

Thesis for the Degree of Doctor of Philosophy

**Robust Speaker Adaptation Framework for
Personalized Emotion Recognition in
Emotionally-Imbalanced Small-Sample Environments**

Jaehun Bang

Department of Computer Science and Engineering
Graduate School
Kyung Hee University
Seoul, Korea

August, 2019

**Robust Speaker Adaptation Framework for
Personalized Emotion Recognition in
Emotionally-Imbalanced Small-Sample Environments**

Jaehun Bang

Department of Computer Science and Engineering
Graduate School
Kyung Hee University
Seoul, Korea

August, 2019

Robust Speaker Adaptation Framework for Personalized Emotion Recognition in Emotionally-Imbalanced Small-Sample Environments

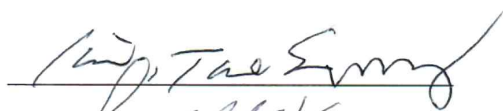
by
Jaehun Bang

Advised by
Professor. Sungyoung Lee

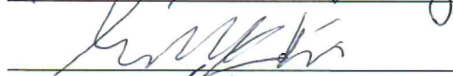
Submitted to the Department of Computer Science and Engineering
and the Faculty of the Graduate School of
Kyung Hee University in partial fulfillment of the requirements
for degree of Doctor of Philosophy

Dissertation Committee:

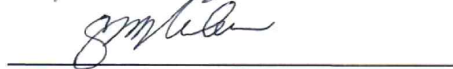
Prof. Tae-Seong Kim



Prof. LokWon Kim



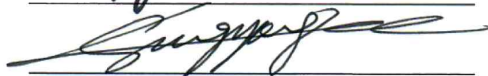
Prof. Sung-Ho Bae



Prof. Jee-In Kim



Prof. Sungyoung Lee



Traditional speech emotion recognition researches collected data from multiple users with various professions, and created a general training model to be used in the recognition process. These traditional speech emotion recognition methods have a critical problem that they show large differences in recognition accuracy from each user. In order to solve the deviation of accuracy differences, a personalized emotion recognition research is actively carried out which provides a customized model for each target user. Existing personalized speech emotion recognition research has a cold start problem that requires sufficient and emotionally balanced personalized data from a target user when creating a customized model. In an environment where a cold start problem occurs, there are 3 kinds of data environments. The first one is having small data in which data is not collected enough to create a personalized model. The second one is having absent data in which a particular emotion is not collected. Finally, the third one is having imbalanced data where the number of emotion data are collected by large difference between minor class and major class. In order to solve these data environment problems, various existing adaptation modeling methods are proposed which incrementally changes the model reflecting the emotional speech of the target user based on the initial model from multiple users. These existing methods are difficult to create a stable personalized training model at the initial stage of emotion recognition because it does not cope well in the environment with having absent data and imbalanced data.

Therefore, a Robust Speaker Adapting Framework solving the cold start problem in three different data environments is proposed in this thesis. The proposed techniques

are consisted of 3 core solutions. The first solution is a similarity data selection technique based on maximum threshold distance, which selects more real case data from existing data sets compared with existing methodologies. The second solution is to find other user who has similar emotional speech with target user based on data distribution factors, and replaces the target user's absent emotion data with this. The third solution is a SMOTE (Synthetic Minority Over-Sampling Technique) based virtual data generation method that generates virtual data by repeatedly using the Oversampling algorithm SMOTE to improve the imbalanced data environment.

The proposed Robust Speaker Adaptation Framework provides a personalized training model for the target user utilizing 3 core solutions by selecting the actual case data useful for the target user from collected target user emotional speech with initial training model and incrementally augmenting virtual data based on the SMOTE. A comparative evaluation is conducted using public databases which are well-known data set such as IEMOCAP (Interactive Emotional Dyadic Motion Capture) and CREMA-D (CRowd-sourced Emotional Multimodal Actors Dataset). In the experimental result, the proposed method has been proved to be able to provide a stable personalization model which is faster than existing techniques in the limited data environment from the whole section by providing a sufficient and balanced personalized training model even in the initial stage.

Keywords: Speaker Adaptation; Machine Learning; Model Adaptation; Data Selection; Data Augmentation; Speech Signal Analysis

Abstract	I
List of Tables	IV
List of Figures	VI
Chapter 1. Introduction	1
1.1. Overview	1
1.2. Motivation	5
1.3. Problem Statement	8
1.4. Proposed Concept	12
1.5. Key Contributions	15
1.6. Thesis Organization	17
Chapter 2. Related Works	18
2.1. Traditional Speech Emotion Recognition	18
2.2. Personalized Emotion Recognition	21
2.2.1 Feature Adaptation	21
2.2.2 Incremental Learning	24
2.2.3 Model Adaptation	26
2.2.4 Deep Domain Adaptation Network	30

Chapter 3. Proposed Robust Speaker Adaptation Methodologies	33
3.1 Robust Speaker Adaptation Framework	33
3.2. Similar data selection based on Maximum Threshold Distance	45
3.3. Other similar user emotional speech mapping based on Data Distribution Factor	53
3.4. Virtual Case Data Augmentation based on SMOTE	59
3.5. Model Creation and Classification	62
Chapter 4. Experiment Result and Discussion	63
4.1. Experimental Setup	63
4.1.1 Dataset	64
4.1.2 Experimental Methodologies	71
4.2. Experimental Results	74
4.2.1 Recognition Accuracy	74
4.2.2 Imbalanced Ratio	86
Chapter 5. Conclusion and Future Direction	88
5.1. Conclusion	88
5.2. Future Direction	89
References	93
Appendix: List of Publication	103
A-1. Journal Papers	103
A-2. Conference Papers	106
A-3. Patents Registration	109
Korean Abstract (국문 초록)	110

Table 3-1. Feature Vector Scheme Description	38
Table 3-2. Correlation Feature Selection Matrix	40
Table 3-3. The Comparison of accuracy between original emotion recognition and applied correlation feature selection - IEMOCAP (Unit %)	42
Table 3-4. Distance measurement performance	48
Table 4-1 Organization of Existing Emotional Speech Database	65
Table 4-2 Original IEMOCAP Dataset Structure	66
Table 4-3 Refined IEMOCAP Dataset Organization	67
Table 4-4. CREMA-D Actors' Age Distribution	68
Table 4-5. CREMA-D semantic contents	69
Table 4-6. Organization of CREMA-D	70
Table 4-7. The confusion matrix of comparison of accuracy (Unit %) (Evaluation Data: IEMOCAP, Initial Model: CREMA-D)	81
Table 4-8. The confusion matrix of comparison of accuracy (Unit %) (Evaluation Data: CREMA-D, Initial Model: IEMOCAP)	83

Figure 1-1 Research Taxonomy	4
Figure 1-2 Personalized Emotion Recognition System Architecture	5
Figure 1-3 Cold start problem environments	6
Figure 1-4 Comparison of existing methods	6
Figure 1-5 Proposed Robust Model Adaptation System Architecture	7
Figure 1-6 Proposed Idea for Small Data Environment	9
Figure 1-7 Proposed Idea for Absent Data Environment	10
Figure 1-8 Proposed Idea for Imbalanced Data Environment	11
Figure 1-9 Abstract Exist method's limitation and proposed solution.	11
Figure 1-10 The concept of the Robust Speaker Adaptation Framework	11
Figure 2-1 MFCC feature based Emotion Recognition with GMM	18
Figure 2-2 SVM and HMM-GMM Hybrid System	19
Figure 2-3 Optimum Hierarchical Decision Tree using best classifier at each decision level	20
Figure 2-4 Traditional speech Emotion Recognition System	20
Figure 2-5 Overview of The Iterative Feature Normalization (IFN) approach ..	22
Figure 2-6 Feature Analysis and Singular Value Decomposition for Speaker State Recognition	22
Figure 2-7 Illustration of baseline and proposed cascaded normalization strategies	23
Figure 2-8 Incremental SVM Architecture	24
Figure 2-9 Example of the data selection by MLLR Adaptation	26
Figure 2-10 Procedure for the conventional MLLR adaptation	27

