A Study on Elderly Speech Emotion Recognition for Low-Resource Languages

Lo Shun Hang Lincoln, 1155176490
Wong Ka Long Myron, 1155181717

Supervisor: Prof. Wu, Xixin

The Chinese University of Hong Kong
SEEM4999 Final Year Project II
Final Report

April 2025

1 Abstract

Speech emotion recognition (SER) is a growing field that utilizes algorithms to identify people's emotions. Most of the studies focus on studying speech emotion recognition for the general population. Not many studies address the more difficult and complicated problem - speech emotion recognition for the elderly. With the growing number of elderly and their need for mental health care, the demand of the elderly for speech emotion recognition must be addressed. Our paper investigates Meta's Wav2Vec 2.0. It is a learning model that allows for a better speech emotion recognition system. Through capturing acoustic features to identify speech emotion patterns for the elderly. We compare the performance of different data sets, including using datasets of different age groups and comparing that with using text-based features. Furthermore, we discuss the possible reason for lower accuracy with text-based features.

2 Introduction

2.1 Background

Hong Kong is a city facing significant population aging problems[10]. Alongside the aging population, Hong Kong does not have enough infrastructure to support the elderly care industry, which is critical to taking good care of the elderly citizens.

Artificial Intelligence can help with a lot of things, including elderly care. Studies have found that for the elderly, loneliness is a significant issue that can lead to various cognitive or mental problems[17]. To maintain the quality of life for the elderly, we can create Artificial Intelligence systems that can analyze whether an older adult is mentally unstable and provide them with the right support.

2.2 Problems

In the modern world where Artificial Intelligence is blooming, human-computer interaction is becoming more and more important[6]. As a result, the popularity of speech emotion recognition is also rising, unfortunately, the majority of studies tend to focus on young adults that speak English[8]. Despite the popularity in speech emotion recognition, there is little research on elderly speech emotion recognition, making it a hard problem to solve[22]. It is even harder for low-resource languages (e.g., Cantonese) to train a powerful enough elderly speech emotion recognition model that can support intensive use in real life with the lack of data.

To train powerful models under such harsh conditions, researchers have derived techniques that can fully utilize the small datasets to train a better model, these methods include automatic feature extraction and cross-domain learning[16, 5].

2.3 Objective

In this paper, two different approaches to low-resource training will be explored: cross-domain learning and concatenating high-level features. For cross-domain learning, we study the possibility of transferring emotion recognition abilities between different age groups (Adult - Elderly). By utilizing pre-trained speech models (e.g., XLSR-Wav2Vec 2.0), a representation of the speech can be acquired, a classifier will then be used to get predictions from the representations. Comparing the results, a conclusion of whether cross-domain learning can be beneficial for elderly speech emotion recognition can then be made. Text-based features will also be added to test the usefulness of text-based features in speech emotion recognition. Texts will be transcribed with a Cantonese fine-tuned Whisper model[1], and then it will be fed into a Cantonese fine-tuned fastText model[27] to extract its text representations.

By testing these approaches, we hope to answer the following questions:

- Whether cross-domain learning is viable in elderly speech emotion recognition
- Whether text-based features contribute to the training of elderly speech emotion recognition models

3 Related Works

3.1 Feature Concatenation

Feature concatenation involves combining two or more distinct feature vectors each representing different aspects of the same speech signal into a single, unified feature vector at the frame level. The goal is to leverage complementary information captured by these features to enhance the robustness and accuracy of the system[2].

Emotional content in speech is conveyed through variations in pitch, energy, speaking rate, and spectral characteristics all of which can be captured by different feature types. For example:

- **Mel-frequency cepstral coefficients (MFCCs)** - Widely used in SER, MFCCs capture the spectral envelope of speech, which reflects timbre and vocal tract characteristics influenced by emotion.
- **Gamma-tone filterbank-based features (e.g., GTF-CC)** - These model the human auditory system more closely and are sensitive to pitch and harmonic structures, which are critical for detecting prosodic cues like intonation patterns tied to emotions.
- **Prosodic features** - Features like fundamental frequency (F0), energy, and duration can directly indicate emotional arousal or valence.

** feature concatenation (1)

3.2 Cross-Domain Learning

Cross-domain learning refers to the process of training a model to recognize emotions from speech across different datasets (or corpora) where the feature distributions vary due to factors like speaker characteristics, recording conditions, languages, or corpora-specific attributes[31]. This is a critical challenge in real-world SER applications because models trained on one dataset often perform poorly when applied to another due to these domain differences, a phenomenon known as domain divergence or domain shift.

3.2.1 Domain Divergence Problem

In SER, datasets (e.g., IEMOCAP, MSP-Improv, SAVEE, Emo-DB) differ in aspects such as language (English vs. German), speaker demographics (gender, number of speakers), recording type (acted vs. hybrid), and emotional annotations. These differences lead to variations in the feature distributions of the source (training) and target (testing) datasets. Traditional SER models, trained and tested on the same dataset, struggle to generalize to unseen datasets because they inadvertently learn domain-specific, non-affective information (e.g., speaker identity, recording environment) alongside emotional cues[31]..

