Flinders UNIVERSITY
ADELAIDE · SOUTH AUSTRALIA

A Bi-Lingual Speech Emotion Recognition Model through Image Processing on Spectral Features

By

Xiaoyu Chen

Thesis
Submitted to Flinders University
for the degree of

<Master of Biomedical Engineering>

<College of Science and Engineering>
<23/10/2023>

TABLE OF CONTENTS

ABSTRACT

Speech Emotion Recognition is an emerging research field due to its potential applications in medical fields, commercial settings, and voice assistance development. This thesis provides a comprehensive investigation into Speech Emotion Recognition (SER) with a focus on the influence of various image processing techniques and feature fusion. The study begins with a Background chapter, introducing the significance of SER in human-computer interaction and its potential applications in diverse fields. In the Literature Review, existing research on SER and its challenges are discussed, leading to the identification of gaps which highlight the need for cross-lingual robustness, feature enhancement, and reduction in model bias toward specific emotions.

The Aim and Hypothesis chapters articulate the study's objectives and the hypothesis that different image processing techniques and feature fusion can enhance SER model performance. The methodology utilized various tools and packages, a diverse dataset, and specific acoustic features, such as Mel Spectrograms and MFCCs, alongside image processing techniques including DoG, Sobel, and CLAHE filters. The SER model in this paper was TIM-Net. A parallel bi-lingual dataset, ESD was also used for training. 10-fold cross validation was used for training. For comparing performances, accuracy, average recall, and confusion matrix were used. Statistical significance was indicated by a p value less than 0.05 between baseline results and other results.

The Results chapter reveals significant findings. Visual analysis highlights spectral differences between emotional and neutral speech, particularly across languages. DoG filtering enhances model accuracy, but also increases confusion between certain emotions. In contrast, Sobel and CLAHE filtering reduce accuracy and increase language-based disparities, which could potentially be due to loss of relevant and subtle spectral information with overprocessing.

In conclusion, the model trained with the fused feature displayed the best average accuracies and more balanced predictions across different emotions and languages. For future direction, spectral features across different emotions could be analysed on a deeper and for each filter could be fine-tuned to balance between the feature enhancement and potential drawbacks. By visually understanding the spectral features, a multi-stage customised adaptive filter could also be developed to enhance the most relevant features. Before applying the model to different applications, the model should be trained with real-world data to further improve the robustness against noises and generalisability for cross-lingual model.

DECLARATION

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signature of student.....Xiaoyu Chen.....

Print name of student.....Xiaoyu Chen.....

Date.....23/10/2023.....

I certify that I have read this thesis. In my opinion it is/is not (please circle) fully adequate, in scope and in quality, as a thesis for the degree of Master of Biomedical Engineering. Furthermore, I confirm that I have provided feedback on this thesis and the student has implemented it.

Signature of Principal Supervisor.....

Print name of Principal Supervisor.....

Date.....

ACKNOWLEDGEMENTS

I would like to express my deep appreciation and gratitude to my supervisor, Russell, for his unwavering support, guidance, and expertise throughout the course of this thesis. His mentorship and commitment played an invaluable role in shaping this research and enhancing its quality.

I am also thankful for the encouragement and wisdom Russell provided, which empowered me to overcome challenges and encouraged my intellectual growth. The insightful feedback, patience, and dedication were instrumental in navigating the complexities of the research process.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

The power of voice assistant transcended its conventional role in a domestic violence incident in New Mexico in 2017, as it played a pivotal part in potentially saving a life. In this case, the virtual assistant was activated unintentionally and called the emergency department during a heated argument between the couple **(Hassan, 2017)**. This case has sparked broader discussions about the importance of virtual assistance being able to detect an emergency by recognising emotions such as distress and other critical emotional cues during the interaction. In the age of Artificial Intelligence (AI), where human-machine interaction (HMI) has become more prevalent, understanding, and recognising human emotions holds significant promise as it enables machines to engage with users on a more profound and empathetic level. This field of study is known as Emotion Recognition in which facial emotion recognition has advanced significantly. Speech Emotion Recognition (SER), on the other hand, has only emerged in the past decade. Speech conveys paralinguistic information and reflects the mental and affective state of an individual. The incident mentioned above highlights the pressing need for accurate SER in virtual assistance technology. Users have often reported frustration when a virtual assistant fails to recognise their distress while seeking assistance in an urgent situation **(Plante, 2022)**. The consequences of this incapability are not only inconvenient but life-altering in some cases. Equipping virtual assistance with SER makes them more reliable and sensitive partners in our daily lives.

The potential impact of SER can be found across multiple domains. It is estimated that 1 in 5 Australians experiences a mental disorder **(Australian Institute of Health and Welfare, 2023)**. Early detection and intervention using SER technology can be helpful for recovery. Autism, which is also referred as autism spectrum disorder, consists of a diverse range of conditions related to brain development. It affects approximately 1 in 100 children worldwide. One characteristic of autism is the lack of ability to effectively identify emotions in communication **(World Health Organization, 2023)**. The development of SER can be incorporated in psychosocial therapy to assist neurodiverse children so that they can interact socially and communicate effectively. In addition to applications in the health sector, SER can be deployed in call centres for providing feedback to operators for monitoring purposes and improving customer experience. It can also be used to process voice message so that those that show signs of an emergency can be prioritised **(Petrushin, 2000)**. Understanding speech emotion can provide valuable insights into cross-linguistic communication and the universal aspects of emotional expression. SER can potentially help bridge communication gaps in multilingual or cross-cultural interactions by detecting and conveying the emotional tone, allowing for better understanding even when language itself might be a barrier. This can be a potential application for real-time translator that also conveys emotions to improve the cross-cultural interactions.

SER has gained increasing attention in recent years driven by its potential applications, which extend beyond the ones discussed above. The development of a SER model includes an emotional speech dataset, pre-processing and feature extraction and emotion classification.

1.1.1 Dataset

Datasets can be divided into three types based on how speech data were collected. The most common type is acted datasets where speech is produced by well-trained actors performing different acted emotions. Elicited datasets are conversations with induced emotions without the actor's knowledge. Natural datasets are speech data collected from podcasts, reality TV shows, call centres and more **(Wani, et al., 2021) (Singh & Goel, 2022)**.

1.1.2 Pre-processing and Feature Extraction

Speech data are pre-processed and relevant features are extracted to be used for training the SER model. Pre-processing include normalisation, augmentation, rotation, noise reduction and data cleaning which aim to reduce the effects of variations in the quality of speech data **(Nema & Abdul-Kareem, 2018)**. The two most predominant acoustic features that have been used in the field of SER are prosodic and spectral features. Prosodic features are high-level features, such as pitch, tone, and speech rate, which can be quantified using parameters, such as root mean square, zero-crossing rate, fundamental frequency and more **(Milton & Tamil, 2013)**. By converting speech signals in time domain to frequency domain, spectral features can be obtained. The most used spectral features can be divided into two types, spectrogram, and Cepstral Coefficient. Spectrograms are visual representations of energy distribution of speech signals across time and frequency. Cepstral Coefficients are higher-order features that are derived from Spectrograms and they can be visually represented as well with x-axis being the time frame and y-axis being the index of coefficients.

1.1.3 Classification

Classifiers are machine learning (ML) or deep learning (DL) models that are trained to predict speech emotions based on selected acoustic features. Complex models combine multiple DL techniques to capture the spatio-temporal features in speech signals **(Zhou, et al., 2018)**. While the performance of the developed SER model increases with the increased complexity, it becomes more difficult to interpret parameters that the model chose based on specific features during training.

1.1.4 Image Processing on Spectral Features

As suggested in section 1.1.2 and 1.1.3, contemporary SER models have leaned heavily on complex classification models, which requires substantial amount of data to train for robust performance. However, they often lack transparency in terms of understanding the decision-making processes and the relative significance corresponding to different acoustic features.

Several studies have incorporated image processing techniques to reduce noise and enhance contrast of spectral features (mainly on spectrograms) in the fields of speech recognition (Cadore, et al., 2011) and acoustic detection (Fang, et al., 2022). Common approaches such as edge detection, contrast enhancement on spectrogram have demonstrated significant improvement on prediction accuracy. This suggests a potential avenue for the application of image processing techniques to improve the robustness of the SER model and reduce the needs for excessive amount of training data for desired performance levels.

1.2 Thesis Outline

A comprehensive literature review is covered in the next section to highlight the developments in the field in terms of emotional datasets, feature extraction and classification. Limitations and gaps are identified by evaluating current trends in the field of SER. A novel approach is introduced based on the gaps identified. Hypothesis, aims, objectives, scopes, and constraints are also outlined at the end of Section 2. Section 3 establishes the methodology and resources used to ensure the validity and repeatability of the study. Results are presented in Section 4 with visual representations and statistical analysis for different approaches. These results are then discussed and analysed in Section 5 which leads to the conclusions and future work in Section 6 and Section 7 respectively.

2. LITERATURE REVIEW

Developing a SER model includes selection of a dataset, speech feature extraction and classification. This section aims to investigate and compare current emotional speech datasets, feature extraction techniques, and classification models; to identify challenges and gaps in the research and propose a potential solution to bridge the gap.

2.1 Speech Emotion Recognition

2.1.1 Dataset

It is essential to select a suitable dataset for training and evaluating the performance of a SER model **(El Ayadi, et al., 2011)**. Based on how speech data was collected, datasets can be categorised into three types – natural, acted and induced emotional speech datasets.

Natural datasets are developed on data of spontaneous speech, which include the data recorded from Podcasts, reality TV shows, call centre conversations and more. The FAU AIBO emotion datasets recorded spontaneous interactions between children and Sony's pet robot AIBO in German. It consists of 51 children speaking 110 dialogues. 11 emotions were identified and labelled by five human labellers **(Steidl, 2009)**.This dataset provided a realistic interaction that reflect natural emotions in a human-machine interaction context. However, the dataset was designed to focus on a specific context which is the interactions with the robotic dogs, and it may not represent emotions in other real-life scenarios. A SER model developed with this dataset may not perform as well in older populations. Belfast is another natural dataset that built with TV interviews in French **(Douglas-Cowie, et al., 2003)**. Vidrascu and Devillers **(2005)** developed a model based on data obtained from call centres that contain real operator-client recordings. The challenges of using these two natural datasets are that there is limited control over variables that could affect the performance of the SER model, such as age, gender, sample duration, recording quality, and more. It is also difficult to ensure data balancing. In other words, these datasets may have imbalanced distributions of emotions, with some emotion being overrepresented and others being underrepresented **(Madanian, et al., 2023)**. This can potentially lead to bias in developed models, as they may perform better on more frequently represented emotions.

Even though acted and induced datasets are unable to fully represent real-life scenarios, they are prevalently used in SER study for several reasons. Contrary to natural datasets, acted and induced datasets allow researchers to have precise control over variables mentioned above, so that all data collected are consistent. Natural databases often contain various background noises, overlapping speech, and confounding factors that can make it more challenging to isolate and analyse the emotional content. Acted datasets, on the other hand, minimises these factors. Additionally, acted databases also provide ground truth labels for emotions. These labels are critical for training and evaluating SER models. As discussed previously, labelling emotions for data of natural

databases like FAU AIBO is subjective (Madanian, et al., 2023). Appendix C Contains a summary of frequently used acted and induced datasets in the field of SER.

Table 1. Acted and Induced Datasets for training SER models

According to the data presented in the table, the majority of the emotional speech datasets are based on English, European languages, and Hindi. Notably, most of these datasets are mono-lingual in nature, focusing on a single language. Emotion Speech Dataset (ESD) has bridged this gap by developing a parallel dataset for English and Chinese (Zhou, et al., 2022). It encompasses 350 different utterances that were collected in a low-noise environment. Ten English actors and ten Chinese actors each speak the 350 utterances delivering five different emotions (anger, happiness, neutral, sadness, and surprise). Referring Figure 1., an utterance sharing the same emotional context was spoken by two different speaks and an utterance with two different emotions was spoken by the same person. The advantage of the parallel nature of the speech data is that for each utterance in one language or emotion, there is a corresponding utterance in another language or emotion. This direct correspondence makes it easier for the model to learn the relationships between the acoustic features and the emotional states.

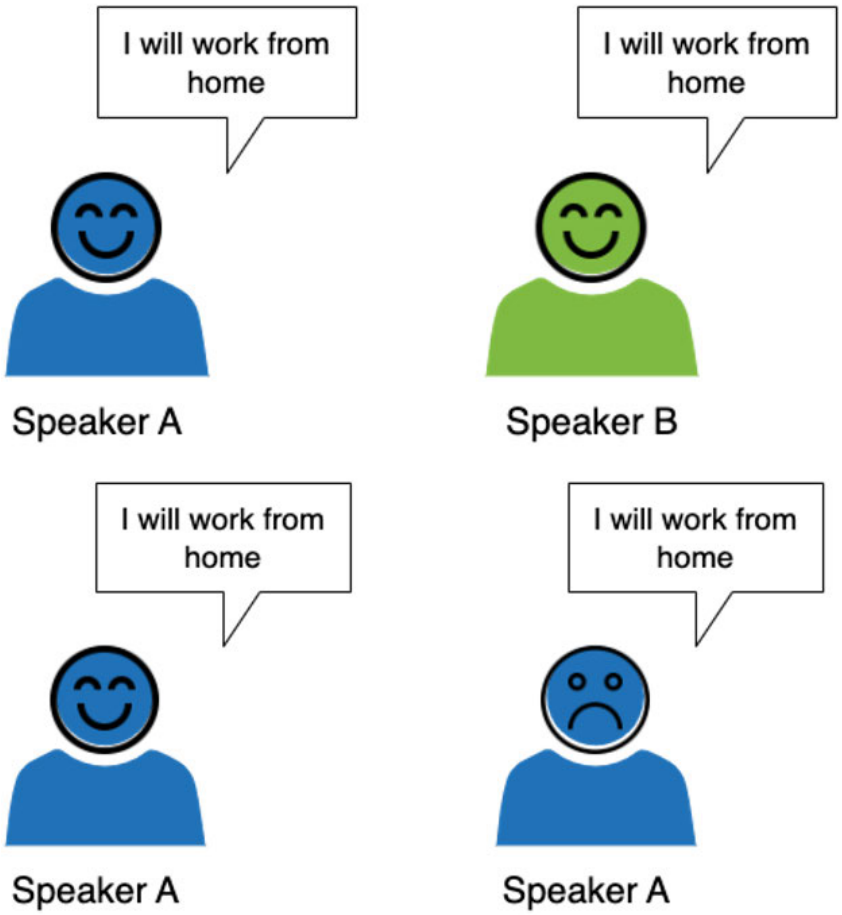


Figure 1. An example of parallel speech data for ESD. Blue and green represent two speakers that speak different languages (Chinese and English). The faces denote emotional states (happiness and sadness)

2.1.1 Feature Extraction

The identification and extraction of acoustic features that are related to different emotions is a challenging task, as emotions can be expressed in various ways, such as changes in volume, rhythm, tone, pitch, and other characteristics. Prosodic features are a type of acoustic features that were traditionally used for speech analysis. They reflect the flow of the speech including duration, intensity, intonation and more (Swain, et al., 2018). Some prosodic characteristics can be quantified by parameters such as spectral centroid, spectral contrast, zero-crossing rate, and fundamental frequency. Descriptions of these features and how they can be quantified are summarised in Table 2. Frick (1985) conducted a thorough review on how prosodic features vary in different emotions. Anger and happiness are reflected by increased loudness, pitch, and faster speech rate, whereas boredom and sadness have slower speech rate and low pitch. Table 3 highlights how emotions differ in prosodic features (Murray & Arnott, 1993) (Anagnostopoulos & Iliou, 2010).

Table 2. Quantifying Prosodic Features

Quantifiable Parameters	Description	Related Prosodic Features	Indication
Spectral Centroid (SC) (Grey & Gordon, 1978)	Frequency distribution trend in each frame	Timbre (quality of voice differs in tone and wave form)	Low SC indicates deeper sound, high SC indicates brisk sound
Spectral Flatness (SF) (Johnston, 1988)	Energy distribution in frequency bands	Tone	Normalised between 0~1, the speech is tonal is the value is closer to 1
Zero-Crossing Rate (ZCR) (Gouyon, et al., 2000)	How many times the signal passes 0 axis	Pitch	Higer ZCR correlated to high-pitch and percussive sound
Fundamental Frequency (F_0) (Dilley & Heffner, 2013)	Perceived pitch of the voice represented by average number of oscillations in a time frame	Pitch and Intonation	F_0 changes more extremely with varying pitch and intonation

Modern approaches have moved away from prosodic features as they are susceptible to noise and lack the discriminative power needed to distinguish subtle variations in emotions. Spectral features such as Mel Spectrogram and MFCC contains rich acoustic information that are more effective in distinguishing between different emotions. They are also robust against noise, making them more suitable for real-life situations (Swain, et al., 2018). Spectral features are derived from original speech signal in time domain, using pre-emphasis, framing, windowing, then a Fourier Transform or filter bank to convert the signal into frequency domain. Figure 2 illustrates a typical process of extracting Mel Spectrograms. It starts with preprocessing the raw data with a high pass filter (Eq. (1)) to empathise high frequency as when the speech signal is typically centred around lower frequencies, resulting in muffled sound. The high-pass filter is expressed below with coefficient α being between 0.9 and 1,

$$y(n) = x(t) - \alpha x(t-1) \qquad (1)$$

Framing blocks, defined by Eq. (2), will then be applied to pre-emphasised signals. Each frame contains 512 sampling points (N), and a Hamming window will be subsequently applied to each frame. S(n) is the speech signal after framing and W(n) is that after windowing.

$$S'(n) = S(N) - W(N) \qquad (2)$$

$$W(n,a) = (1-a) - a\,cos\left[\frac{2\pi n}{N-1}\right], 0 \leq n \leq N-1$$

The signal then will be converted into frequency domain by applying Fourier Transform. A bank of triangular filters will then be applied to the signal to produce a spectrogram with Mel scale (short for melody scale), which is a perceptual frequency scale to represent how humans perceive sound (Lee, et al., 2022). Mel Spectrogram has been widely used as it is an emulation of human auditory system, which uses non-linear frequency mapping and allocates more space to lower frequencies, where humans are more sensitive to changes in pitch, and compresses the space for higher frequencies. It has proven to achieve high performance combined with convolutional neural network classification models (Pandey, et al., 2019) (Toyoshima, et al., 2023) (Bhangale & Mohanaprasad, 2021).