Figure 2-11 Multistage data selection based on LDM-MDT MLLR Algorithm	28
Figure 2-12 The example of data selection based on LDM-MDT MLLR	
Algorithm	29
Figure 2-13. Outline of the SSPP-DAN	30
Figure 2-14. Architecture of the domain adversarial neural network (DANN) ..	31
Figure 3-1 Proposed Robust Speaker Adaptation Framework	34
Figure 3-2 Waveform of Original Raw Speech Signal	36
Figure 3-3 Waveform of Converted STE Signal	36
Figure 3-4 Waves of (a) before and (b) after preprocessing module in a sentence	37
Figure 3-5 Feature Extraction Procedure	39
Figure 3-6. Similar data selection based on Maximum Threshold Distance	45
Figure 3-7 Unlabeled transformation	46
Figure 3-8 Unlabeled transformation in feature space	47
Figure 3-9 MLE value calculation based on target user data	47
Figure 3-10 MTD based Similar Data Selection Process	50
Figure 3-11 Similar Data Selection Algorithm	51
Figure 3-12 The Proposed Algorithm of Similar Speech Data Selection based on Maximum Threshold Distance	51
Figure 3-13 Output comparison of the proposed method and existing method	52
Figure 3-14 Absent Emotion Data Reinforcement Workflow	53
Figure 3-15 Other similar user emotional speech mapping based on Data Distribution Factor	57
Figure 3-16 Output comparison of the proposed method and existing method ·	58
Figure 3-17 The Flowchart of SMOTE Algorithm	60
Figure 3-18 The Proposed Algorithm of Virtual Case Data Augmentation based on SMOTE	60
Figure 3-19 Final output comparison of the proposed method and existing	

method	61
Figure 3-20 Simplified Random Forest	62
Figure 4-1. Valence-arousal Dimensional Model	67
Figure 4-2 Refined IEMOCAP Dataset Represented by Each User	68
Figure 4-3 The concept of the experimental methodologies	72
Figure 4-4 Experimental Results for Each Classifier (Unit %) (Initial Model: CREMA-D, Evaluation Dataset: IEMOCAP)	75
Figure 4-5 Experimental Results for Each Classifier (Unit %) (Initial Model: IEMOCAP, Evaluation Dataset: CREMA-D)	76
Figure 4-6 Experimental Results for Each Classifier (Unit %) (Initial Model: CREMA-D, Evaluation Dataset: IEMOCAP)	77
Figure 4-7 Detailed experimental results (Initial Model: CREMA-D, Evaluation Dataset: IEMOCAP)	79
Figure 4-8 Detailed Imbalanced Ratio experimental results	87
Figure 5-1 Concept of the proposed personalized modeling using DNN based on target user speech synthesis	90
Figure 5-2. Application of proposed framework in face recognition	91
Figure 5-3. Concept of the proposed multi-modal fusion	92

1.1 Overview

Emotion recognition is a technology that recognizes and processes human emotions using video, voice, text, and biological signals. Emotion recognition technology makes human life comfortable by quantitatively and qualitatively measuring human emotion and applying it to product and environment design through scientific analysis and evaluation.

Speech emotion recognition automatically recognizes a user's emotions by analyzing the user's voice signal. Various technologies such as audio preprocessing, feature extraction, model creation, feature/ decision level fusion, and adaptation has been researched in the field of speech emotion recognition recently [1-2].

Traditional speech emotion recognition studies aim at improving the feature extraction and classification methodologies in order to improve the accuracy of various amounts of recorded emotional speech from multiple users. Such feature extraction studies consist of filter-bank algorithm improvements and statistical feature discoveries [3-4]. On the other hand, the classification studies include a hierarchical classification methodology [5], a mixture of two classifiers [6], and the creation of training models of males and females [7]. These previous studies achieved high accuracy based on Speaker-Dependent (SD) model experiments, where the users participated in the training process. However, the accuracy is significantly lowered when the target user's speech does not participate in the training [8].

Therefore, the speech emotion recognition studies have been conducted to create training model that achieves high accuracy in Speaker Independent (SI) experiments. SI models studies have also been researched to create high accurate predictive model for every user. The accuracy of an SI model was 2 – 3 times lower than that of existing SD models [9]. Nowadays, the gap of accuracy difference has been reduced with SD model by introducing many machine learning techniques and strategies such as the Deep Learning [10-13], Extreme Learning Machines [14-17], Classification fusions [18-19], AdaboostMH [20] and Ensemble Classifier [21]. These method has the advantage of providing recognition service to users with reasonable models immediately in initial stage. However, it still does not guarantee high accuracy level for every user. And it also requires sufficient training dataset in training phases. Additionally, it is difficult to improve the accuracy level due to static training model.

As such, the speech emotion recognition research in recent year has focused on creating personalized emotion recognition to be able to process the recognition immediately such as SI model and high accuracy level of SD model. In personalized emotion recognition, it is very important to establish an efficient machine learning strategy. There are 3 typical machine learning strategies including convolutional learning, self-learning and adaptive Learning. The convolutional Learning strategy is to improve the accuracy level by extracting various and sophisticated features set from the obtained data. It requires a lot of data in feature vector extraction process. The self-learning strategy is a way to improve the accuracy by updating the existing training model continuously through many iterations without user intervention on the new inputs. It does not only take a long time to change to customized model but it also requires a lot of data. The convolutional and self-learning strategies face the underfitting problem in small data set. However, the adaptive Learning strategy can avoid the underfitting problem through combining existing dataset with target user dataset from feedback. And also, the adaptive learning strategy is an efficient methodology to create personalized emotion recognition models, and is able to show a

high accuracy with only a small amount of data [22]. Therefore, most of the personalized emotion recognition has researched on Speaker Adaptation (SA) Model using adaptive learning strategies when considering the amount of limited data and the duration of the training process.

SA models are dynamic training models for target user created by combining the target user speech and speech from multiple users. SA model researches consists of feature normalization, supervised adaptation and unsupervised adaptation. Feature normalization studies [23-24] have created personalized models through iterative feature value normalization processes. In particular, these models can create individual models for target users by controlling the overall ranges of the feature values of the training dataset. However, in small-sample environments, these studies have not achieved high accuracy, as it is difficult to estimate target user speech characteristics. Supervised adaptation studies [25-28] consist of individual model creation utilizing only the target user speech and incremental learning [29-31], which adds target user speech to existing multiple-user training models. However, these methods require large amounts of data to create personalized models that are dependent on target user speech. Unsupervised adaptation [32-33] has an advantage in easily constructing SA models via cluster models of the target user speech without any emotional speech annotation processes. However, this leads to low accuracy when using small amounts of samples, making it difficult to predict the probability distribution of clustering.

In other words, the experimental results of existing SA studies have considered numerous target user samples and balanced data for each emotion. In real environments, the acquired target user speech in the initial stage cannot guarantee a large number of samples with balanced emotion due to imbalanced emotion expression as seen in daily life. Regarding the small amount of imbalanced data at the initial stage, the experimental results indicate that no reinforcement methods have been conducted due to the lack of emotional speech cases. This is known as a cold-start problems, which can be overcome by constructing personalized training datasets using real data selection and

virtual data augmentation.

Therefore, this thesis proposes Robust Speaker Adaptation Framework to deal with the cold-start problems in small, absent and imbalanced datasets during the initial stage and implement the Robust Personalized Speech Emotion Recognition Framework. The proposed Robust Speaker Adaptation reinforces the training dataset with a similar real training data when there is an insufficient amount or absence of emotion data. This process is conducted by constructing a similarity of speech feature vector by comparing the acquired target user speech with the initial multiple-user database. Further, the system also augment virtual data using the SMOTE (Synthetic Minority Over-sampling Technique) algorithm to create a robust model considering the new data. The proposed Personalized Speech Emotion Recognition Framework incrementally provides personalized models for target users through a retraining process via a machine learning algorithm based on the boosted personalized data from Robust Speaker Adaptation. Figure 1-1 shows the taxonomy of this thesis.

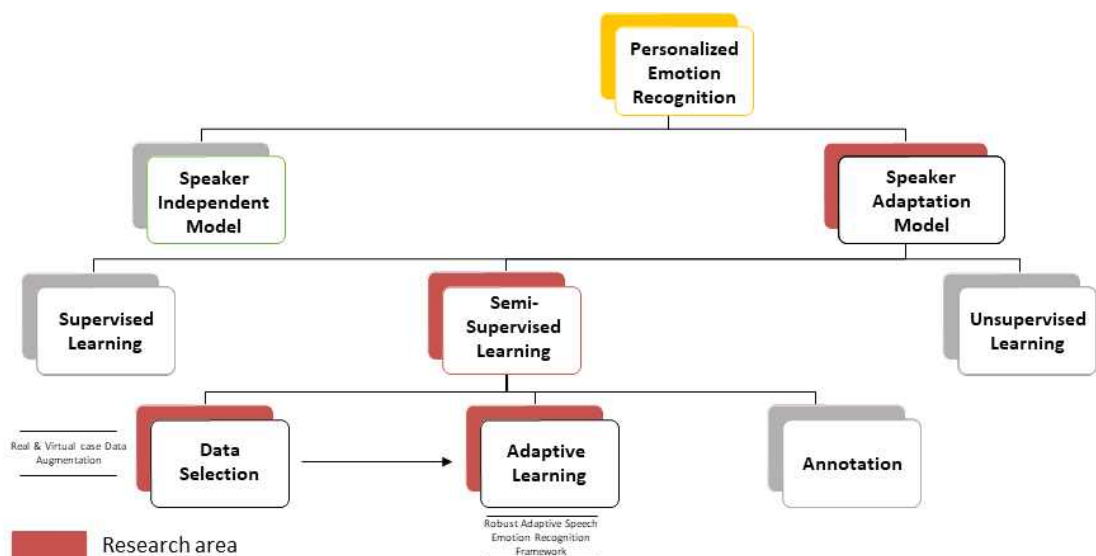


Figure 1-1. Research Taxonomy

1.2 Motivation

The Personalized Emotion Recognition System is a technique for incrementally creating a customized model by collecting the data of the labeled target users from the initial models built from various users. Figure 1-2 is shown the Personalized Emotion Recognition System Architecture.