3.2.2 Domain Adversarial Neural Networks (DANN)

$$L = L_E(G(x, \theta), y) + \gamma L_D$$

Figure 1: Loss Function of Emotion Classifier [31].

1. **Feature Extractor** - A deep convolutional neural network (CNN) combined with a bidirectional LSTM (BLSTM) extracts features from speech spectrograms.
2. **Emotion Classifier** - Predicts emotions (e.g., arousal and valence) using these features.
3. **Domain Classifier** - Attempts to identify domain-specific attributes (e.g., corpus, language, gender).
4. **Gradient Reversal Layer (GRL)** - Inserted between the feature extractor and domain classifier, this layer reverses the gradient during backpropagation. This forces the feature extractor to maximize the domain classifier's loss (making domain features indistinguishable) while minimizing the emotion classifier's loss, thus learning domain-invariant emotional features.

3.2.3 Center Loss Integration

$$\begin{aligned} L_E(G(x, \theta), y) = \lambda Softmax(G(x, \theta), y) \\ +(1 - \lambda) Center(G(x, \theta), c) \end{aligned}$$

Figure 2: Overall objective function [31].

1. **Softmax Loss** - Separates different emotion classes by finding decision boundaries
2. **Center Loss** - Minimizes the Euclidean distance between feature representations and their corresponding emotion class centers, reducing intraclass variation (e.g., ensuring features of "happy" from different datasets cluster together)

3.3 Low-Level Features

Low-level features in SER are typically time- or frequency-domain descriptors of the audio signal that reflect acoustic properties such as pitch, energy, and spectral characteristics. These features are often hand crafted or extracted using well established signal processing techniques, making them ”knowledge-based” as they rely on prior understanding of how sound relates to human perception. In the paper, MFCC is highlighted as a key low-level feature due to its widespread use in speech processing and its ability to mimic the auditory characteristics of the human ear.

3.3.1 MFCC (Mel-frequency Cepstral Coefficients)

MFCCs are derived from the audio signal by applying a series of transformations, including the Fourier transform, Mel-scale filtering, and discrete cosine transform[21]. They represent the short-term power spectrum of sound on a perceptually relevant scale, capturing how humans perceive frequencies.

3.3.2 Alternative

ComParE feature set and eGeMAPs, which include additional low-level descriptors such as pitch, jitter, shimmer, and formants. However, MFCC is chosen as the target low-level feature for its complementary nature to high-level features[21].

3.3.3 Complementary Nature

Low-level features are significant in SER because they provide a foundational layer of information about the emotional content of speech that might not be fully captured by high-level features alone. While pre-trained models like wav2vec 2.0 extract high-level, context-rich representations encompassing acoustic cues and semantic information, they may overlook subtle, emotion-specific nuances present in the raw signal. The research argues that combining these low-level features with high-level features can improve emotion recognition accuracy, particularly in speaker-independent scenarios where the model must generalize across different voices[21].

MFCC and W2V2 are described as a complementary pair because they reflect emotional properties in different domains MFCC in the frequency domain and W2V2 in the time domain. This duality allows the model to capture a broader range of emotional cues.

3.4 Semi-Supervised Models

3.4.1 Dual Learning Objectives

Supervised task involves training a model to predict emotional categories (e.g., angry, happy, sad, neutral) using a limited set of labeled speech samples. For instance, the study uses 300 to 2400 labeled utterances from the IEMOCAP database to classify four emotions. Unsupervised task utilizes unlabeled data to learn robust feature representations by reconstructing the input signals. This auxiliary task acts as a regularizer, ensuring the model captures intrinsic data structures beyond just the labeled emotional cues. In the ladder network, this is achieved by reconstructing hidden representations at each layer[20].

3.4.2 Benefits

The model minimizes a combined loss - a supervised loss and an unsupervised cost. It addresses data scarcity, generalization, feature relevancy. Semi-supervised models exploit abundant unlabeled speech to enhance feature learning, reducing reliance on extensively labeled datasets. By regularizing the supervised task with unsupervised reconstruction, the model avoids overfitting to specific labeled samples, performing better across varied speakers and conditions. Unlike pure unsupervised methods (e.g., DAE, VAE) that may retain irrelevant information, the simultaneous training aligns unsupervised features with the supervised emotional classification goal, as seen in the ladder network outperforming DAE (56.3%) and VAE (58.0%) at 59.4% UAR with all samples using CNN[20].

3.5 Speech Emotion Recognition

3.5.1 Clinical Application

The primary application of SER is to improve the quality of doctor to patient communication by identifying and measuring the emotional states of patients during medical procedures. According to the world health organization (WHO), mental state refers to the state of being that a person can perform, communicate positively, face difficulties and experience life fruitfully. Study show that up to 60 percent of medical patients experience mental distress, including anxiety and depression [26]. Therefore, by automating emotion detection, the health care profession can monitor all patients emotion state at all time. SER can identify issues that otherwise maybe unnoticed. Furthermore, it can help improve the physician to patient interaction, potentially improving patient trust and treatment adherence. It helps objectify emotion state rather than subjective judgement by human. With a mature model, computational analysis can reduce bias and increase consistenc overtime [19].