Figure removed due to copyright restriction.

Figure 2. The process of Mel Spectrogram extraction (Habib, et al., 2021)

Mel Frequency Cepstral Coefficients (MFCCs) can be extracted from applying logarithmic filter bank and Discrete Cosine Transform (Eq. (3)). L is the number of filters applied, which determines the number of resulting coefficients (Li, et al., 2005). The input is denoted by m(l) which is the Mel-scaled output from Mel Spectrogram extraction.

$$c_{MFCC}(i) = \sqrt{\frac{2}{L}} \sum_{l=1}^{L} \log m(l) \cos\left[\frac{\left(l - \frac{1}{2}\right) i\pi}{L}\right] \quad (3)$$

Using MFCCs has the advantage reduced dimensionality of the input which can lead to faster training and more efficient modelling. They can also be helpful for capturing temporal patterns as they are computed as a sequence of feature vectors over time, allowing for the capture of statistical information about the distribution of acoustic features (Madanian, et al., 2023). 39 MFCCs are often used and consists of the following information (Cen, et al., 2016),

- Total Energy – MFCC 0
- Spectral Shape – MFCC 1–12
- Spectral Dynamics – MFCC 13-24
- Derivatives of Spectral Dynamics – MFCC 25-36
- Higher-Order Information – MFCC 37&38

Unlike Mel Spectrogram, MFCCs can be trained with both traditional machine learning models and deep learning models due to its reduced dimensionality. Machine learning SER models trained with MFCCs have shown promises in accurately predicting emotions. These models include variations of Support Vector Machine (SVM) (Gao, et al., 2017) (Kerkeni, et al., 2019), Multilayer Perceptron (MLP) (Chen, et al., 2012), and k Nearest Neighbours (kNN) (Rieger, et al., 2014).

There has not been a consensus on if MFCCs performs better than Mel Spectrogram, or vice versa. However, fusing the two features along with prosodic features has shown improved performance (Muljono, et al., 2023). On the other hand, with increased input size and dimensionality, it may require more complex models and higher computational cost for training.

An example for a pair of Mel spectrogram and MFCCs are shown below in Figure 3. With converting the signal from time domain to frequency and cepstral domain, and being represented visually, these features can potentially be treated as 2-dimensonal grey-scale pictures for further processing with image processing filters to enhance certain features.

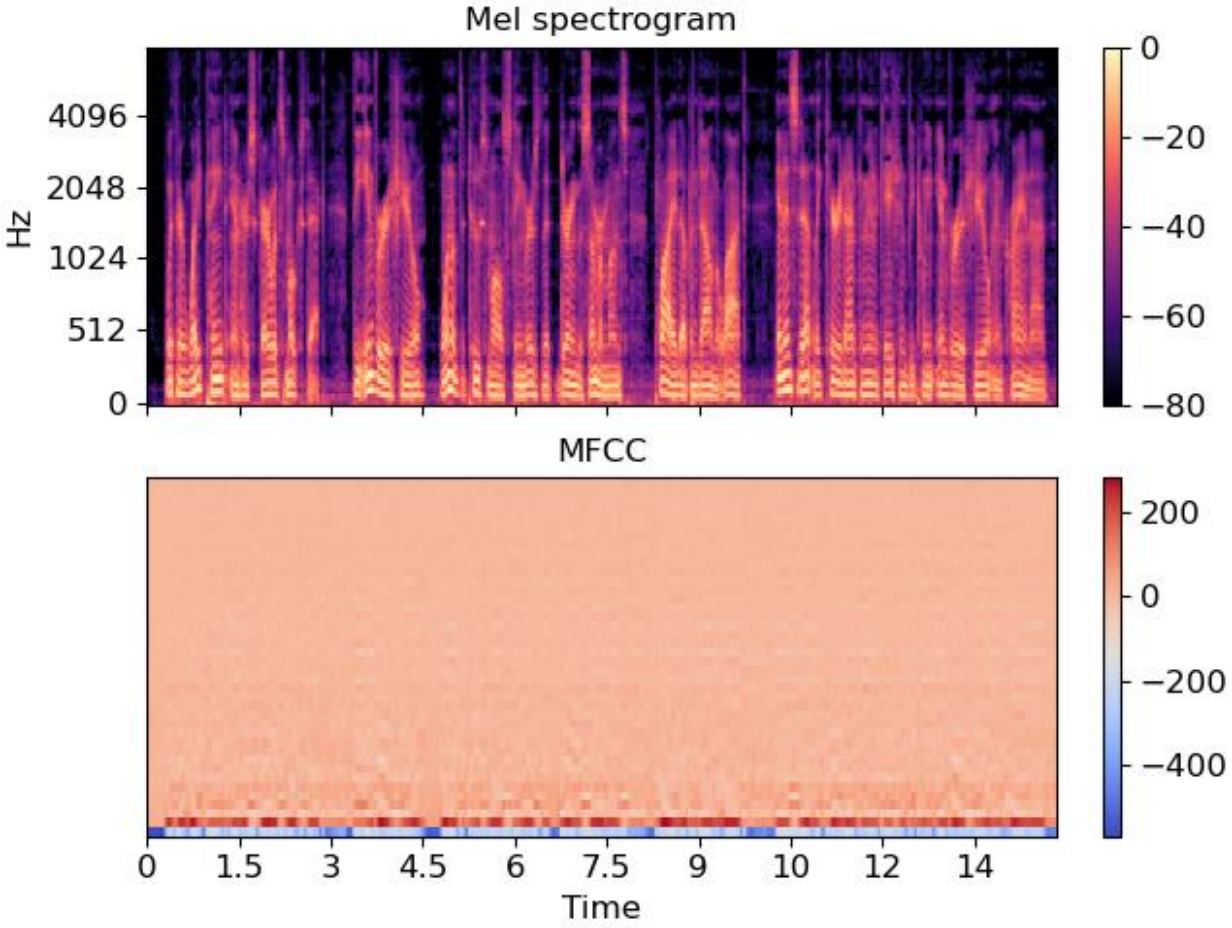


Figure 3. Examples of Mel Spectrogram and MFCCs generated by Librosa Toolbox (librosa, 2023). X-axis represents time frames while y-axis represents frequency bins on the left-hand side and decibels on the right-hand side for Mel Spectrogram. That for MFCC represents the index of cepstral coefficients on the left-hand side and values on the right-hand side. Note that the colours are only for the purpose of representing values. They are not 3-dimensional as coloured images. Darker colour represents lower energy while lighter colour represents higher energy for spectrogram.

2.1.1 Classification

Support Vector Machine (SVM) is a conventional machine learning model and is one of the most cited ML models. It is a supervised classification model that is used for linearly separable and multi-dimensional data using a hyperplane. SVM models are trained with small datasets and the computational cost increases significantly with larger and non-linear datasets (Pan, et al., 2012) (El Ayadi, et al., 2011) (Koolagudi & Rao, 2012).

Modern classification models are focused on deep learning techniques such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and multi-model classifiers. CNN has multiple sequential connected layers, include filtering layer, a ReLU

layer, a pooling layer, and a fully connected layer (see Figure 3.) (Badshah, et al., 2017). It is biologically inspired by neurons in brains which reduce network parameters, extract high-level features, produce feature maps, and classify based on trained parameters (Alzubaidi, et al., 2021). CNN and its variations have been predominantly used for image and speech classification for the following benefits,

- CNNs reduce the trainable parameters, which improves the generalisation of the model to avoid overfitting.
- CNNs employ convolutional layers that scan the input in a spatially hierarchical manner. This allows them to capture both local and global patterns.
- They are computationally efficient to train with large dataset compared to conventional ML models as they can automatically learn meaningful features from raw data, whereas traditional ML models often rely on handcrafted feature engineering, which can be time-consuming and domain dependent. DL also implement parallel processing to accelerate training on larger dataset.

Figure removed due to copyright restriction.

Figure 4. Example architecture of CNN (Alzubaidi, et al., 2021)

LSTMs is another DL model that are capable of learning long-term dependencies. Based on Figure 4, the core of the LSTM model is its memory cells that can store and access information over long sequences. The input gate regulates what information is allowed to enter the memory cell and decides which new information should be stored by using a sigmoid activation (σ). The forget gate decides what information should be discarded from the memory by using another sigmoid activation. The output gate determines what information stored in the memory cell should be used to produce the output using both the current input and the previous hidden state which is the output of the LSTM for a given time step (Hochreiter & Schmidhuber, 1997). The advantages of using LSTM for SER are that (Swain, et al., 2018) (Madanian, et al., 2023),

- LSTMs are designed to mitigate the vanishing gradient problem in standard Recursive Neural Network, making them a better option for training on longer sequences without the loss of information.

- They use memory cells to allow storing and retrieving information so that crucial information can be retained overtime.

Figure removed due to copyright restriction.

Figure 5. Example architecture of LSTM (Eswarsai, 2021)

More complex DL models combine various DL techniques, for example, combining CNN and Bi-directional LSTM. Complex DL models can capture subtle differences in patterns and spatio-temporal features (Zhou, et al., 2018). However, to further improve the generalisability of the SER model, TIM-Net has been proposed to capture contextual information with both local and global temporal dependencies (Ye, et al., 2023). The framework (see Appendix A) contains temporal aware blocks (TABs) in both forward and backward directions. Dilated Casual Convolution layers (DC Conv) are included in the sub-block to ensure that there is no future information is present in the past by refining the receptive fields. A casual convolution enforces the causality property, meaning that the output at any time step depends only on the current and past time steps. It prevents information from future time steps from leaking into the current time step. A dilated convolution involves skipping input values with a fixed step size when applying the convolution filter. It increases the receptive field of the convolution without adding more parameters or significantly increasing computational cost. Combining the two types of Conv layers is used to capture both the temporal dependencies in the data and a larger context of information. The output of both directions is fused to produce the contextual representation with multi-scale dynamic fusion.

The comparisons among the models discussed are summarised in the table below. Their performance levels were judged based on availability, computational efficiency, interpretability, dataset requirement, simplicity of training, simplicity of use, accuracy as well as generalisability. Based on this comparison matrix, TIM-Net has better performance overall than the rest of the models presented.

Table 3. Comparison Matrix for Classification Models

Parameters	TIM-Net	SVM	CNN	CNN+Bi-LSTM
Related Studies	(Ye, et al., 2023)	(Dahake, et al., 2016) (Milton, et al., 2013) (Sinith, et al., 2015)	(Lim, et al., 2016) (Anrarjon & Mustaqeem, 2020)	(Zhou, et al., 2018) (Meng, et al., 2019) (Zhao, et al., 2019)
Availability (5%)	Open source (5)	Open source (4)	Open source (4)	Open source (4)
Computational Efficiency (10%)	High (4)	Low (2)	High (4)	High (4)
Interpretability (10%)	Moderate (3)	High (4)	Moderate to High (3)	Moderate (3)
Dataset Requirement (10%)	Moderate to Large (3)	Small (2)	Moderate to Large (3)	Moderate to Large (3)
Simplicity of Training (15%)	Moderate (3)	High (4)	High (4)	Moderate (3)
Simplicity of Use (15%)	High (4)	High (4)	High (4)	Unknown (2)
Accuracy (15%)	High (5)	Moderate (3)	Moderate (3)	Moderate to High (4)
Generalisation (15%)	High (5)	Low to Moderate (2)	Moderate (3)	Moderate to high (4)
Total	4.05	3.05	3.45	3.35

Note: Each parameter is graded 1-5. 1: Poor Performance 2: Unknown/below average 3: Satisfactory 4: Above Average 5: Excellent

2.2 Identifying the Gap

Based on the literature review, there are challenges remaining in the field of SER, including

1. Lacking natural datasets.
2. Lacking bi- or multi-lingual datasets.
3. High computational cost, and longer training time for complex multi-modal deep learning models which require large dataset to train.

4. Limited Interpretability of decision-making process based on extracted features; therefore, there has not been a consensus on what features are more relevant than others.

In the context of transferability of a SER model trained with one language to another, reduced performances were observed in recognising emotions compared to that the model was trained with (Iosifov, et al., 2022). On the contrary, another study suggests that emotions are expressed similarly across different languages, supported by accuracy difference within 10% across 4 corpuses (Al-onazi, et al., 2022). Whether language differences affect the performance of multi-lingual SER model is unknown.

2.3 A Novel Approach

To investigate the effect of different language (between Chinese and English), the parallel dataset ESD will be used as it is a bi-lingual database with limited confounding factors that could contribute to the performance of the SER model.

To address the gap in areas of computational cost of training and limited interpretability of decision-making process (3. and 4.), different image processing techniques will be applied to extracted features. Applying image processing principles such as edge and contrast enhancements has been proven to be effective in improving predicting performance of developed model in the fields of speech recognition (Cadore, et al., 2011) (Asahi & Ogawa , 2006) and acoustic detection (Fang, et al., 2022). This proposed approach assumes that if differences in spectral features can be spotted by human, these features can be further enhanced to be easily picked up by the Neural Network. Figure 5 demonstrates how different speech emotions can be visually identified from Mel spectrograms (Jiang, et al., 2019).

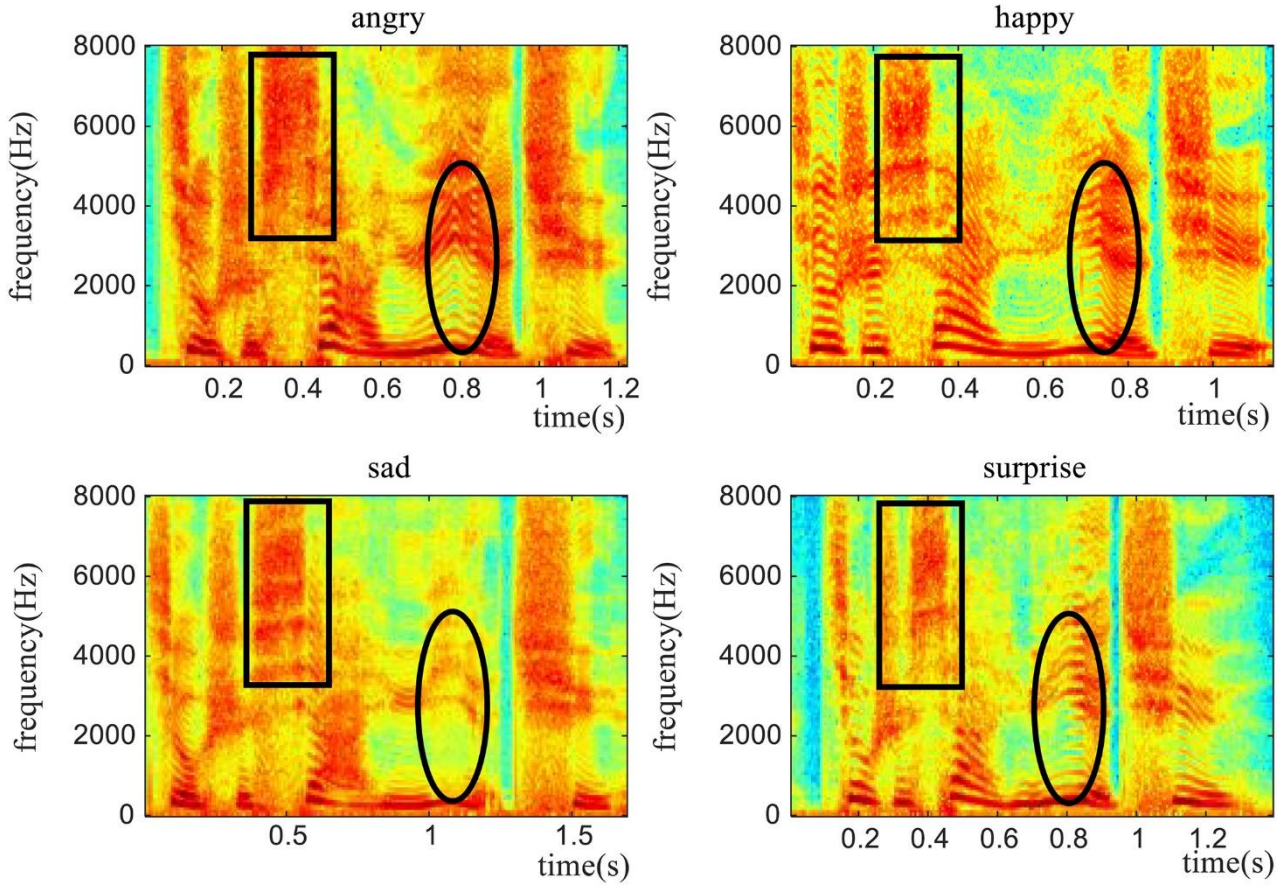


Figure 6. Spectrograms for Different Emotions. This Figure shows an example of variations in spectrogram for anger, happiness, sadness, and surprise. The energy differences are shown in the rectangular areas while the harmonic differences are shown in the ellipse areas. Harmonic structures reflect the energy distribution across frequency bands. Anger and Surprise share the same harmonic structure which reflects high-pitched tonality. The sound intensity if indicated by spectrum energy across the frequency bands, which high if the emotion is more intense, such as ager and happiness.