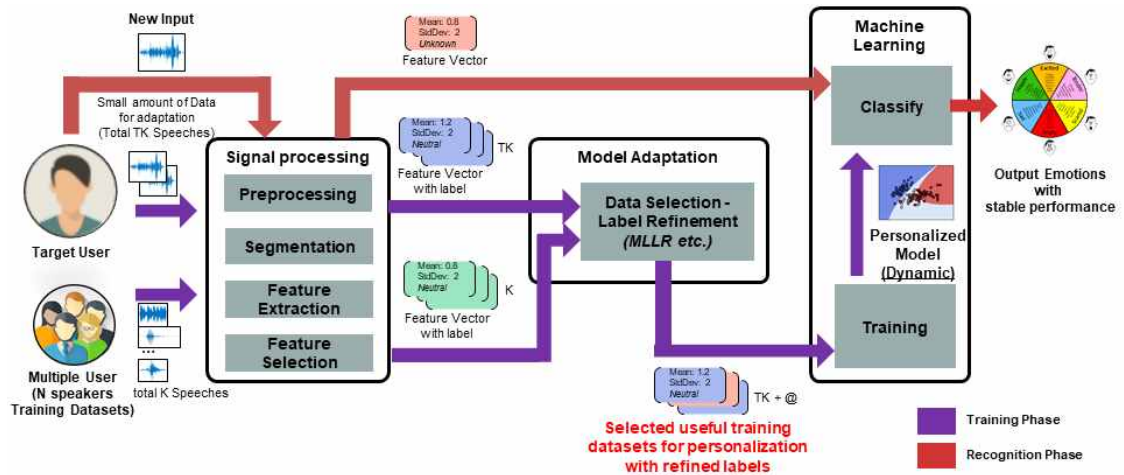


Figure 1-2. Personalized Emotion Recognition System Architecture

Personalized emotion recognition is a technique to maintain a high recognition accuracy by generating a model suitable for individuals to collect and it is very important to acquire data of target users. In real environments, the acquired target user speech in the initial stage cannot guarantee a sufficient number of samples with balanced emotion due to imbalanced emotion expression as seen in daily life. In other words, there is a cold start problem that it is difficult to create highly accurate personalized models in such a emotionally imbalanced small data environments.

The cold-start problem is occurred 3 kinds data environment in data acquisition process such as small data, absent data, imbalanced data. The small data environment is

that Target user emotional speeches are collected in limited numbers in the initial stage of personalized emotional speech acquisition. The absent data environment is that impossible to reflect personalized model about absent emotion. The imbalanced data environment is that the difference is ver high between the major class and minor class.

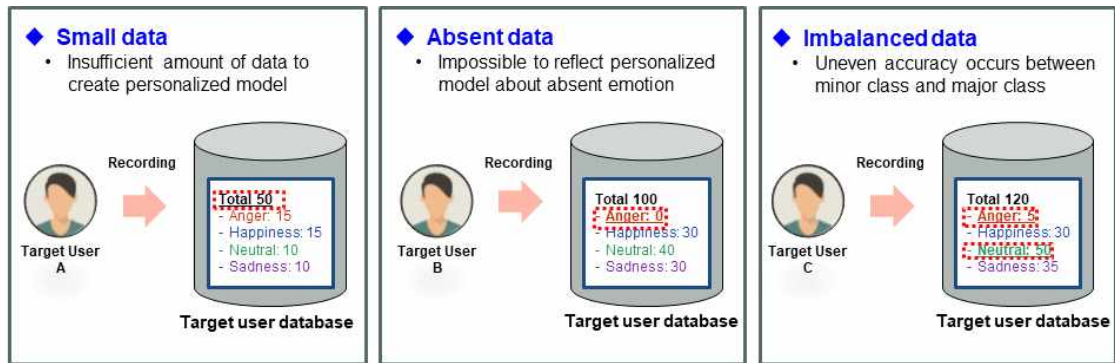


Figure 1-3. Cold start problem environments

Various studies have been conducted to solve these three cold start environments, but most of the studies do not correspond to absent data and imbalanced data environments. Even small data environments require as many as 200 to 700 target user data. Therefore, existing methods are difficult to solve these three cold start problems. Figure 1-4 is shown the comparison of existing methods.

Categories	Methodologies	Small Data Environment	Absent Data Environment	Imbalanced Data Environment	Emotions
Small & Absent Data	conventional MLLR	X (about 700 data required)	△ (Utilize Initial Model)	X	Neutral, Anger, Happiness, Sadness
	MLLR-SLR	X (about 700 data required)	△ (Utilize Initial Model)	X	Neutral, Anger, Happiness, Sadness
	LDM-MDT MLLR	△ (about 360 data required)	△ (Utilize Initial Model)	X	Neutral, Anger, Happiness, Sadness
	Incremental Adaptation	△ (300 data required)	X	X	Neutral, Anger, Happiness, Sadness
	Domain Adaptation	△ (Over 200 data required)	X	X	Arousal, Valance
Small & Imbalanced Data	Iterative Feature Normalization	△ (Over 400 data required)	X	△	Neutral, Emotional
Imbalanced Data	SMOTE	X (Over 500 data required)	X	O	Negative, Positive
Small & Absent & Imbalanced Data	Proposed method	O (Real case data selection & virtual case data augmentation)	O (Replacing similar user emotional speech)	O (Virtual case data augmentation)	Neutral, Anger, Happiness, Sadness

Figure 1-4. Comparison of existing methods

To overcome these existing limitations, this thesis proposes a Robust Speaker Adaptation Framework that augment real-case and virtual-case data. This hybrid data augmentation approach is can provide balanced sufficient data samples to create stable personalized emotion recognition model. This thesis gaol is to research the process and methodologies to create personalized emotion model to solve the cold-start problems. And the main challenges and proposed solutions are as follows.

- Challenge 1 - Increasing target user oriented training data set for small data
→ Solution 1 - Similar Data Selection based on Maximum Threshold Distance.
- Challenge 2 - Reinforcing absent data to target user relevant data
→ Solution 2 – Similar Other Similar User Emotional Speech Mapping.
- Challenge 3 - Solving imbalanced data problem from selected real-case dataset
→ Solution 3 - Virtual Case Data Augmentation based on SMOTE.

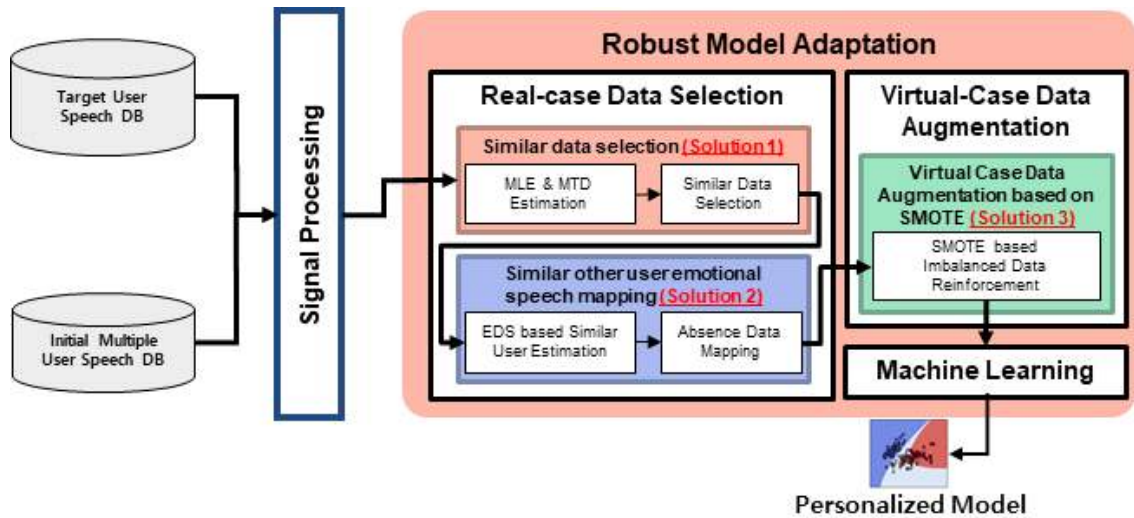


Figure 1-5. Proposed Robust Model Adaptation System Architecture

1.3 Problem Statement

There are 3 kinds of cold start problem to create personalized emotion recognition model is very difficult in limited data environments such as having ① small data, ②absent data and ③ imbalanced data. Existing methods also have a limitation for these cold start problems.

In the small data environment, the existing method calculate the Maximum Likelihood Estimation (MLE) value integrating target user speech data and existing initial data. This is a method to change the parameters of the model by reflecting the target user data from the existing data so that the accuracy of the initial model can be maintained in the initial stage when there is not enough data. However, this approach has the disadvantage that it does not accurately reflect the characteristics of the target user speech in a small data environment. In other words, if the target user data is not sufficiently acquired, it take a long time to change to a personalization model. Therefore, this method requires sufficient target user data to create personalized model with stable recognition accuracy. And it also has the disadvantage of selected data is small amount for personalization by setting a low range of threshold for the data selection. That means, this approaches is not suitable to solve the small amount data environment to create personalized model. Therefore, the solution for small data environment propose the approach that selecting a lots of similar data by setting a larger range of threshold values based on calculating MLE values with only the target user's emotional speeches. This approach can effectively create a personalized emotion model in a small data environment by selecting more data that is more similar to the target user.