4 Benchmark and Dataset

4.1 Benchmark

Our main objective is to test the usefulness of various techniques in low-resource learning, and the benchmarking will be done by comparing each model's:

- Accuracy
- F1 Score

For the classification of elderly and adult, the elderly are ≥ 55 years old, and adults are ≤ 54 years old. A vanilla version of the data will be used to train a baseline model, which means that a baseline model will be trained only with data from the elderly, and others will adopt the techniques that we have mentioned. Then, we will compare the accuracy and the F1 score of each model to the baseline model and see if there is a positive impact on the performance of a model after applying the techniques.

4.2 Dataset

Our dataset comprises 1 publicly available dataset (YueMotion[9]) and 1 dataset that we have gathered ourselves. YueMotion is a Cantonese speech emotion recognition database that has 18 speakers (11 adults, 7 elderly) with 6 emotion labels, consisting of 420 records for the elderly and 660 for adults. Our dataset is composed of audios clipped from TV shows, podcasts, documentaries, etc., it has 176 records and 5 emotion labels. As for this study, we are only using 5 labels for simplicity: anger, happy, sad, neutral, fear.

An important thing to note is that the sentimental and linguistic structure of our dataset and YueMotion is vastly different. The linguistic content in YueMotion is repetitive. A sentence like "你宜家幾多歲" will be repeated with different emotions and with different speakers. For our dataset, the sentence in the audio will be random and will not be repeated. The emotions of the audio will also be related to the speech being said.

5 Model Explanations

Our model consists of various parts:

- Feature Extraction Models
  - XLSR-Wav2Vec 2.0 (Speech Representations)
  - Whisper (Transcription)
  - fastText (Text Representations)
- Classification Model

The details of the models will be explained in the following subsections.

5.1 Wav2Vec 2.0

Wav2Vec 2.0 is a powerful semi-supervised model developed by Facebook AI that outputs representations from speech audio[4]. The architecture of the model can be seen in the following figure:

The model consists of a few parts: a feature encoder, a transformer, and a quantization module. A raw audio X is inputted into the feature encoder. Then, a latent speech representation is fed into both the transformer and the quantization model. The transformer tries to capture information in the speech in the form of representation vectors. The quantization module discretizes

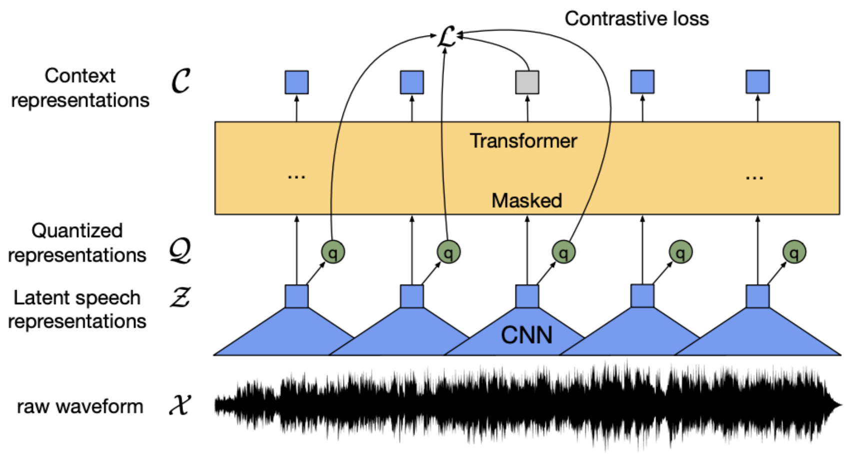

Figure 3: Architecture of Wav2Vec 2.0[4]

the latent speech representations to represent the targets in the self-supervised objective[4]. After training on relevant speech data, the model can then output contextualized representations of an audio that contains high-level features that are vital to speech emotion recognition tasks. A representation L is formed by a true quantized latent speech representation by solving contrastive task L_m and a codebook diversity loss L_d.

$$L = L_m + \alpha L_m$$

Where α is a hyperparameter that is tuned.

The contrastive loss is calculated as:

$$L_m = -\log \frac{\exp(sim(c_t, q_t)/k)}{\sum_{\tilde{q} \sim Q_t} \exp(sim(c_t, \tilde{q})/k)}$$

and the cosine similarity as $sim(a, b) = a^T b / \|a\| \|b\|$.

Where c_t is a output from the transformer centered over masked time step t, and Q_t is a set of $K + 1$ quantized representations $\tilde{q} \in Q_t$, which includes a true quantized representation and K other distraction representations. The model needs to identify q_t for a masked time step within the set Q_t

The Diversity loss is calculated as:

$$L_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

Where there are V entries in each of the G codebooks. And it wants to maximize the entropy of averaged softmax distribution over the codebook entries for each codebook $\bar{p}_g$.

In our previous study, Wav2Vec 2.0 was used, but as Wav2Vec 2.0 is trained on monolingual speech corpora (English), it is not suitable for speech emotion detection for other languages. Thus, we have chosen XLSR-Wav2Vec 2.0, a variation of Wav2Vec 2.0, as our model. It is trained on multilingual speech corpora, and its ability to learn multilingual representations will be crucial to our task.

For convenience, we have sourced a pre-trained model of XLSR-Wav2Vec 2.0 that is fine-tuned on Cantonese[12].