There are two types of enhancements that are widely used for image processing, edge detection and contrast enhancement. For edge detection, the Difference of Gaussian (DoG) and Sobel filter on X and Y axis are the two most common ones. DoG is equivalent to a band-pass filter which removes frequencies except specified range. It is obtained by subtracting one blurred version of an image from another, effectively enhancing the texture details. In the context of enhancing a spectrogram, the DoG filter may be able to highlight important spectral features, making them more distinguishable. The mathematical derivation of the DoG filter starts with the gaussian function, denoted as

$$G(x, y, \sigma) = \left(\frac{1}{(2 * \pi * \sigma^2)}\right) * e^{-\left(\frac{(x^2+y^2)}{(2*\sigma^2)}\right)} \qquad (4)$$

,where (x, y) represents the spatial coordinates and sigma represents the standard deviation for normal distribution (Fu, et al., 2018).

Therefore, this study hypothesises that by enhancing the edge of the harmonic structure and contrast between high energy and low energy area, the classification accuracies will be improved for both Chinese and English. We also hypothesise that with enhanced features, the generalisability of the model will be improved reflected by decreased accuracy difference between Chinese and English.

Sobel filter is another edge detection operation that is performed on a blurred image. It is applied on both horizontal and vertical directions respectively, by convolving with two 3x3 kernels, one for detecting changes in pixel values along the horizontal direction and another for the vertical direction. The results obtained from the two directions are combined to create a single image that emphasizes both horizontal and vertical edges. A common approach is to take the magnitude of the gradient, which combines the information from both directions using the Pythagorean theorem (Jana, et al., 2021),

$$Gradient\ Magnitude\ (M) = \sqrt{\big(sobel(x)\big)^2 + \big(sobel(y)\big)^2}$$

Image can also be enhanced by increasing contrast between darker areas and brighter areas. Contrast Limited Adaptive Histogram Equalisation (CLAHE) is a non-linear adaptive histogram equalisation filter. It divides an image into small tiles, computes the histogram for each tile, and then redistributes the pixel values within each tile based on these local histograms. Image processing with CLAHE has shown promising results for many image classification applications such as object detection (Rodriguez-Rodriguez, et al., 2020).

Aims and Objectives

This study aims to evaluate the effect of different image processing techniques on the performance of SER model and to develop a bi-lingual SER model, capable of recognising five emotions (Angry, Happy, Neutral, Sad, and Surprise) in both English and Chinese. The target is to achieve a minimum average accuracy of 75% and a maximum accuracy difference of 5% between the two languages. The following objective were defined to achieve this aim,

- Identify visual differences within the spectral features that correspond to the five emotions and between the two languages. Subsequently, choose suitable processing techniques based on the observations.

- Establish a baseline for the performance of the chosen SER model based on unprocessed spectral features. The overall accuracy should be at least 60%.

- Evaluate and compare performances of the SER model based on processed spectral features with different processing techniques and identify the best approach based on the performance metrics. For each processing technique, compare performances (average accuracy, average recall for each emotion as well as confusion matrix) between English and Chinese.

- Obtain the performance of the SER model when employing fused and processed spectral features with best-performing approach determined in the previous objective. Compare performances between English and Chinese. At this final step, the developed model will have minimum accuracies of 75%, and maximum accuracy difference of 5% between Chinese and English.

1.3 Scope and Limitations

There are various spectral features that can be used for training the SER model; however, this research only focuses on Mel-spectrogram and MFCC. Though variations in age and gender can potentially affect the performance, only the effects of different languages are considered. Additionally, this study only includes two languages (Chinese and English) and five emotions (Angry, Happy, Neutral, Sad and Surprise). The classification model is adopted from existing literature that was proven to be effective on multiple datasets. The development of the classifier or improving the classifier by adjusting hyperparameters of the classifier is out of scope.

Several limitations of this study are identified,

- This study uses an acted dataset, where emotions could be exaggerated.
- The resource-intensive nature of implementing various processing techniques may pose a potential constraint, as it may demand substantial computational resources.
- Multiple emotions can intertwine within a single sentence real-life scenario. However, this study assumes isolated emotions, potentially diverging from the intricacies of emotional expression in everyday conversations.
- The speech data that are used to train the model were collected under pristine studio conditions, whereas real-world scenarios often involve background noise and interferences that the model may be robust against.

3. METHODOLOGY

3.1 Software and Packages

The methodology employed in this study encompassed various components and tools. The software and packages utilized in the research included Visual Studio Code (VS code) for local image processing tests, Google Colab with Jupyter Notebook for feature extraction, training, and evaluation, as well as Excel for statistical analysis. The programming language used was Python, and the main modules that were used include,

- Numpy: it is a package for scientific computing and data manipulations with Python which provides support for arrays and matrices, as well as a large library of mathematical functions to operate on these arrays.
- Tensorflow: it is an open-source ML framework that is commonly used for deep learning tasks, including NN modelling.
- Matplotlib.pyplot: it is a data visualisation library for creating static, animated and interactive plots in Python.
- Sklearn: it is a popular machine learning library that provides tools for preprocessing, classification, regression, clustering and more.
- Pandas: it is a library for data analysis and structuring data.
- OpenCV: it is a computer vision library that provides tools for image processing.
- Librosa: it is a package used for audio processing and analysis including loading files, extracting features, and analysing acoustic features.

Most of the tasks were done in Google Colab (see detailed architecture in Appendix C). It is a cloud-based platform provided by Google for developing and running Python code, with a focus on machine learning and data analysis. It provides a free online Python environment with pre-installed packages mentioned above. Colab is build based on Jupyter Notebook, which can contain code, text, or visual elements. The GPUs used for this thesis were NVIDIA V100 Tensor Core and NVIDIA T4 Tensor Core.

3.2 Experiment Set Up

3.2.1 Dataset

The dataset used was the Emotional Speech Dataset that contains both Chinese and English emotional speech data. In summary, speech data were developed with 10 Chinese speakers and 10 English speakers performing with five emotions, anger, happiness, neutral, sadness and surprise. Each speaker performed the same 350 utterances for each emotion, which resulted in 1750 utterances for each language. The

average length of each speech sample was 2.76 s. All speech data were recorded in an environment with signal-to-noise ration of above 20 dB and a sampling frequency of 16 kHz (see Appendix-B for detailed information of ESD). 300 utterances for each speaker were used for training, and the remaining were used for evaluation.

3.2.2 Preprocessing and Feature Extraction

The two acoustic features that were used for training and SER modelling were Mel Spectrogram as well as MFCCs. They were extracted using the Librosa Python toolbox with the following parameters,

- Mel Spectrogram
    - Sample rate = 16 kHz
    - Frame length = 50 ms
    - Frame shift = 12.5 ms
    - Hamming window
- MFCCs
    - Sample rate = 16 kHz
    - Frame length = 50 ms
    - Frame shift = 12.5 ms
    - Hamming window
    - Number of coefficients: 39

39-coefficient were adopted from Ye, et al. (2023). Features extracted were then stored in csv files for further processing in the next step.

3.2.2 TIM-Net framework

The deep learning technique was directly adopted from the TIM-Net framework (see Appendix – A for details) which is based on a CNN model with bi-directional LSTM with dynamic reception field to capture temporal and contextual features. To compare the results of this thesis to that obtained by the original literature, the implementation details and hyperparameters were identical to the one mentioned by Ye et al. (2023) which were,

- Epoch: 500
- Learning rate: 0.001
- Batch size: 64
- Label smoothing factor: 0.1
- Dropout rate: 0.1
- Kernal Size: 39 Kernels with Size 2 for Convolution layers
- 10 folds cross-validation
- Random seed: 64

3.2.2 Image Processing

The image processing techniques that were used in this thesis include Difference of Gaussian filter, Sobel filter, as well as CLAHE filter. The parameters used for each filter are shown below,

- DoG: Sigma 1 = 0.5; Sigma 2 = 0.7
- Sobel: Gaussian blur (sigma = 0.5): Kernel size = 3
- CLAHE: Clip limit = 1.2; Tile gride size = 4x4

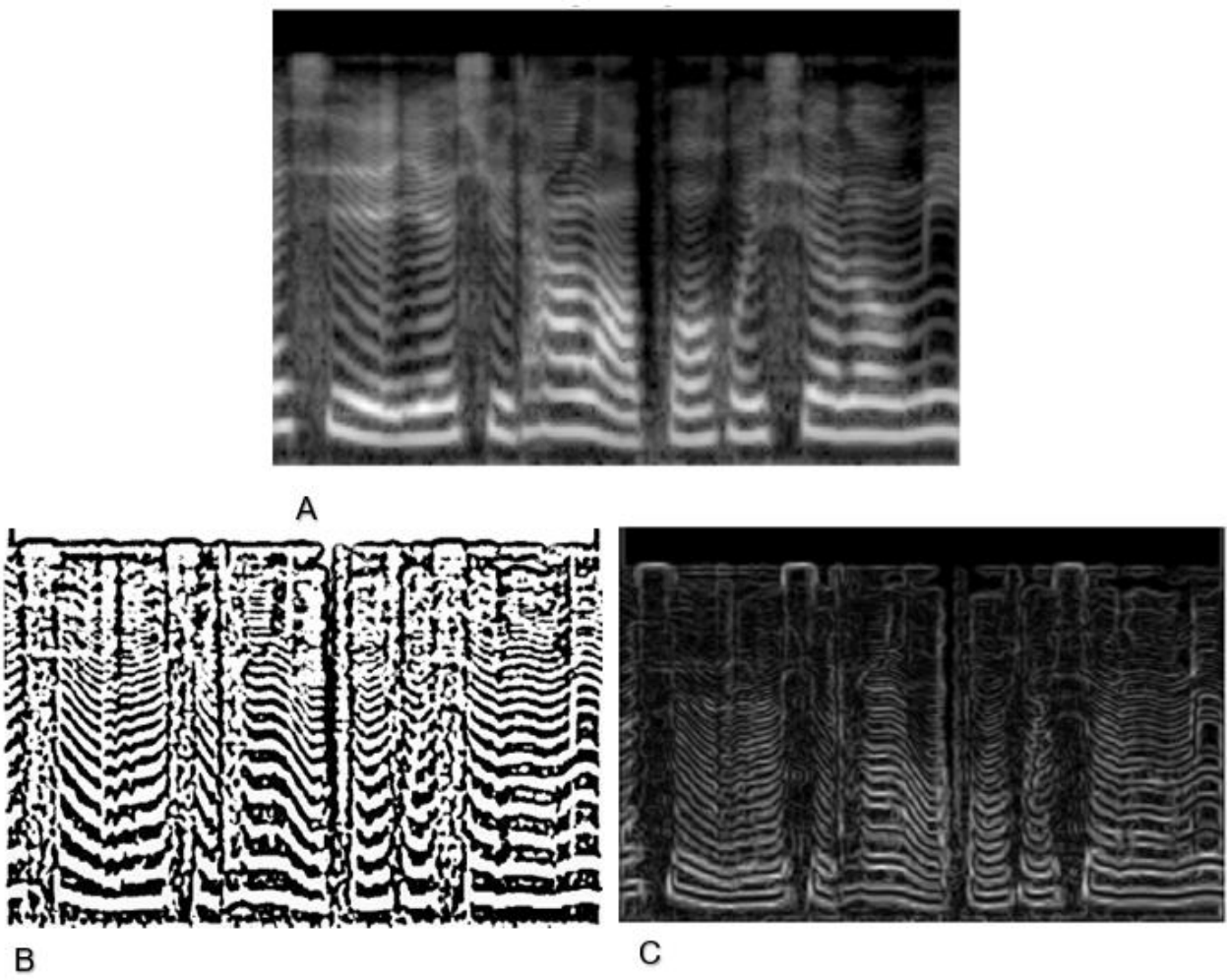


Figure 7. Examples of processed Mel Spectrograms. The y-axis represents the frequency bins and the axis represents the time frame. (A: Unprocessed Mel Spectrogram; B: With DoG filter; C: With Sobel filter)

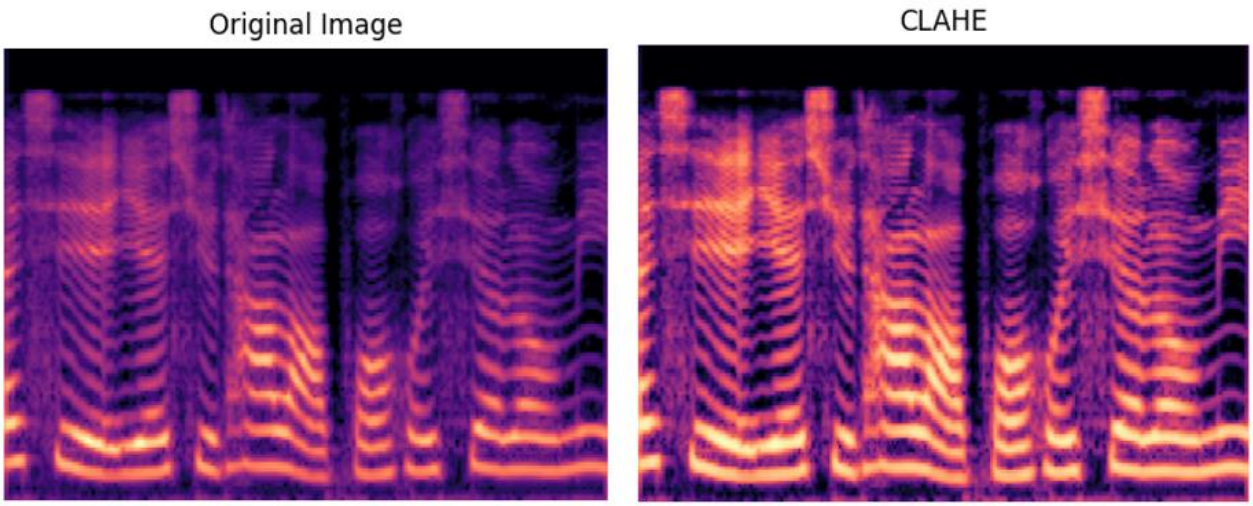


Figure 8. Example Mel Spectrogram processed with CLAHE filter. The y-axis represents the frequency bins and the axis represents the time frame. Higher energy is represented by brighter colour while lower energy is represented by darker colour.

3.2.3 Experiment Structure

To compare the performance of different image processing techniques, a baseline result was first obtained by training the model with Mel Spectrogram and MFCCs with no further processing. For each image processing technique, parameters were fine-tuned manually to satisfaction based on visual feedback. Each Processing technique was applied to the unprocessed Mel Spectrogram and MFCCs respectively. The model was trained with processed Mel Spectrogram and MFCCs. Evaluation set was separated into English and Chinese to be evaluated separately. Performance results including overall accuracies, weighted and unweighted average recalls, and confusion matrix were obtained for each trained model. The performance of each processing techniques was compared based on the parameters mentioned above. The best-performing processed Mel Spectrogram was then be fused with the best-performing processed MFCCs to further investigate the effect of feature fusion on the performance of trained SER model. The two features were fused to form a single input by concatenating along horizontal axis. Due the difference on input sizes along vertical axis, Mel Spectrogram was zero-padded to match the number of columns of MFCCs.