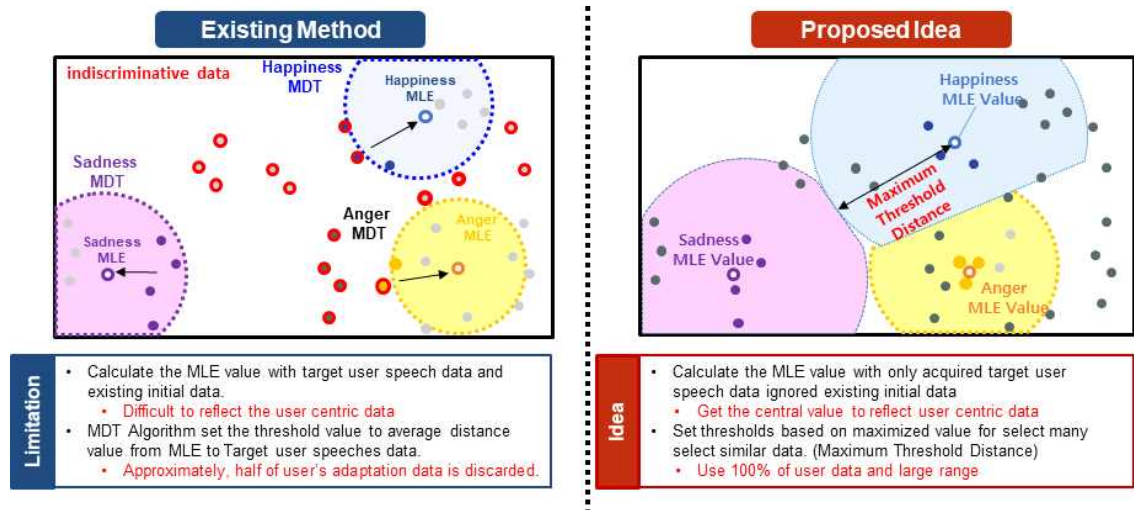


Figure 1-6. Proposed Idea for Small Data Environment

In the absent data environment, when the personalization model is generated based on the collected data, there is a problem that the recognition accuracy of the absent label is 0%. Therefore, the existing method deal with non-existent emotional data in a way that utilizes existing models. This method has been utilized in many methodologies as a way to maintain the accuracy of the initial model in the absence data environment. However, this method causes a problem of data imbalance because it utilizes a large number of existing data of the absence label. In addition, when the data of the initial model is much larger than the data collected from the target user, there is a disadvantage that more target user data is required to show high recognition accuracy in addition to the absence label. Therefore, the solution for absent data environment propose the approach that utilizes emotional voice data of the user most similar to the absent emotional label. This approach considers only data that is more similar to the target user, and can solve some of the imbalanced data environment that is generated when using all of the existing data.

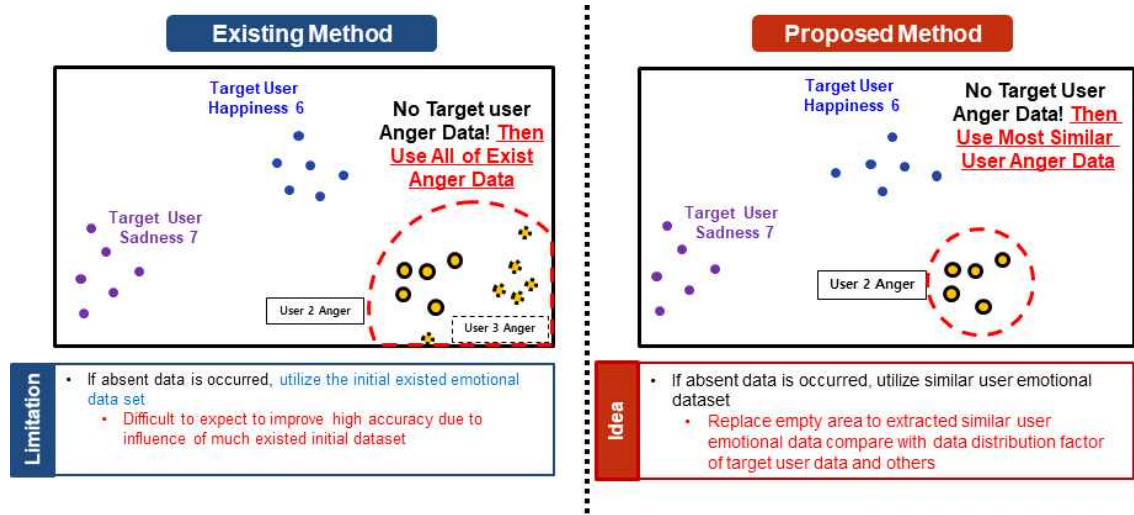


Figure 1-7. Proposed Idea for Absent Data Environment

As is well known, in imbalanced data environments it is difficult to show even accuracy for each class. Especially in environments where training data sets are not well balanced, the amount of data collected does not show high accuracy in the actual field. To solve these problems, various studies have been conducted in various machine learning fields. In the personalized emotion recognition, the target user's data can not be received in a balanced manner for each emotion, and the difference between the minor class data and the major class data is large. In addition, existing personalized emotion recognition studies do not provide any process considering imbalanced elements because most of them conduct experiments using balanced data sets. Therefore, the solution for imbalanced data environment proposes the virtual case data augmentation approach that utilizes SMOTE oversampling algorithms.

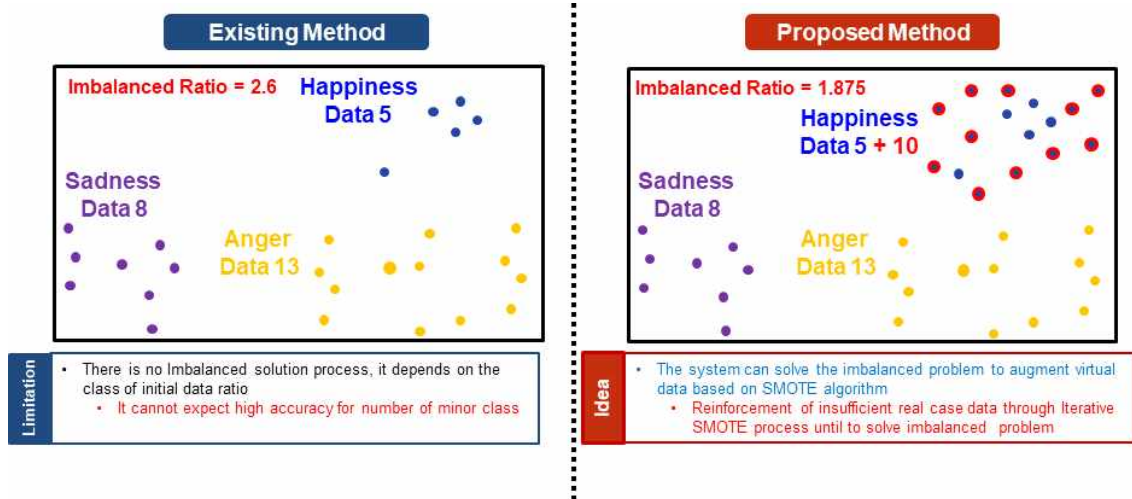


Figure 1-8. Proposed Idea for imbalanced data environment

Existing methods can not solve effectively about 3 kinds of cold-start problem. These limitations are the reason that the accuracy of personalized emotion recognition model to be lowered. Therefore, I proposed a personalized model creation with 3 solutions to solve each challenge. The description of the proposed method concept is presented in the following section.

1.4 Proposed Concept

To provide a personalized model for the target user, it is important to collect a varied amount of target user speech in a balanced manner. However, the target user's speech may not exist when using the recognition process for the first time, and it is impossible to collect emotion data if the user does not appropriately express themselves during the data collection period. In this initial stage, it is difficult to create a personalized model with high accuracy since there is no speech dataset that includes various cases, thus making it impossible to predict the data distribution of the target user. In order to create a highly personalized training model, it is necessary to reinforce and augment various speech data.

Therefore, this thesis propose Robust Speaker Adaptation methods to acquire an initial dataset through data reinforcement and data augmentation to create a personalized model with high accuracy with a emotionally imbalanced minimal number of samples.

Robust Speaker Adaptation reinforces and augments real and virtual data to provide a customized model for target users. Robust Speaker Adaptation consists of 3 kinds solutions to resolve the cold-start problem such as small data environment and absent data environment, imbalanced data environment. The proposed framework designed with solution 1, 2 and 3. The below figure shown the concept of the Robust Speaker Adaptation Framework. The descriptions of the detailed methodologies are given in the Chapter 3.

- **Solution 1: Similar Data Selection based on Maximum Threshold Distance.**

Reinforces the insufficient target emotional samples from an initial constructed multiple user speech dataset to increase the data samples to solve the small data environment.

- **Solution 2: Similar Other Similar User Emotional Speech Mapping.**

Replaces the dataset of empty target emotional samples through similar user

emotional speeches from another user speech dataset to solve the absent data environment.

● **Solution 3: Virtual Case Data Augmentation based on SMOTE.**

Augment virtual case data using oversampling algorithm to solve the imbalanced data environment on selected real-case data.