5.2 Whisper

Whisper is an OpenAI developed multilingual speech model that is capable of many tasks, including translation and transcription. Whisper's approach to training a multilingual speech recognition model can be an example of our cross-domain learning techniques, as it is also trained on multilingual speech corpora. The approach of Whisper can be seen in the following figure:

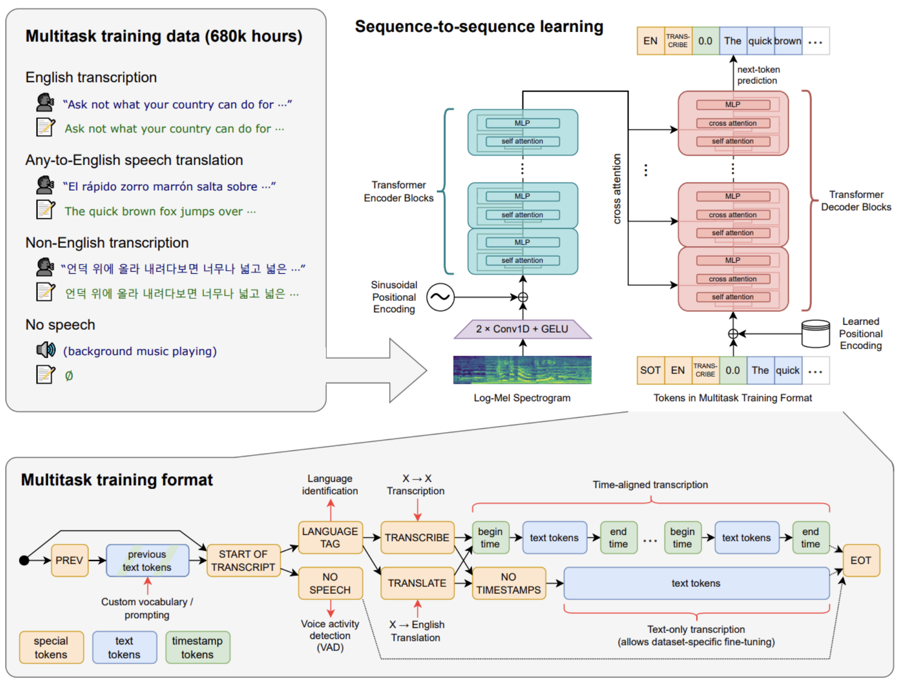

Figure 4: Overview of Whisper's approach[25]

It is a model adopting an encoder-decoder Transformer architecture, 2 convolutional layers with a filter and a GELU activation function. For the tokenizer, Whisper refits multilingual vocabulary for multilingual models to maintain accuracy for all languages.

For the training of Whisper, although it considers the history of the text being predicted, the tokens of the previous text will be masked, allowing it to resolve more ambiguous audio. Then, the model identifies if speech is present in the audio. If speech is present, transcription and translation are trained, where transcription is just time-aligned transcription in different languages.

The difference between transcription and translation in Whisper is that translation is limited to X to English instead of X to X. Transcription can be done in most languages and can be further fine-tuned for a chosen set of languages.

For convenience, we have also used a Whisper model fine-tuned on Cantonese that is available online[1].

5.3 fastText

fastText is an unsupervised model developed by FacebookAI that learns representations of words while considering the morphology of words [7].

The model employs a skipgram model and tries to maximize the objective function:

$$\sum_{t=1}^{T} \sum_{c \in C_t} \log p(w_c | w_t)$$

Where C_t is a set of indicies of words surrounding word w_t.

The model supports fine-tuning and training of multiple languages, instead of being English only. Which makes it a suitable choice for this work.

For convenience, a publicly available Cantonese fine-tuned fastText[27] was used in this work.

5.4 Classifier

The classifier model contains 3 layers: 2 linear activated layers, a ReLU layer, and a softmax layer that outputs the probability density for the emotions. During training, the model will be trained on the cross-entropy loss between the features and the labels.

The architecture of the classifier can be seen in the following figure:

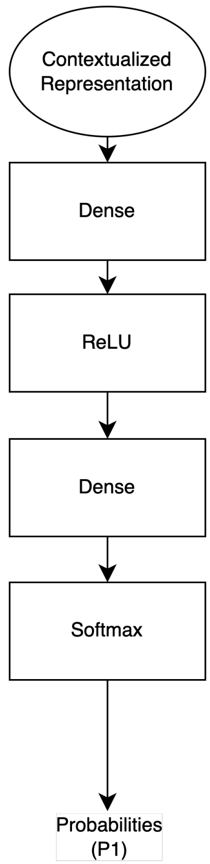

Figure 5: Architecture of classifier

6 Methodology

As mentioned in the introduction, we will adopt two different approaches to see if they can improve a model's accuracy.

The details of each low-resource training technique will be explained in the following subsections.

6.1 Cross-Domain Learning

Cross-domain learning is a significant problem in speech emotion recognition because, to build multilingual speech recognition models, we must first solve the problem of domain mismatch between different languages. In this study, we have gathered a dataset that contains speakers of different sexes, and different ages. These speakers possess different acoustic features, such as the intensity and pitch of their speech[28].