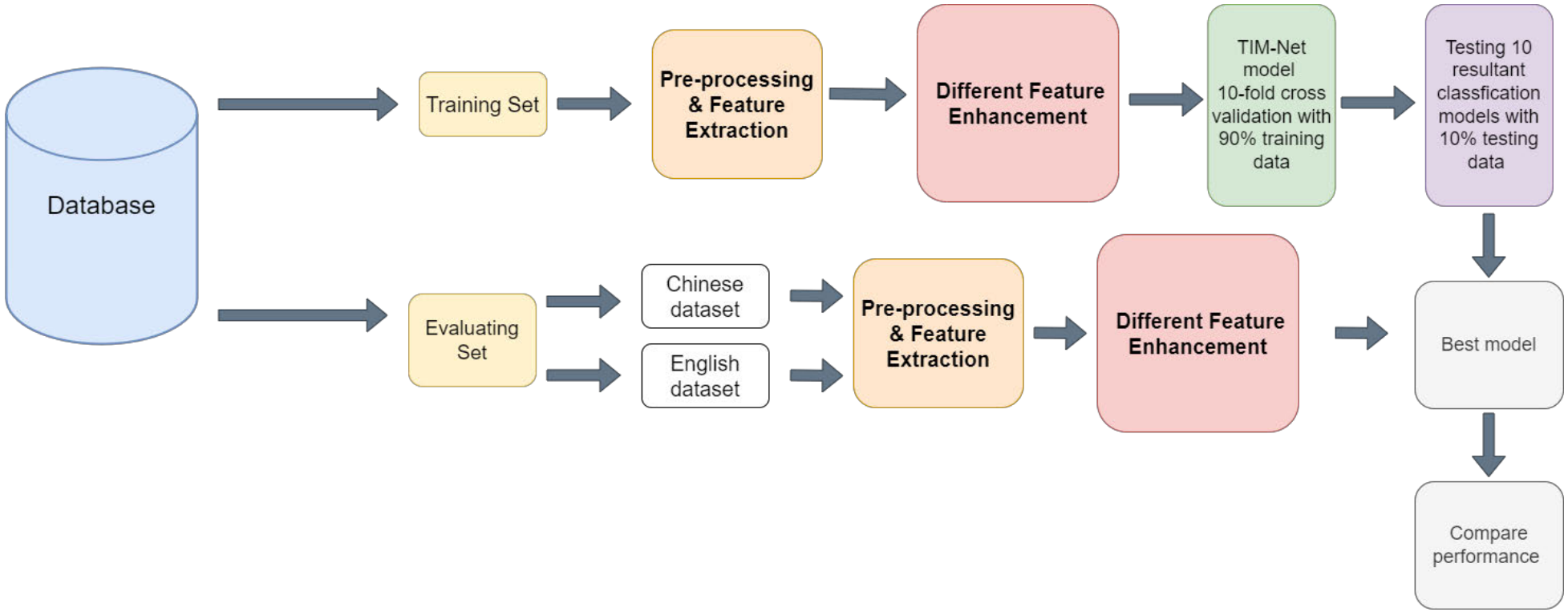


Figure 9. Experiment Framework (The overall flow of the development of a SER model with different techniques include splitting the database into training and evaluating sets. The training set went through preprocessing and feature extraction. The extracted features went through one of the following processing methods, no processing, DoG, Sobel, and CLAHE. The model was trained with 10-fold cross validation. The evaluation set were separated into Chinese and English subsets to be used to evaluate the developed model. The performances of all models were compared based on accuracies, average recalls, and confusion matrices)

3.3 Statistical Analysis

The parameters that were used for statistical analysis are shown below,

- Average accuracies were calculated displayed in the format of $\bar{x} \pm \sigma$, where $\bar{x}$ is the mean accuracy σ is the standard deviation to show variance in predictions across different emotions. It is represented in percentage ($\bar{x} \pm \sigma\%$).

- Both weighted average recall and unweighted average recall were used. To limit the effect of potential emotional class imbalance, weighted average recall was used which provides a more balanced view of model performance. They are represented in probabilities in range 0 – 1.

- Confusion matrices were constructed based on true emotion against predicted emotions, including true positives, true negatives, false positives, and false negatives.

- P-values was used for the same purpose as the confidence level. A p-value less than 0.05 is considered statistically significant. Two-tailed paired t-test was used for comparing the average accuracies among different processing techniques. Two-tailed unpaired t-test was used to determine any significant differences between the performances of Chinese and English.

4 RESULTS

4.1 Identifying Differences Visually

Upon the extraction of Mel Spectrograms and MFCCs a comparative analysis based on languages and emotions was conducted to visually identify general trends in disparities. Figure 7. illustrates an example of differences observed between English and Chinese, and between happiness and neutrality. In this analysis, A and C were compared to identify the difference between the two emotions in Chinese. Illustrated by the green rectangles, speech in happiness showed intense energy across less frequency bins while speech in neutral emotion showed more evenly distributed energy across more frequency bins at lower frequencies. For both emotions, the energy distribution became less distinct with increasing frequencies. Along the x-axis, the energy fluctuated more with speech in happiness compared to neutral speech. Indicated by the yellow rectangles, the speech in neutral emotion showed a more distinguishable and longer pause than speech in happiness. These observations were consistent with those identified in English when comparing B and D. A and B were compared to determine variations in happy speech for different languages. It showed that, English speech had a less distinctive harmonic structure across the frequency bins when compared to Chinese speech. It was also observed that the change in energy along x-axis was greater when compared to speech in English. Neutral speech showed similar characteristics between Chinese and English.

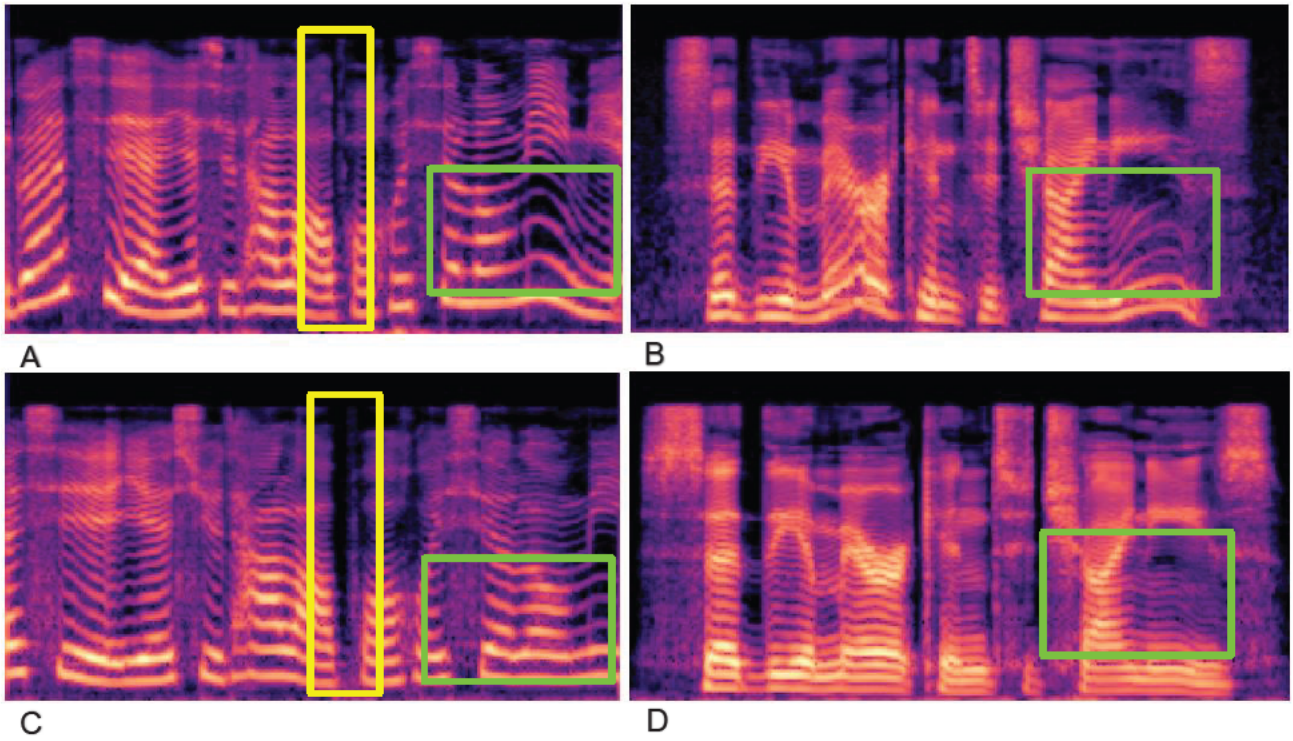


Figure 10. Comparison of Mel Spectrograms based on languages and emotions (A: Happiness/Chinese; B: Happiness/English; C: Neutral/Chinese; D: Neutral/English). The green rectangles indicate the harmonic structure and the yellow rectangles indicate the pauses in speech. Y-axis represents the frequency bins and x-axis represents the time frame. Higher energy corresponds to brighter colour while low energy is associated with darker colour.

Comparisons based on MFCC for different languages and emotions are shown in Figure 8. A and C were compared to identify differences between happiness and neutral in Chinese speech. The values of coefficients varied more for speech in happiness along x axis when compared to neutral speech. Visually, the fluctuation was present across wider range of coefficients along y axis for speech in happiness. Comparing A and B, the variation in values of coefficients was less noticeable for English speech in happy and neutral speech.

These findings confirmed that there was a general trend in differences across languages and emotions. These variations were then enhanced by the three image processing techniques to obtain performance results.

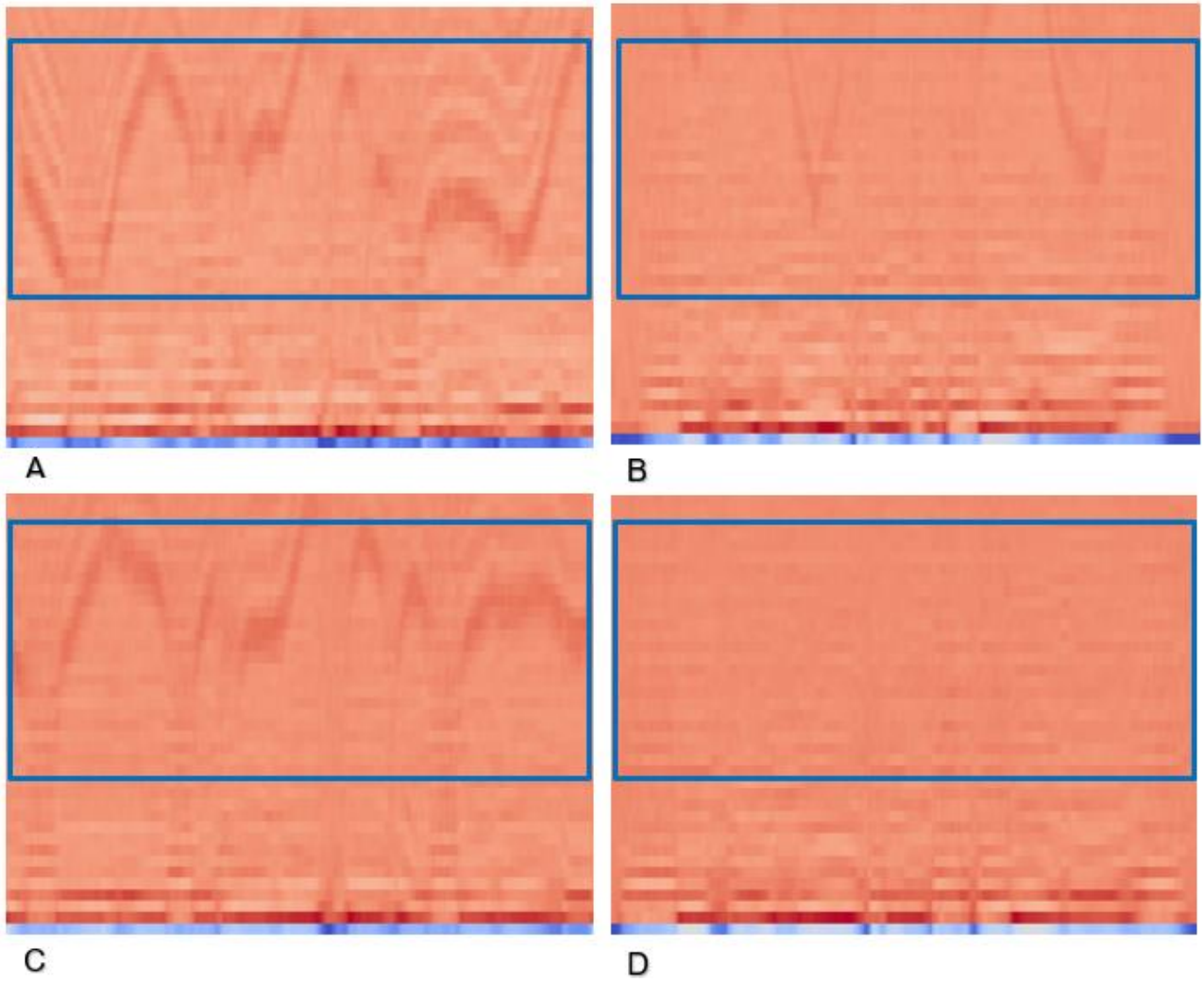


Figure 11.Comparison of MFCCs based on languages and emotions (A: Happiness/Chinese; B: Happiness/English; C: Neutral/Chinese; D: Neutral/English). The blue rectangles indicate differences in features along the x-axis. Y-axis represents the MFCC coefficients while x-axis represents the time frames. Coefficients with greater values are represented by deeper colour, whereas those with smaller values are represented by lighter colour. Higher values of MFCCs indicate stronger spectral features and spectral contrasts.

4.2 Results for Different Processing Techniques

Average accuracies for models trained with unprocessed features, and processed features are summarised in the Table below. For MFCCs, the model trained with unprocessed MFCCs showed a difference in accuracies of 7.4% between the two languages. This difference was reduced significantly by training the model with DoG-filtered MFCCs ($p = 0.001 < 0.05$). As shown in Figure 12, the variation of MFCC values became more distinctive after processing across lower and higher order coefficients.

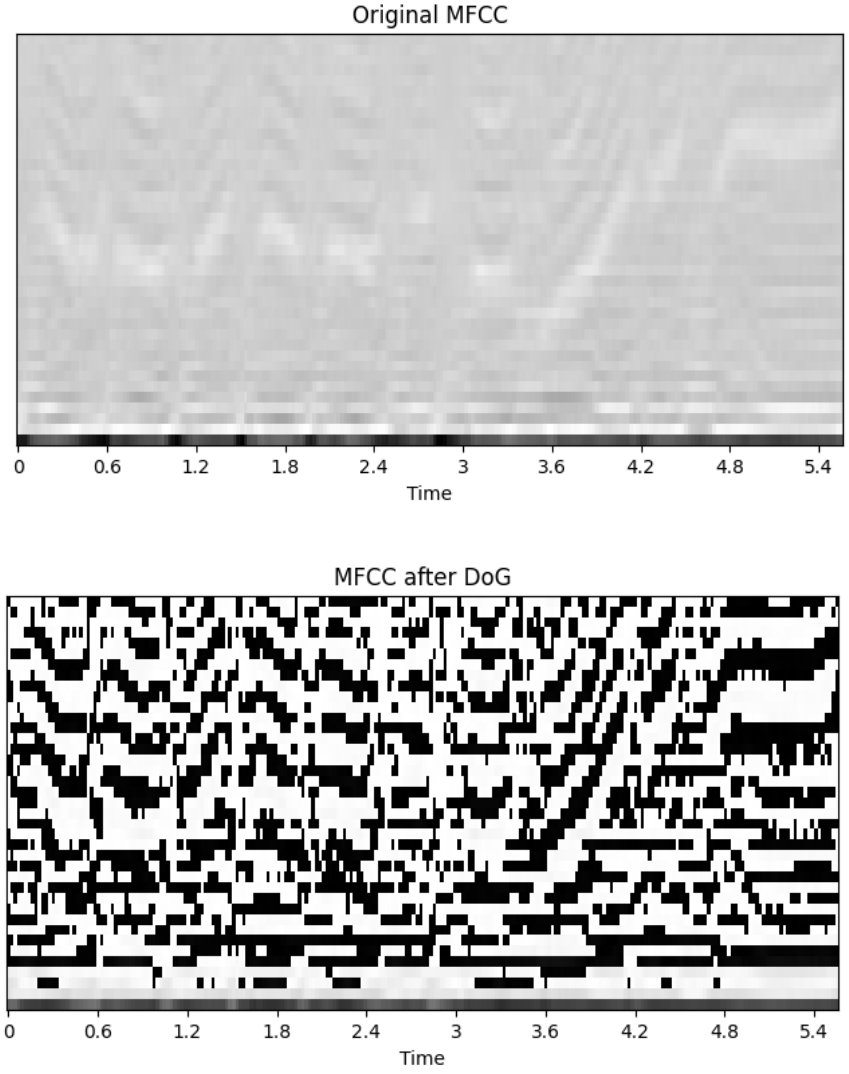


Figure 12. Example of Unprocessed MFCCs and DoG-filtered MFCCs. Y-axis represents the index of coefficients while x-axis represents the time frame. The speech sample was in Chinese with angry emotion.

The average recalls were improved by 0.22, 0.21, 0.28, 0.13, and 0.03 for anger, happiness, neutral, sad, and surprise respectively. For the baseline model, the probabilities for the model to false predict anger and happiness as surprise were 0.21 and 0.13 respectively. These were reduced to 0.07 and 0.05 with the model trained with DoG-filtered MFCCs. The confusion between neutral and sadness were significantly improved as well ($p = 0.001 < 0.05$)., indicated by probabilities of 0.00 and 0.01 (Baseline: 0.15 and 0.16) respectively.

On the other hand, by applying Sobel or CLAHE filter to MFCCs, the trained model displayed reduced accuracies and increased difference in accuracies between Chinese and English. This could be due to possible overprocessing with these filters, resulting in loss of information and introducing additional noise to the feature. The increased standard deviations (6.43%, 11.08% and 8.01% for English with unprocessed MFCCs, Sobel-filtered MFCCs, and CLAHE-filtered MFCCs respectively; 4.15, 6.85 and 5.93 for Chinese with unprocessed MFCCs, Sobel-filtered MFCCs, and CLAHE-filtered MFCCs respectively) indicated that by applying these two filters, the abilities for the trained model to predict varied more for different emotions. Table 7. shows detailed breakdown of performance for predicting each emotion between the base line model and the model trained with Sobel-filtered MFCCs. The average recalls for predicting neutral and sad speech did not reduce significantly (0.65 and 0.82 vs. Baseline: 0.66 and 0.81, respectively). However, those for predicting angry, happy, and surprised speech were decreased significantly (.055, 0.49 and 0.65 vs. Baseline 0.66, 0.7 and 0.79 respectively, ($p = 0.001 < 0.05$) for English speech. The confusion matrices shown in Figure 12. shows more details on the effect of training the model with Sobel-filtered MFCCs. There were increased confusions between happiness and surprise, indicated by increased false predictions. The probability for the baseline model to falsely predict surprise as happiness was 0.15; however, it increased to 0.18 with the model trained with Sobel-filtered MFCCs. Similarly, the probability for the model to falsely predict happiness as surprise was increased from 0.13 to 0.19. Based on Figure 13, visually, it is difficult to identify differences between these two emotions for the same utterances. After processing, even though the distinctions along the x-axis were improved, those along the y-axis became less distinctive.