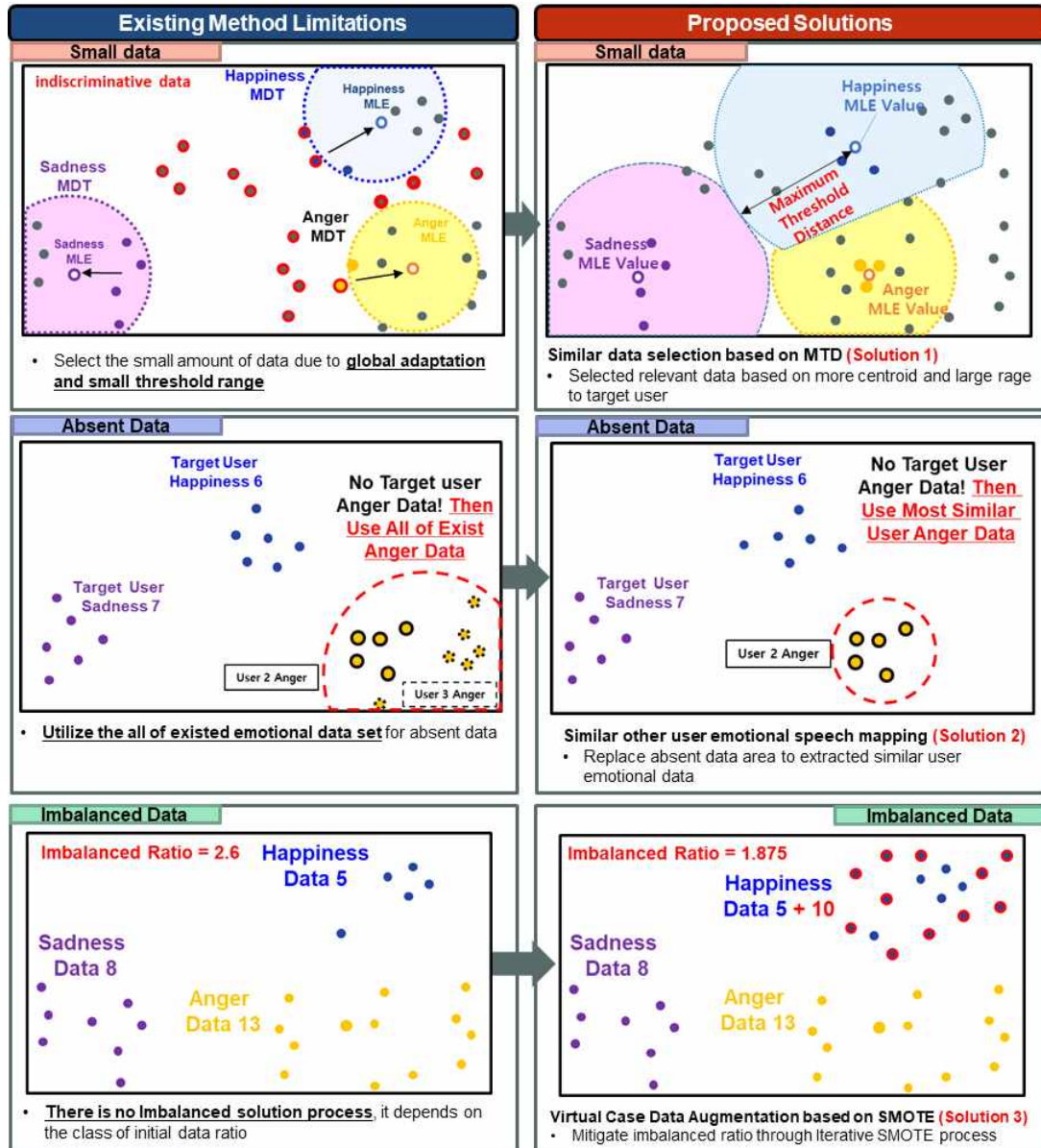


Figure 1-9. Abstract Exist method's limitation and proposed solution.

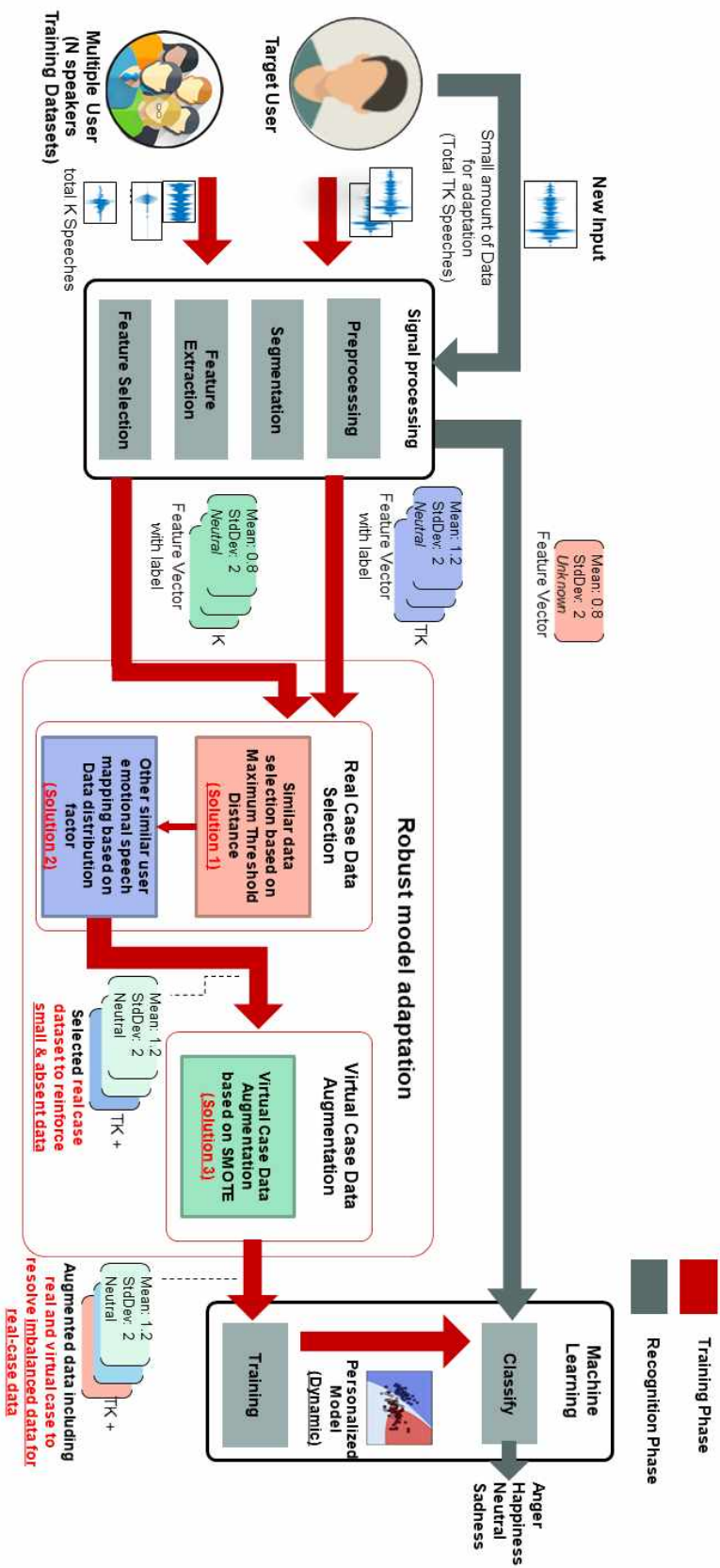


Figure 1-10 The concept of the Robust Speaker Adaptation Framework

1.5 Key Contributions

In this thesis, I proposed Robust Speaker Adaptation Framework to solve the cold-start problem in initial stage of personalized emotion recognition. The proposed Robust Speaker Adaptation Framework select similar real-case data for small and absent data environment and augment virtual data for imbalanced data environment.

The main work of this study can be divided 3 parts to solve the cold-start problem. First, target user oriented similar user speech selection to insufficient data reinforcement. Second, I proposed similar other user emotional mapping to reinforce absent data. Third, I proposed the virtual data augmentation using the oversampling algorithm to reduce the imbalanced ratio.

- **Solution 1: Similar Data Selection based on Maximum Threshold Distance for Small Data Environment. (Real-case Data Selection)**

In the first study, I propose the data selection algorithm based on Maximum Threshold Distance. This is a new approach that select the similar data based on target user centric value and maximized range. It can select more target user relevant speech data than exist method.

- **Solution 2: Similar Other Similar User Emotional Speech Mapping based on Data Distribution Factor for Absent Data Environment. (Real-case Data Selection)**

Second, I proposed a process of replacing absent data to other similar user emotional speech based on 4 kinds of data distribution factors as median, variance, skewness and kurtosis. This is a new approach to reinforce absent data in the process of target user data acquisition. The exist approach can not improvement of accuracy for absent data due to utilize the initial model. However, this proposed approach can improve the accuracy for absent data due to utilize similar user speech data.

- **Solution 3: Virtual Case Data Augmentation based on SMOTE for Imbalanced Data Environment. (Virtual-case Data Selection)**

Finally, I proposed the iterative virtual case data augmentation to solve the imbalanced data environment for selected real-case data by solution 1 and 2. This approach employed SMOTE algorithm which is most popular oversampling techniques to reduce the imbalanced ratio between major class and minor class. It can create personalized training dataset effectively through virtual-case data augmentation repeatedly until the number of the data of imbalanced ratio is low.

1.6 Thesis Organization

This thesis is organized into chapters as following.