The logic of cross-domain learning is that we believe a source dataset X_s has some hidden relationship with a target dataset X_t, and this relationship can be learned through learning X_s[11]. That is, we expect that there is some intersection between the domain of X_s and X_t, and by learning X_s, we can have a good enough approximation of the feature in X_t. For example, there is a significant relationship between Mandarin and Cantonese, so we train our transcribing model on Mandarin, hoping that it learns this relationship and performs well on Cantonese speech emotion recognition as well. A figure of the relation can be seen in the following figure:

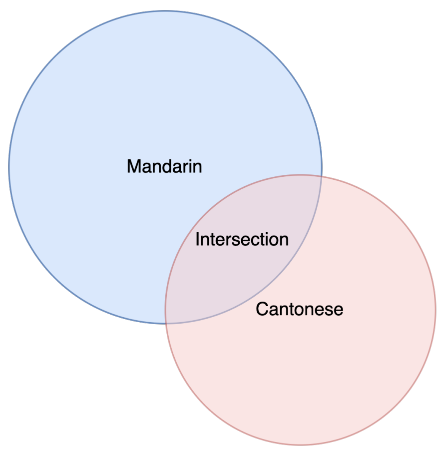

Figure 6: Relation between domains

In the context of our study, because we lack high-quality speech data from the elderly. To train a robust speech emotion recognition model that can accurately identify emotions from speech, we would need to adopt cross-domain learning. That is, we try to transfer the ability to recognize speech emotions in young adults to the elderly.

This will be done by only adding speech audio of young adults to the training data. Then, we will conduct the training as usual, that is, after extracting the representation vector, we train the classifier and try to predict the test set that consists of only elderly speakers. More details on the exact approach will be mentioned in the training approach section.

6.2 Text-Based Features

Text-based features include many categories, text, word embeddings, and others. In the context of this study, text-based features are contextualized word embeddings that are learned by pre-trained models.

Word embeddings, also known as word vectors, are represented as a series of numbers, hence the name word vectors. These vectors allow models to understand the meaning of the words in a way that humans do not. Similar words should ideally be closer to each other. Meaning:

$$\|\overrightarrow{Food} - \overrightarrow{Ramen}\|_2 < \|\overrightarrow{Food} - \overrightarrow{Door}\|_2$$

should hold[3].

Word vectors have been used in recent studies in many fields: social science[3], sentiment analysis[14], and speech emotion recognition[24]. The number of studies embracing the use of word vectors can be an indicator of how powerful word vectors are in capturing high-level features of words.

In this study, word vectors will be extracted from words by a variation of fastText that is pre-trained on Cantonese data. To ensure the accuracy and the robustness of the vectors, we will be using a 300-dimension embedding. The word vectors are supplementary features to the acoustic features (speech representation) in hopes of boosting the accuracy of the model that is lacking training resources. More details about the training method will be mentioned in the training section below.

6.3 Training Approach

In order to extract high-level features for the classifier to learn, pre-trained models that can extract contextualized features will be utilized. For feature extraction, XLSR-Wav2Vec 2.0, Whisper, and fastText fine-tuned on Cantonese will be used.

Raw audio will be passed to the model, then be transformed into a representation of the audio with a dimension of 500.

In the case that both speech representations s_i and text vectors t_i will be used, Whisper will be used to transcribe the audio into text. Then, the text will be fed to fastText for feature extraction. Then the speech representation and the text embedding will be concatenated together as a new vector r_i in the form:

$$r_i = [x_i^T, t_i^T]^T$$

The resulting vector from this operation will have 900 dimensions.

The classifier model will be trained by trying to solve the following minimization problem (Cross Entropy Loss):

$$\min \frac{\sum_{n=1}^{N} l_n}{N}$$

where

$$l_n = -\sum_{c=1}^{C} w_c \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} y_{n,c}$$

and x is the input vector, y is the target, w is the weight, C is the number of classes, and N is the number of minibatches of data. Where x is a vector with 500 or 900 dimensions and y is an integer.

The vectors pass through two linearly activated layers and a ReLU layer, then will finally be passed to a softmax layer to generate the probability density of the emotions.

3-Fold Cross-Validation will be used in training and evaluating the model, the use of cross-validation is to ensure that the results are consistent and reliable. We can test the consistency of the results by calculating its variance and check whether the variance is too high for the model to be consistent. Stratified splitting will be used in splitting the data. This is to preserve the distribution of data within each sub-dataset and ensure that for each fold of training, the results are reliable and consistent.

The training hyperparameters are:

- Learning Rate: 5e-5
- Number of Epochs: 10
- Weight Decay: 0.01

The hyperparameters are carefully tested and selected to prevent overfitting of the classifier. It is also important to note that the feature extraction models will not be trained or modified in the course of the training.

7 Results and Discussions

7.1 Training Results

We have in total 6 variations of the dataset. The training results of each dataset can be seen in the following table:

YueMotion	Cross-Age?	Text-Based Features?	Average Accuracy	Average F1-Score	Accuracy Variance
Not Used	No	No	32.38%	0.2973	0.0002
Not Used	No	Yes	35.79%	0.3356	0.0007
Used	No	No	37.53%	0.3750	0.0009
Used	No	Yes	34.90%	0.3404	0.0000
Used	Yes	No	49.39%	0.4933	0.0015
Used	Yes	Yes	51.03%	0.5129	0.0011

Table 1: Training Result of Classification Model

The statistics of each dataset is an average of a 3-fold cross-validation to ensure consistency and unbiasedness of the results. And no in cross-age means only elderly data is used, yes means both elderly and adult data are used.