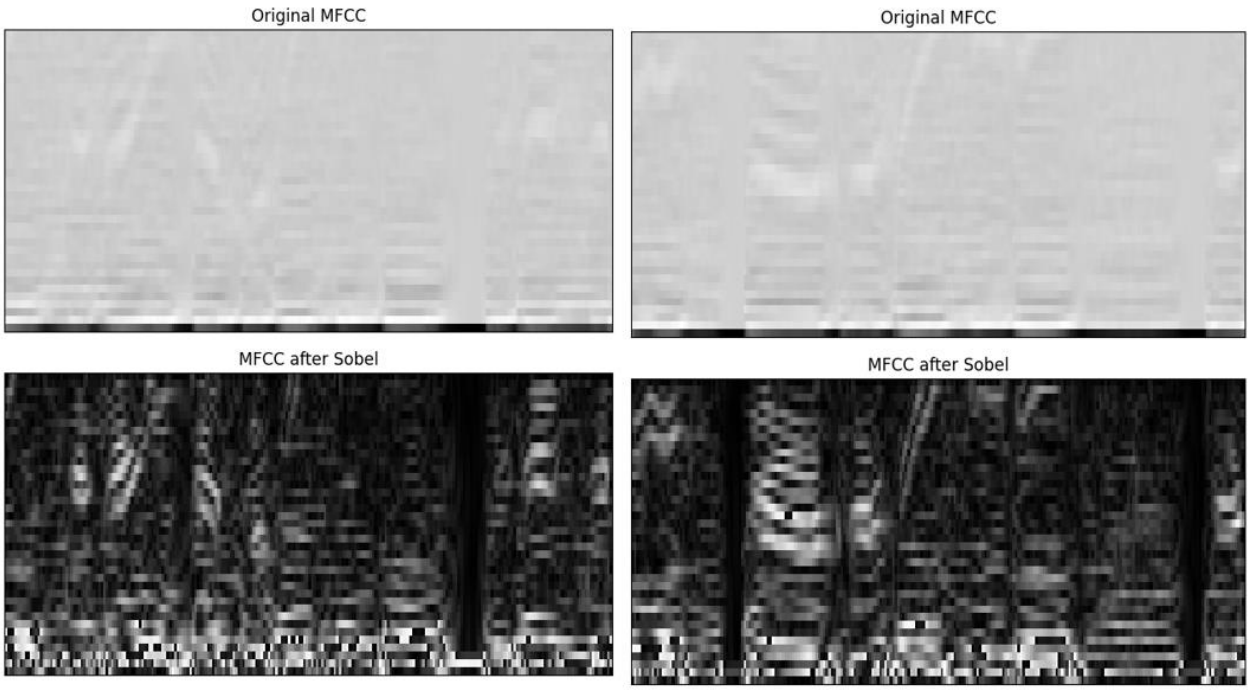


Figure 13. An example of happy speech in Chinese before (top-left) and after (bottom-left) applying Sobel filter and surprise speech in Chinese before (top-right) and after (bottom-right) with Sobel filter. Y-axis represents the index of coefficients while the x-axis

Table 4. Average accuracy for each processing technique

	MFCC - Baseline		MFCC - DoG		MFCC - Sobel		MFCC - CLAHE	
	English	Chinese	English	Chinese	English	Chinese	English	Chinese
Overal Accuracy	72.33 ±6.43	79.73 ±4.15	90.13 ±4.48	89.73 ±5.33	63.2 ±11.08	72.67 ±6.85	77.87 ±8.01	85.07 ±5.93
Difference	7.4		0.4		9.47		7.2	
	Mel Spectrogram - Baseline		Mel Spectrogram - DoG		Mel Spectrogram - Sobel		Mel Spectrogram - CLAHE	
	English	Chinese	English	Chinese	English	Chinese	English	Chinese
Overal Accuracy	84.4±7.03	90.4±3.21	63.07±10.10	71.93±9.54	60.93±8.91	74.13±9.62	86.67±5.79	89.87±5.28
Difference	6		8.86		13.2		3.2	

i. results are displayed in the format of mean±standard deviation.

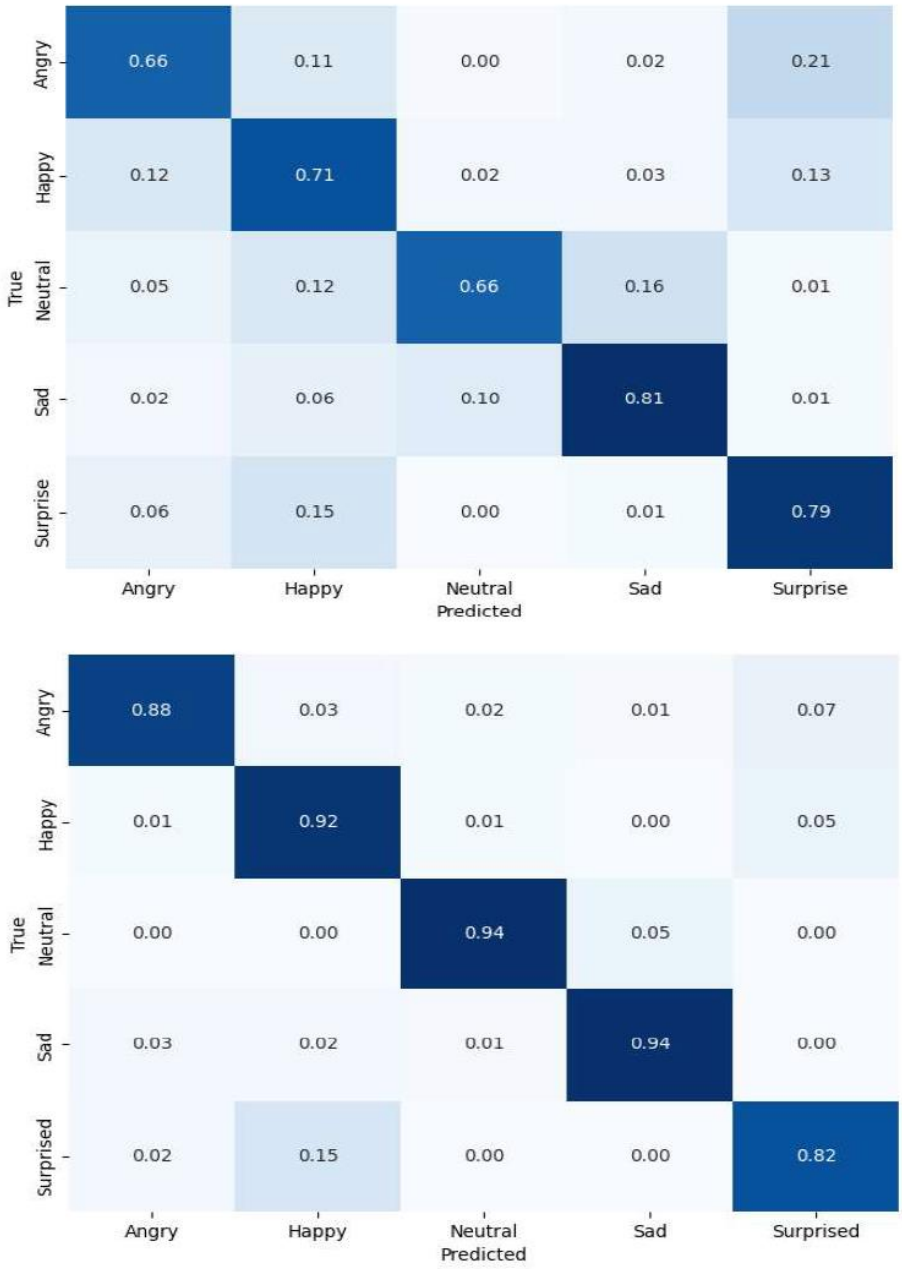


Figure 14.Confusion Matrices for baseline model (top) and the model trained with DoG-filtered MFCCs (bottom). Horizontal axis represents the predicted emotions while the vertical axis represents the true emotions. Higher average recall is indicated by darker blue.

Table 5. Comparison between baseline model and model trained with Sobel-filtered MFCCs

	MFCC - Baseline				MFCC - Sobel			
	English		Chinese		English		Chinese	
Emotion Classes	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)
Angry	65.67	68.76	77.67	79.25	54.67	57.34	73	73.86
Happy	70.67	65.84	78.33	71	49.33	53.14	71	67.51
Neutral	65.67	74.06	77	84.93	65.33	67.59	74.67	77.37
Sad	81	80.07	88	85.44	81.67	72.38	83	78.92
Surprised	78.67	73.29	77.67	79.25	65	63.52	61.67	65.37
Overall Accuracy	72.33		79.73		63.2		72.67	
Difference	7.4				9.47			

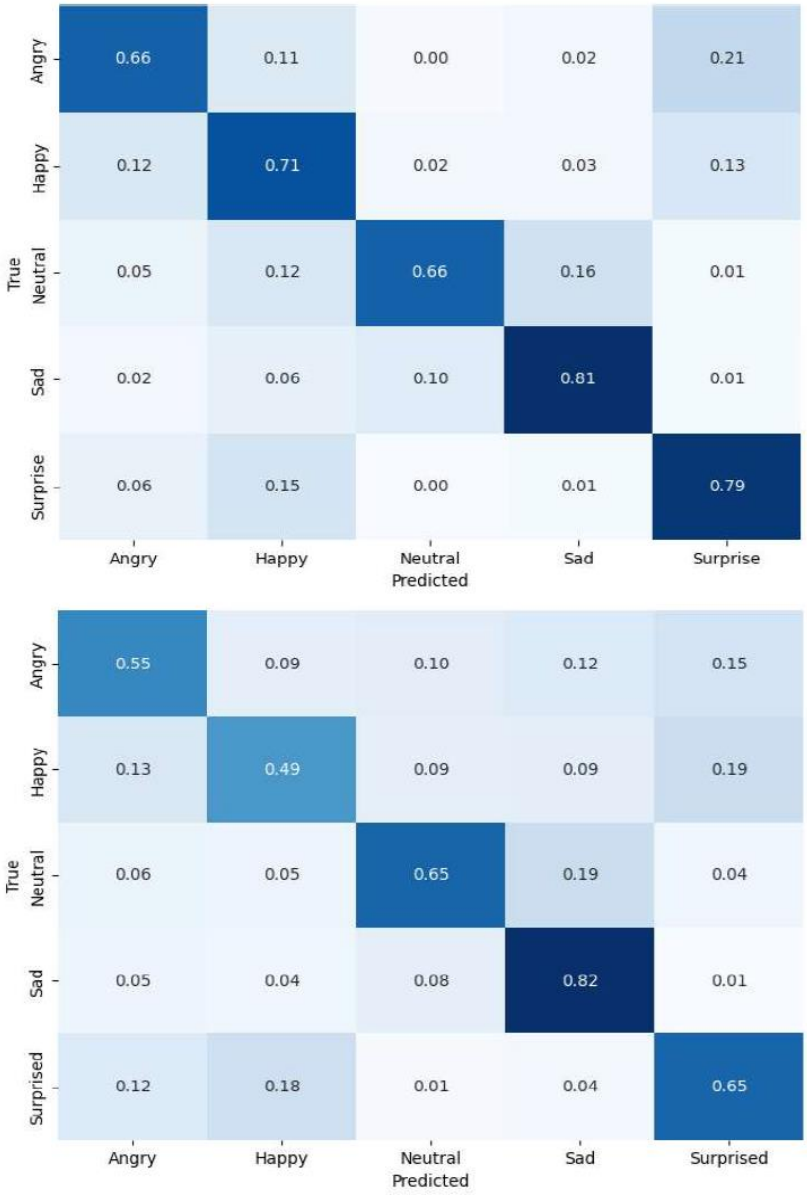


Figure 15. Confusion Matrices for baseline model (top) and the model trained with Sobel-filtered MFCCs (bottom). Horizontal axis represents the predicted emotions while the vertical axis represents the true emotions. Higher average recall is indicated by darker blue.

For results related to training with Mel Spectrogram, there was no significant ($(p = 0.06 > 0.05\ for\ CLAHE)$ improvements on overall performance. However, by applying CLAHE to Mel Spectrogram, the difference in accuracies between Chinese and English was improved by 2.8%. Table 8 shows detailed performance comparison between the baseline model and the model trained with CLAHE-filtered Mel Spectrograms. The average recalls for predicting anger and surprise were decreased for Chinese speech, while those for predicting happiness and neutral were increased. For English emotional speech, the predictions were improved for all emotions based on the WAR.

Table 6. Comparison between baseline model and model trained with CLAHE-filtered Mel Spectrograms

	Mel Spectrogram - Baseline				Mel Spectrogram - CLAHE			
	English		Chinese		English		Chinese	
Emotion Classes	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)
Angry	75.67	79.09	90	90.15	78.67	81.52	85.33	89.51
Happy	76.67	78.1	87	86.14	85.33	83.12	91.33	96.3
Neutral	90.67	90.37	96	95.68	95.33	91.08	98	96.08
Sad	92.67	92.67	91.33	94.32	90.67	92.83	91.67	93.7
Surprised	86.33	81.57	87.67	85.95	83.33	84.6	83	83.84
Overal Accuracy	84.4		90.4		86.67		89.87	
Difference	6				3.2			

Both DoG and Sobel filters reduced the performance of trained models. To investigate further, the confusion matrices for these two models were compared to the baseline model. Referring Appendix-D, the confusion matrices showed that by applying DoG filter, the probabilities for the trained models to falsely predict anger/surprise as happiness were significantly increased by 0.09/0.2 and 0.11/0.14 for English and Chinses respectively ($(p = 0.0003,0.001 < 0.05\ respectively)$. There were also increased confusions for the trained model to falsely predict anger and sadness as neutral, particularly for English speech. with applying Sobel filter to Mel Spectrograms, similar effects were observed for Chinese emotional speech. Furthermore, the performances were decreased across all emotions for English.

4.3 Results for Feature Fusion

Based on the results discussed above, DoG-filtered MFCCs and CLAHE-filtered Mel Spectrogram were fused to determine if the fused feature could improve the performance. Based on Table 9, the overall accuracy for predicting emotions in English (87.75±3.67%) was improved by 3.3% and that for predicting emotions in Chinses (90.3±3.88%) remained the same when comparing to the baseline model trained with Mel Spectrogram. The difference in accuracies between the two languages was improved to 2.55%. Compared to the results presented in Table 6., the standard deviations were the smallest and second smallest for English and Chinese respectively. This suggests that the developed model was less biased toward some emotions than others.

Table 7. Average recall and accuracy for the model trained with fused features.

	Fused Feature - Mel Spec + MFCC			
	English		Chinese	
Emotion Classes	UAR(%)	WAR(%)	UAR(%)	WAR(%)
Angry	85.33	86.63	83.67	89.38
Happy	82.67	82.94	93	88.01
Neutral	92.33	90.82	95	94.84
Sad	97.33	92.85	93.33	93.65
Surprised	81	84.97	84	84.14
Overall Accuracy	87.75		90.3	
Difference	2.55			

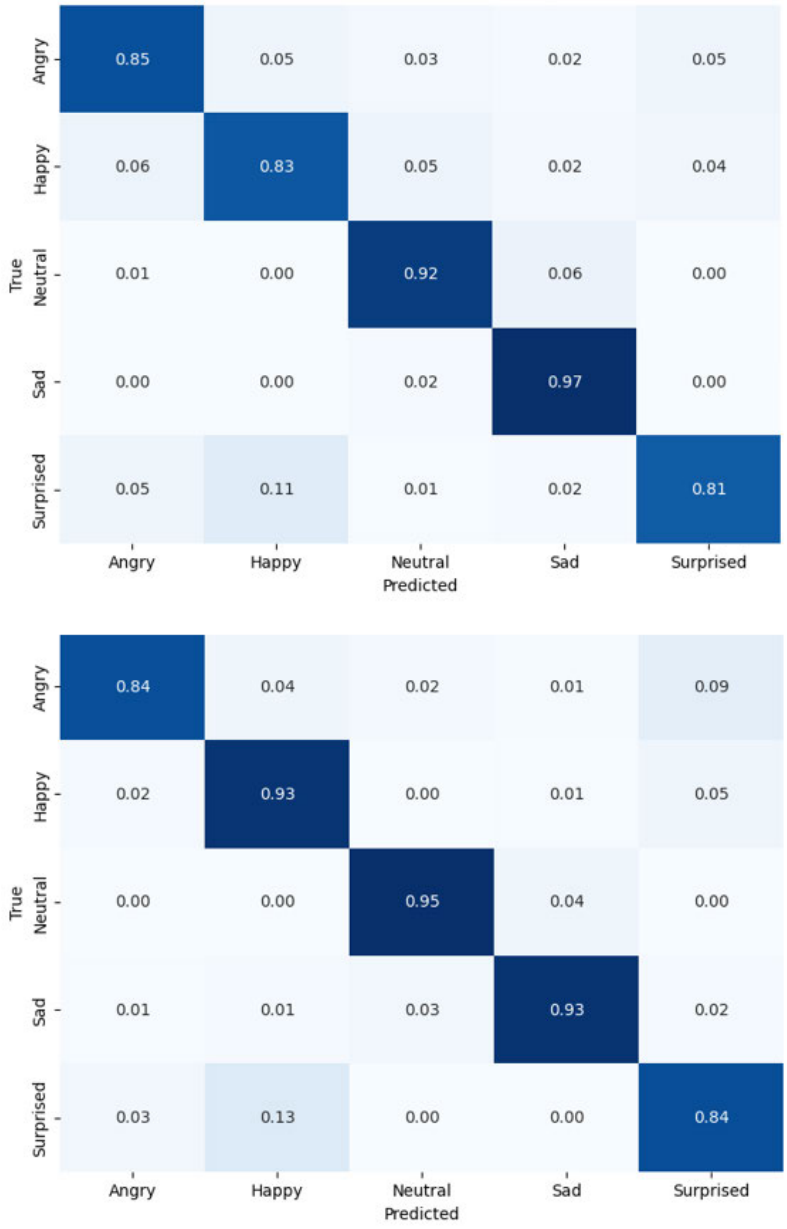


Figure 16.Confusion Matrices for the model trained with fused features for English (top) and Chinese (bottom). Horizontal axis represents the predicted emotions while the vertical axis represents the true emotions. Higher average recall is indicated by darker blue.