- **Chapter 1: Introduction.** This Chapter provides a brief introduction to the research work of personalized speech emotion recognition. The chapter further summarized several major problems in the area of cold-start problem and the limitations of current approaches. After that, the goal and the overview of the contribution of the dissertation are presented.
- **Chapter 2: Related Work.** This chapter provides the detail review and discussion on previous approaches of personalized speech emotion recognition to provide highly accurate recognition model and their corresponding limitations. In addition, traditional speech emotion recognition limitations also given.
- **Chapter 3: Proposed Robust Speaker Adaptation Methodologies.** This chapter, I present the robust speaker adaptation framework with 3 core solutions. This work mainly propose real-case and virtual-case data augmentation to solve the cold-start problem in personalized emotion recognition.
- **Chapter 4: Experimental results and Discussions.** This chapter provide the description of testing datasets, currently available publicly, which are utilized for evaluating the performance of our proposed robust speaker adaptation algorithm on results of 2 kinds evaluation as accuracy and imbalanced ratio.
- **Chapter 5: Conclusion and Future Directions.** This chapter concludes the dissertation with some discussions of limitations and also provides future directions for performance improvement.

This chapter presents the limitations of traditional speech emotion recognition approaches and why the emotion recognition system need the personalization. in addition, this chapter also introduces state-of-the-art Personalized Speech Emotion Recognition methods.

2.1 Traditional Speech Emotion Recognition

Traditional speech emotion recognition aim to improve accuracy thorough extract new features [3, 34] or apply classification methodologies [35-38] based on generalized model [34]. As a representative new feature extraction study, it is a technique that recognizes emotions without specifying a window size because it uses features that reflect the characteristics of each individual vocalization. [3] The features used in this study are 14 MFCC, Delta 14 MFCC, Delta-Delta 14 MFCC, Log-Energy features.

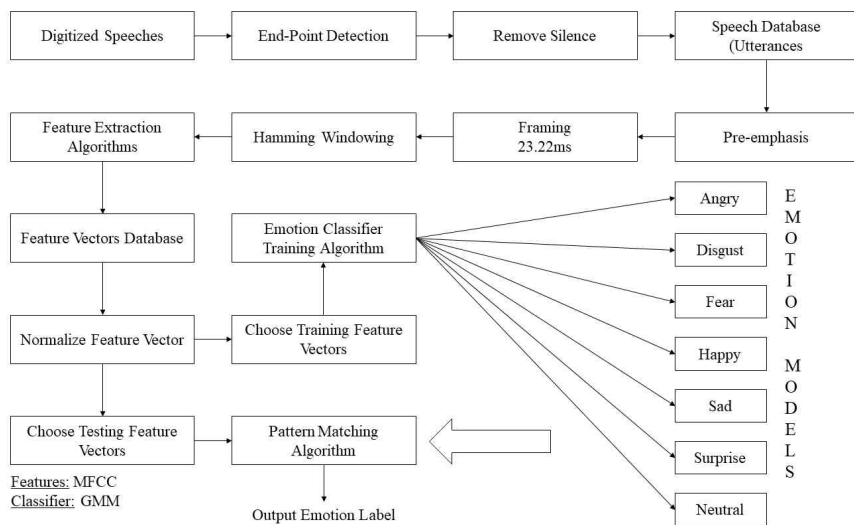


Figure 2-1. MFCC feature based Emotion Recognition with GMM

A representative study that applies the classification methodology is emotion recognition based on SVM and HMM-GMM Hybrid System. [37] This approach utilized GMM-HMM model fed with MFCC feature to exploit the dynamics of emotional signals. Then a balanced SVM classifier was applied on the static Low-Level Descriptors (LLD) feature as Zero Crossing Rate (ZCR), Root Means Square (RMS) Energy, F0-frequency(pitch), Harmonics to Noise and 1-12 MFCC features. In order to take advantage of generative model and discriminate model, these two methods were combined by taking the probability representation of the GMM-HMM model as the additional features for the SVM classifier.

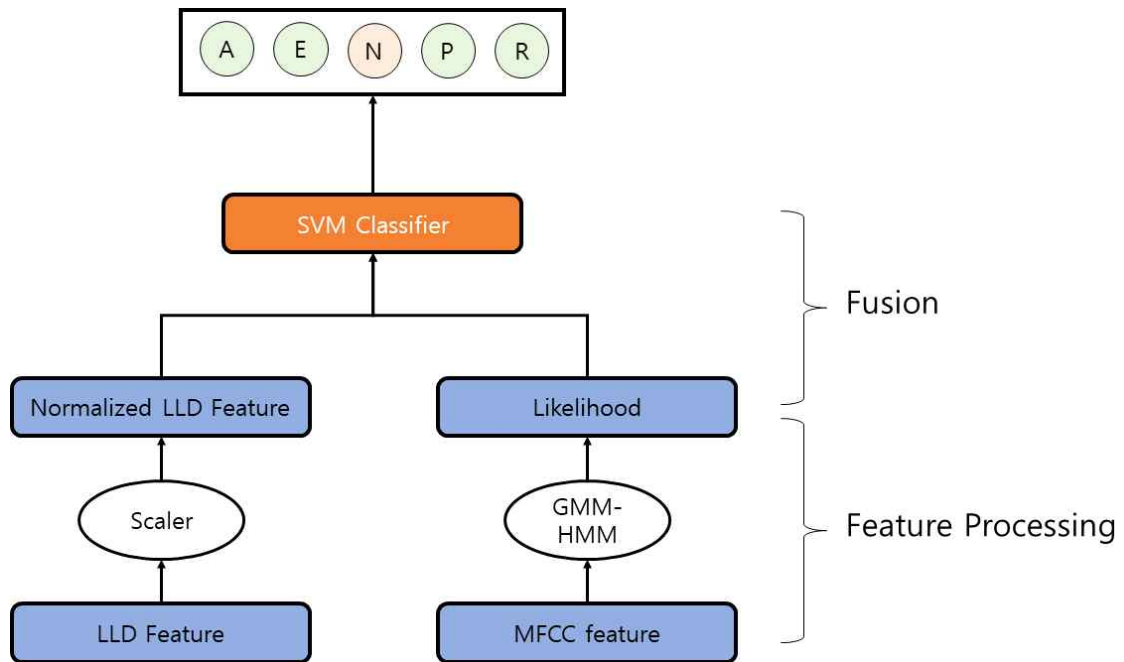


Figure 2-2. SVM and HMM-GMM Hybrid System

Another study applying the classification methodology has applied the hierarchical classification methodology [38]. This study divides and classifies similar emotional factors in speech by using various classifiers as SVM, BLG and SVR.

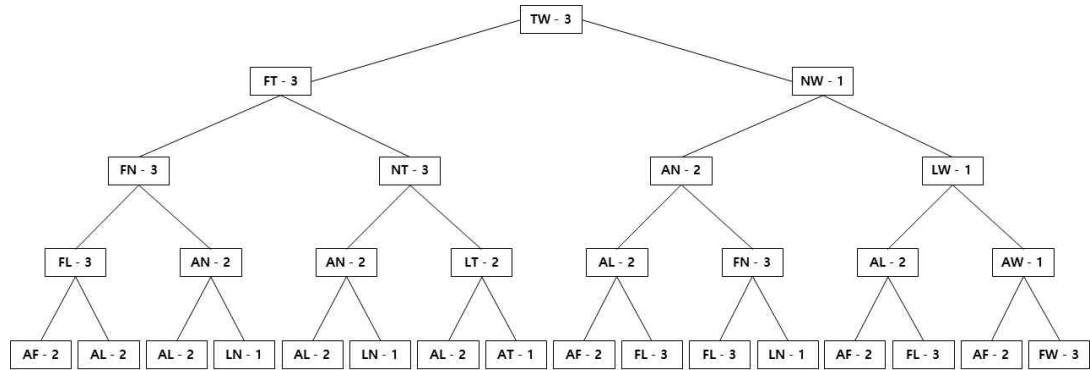


Figure 2-3. Optimum Hierarchical Decision Tree using best classifier at each decision level

These traditional Speech Emotion Recognition system generate static model based on training data collected from various users. And then, the system provide recognition process to every user based on generated static model. This system Performed low accuracy in speaker independent evaluations and Impossible to modify training model due to implement by static model. These techniques have a limitation in user independent evaluation. That means they don't guarantee same accuracy on all of user. And these traditional speech emotion recognition system can't reflect user voice characteristic. In other words, traditional speech emotion recognition framework have very different accuracy in user independent evaluation. Therefore, recent emotion recognition technologies have been research personalized emotion recognition framework. Figure 2-4 shows the structure of a traditional speech emotion recognition system.

