ± 4.25	41.65 ± 4.88	63.70 ± 3.69

(b) Evaluation metrics from Experiment 2 with initial models trained on Mandarin data