5 DISCUSSION

5.1 Visual Differences in Mel Spectrograms and MFCCs

A visual analysis of Mel Spectrograms and MFCCs was first conducted to identify differences based on languages and emotions. Figure 9 and Figure 10 provided clear illustrations of these differences. In Figure 9, Mel Spectrograms are compared based on languages and emotions. It is observed that the energy distribution in Mel Spectrograms differed between happiness and neutral emotions, as well as between English and Chinese. This finding is consistent with previous research in speech emotion recognition, which has identified that emotional speech tends to exhibit distinct spectral characteristics in comparison to neutral speech. Specifically, the differences in energy distribution and frequency bands between happiness and neutral emotions can be attributed to the varying prosodic features and pitch patterns associated with different emotional states (Jiang, et al., 2019). Comparing happiness in Chinese and English, the study found that English speech displayed a less distinctive harmonic structure, indicating that emotional speech characteristics can differ between languages. This observation is consistent with existing research that has emphasized the role of language-specific phonetic and prosodic patterns in emotional speech expression (Li & Akagi, 2019). The distinctiveness of emotional speech across languages underscores the importance of training emotion recognition models on diverse datasets to ensure cross-lingual robustness.

The study notes that the variation in MFCC coefficients was more noticeable for speech in happiness compared to neutral speech. This visual difference is consistent with the fact that emotional speech typically exhibits greater spectral variations in comparison to neutral speech (Anagnostopoulos & Iliou, 2010). Chinese and English displayed different characteristics in MFCCs for happy speech. This is in line with the notion that languages can influence the acoustic features of emotional speech, such as the spectral envelope and cepstral coefficients. In the current literature, researchers have shown that MFCCs are effective features for speech emotion recognition due to their ability to capture spectral dynamics, formant information, and prosodic cues associated with emotional expressiveness (Chen, et al., 2016).

5.2 The effects of Different Processing Techniques

The impact of various processing techniques, such as DoG filtering, Sobel filtering, and CLAHE filtering, on the accuracy of emotion recognition models was evaluated. The findings revealed both improvements and drawbacks in the performance of these models.

5.2.1 DoG Filtering

DoG filtering applied to MFCCs led to a significant improvement in model accuracy. This improvement can be attributed to the filtering process enhancing the discriminative features of MFCCs for emotion recognition. By sharpening the spectral features, DoG filtering makes it easier for the model to distinguish between emotional and neutral speech. However, the study also noted increased confusion between certain emotions,

such as anger/surprise and anger/sadness, indicating that while DoG filtering can enhance overall accuracy, it may introduce some trade-offs in emotion discrimination.

techniques aim to emphasize relevant spectral details and suppress noise, contributing to more effective feature representations for emotion classification. The trade-offs observed in the study align with the broader discussions in the field about the challenges of balancing emotion recognition accuracy and the potential for increased confusion between similar emotions.

5.2.2 Sobel and CLAHE Filtering

In contrast, applying Sobel and CLAHE filtering to MFCCs resulted in reduced accuracy and increased differences in accuracy between Chinese and English. These filters seemed to introduce more noise or distortions into the feature representations, which was also observed in the study by Geleijnse & Rieger (2022), making it more challenging for the model to correctly identify emotions. The results also suggested increased confusion between happiness and surprise, as well as between other emotions. The negative impact could be due to excessive filtering or pre-processing that could harm the discriminative power of features. While some level of noise reduction can be beneficial, overprocessing may lead to the loss of essential emotional cues and introduce unintended artifacts.

5.2 Feature Fusion

The fusion of Mel Spectrograms and MFCCs was a crucial aspect of this study, aimed to assess whether the fused feature can enhance the performance of emotion prediction. As discussed in the literature review, fusing features has proven to be improving the performance of the developed SER model.

The improvements shown in this study was not as significant compared to the study done by Peng, et al (2020), where Mel Spectrograms were also fused with MFCCs to improve the environmental sound classification. However, the modest but consistent improvement in overall accuracy in English suggests that these specific feature combinations provided complementary information for emotion classification, which aligns with the broader literature (Toyoshima, et al., 2023) (Jothimani & Premalatha, 2022) (Meng, et al., 2019), where feature fusion has been applied to capture different aspects of acoustic and facial features for emotion recognition. The reduction in bias toward certain emotions as reflected in the smaller standard deviations for both English and Chinese, is a promising finding as many emotion recognition systems tend to perform better on certain emotions while struggling with others.

The results also highlight the importance of cross-lingual emotion recognition, as indicated by the reduction in the performance gap between English and Chinese when fused features were employed. Feature fusion as observed in this study, can potentially serve as a valuable technique for mitigating this gap and developing a more generalised model that can achieve more consistent results across different languages.

6 CONCLUSION& FUTURE DIRECTION

6.1 Conclusion

In conclusion, this study aimed to investigate and enhance speech emotion recognition (SER) models, particularly focusing on the impact of different processing techniques and feature fusion on the performance of a bi-lingual model. The SER model trained with fused feature displayed average accuracies of 87.75±3.67% and 90.3±3.88% for English and Chinese respectively, with an accuracy difference of 2.55%. The key findings and their contributions to the field are as follows.

Visual analysis of Mel Spectrograms and MFCCs unveiled pronounced spectral differences in emotional speeches, which aligns with existing research. These visual disparities signify the importance of distinct spectral characteristics in recognizing emotions. Notably, the variations in emotional speech characteristics across languages underscore the significance of training SER models on diverse datasets to achieve cross-lingual robustness. It suspected that language-specific phonetic and prosodic patterns in emotional speech could play a pivotal role in these differences.

DoG (Difference of Gaussian) filtering, when applied to MFCCs, led to a substantial improvement in model accuracies for both English and Chinese. This enhancement can be accredited to the filtering process sharpening the spectral features, making it easier for the model to differentiate between emotional and neutral speech. The improved accuracy was, however, accompanied by increased confusion between specific emotions, such as anger/surprise and anger/sadness. This suggested that DoG filtering enhances overall accuracy but might introduce trade-offs by affecting the model's ability to discriminate between closely related emotions. One speculation is that the sharpening of spectral features might lead to subtle variations being magnified, causing confusion. In contrast, Sobel and CLAHE filtering applied to MFCCs resulted in reduced accuracy and amplified differences in accuracy between Chinese and English. These filters appeared to introduce more noise or distortions into the feature representations. Speculatively, over-processing or aggressive noise reduction could result in the loss of essential emotional cues or even the introduction of unintended artifacts, making similar emotions less distinguishable from each other. These filters might strip away critical emotional information, leading to the reduced accuracy and increased differences in performance between languages.

The fusion of Mel Spectrograms and MFCCs was a crucial aspect of this study, aimed at assessing whether the fused features could enhance the performance of emotion predictions. The improvements observed in this study, although not as significant as in prior research, were particularly obvious for English speech emotion recognition. This modest but consistent improvement in overall accuracy suggested that the specific feature combinations of Mel Spectrograms and MFCCs provide complementary information for emotion classification. Importantly, the results showed a reduction in bias toward certain emotions, which is a promising finding. This is significant in the field of SER since many emotion recognition systems tend to perform better on certain emotions

while struggling with others. The reduction in the performance gap between English and Chinese, when fused features were employed, underscores the potential of feature fusion as a valuable technique for mitigating this gap and developing a more robust generalised model that consistently performs well across different languages. Feature fusion could bridge the performance variations between languages and contribute to more consistent results in cross-lingual emotion recognition.

This research contributes to the field by providing insights into why certain filters improved SER model performance while others did not. The findings offered valuable guidance on the trade-offs associated with different processing techniques, highlighting the need to balance between feature enhancement and potential drawbacks such as noise introduction.

Despite these contributions, this study acknowledges that further research is needed to optimise the performance of processing techniques and feature fusion. The study speculates that confusion between similar emotions might be mitigated through more tailored processing approaches. Additionally, there is room for refining the use of filters to prevent feature distortion, which could enhance SER model accuracy. The study also recognises the challenges associated with evaluating model robustness across different languages and the need for more extensive cross-lingual testing.

In summary, this study provides valuable insights into the optimisation of SER models through image processing techniques and feature fusion, offering a pathway toward more generalised and effective emotion recognition models while recognising the need for continued research in this evolving field.

6.2 Future Directions

Future research could dive deeper into optimising processing techniques. This could include fine-tuning the parameters of filters like DoG, Sobel, and CLAHE to strike a balance between feature enhancement and noise as well as artefact introduction. Understanding how these techniques affect the spectral characteristics of emotional speech could lead to developing a customised adaptive filter that enhance the relevant spectral features. Other more complex filters or combining multiple filters to build a multi-stage adaptive filtering process could be the answer to improve the model performance.

Expanding research into the practical applications of SER is crucial. Emotion recognition has significant potential in industries like customer service, mental health, and entertainment. Understanding the real-world challenges and requirements of these applications will be essential for the field's growth. Before the practical application, the model should be trained with larger dataset with real-life speech data to increase the robustness against noise and the generalisability of the model. Mixed emotions could also be explored for the same purpose.

BIBLIOGRAPHY

Al-onazi, B. B. et al., 2022. Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion. *Applied Sciences,* 12(18), p. 9188.

Alzubaidi, L. et al., 2021. Review of Deep Learning: Concepts, CNN architectures, Challenges, Applications, Future Directions. *Journal of Big Data,* 8(53).

Anagnostopoulos, C. N. & Iliou, T., 2010. Towards Emotion Recognition from Speech: Definition, Problems and the Materials of Research. In: *Semantics in Adaptive and Personalized Services.* Berlin: Springer, pp. 127-143.

Anrarjon, T. & Mustaqeem, K., 2020. Deep-net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep System. *Sensors,* Volume 20, p. 5212.

Asahi, K. & Ogawa , A., 2006. Reduction of Noise in Speech Signals through Image Processing Using Spectrogram. *IEEJ Transactions on Electronics Information and Systems,* 126(12), pp. 1483-1489.

Australian Institute of Health and Welfare, 2023. *Prevalence and Impact of Mental Illness.* [Online] Available at: https://www.aihw.gov.au/mental-health/topic-areas/mental-illness [Accessed 10 October 2023].

Badshah, A. M., Ahmad, J., Rahim, A. & Baik, S. W., 2017. *Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network.* Busan, IEEE.

Bhangale, K. & Mohanaprasad, K., 2021. *Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network.* Singapore, Springer.

Burkhardt, F. et al., 2005. A Database of German Emotional Speech. *Proceedings of 9th European Conference: Speech Communication and Technology,* pp. 1517-1520.

Busso, C. et al., 2008. IEMOCAP: Interactive Emotional Dyadicmotion Dyadic Motion Capture Database. *Language Resources and Evaluation,* 42(4), pp. 335-359.

Cadore, J., Gallardo-Antolin, A. & Pelaz-Moreno, C., 2011. *Morphological Processing of Spectrograms fro Speech Enhancement.* Berlin, Springer.

Cen, L., Wu, F., Yu, Z. & Hu, F., 2016. A Real-Time Speech Emotion Recognition System and its Application in Online Learning. In: *Emotions, Technology, Design, and Learning.* s.l.:Academic Press, pp. 27-46.

Chen, C., Bunescu, R., Xu, L. & Liu, C., 2016. Tone Classification in Mandarin Chinese using Convolutional Neural Networks. *Interspeech,* pp. 2150-2154.

Chen, L., Mao, X., Xue, Y. & Cheng, L. L., 2012. Speech Emotion Recognition: Features and Classification Models. *Digital Signal Processing,* Volume 22, pp. 1154-1160.

Costantini, G., Iaderola, I., Paoloni, A. & Todisco, M., n.d. EMOVO Corpus: an Italian Emotional Speech Database. *International Conference on Language Resources and Evaluation, European Language Resources Association,* pp. 3501-3504.

Dahake, P. P., Shaw, K. & Malathi, P., 2016. *Speaker Dependent Speech Emotion Recognition Using MFCC and Support Vector Machine.* Pune, IEEE.

Dair, Z., Donovan, R. & O'Reilly, R., 2022. Linguistic and Gender Variation in Speech Emotion Recognition using Spectral Features. *arXiv.*

Dilley, L. C. & Heffner, C. C., 2013. The Role of f0 Alignment in Distinguishing Intonation Categories: Evidence from American English. *Journal of Speech Sciences,* 3(1), pp. 3-67.

Dong, G., Pun, C. & Zhang, Z., 2022. Temporal Relation Inference Network for Multi-Modal Speech Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology.*

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P., 2003. Emotional Speech: Towards a New Generation of Databases. *Speech Communication,* Volume 40, pp. 33-60.

El Ayadi, M., Kamel, M. & Karray, F., 2011. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition,* 1(44), pp. 572-587.

Engberg, I. & Hansen, A., 1996. *Documentation of the Danish Emotional Speech Database.* [Online]
Available at: http://cpk.auc.dk/tb/speech/Emotions/
[Accessed 15 September 2023].

Eswarsai, 2021. *Exploring Different Types of LSTMs.* [Online]
Available at: https://medium.com/analytics-vidhya/exploring-different-types-of-lstms-6109bcb037c4
[Accessed 9 October 2023].

Fang, J., Finn, A., Wyber, R. & Brinkworth, R. S. A., 2022. Acoustic Detection of Unmanned Aerial Vehicles using Biologically Inspired Vision Processing. *The Journal of the Acoustical Society of America,* Volume 151, pp. 968-981.

Frick, R. W., 1985. Communicating Emotion: The Role of Prosodic Features. *Psychological Bulletin,* Volume 97, pp. 412-429.

Fu, Y. et al., 2018. Screen Content Image Quality Assessment Using Multi-Scale Difference of Gaussian. *IEEE Transmission of Circuits System Video Technology,* 28(9), pp. 2428-2432.

Gao, Y., Li, B., Wang, N. & Zhu, T., 2017. *Speech Emotion Recognition Using Local and Global Features.* Beijing, Springer.

Geleijnse, G. & Rieger, B., 2022. Influence of Edge Enhancement Applied in Endoscopic Systems on Sharpness and Noise. *Journal of Biomedical Optics,* 27(10).

Gouyon, F., Pachet, F. & Delerue, O., 2000. *On the Use fo Zero-Crossing Rate for an Application of Classification of Percussive Sounds.* s.l., s.n.

Grey, J. M. & Gordon, J. W., 1978. Perceptural Effects of Spectral Modifications on musical Timbres. *Journal of the Acoustical Society of America,* Volume 63, pp. 1493-1500.

Habib, M. et al., 2021. Towards an Automatic Quality Assessment of Voice-Based Telemedicine Consultations: A Deep Learning Approach. *Sensors,* Volume 21, p. 3279.

Hassan, C., 2017. *Voice-Activated Device Called 911 During Attack, New Mexico Authorities Say.* [Online]
Available at: https://edition.cnn.com/2017/07/10/us/alexa-calls-police-trnd/index.html
[Accessed 18 October 2023].

Heredia, J. et al., 2022. Aaptive Multimodal Emotion Detection Architecture for Social Robots. *IEEE Access,* Volume 10, pp. 20727-20744.

Hochreiter, S. & Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computing,* 9(8), pp. 1735-1780.

Huang, Z., Dong, M., Mao, Q. & Zhan, Y., 2014. Speech Emotion Recognition using CNN. *Proceedings of the 22nd ACM International Conference on Multimeadia,* pp. 801-804.

Iosifov, I. et al., 2022. Transferability Evaluation of Speech Emotion Recognition Between Different Languages. *International Conference on Computer Science, Engineering and Education Applications,* pp. 413-426.

Jackson, P. J. B. & Haq, S., 2010. Multimodal Emotion Recognition. *Machine Audition: Principles, Algorithms and Systems,* pp. 398-423.