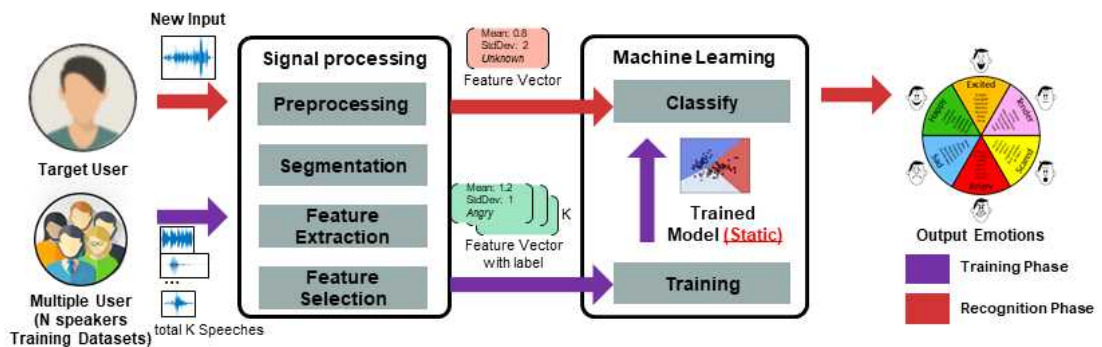


Figure 2-4. Traditional speech Emotion Recognition System

2.2 Personalized Emotion Recognition

A personalized emotion recognition system aims to tune the initially constructed model to recognize accurately the targeted user's emotion. Personalized emotion recognition researches are divided into three categories: feature adaptation, incremental learning, and model adaptation. Following sections introduce the state-of-art personalization methods.

2.2.1 Feature Adaptation

Feature adaptation is a technique to find a suitable feature scheme for users by using unsupervised learning methodology. These approach can make the personalized model.

A typical study of feature adaptation is the Iterative Feature Normalization (IFN) framework [23]. This approach proposed to eliminate user intervention for the annotation process in a personalized emotion recognition process. Adapting any general emotion recognition system for a particular individual requires speech samples and prior knowledge about their emotional content. These assumptions constrain the use of these techniques in many real scenarios in which no annotated data is available to train or adapt the models. To address this problem, this paper introduces an unsupervised feature adaptation scheme that aims to reduce the mismatch between the acoustic features used to train the system and the acoustic features extracted from the target user [24].

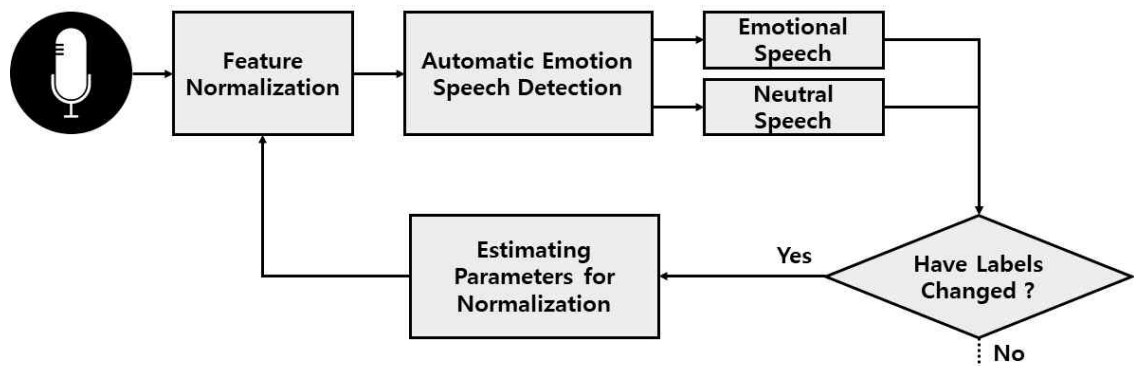


Figure 2-5. Overview of The Iterative Feature Normalization (IFN) approach

Another study is the Robust Feature Normalization and Selection [39]. This approach focus on reducing the feature mismatch caused by the variability of speakers and channels. The histogram equalization normalization is used to normalize each feature component. In addition, an eigen feature selection is performed for the discriminative representation. In this representation, the meaningful features in a reduced feature dimension are obtained via subspace projection to eliminate the noise features. Finally, Gaussian Mixture Models (GMM) are adopted to classify the emotional states.

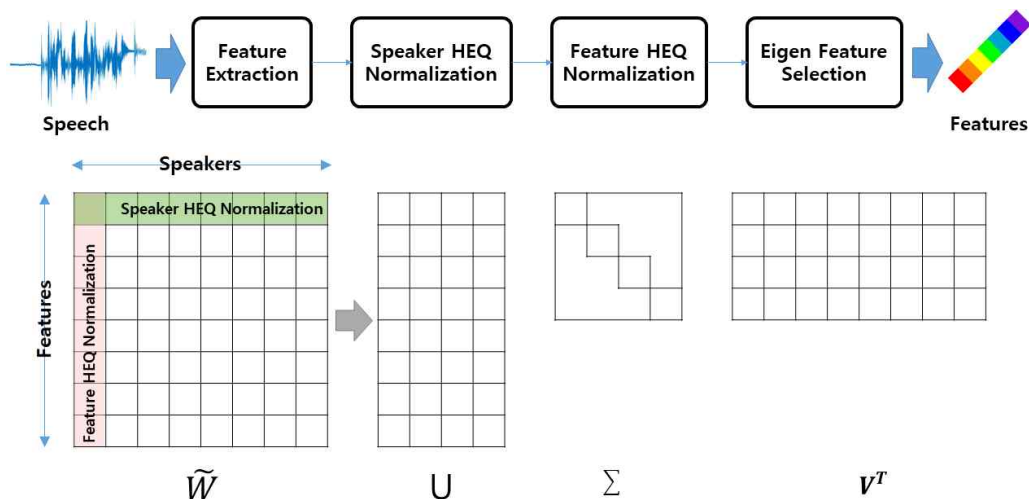


Figure 2-6. Feature Analysis and Singular Value Decomposition for Speaker State Recognition

The most recent feature normalization study is the Cascaded Normalization approach. This approach is focus on the robustness for cross-corpus and corps-language emotion recognition. This paper finds the best feature normalization strategies deploys a cascaded normalization approach, combining linear speaker level, nonlinear value level and feature vector level normalization to minimize speaker- and corpus-related effects as well as to maximize class separability with linear kernel classifiers.

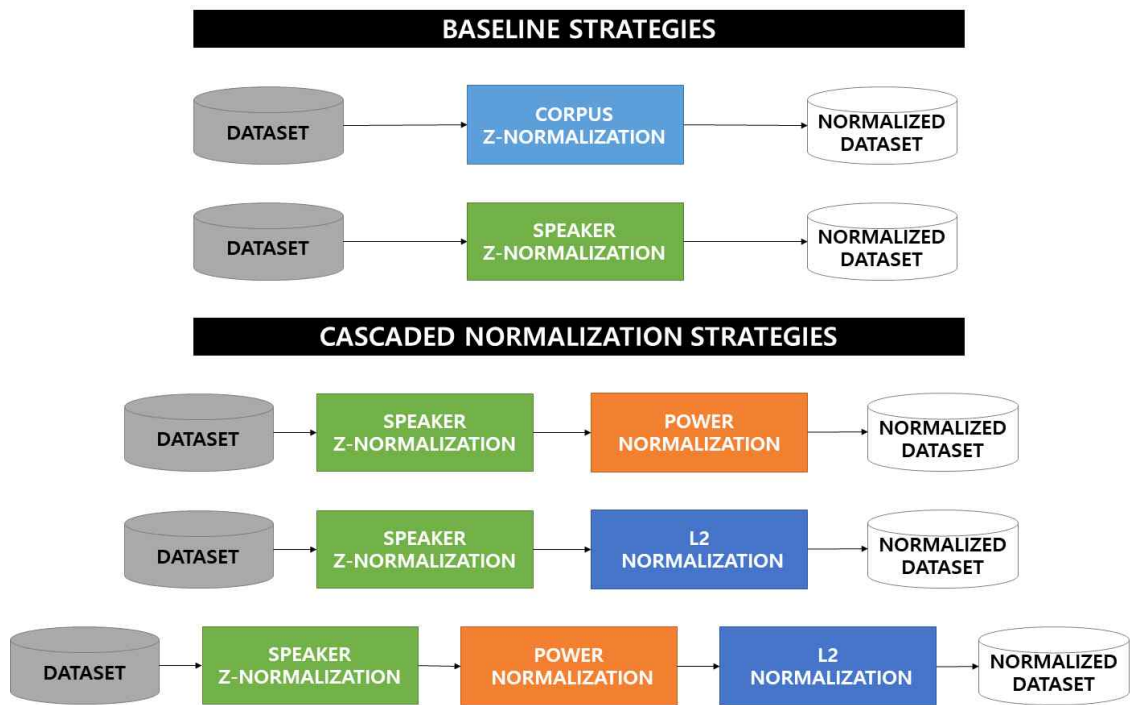


Figure 2-7. Illustration of baseline and cascaded normalization strategies

These approaches are convenient for the user, but there is no verification of the user's data and it is difficult to make a proper personalized model in a short time. That means, it require large amount of your target user data and time to create a stable, personalized model.

2.2.2 Incremental Learning

The incremental learning methodology provides a personalized model by retraining process with machine learning algorithm continuously when the target user's data is input in the existing model. These approach can make the personalized emotion recognition model by supervised or unsupervised manner.

A typical research of incremental learning methodology in speech emotion recognition is the active learning approach using an Incremental Support Vector Machine (SVM) [40]. Incremental SVM classifiers were introduced to reduce batch SVM memory and computational requirements, especially for very large data sets. One of the useful features of incremental learning is the ability to add more training data. These approaches employ incremental learning techniques, where only a subset of the data is considered at each step of the learning process, discarding old data while maintaining the support vectors learned in previous steps [41, 43]. Shalev et al [42] proposed and analyzed a simple and effective stochastic sub-gradient descent algorithm for solving the optimization problem imposed by SVMs. Each iteration of the algorithm operates on a single training example selected at random. By selecting the training examples at random, the authors demonstrated that the solution converges in probability regardless of the data used in the classification problem.