From the result table, we can see that both cross-domain learning and text-based features improved the accuracy of the model. Proving the feasibility of cross-domain learning in elderly speech emotion recognition. This is significant because in Cantonese elderly speech emotion recognition, there is very little labeled data publicly available[9]. But if we can use data from closely related domains, we can use datasets of adults or even datasets in Mandarin. These extended data would be excellent resources to train elderly speech emotion recognition models on.

Also, text-based features are increasing the accuracy of the model for 2 of the 3 datasets. This can be an indication of how powerful high-level representations can be, they can operate many tasks, such as transcribing, translating, and even speech emotion recognition. In the future, this study may continue to optimize the use of high-level features in elderly speech emotion recognition to train more robust models in a low-resource setting.

7.2 Discussions

There are a few interesting things that were observed in the process of training the model:

- Relationship between accuracy and dataset size
- For 2 datasets, text-based features caused a drop in accuracy
- The Variance of accuracy between different folds of training lowers when text-based features and cross-domain learning are both used

These are all interesting phenomena that occurred in the training process of the classifier, and in the following sub-sections, we will try to explain why these phenomena occurred.

7.3 Relationship of Accuracy and Dataset Size

One interesting thing that happened during training is that we thought the accuracy of the model would grow somewhat linearly. After doing some exponential regression and plotting some graphs, we discovered the following:

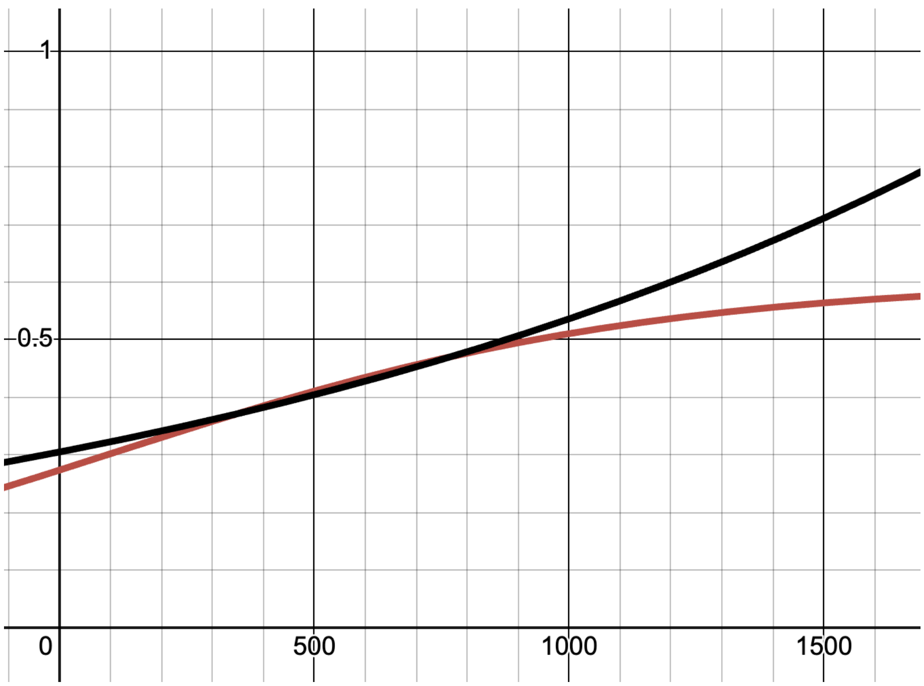

Figure 7: Logistic Regression of Two Groups

The black line in the graph is the line of the logistic regression for the accuracy of $O^{d_{elderly}}, X^{d_{elderly}}, X^{d_{elderly}, d_{adult}}$ (Datasets without text-based features), and the red line is the logistic regression line for the accuracy of $O_{text}^{d_{elderly}}, X_{text}^{d_{elderly}}, X_{text}^{d_{elderly}, d_{adult}}$ (Datasets with text-based features).

It can be seen that there is some form of linearity in the growth of accuracy in the first graph, while in the second graph, the relationship looks more non-linear. While the sample size of this observation is very small, we still think that this is an interesting phenomenon because this may be an important step to improving prediction accuracy in low-resource domains. There can be many factors contributing to the non-linearity of the accuracy growth.

One thing that comes to mind is the power of word embeddings. Word embeddings are powerful in the sense that they capture a lot of low-layer or hidden features that humans can't comprehend. Even with some transcription errors (0.0972 CER), it can still harness a great portion of the sentimental information available in the text. When more similar word embeddings are fed to the classifier, it starts to generalize and apply the high-level features to future predictions, thus resulting in the non-linear growth. This might be a reminder to use related deep-learning-based features, as their correlation might be greater than we have imagined.