Jana, S., Parekh, R. & Sarkar, B., 2021. A Semi-Supervised Approach for Automatic Detection and Segmentation of Optic Disc from Retinal Fundus Image. In: *Handbook of Computational Intelligence in Biomedical Engineering and Healthcare.* Bengal: Academic Press, pp. 65-91.

Jiang, L. et al., 2019. Speech Emotion Recognition Using Emotion Perception Spectral Feature. *Concurrency and Computation,* 33(11).

Johnston, J. D., 1988. Transform Coding of Audio Signals Using Perceptual Noise Criteria. *The IEEE Journal on Selected Areas in Communications,* Volume 6, pp. 314-323.

Jothimani, S. & Premalatha, K., 2022. MFF-SAug: Multi Feature Fusion with Spectrogram Augmentation of Speech Emotion Recognition Using Convolution Neural Network. *Chaos, Solitions & Factals,* Volume 162.

Kerkeni, L. et al., 2019. Automatic Speech Emotion Recognition Using Machine Learning. *Soical Media and Machine Learning.*

Koolagudi, S. G. & Rao, K. S., 2012. Emotion Recognition from Speech: Review. *International Journal on Speech Technology Springer,* pp. 99-117.

Lee, M., Yeh, S., Chang, J. & Chen, Z., 2022. Research on Chinese Speech Emotion Recognition Based on Deep Neural Network and Acoustic Features. *Sensors,* 22(4744).

librosa, 2023. *librosa.feature.mfcc.* [Online]
Available at: https://librosa.org/doc/main/generated/librosa.feature.mfcc.html
[Accessed 18 Oct 2023].

Li, F., Ma, J. & Huang, D., 2005. MFCC and SVM Based Recognition of Chinese Vowels. *Computational Intelligence and Security,* Volume 3802, pp. 812-819.

Lim, W., Jang, D. & Lee, T., 2016. *Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks.* Jeju, IEEE.

Livingstone, S. R. & Russo, F. A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS): A Dynamic, Multimodel Set of Facieal and Vocal Expressions in North American English. *PLoS ONE,* Volume 13, p. e0196391.

Li, X. & Akagi, M., 2019. Improving Multilingual Speech Emotion Recognition by Combining Acoustic Features in a Three-Layer Model. *Speech Communication,* Volume 110, pp. 1-12.

Madanian, S. et al., 2023. Speech Emotion Recognition Using Machine Learning - A Systematic Review. *Intelligent Systems with Applications,* Volume 20, p. 200266.

Mao, S., Ching, P. C. & Lee, T., 2021. Enhancing Segment-Based Speech Emotion Recognition by Deep Self-Learning. *arXiv,* 14(8).

Meng, H., Yuan, F. & Wei, H., 2019. Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network. *IEEE Access,* Volume 7, pp. 125868-125881.

Milton, A., Roy, S. S. & Selvi, S. T., 2013. SVM Scheme for Speech Emotion Recognition using MFCC Feature. *International Journal of Computer Applications,* 69(9), pp. 34-39.

Milton, A. & Tamil, S. S., 2013. SVM Scheme for Speech Emotion Recognition using MFCC Feature. *International Journal of Computer Applications,* 69(9), pp. 34-39.

Muljono, J. G., Pujiono, E. N. & Setiadi, D. R. M., 2023. Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning. *International Journal of Advanced Computer Science and Applications,* 14(6), pp. 198-206.

Murray, I. R. & Arnott, J. L., 1993. Towards a Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of Acoustic Society America,* 93(2), pp. 1097-1108.

Muthuvel, P. et al., 2012. *Emotion Recognition in Speech Signals using MFCC and Mel-Spectrogram Analysis.* Chennai, IEEE.

Nema, B. & Abdul-Kareem, A. A., 2018. Preprocessing Signal for Speech Emotion Recognition. *Journal of Science,* Volume 28, pp. 157-165.

Pandey, S. K., Shekhawat, H. S. & Prasanna, S. R. M., 2019. *Deep Learning Techniques for Speech Emotion Recognition: A Review.* Pardubice, IEEE.

Pan, Y., Shen, P. & Shen, L., 2012. Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home,* 6(2), pp. 101-108.

Pan, Y., Shen, P. & Shen, L., 2012. Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home,* 6(2), pp. 101-108.

Peng, N., Chen, A., Zhou, G. & Chen, W., 2020. Environment Sound Classification Based on Visual Multi-Feature Fusion and GRU-AWS. *IEEE Access,* Volume 8, pp. 191100-191114.

Petrushin, V. A., 2000. *Emotion in Speech: Recognition and Application to Call Centres.* s.l., Artificial Neural Networks in Engineering.

Plante, O., 2022. *5 Major Flaws of Voice Assistant Technology in 2022.* [Online]
Available at: https://www.fleksy.com/blog/5-major-flaws-of-voice-assistant-technology-in-2022/

Ramakrishnan, S. & El-Emary, I. M., 2013. Speech Emotion Recognition Approaches in Human Computer Interaction. *Telecommunication Systems,* 52(3), pp. 1467-1478.

Rieger, S. A., Muraleedharan, R. & Ramachandran, R. P., 2014. *Speech Based Emotion Recognition Using Spectral Feature Extraction and an Ensemble of kNN Classifiers.* Singapore, IEEE.

Rodriguez-Rodriguez, J. A., Molina-Cabello, M. A., Benitez-Rochel, R. & Lopez-Rubio, E., 2020. *The Effect of Image Enhancement Algorithms on Convolutional Neural Networks.* Milan, IEEE.

Singh, Y. B. & Goel, S., 2022. A Systematic Literature Review of Speech Emotion Recognition Approaches. *Neurocomputing,* Volume 492, pp. 245-263.

Sinith, M. S. et al., 2015. Emotion Recognition from Audio Signals Using Support Vector Machine. *IEEE Recent Advances in Intelligent Computational Systems,* pp. 139-144.

Steidl, S., 2009. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. *Studien zur Musterekennung,* Volume 28.

Swain, M., Routray, A. & Kabisatpathy, P., 2018. Databases, Features and Classifiers for Speech Emotion Recognition: A Review. *International Journal of Speech Technology,* Volume 21, pp. 93-120.

Toyoshima, I. et al., 2023. Multi-Input Speech Emotion Recognition Model Using Mel Spectrogram and GeMAPS. *Sensors,* 23(3), p. 1743.

Vidrascu, L. & Devillers, L., 2005. Detection of Real-life Emotions in Call Centers. *Interspeech,* pp. 1841-1844.

Wang, C. et al., 2022. Speech Emotion Recognition Based on Multi-Feature and Multi-Lingual Fusion. *Multimedia Tools and Applications,* Volume 81, pp. 4897-4907.

Wani, M., Routray, A. & Kabisatpathy, P., 2021. Databases, Features and Classifiers for Speech Emotion Recognition: A Review. *International Journal of Speech Technology,* Volume 9, pp. 47795-47814.

World Health Organization, 2023. *Autism.* [Online]
Available at: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders#:~:text=It%20is%20estimated%20that%20worldwide,figures%20that%20are%20substa

ntially%20higher.
[Accessed 28 September 2023].

Ye, J. et al., 2023. Temporal Modelling Matters: A Novel Temporal Emotional Modelling Approach for Speech Emotion Recognition. *arXiv.*

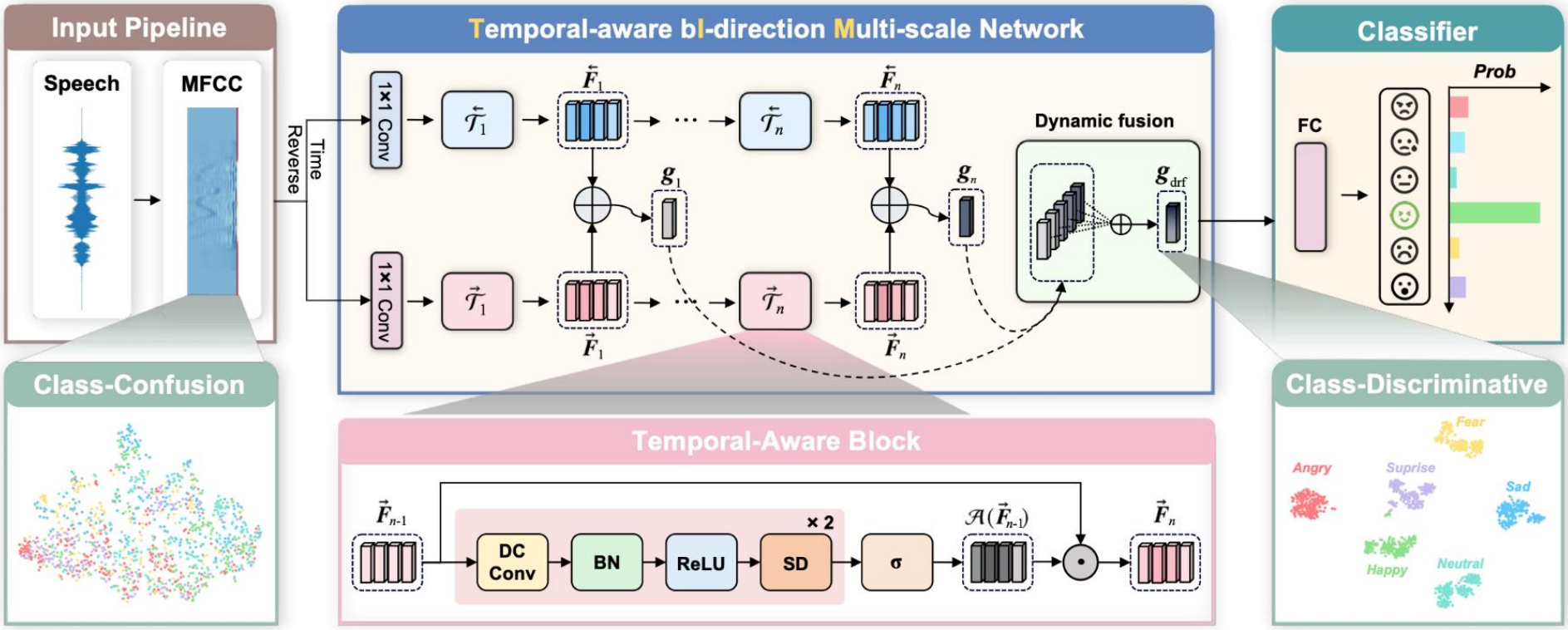
Zhao, J., Mao, X. & Chen, L., 2019. Speech Emotion Recognition Using Deep 1D&2D CNN LSTM Networks. *Biomedical Signal Processing and Control,* Volume 47, pp. 312-323.

Zhou, K., Sisman, B., Liu, R. & Li, H., 2022. Emotional Voice Conversion: Theory, Databases and ESD. *Speech Communication,* Volume 137, pp. 1-18.

Zhou, Z. et al., 2018. Exploring Spatio-Temporal Representations by Integrating Attention-based Biodirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition. *Interspeech.*

APPENDICES

Appendix A - TIM-Net architecture



Appendix B - ESD Database Statistics

Parameter	Chinese						English					
	Neu	Ang	Sad	Hap	Sur	All	Neu	Ang	Sad	Hap	Sur	All
# speakers	10	10	10	10	10	10	10	10	10	10	10	10
# utterances per speaker	350	350	350	350	350	1750	350	350	350	350	350	1750
# unique utterances	350	350	350	350	350	350	350	350	350	350	350	350
# characters/words per speaker	4005	4005	4005	4005	4005	20,025	2203	2203	2203	2203	2203	11,015
# unique characters/words	939	939	939	939	939	939	997	997	997	997	997	997
Avg. utterance duration [s]	3.23	2.68	4.04	2.84	3.32	3.22	2.61	2.80	2.98	2.70	2.73	2.76
Avg. character/word duration [s]	0.28	0.23	0.35	0.25	0.29	0.28	0.41	0.44	0.47	0.43	0.43	0.44
Total duration [s]	11,305	9380	14,140	9940	11,620	56,385	9135	9800	10,430	9450	9555	48,370

Emotion abbreviations are used as follows: *Neu* stands for neutral, *Ang* stands for angry, *Sad* stands for sad, *Hap* stands for happy and *Sur* stands for surprise. The statistics of characters are reported for Chinese, and the statistics of words are reported for English.

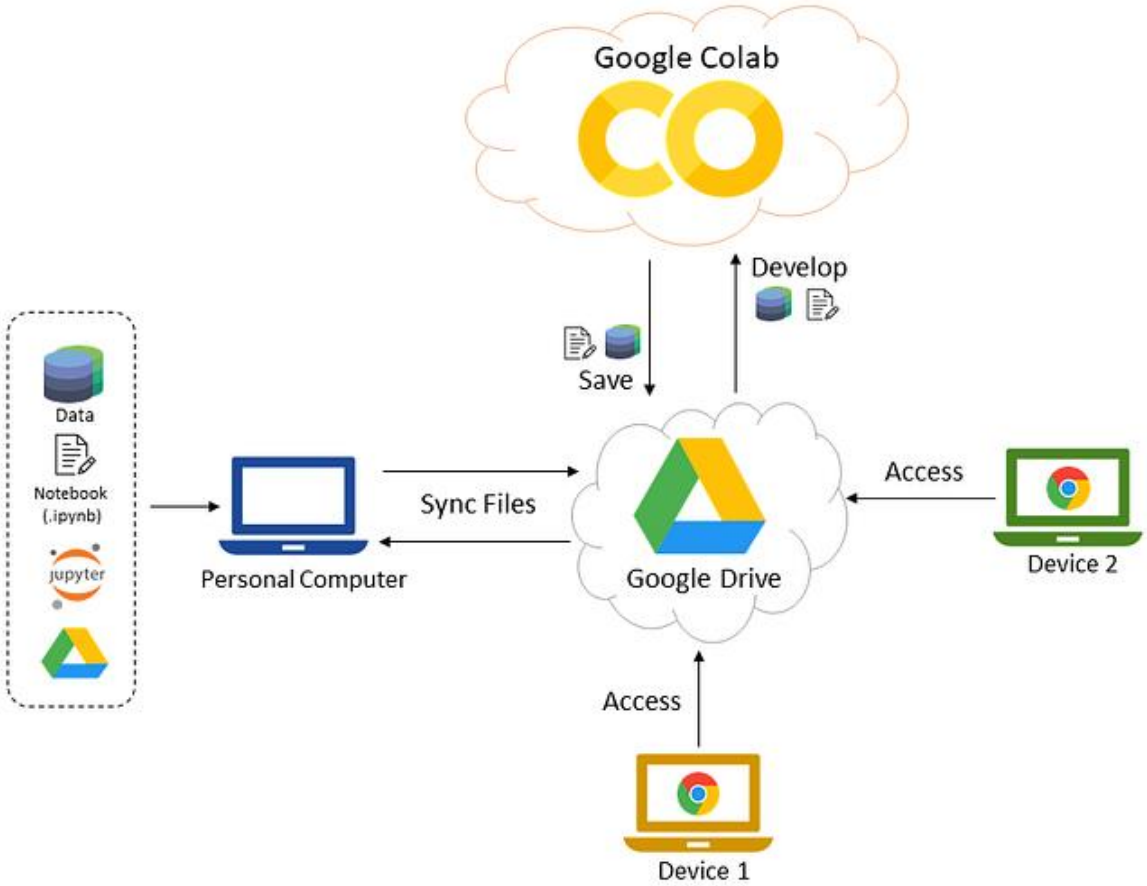
Appendix C - Acted and Induced Datasets for training SER models

Dataset	Type	Language	Emotion (#)	Size	SER models based on this dataset
Danish Emotional Database (Engberg & Hansen, 1996)	Simulated	Danish	5	Two male actors, two female actors, two isolated words, nine sentences and two passages	(Huang, et al., 2014) (Ramakrishnan & El-Emary, 2013)
EMODB (Burkhardt, et al., 2005)	Acted	German	7	10 speakers, 10 utterances each	(Badshah, et al., 2017) (Ye, et al., 2023) (Gao, et al., 2017) (Sinith, et al., 2015) (Milton, et al., 2013)
IEMOCAP (Busso, et al., 2008)	Acted	English	10	5 male actors, 5 female actors	(Zhao, et al., 2019) (Dong, et al., 2022) (Heredia, et al., 2022) (Ye, et al., 2023)
SAVEE (Jackson & Haq, 2010)	Acted	English	7	4 male actors with 120 utterances	(Huang, et al., 2014) (Li & Akagi, 2019) (Ye, et al., 2023)
RAVDESS (Livingstone & Russo, 2018)	Acted	English	8	12 male and 12 female speakers with 2 utterances each	(Ye, et al., 2023) (Gao, et al., 2017) (Mao, et al., 2021)
CASIA	Simulated	Chinese	6	4 speaker, 500 utterances in total	(Ye, et al., 2023)
EMOVO (Costantini, et al., n.d.)	Acted	Italian	7	6 speakers, 588 utterances in total	(Ye, et al., 2023) (Dair, et al., 2022)

Appendix D - Prosodic features in different emotions (Anagnostopoulos & Iliou, 2010)