7.4 Accuracy Drop for Text-Based Features

In the dataset where YueMotion is used and no cross-age learning is adopted, the accuracy of the model drops after using text-based features in the training process of the model. This can be caused by a few reasons: 300 dimension word embeddings may be too hard to learn with small datasets, and the linguistic content of the speech might not match the sentiment of the speech.

For example, speakers in the YueMotion dataset will say "你宜家幾多歲" in fear, which, in general, will not be the case. The model may be distracted by these misleading audio in the dataset and learn unrelated features. In our database, the linguistic content of the speech often highly correlates with the sentiment of the audio, and that is maybe why, in our own dataset, the text-based feature performs better. Some examples of the lexical content and sentiment can be seen in the following table.

Linguistic Content	Translation	Linguistic Sentiment
你而家幾多歲	How old are you?	Neutral
今年係咩年份呀	What year is it?	Neutral
而家香港特首係叫做乜名	What is the name of the current Chief Executive of Hong Kong?	Neutral

Table 2: Linguistic Content and Sentiment of Data in YueMotion

We can see that all of these are neutral sentences that do not carry any sentimental information on it's own, but in the YueMotion dataset, speakers will add their own interpretation of the chosen emotion. Although distractors are used in semi-supervised learning[4], it is bad for supervised learning as it overloads the model with irrelevant data[15, 29]. To achieve a better performance for the classification models, we would have to mitigate the effect of these distractors, which we will discuss in the future works section.

Also, word embeddings with higher dimensions are harder to learn. In low-resource settings, high-dimension word embeddings can encounter the problem of data sparsity, meaning they are not correctly representing a word[23] due to lack of data to accurately describe in higher dimensions. Cantonese, being a low-resource language, just as we have mentioned many times in this study, is prone to this problem of data sparsity. The data used to train the embedding model might not be dense enough to correctly transform words into high-dimensional vectors. Hence, the accuracy drops when text-based features are used.

Our classifier might also not be powerful enough to learn the features fully.

For the concatenated features, there are 900 dimensions, but there are only 362 records in this dataset. Which is less than the number of dimensions in the feature vector, and in non-cross-domain learning datasets, the number of dimensions is even larger than the number of data. This scarcity in data may cause a low accuracy in the classification task because the model could not fully learn the high dimension features. To solve this problem, lower dimension features can be used to allow a better interpretation of the features by the model, which we will further discuss in the future works section.

7.5 Variance in Accuracy

By looking at Table 1, we can see that the variance of accuracy for almost all text-based features included datasets are smaller than those that did not include text-based features. This may indicate that by supplying more features, the model becomes more stable.

One of the causes of this consistency can be the large number of patterns and relations given to the model. Intuitively, a model would require more data to become robust and consistent. This is logical because all that a model does is make educated guesses based on the information that is fed to it, more high-quality data would surely make it more powerful and robust. But the amount of information does not only rely on the size of a dataset, a dataset with high-level features can also train powerful models. A study about human cognitive theories of attention found that both high-level and low-level data are important in machine learning[18]. Pointing out the point that not only low-level data are needed but also high-quality high-level data.

This observation can be an approach to building consistent models in the future, as consistency is also a very important metric in speech emotion recognition[13].

8 Future Work

A lot of work to be done has been mentioned in the discussions section, and in this section, more details on the exact approaches for these works will be provided.

8.1 Reducing the Dimension of Features

The intuition with dimensions in features might be that higher dimensions lead to better results in classification tasks because they capture more information. While this may be true in some cases, in the context of low-resource learning and in the context of this study, features with high dimensions may lead to a drop in accuracy. Simply because the model could not fully learn the information captured inside the feature.

Reducing the dimension of the vectors can be done in a few ways: vector quantization or shortening the length of audio. These methods will be briefly explained in the following subsections.

8.1.1 Vector Quantization

Vector quantization is a well-known technique in machine learning. It is used in many areas of machine learning, such as dimension compression, vector search optimization, and many others.

The main result of vector quantization is the compression of dimensions and speed increase in vector comparisons. By compressing the feature that we currently have into a compact and simple representation, we can maintain the essential patterns and relationships that the feature vector currently holds[30]. At the same time, it lowers the number of dimensions of the vector, allowing the model to fully learn the representation with less amount of data.

The reduction in dimensions can also speed up training. Allowing more flexible development and evaluation of the model. Currently, the feature has at least 500 dimensions and at most 900 dimensions. Multiplications are inevitable in training the classification model as it needs to pass through two linear layers

$$\theta X + c = Y$$

The dot product of vectors has a time complexity of $O(n^2)$, and these calculations take up a lot of computational resources. By reducing the dimensions, we would significantly lower the time needed to train our models and be more flexible in testing other architectures and approaches.

8.1.2 Reducing Length of Audio

Reducing the length of the audio while preserving the essential information can be the future approach for this work. The problem with our current approach is that the features are too long for the model to fully understand with such limited data size. Aside from reducing the dimensionality of the feature representations, we can also choose to reduce the size of the audio to generate shorter vectors.

Reducing the length of the audio can bring many benefits under low-resource settings, but it can also bring many negative impacts to the performance of the model. Reducing audio length means that some relations or patterns will also be left out, lowering the chance for the model to fully learn the patterns in elderly speech. Also, the speed of elderly speech is often slow, it will be a challenge to reduce the audio length for elderly speech because of this. One approach that we can try is to use only fragments of their speech, train the model to recreate their speech based on some constrastive loss. This approach would somewhat resemble Wav2Vec 2.0[4] but might just work for low-resource settings as it does not require labeled information to work.