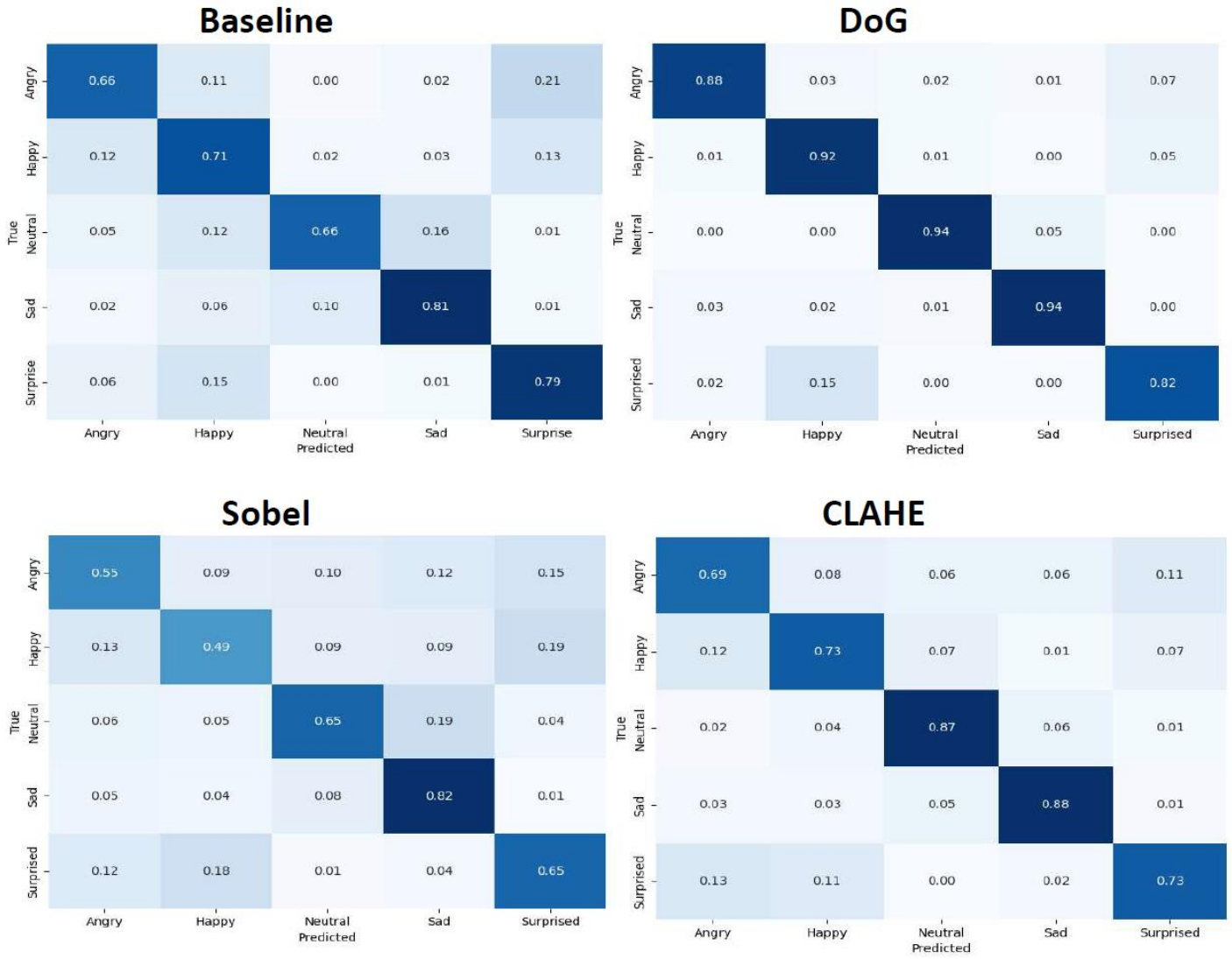
	Anger	Happiness	Sadness	Fear	Disgust
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much higher	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tonic	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

Appendix E – Architecture of Google Colab

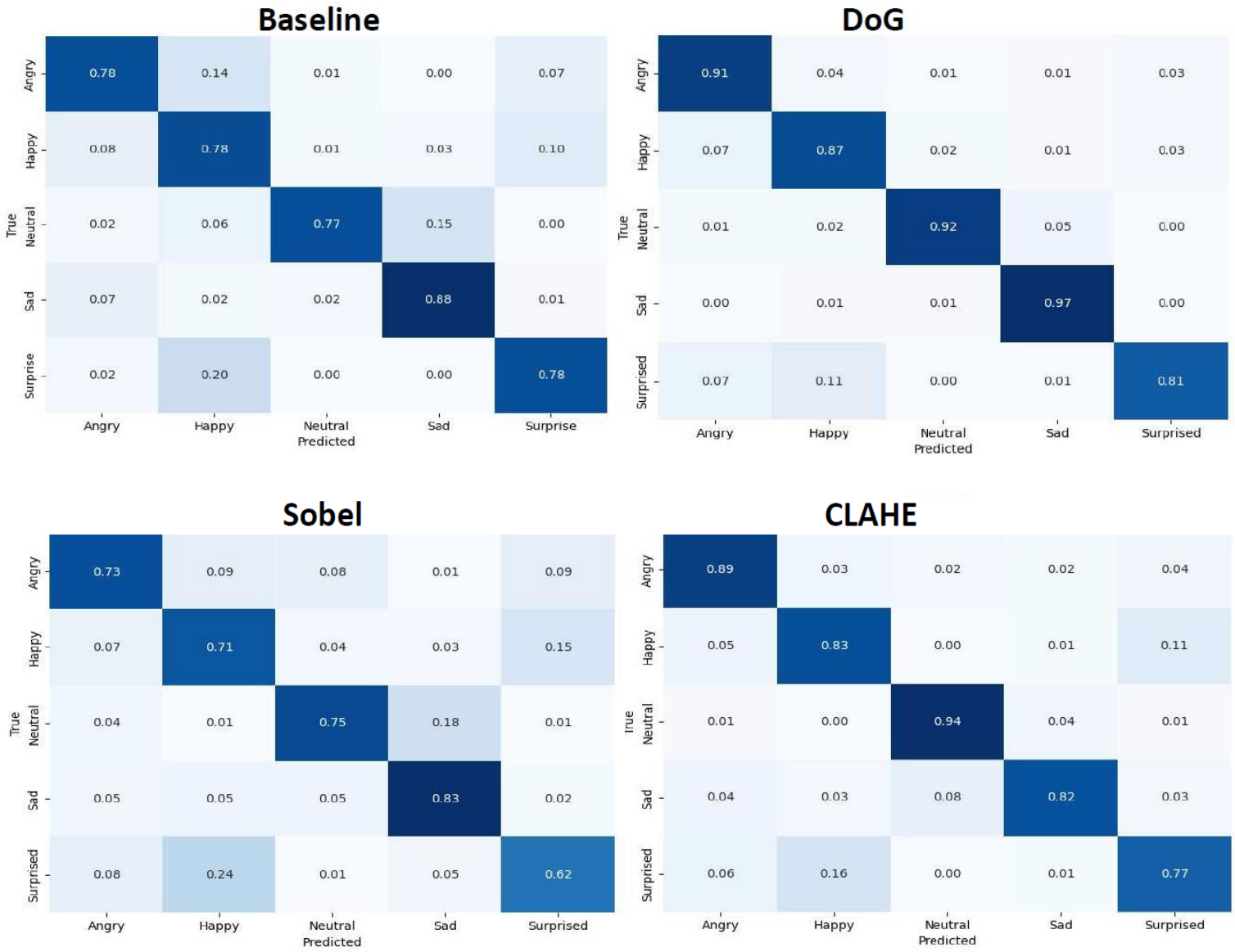


Appendix F Results: Confusion Matrices

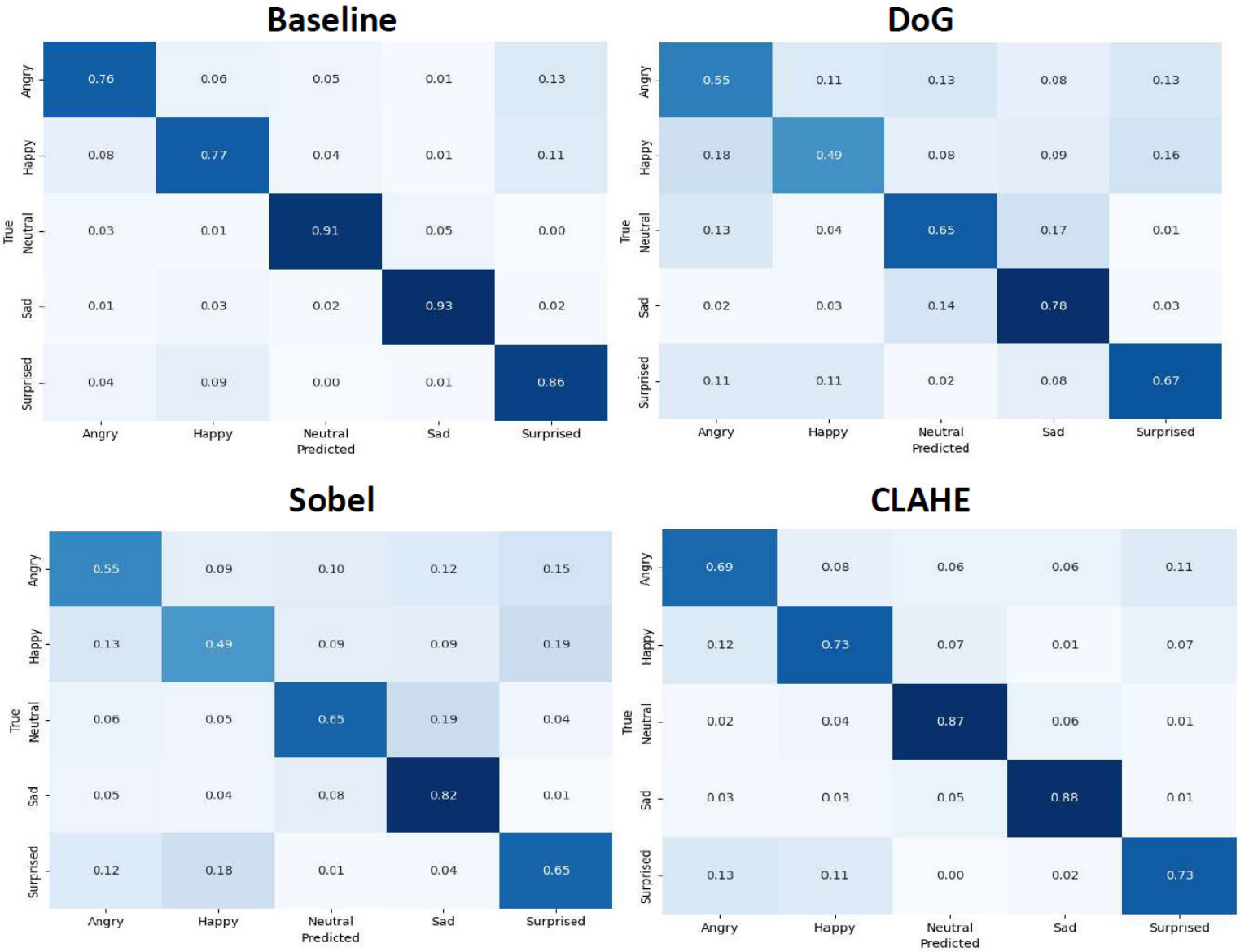
MFCC – English



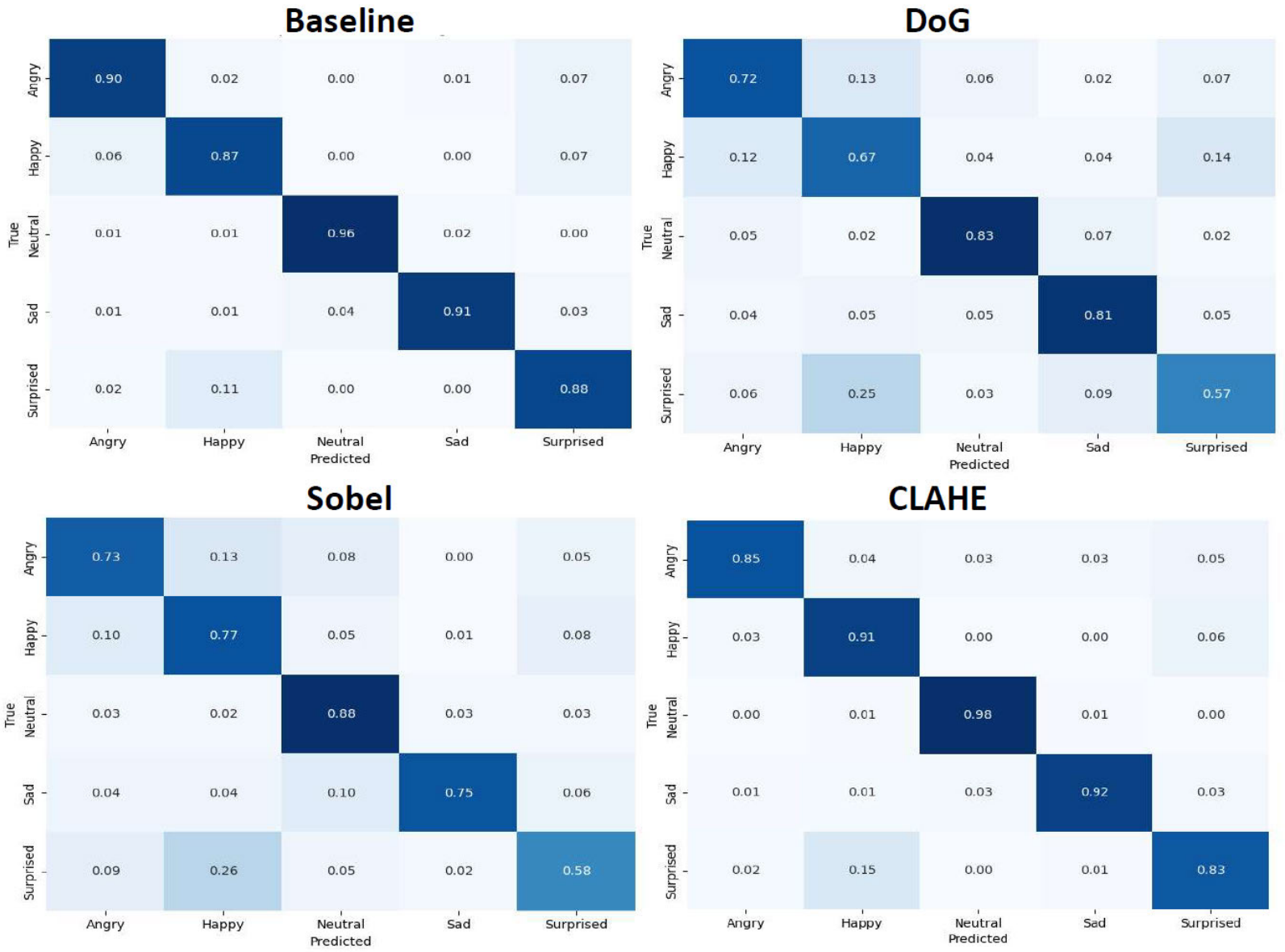
MFCC – Chinese



Mel Spectrogram - English



Mel Spectrogram - Chinese



Appendix G - Results: Average Accuracy and Average Recall

	MFCC - Baseline				MFCC - DoG				MFCC - Sobel				MFCC - CLAHE			
	English		Chinese		English		Chinese		English		Chinese		English		Chinese	
Emotion Classes	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)
Angry	65.67	68.76	77.67	79.25	88	90.57	91.33	88.39	54.67	57.34	73	73.86	68.67	69.13	88.67	86.79
Happy	70.67	65.84	78.33	71	92.33	86.83	87	85.15	49.33	53.14	71	67.51	73	73.49	83	80.84
Neutral	65.67	74.06	77	84.93	94.33	95.13	91.67	93.7	65.33	67.59	74.67	77.37	87.33	84.65	94.33	92.03
Sad	81	80.07	88	85.44	93.67	93.67	97.33	95.11	81.67	72.38	83	78.92	87.67	86.23	82.33	86.67
Surprised	78.67	73.29	77.67	79.25	82.33	84.59	81.33	86.22	65	63.52	61.67	65.37	72.67	75.3	77	78.84
Overal Accuracy	72.33		79.73		90.13		89.73		63.2		72.67		77.87		85.07	
Difference	7.4				0.4				9.47				7.2			
	Mel Spectrogram - Baseline				Mel Spectrogram - DoG				Mel Spectrogram - Sobel				Mel Spectrogram - CLAHE			
	English		Chinese		English		Chinese		English		Chinese		English		Chinese	
Emotion Classes	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)
Angry	75.67	79.09	90	90.15	55	55.46	72	72.12	52	54.26	72.67	72.91	78.67	81.52	85.33	89.51
Happy	76.67	78.1	87	86.14	49.33	55.22	66.67	62.7	60.67	56	76.67	68.76	85.33	83.12	91.33	96.3
Neutral	90.67	90.37	96	95.68	65.33	64.69	83	82.72	76.33	65.9	88	81.48	95.33	91.08	98	96.08
Sad	92.67	92.67	91.33	94.32	78.33	71	81	80.07	63.33	69.47	75.33	82.78	90.67	92.83	91.67	93.7
Surprised	86.33	81.57	87.67	85.95	67.33	67.22	57	61.73	52.33	58.91	58	64.56	83.33	84.6	83	83.84
Overal Accuracy	84.4		90.4		63.07		71.93		60.93		74.13		86.67		89.87	
Difference	6				8.86				13.2				3.2			