9 References

References

[1] alvanlii. whisper-small-cantonese. https://huggingface.co/alvanlii/whisper-small-cantonese. Accessed: 2025-04-07.

[2] Ankita and S. Shahnawazuddin. Exploring the role of data augmentation and acoustic feature concatenation in the context of zero-resource children's asr. In *Circuits Syst Signal Process*, pages 1914–1937, 2025.

[3] Alina Arseniev-Koehler. Theoretical foundations and limits of word embeddings: what types of meaning can they capture?, 2021.

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[5] Imen Baklouti, Olfa Ben Ahmed, and Christine Fernandez-Maloigne. Cross-lingual low-resources speech emotion recognition with domain adaptive transfer learning. In *Proceedings of the 13th International Conference on Data Science, Technology and Applications - Volume 1: DATA*, pages 118–128. INSTICC, SciTePress, 2024.

[6] Russell Beale and Christian Peter. *The Role of Affect and Emotion in HCI*, pages 1–11. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2017.

[8] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, Nov 2008.

[9] Samuel Cahyawijaya, Holy Lovenia, Willy Chung, Rita Frieske, Zihan Liu, and Pascale Fung. Cross-lingual cross-age group adaptation for low-resource elderly speech emotion recognition, 2023.

[10] Census and Statistics Department of Hong Kong. Hong kong population projections 2020-2069. https://www.statistics.gov.hk/pub/B1120015082020XXXXB0100.pdf.

[11] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning, 2023.

[12] ctl. wav2vec2-large-xlsr-cantonese. https://huggingface.co/ctl/wav2vec2-large-xlsr-cantonese. Accessed: 2025-04-07.

[13] Anna Derington, Hagen Wierstorf, Ali Özkil, Florian Eyben, Felix Burkhardt, and Björn W. Schuller. Testing correctness, fairness, and robustness of speech emotion recognition models, 2025.

[14] Xian Fan, Xiaoge Li, Feihong Du, Xin Li, and Mian Wei. Apply word vectors for sentiment analysis of app reviews. In *2016 3rd International Conference on Systems and Informatics (ICSAI)*, pages 1062–1066, 2016.

[15] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. Finding distractors in images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1703–1712, 2015.

[16] Fasih Haider, Senja Pollak, Pierre Albert, and Saturnino Luz. Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. *Computer Speech Language*, 65:101119, 2021.

[17] Abid Haleem, Mohd Javaid, Ravi Pratap Singh, and Rajiv Suman. Telemedicine for healthcare: Capabilities, features, barriers, and applications. *Sensors International*, 2:100117, 2021.

[18] Taylor R. Hayes and John M. Henderson. Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific Reports*, 11(1):18434, Sep 2021.

[19] Li Huan-Chung, Pan Telung, Lee Man-Hua, and Chiu Hung-Wen. Make patient consultation warmer: A clinical application for speech emotion recognition. In *Applied Sciences*, 2021.

[20] Tao Jian-Hua, Huang Jian, Li Ya, Lian Zheng, and Niu Ming-Yue. Semi-supervised ladder networks for speech emotion recognition. In *International Journal of Automation and Computing*, pages 437–448, 2019.

[21] Liu Ke, Wei Jiwei, Zou Jie, Wang Peng, Yang Yang, and Heng Tao Shen. Improving pre-trained model-based speech emotion recognition from a low-level speech feature perspective. In *IEEE*, pages 10623–10636, 2024.

[22] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M Girard, and Louis-Philippe Morency. Elderreact: A multimodal dataset for recognizing emotional response in aging adults. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 349–357, New York, NY, USA, 2019. Association for Computing Machinery.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[24] Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustín Gravano. Fusion approaches for emotion recognition from speech using acoustic and text-based features. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6484–6488, 2020.

[25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[26] Tesfa Shegaw, Wordofa-Giru Berhanu, Bedada Tadesse, and Gela Debela. Mental distress and associated factors among hospitalized medical-surgical adult inpatients in public hospitals, addis ababa, ethiopia, 2020: Cross-sectional study. In *Psychology Research Behavior Management*, 2021.

[27] toastynews. hong-kong-fasttext. https://github.com/toastynews/hong-kong-fastText. Accessed: 2025-04-07.

[28] Peter Torre and Jessica A. Barlow. Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders*, 42(5):324–333, 2009.

[29] Shenghua Wan, Yucen Wang, Minghao Shao, Ruying Chen, and De-Chuan Zhan. Semail: Eliminating distractors in visual imitation via separated models, 2023.

[30] Zhou Yang, Weisheng Dong, Xin Li, Mengluan Huang, Yulin Sun, and Guangming Shi. Vector quantization with self-attention for quality-independent representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24438–24448, 2023.

[31] Longbiao Wang Jiaxing Liu Jianwu Dang Yuan Gao, Shogo Okada. Domain-invariant feature learning for cross corpus speech emotion recognition. In *ICASSP*, pages 6427–6431, 2022.