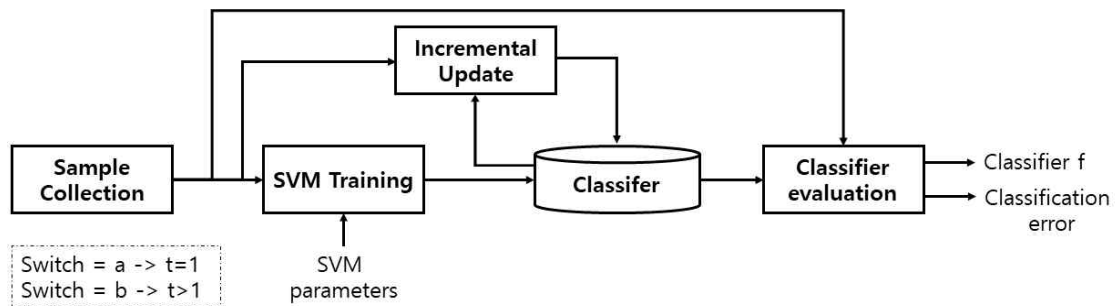


Figure 2-8. Incremental SVM Architecture

Another incremental learning study is the SVM Adaptation. This work uses the adaptive SVM algorithm proposed by Yang et al. [44] in an attempt to transform existing SVM classifiers into a new effective SVM classifier that would work on a new dataset with limited number of labeled data. The approach aims to minimize both the classification error over the training examples, and the discrepancy between the originals and adapted classifier. The new optimization problem seeks a decision boundary close to that of the classifier trained from the source domain, while managing to separate the new labeled data from the target domain.

These approaches require large amounts of data to create stable personalized models that are dependent on target user speech. And also, these approaches are less likely to be an imbalanced data environment if the number of initial data is balanced. However, there is a disadvantage that it is not possible to create a personalized model properly in a small data set because it requires a lot of data. Recently, Model adaptation technology which can give a quick change of model to overcome the limitation of incremental learning methods is mainly studied, and detailed description is given in the next section.

2.2.3 Model Adaptation

The model Adaptation technology constructs training dataset by collecting a small amount of adaptive data from the target user and selecting similar user data from the initial model based on the target user speeches. This approach only requires a relatively small amount of data from the target user through the target user customized emotional speech training model, it can nearly achieve the performance of the SD model [45].

The typical personalized emotion recognition method is Maximum Likelihood Linear Regression (MLLR) based speaker adaption technique [46]. The conventional MLLR adaptation technique modify the initial SI models parameters, i.e. Gaussian means and variances, according to transformation matrices. Given adaptation data collected from target users and their labels, the transformation matrices are estimated to maximize the likelihood of the adapted models observing the adaptation data, using the expectation-maximization (EM) algorithm [33]. However, this approach still require sufficient target user speeches of adaptation data are necessary in order to reliably calculate the transformation matrices [47]. Therefore, this approach is difficult to solve cold start problem in initial stage of personalized emotion recognition. Figure 2-10 represents a general procedure for the conventional MLLR adaptation.

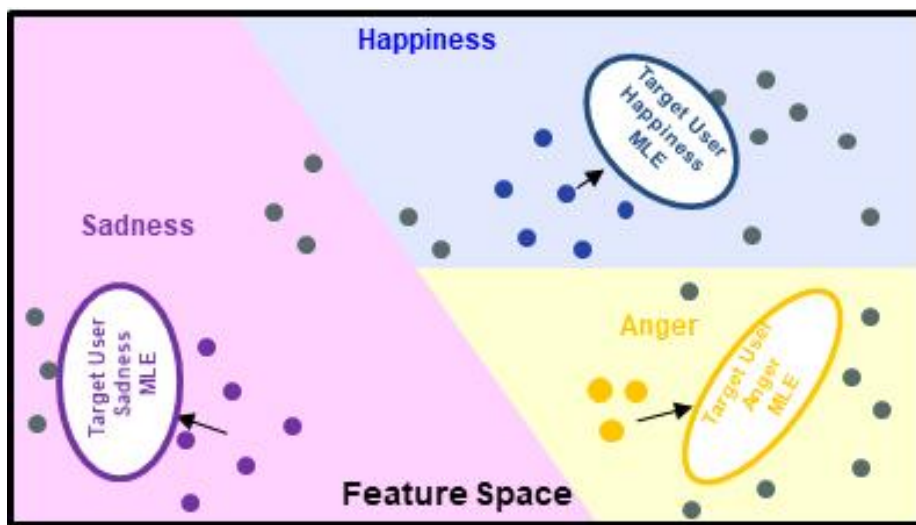


Figure 2-9. Example of the data selection by MLLR Adaptation

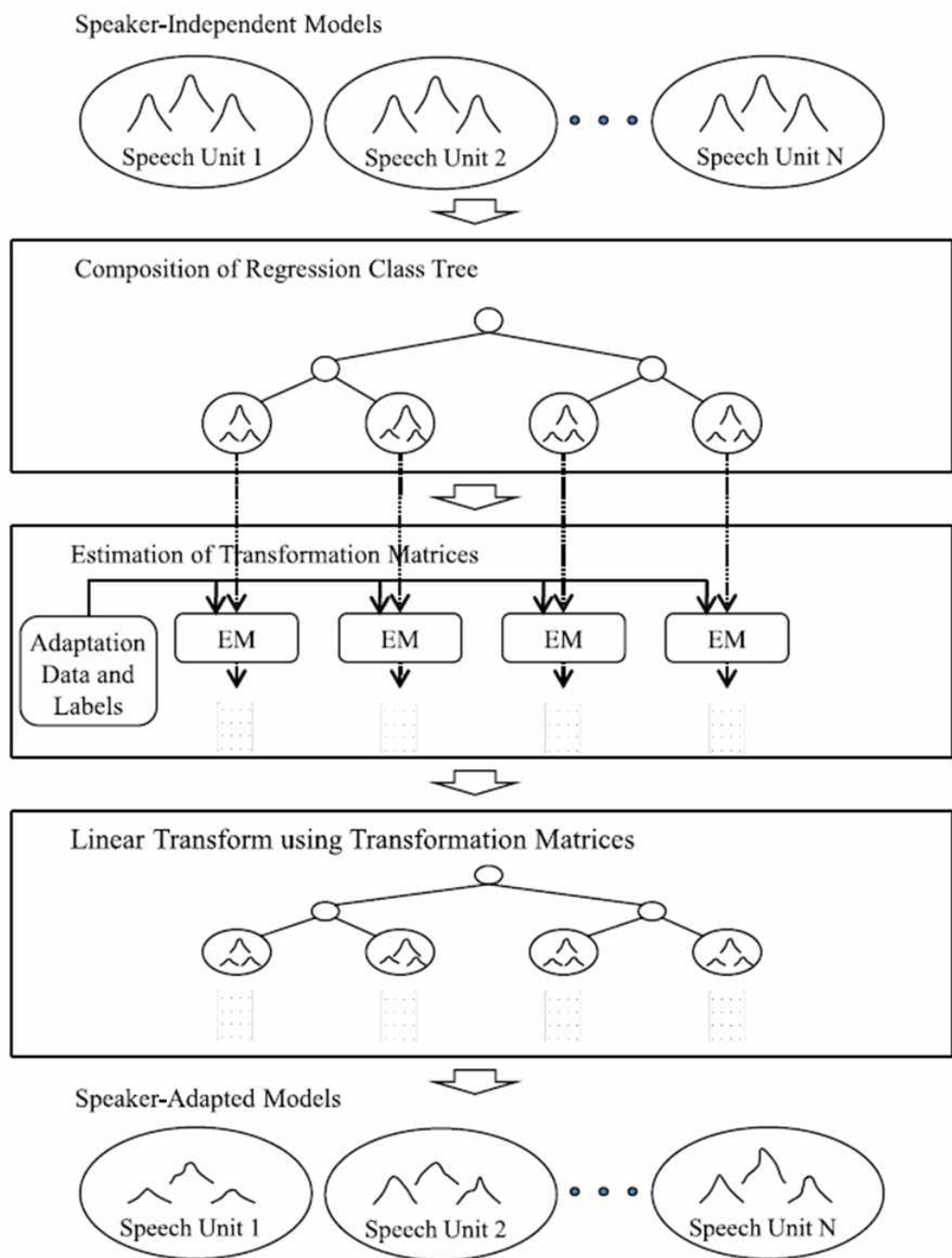


Figure 2-10. Procedure for the conventional MLLR adaptation [33]

To resolve the MLLR problems, There is LDM-MDT(Log-likelihood Distance based confidence Measure - Model based Dynamic Threshold) MLLR based Data Selection techniques [24]. This approach solved conventional MLLR adaptation problem to select useless data from the initial model with discarding indiscriminative emotional speech data based on MDT. However this approach still require sufficient data environment (At least 18min data, about 360 samples need) and approximately, half of all of target user adaptation data are determined to be indiscriminative and are disregarded. And also If absence data is exist, utilize Initial model.

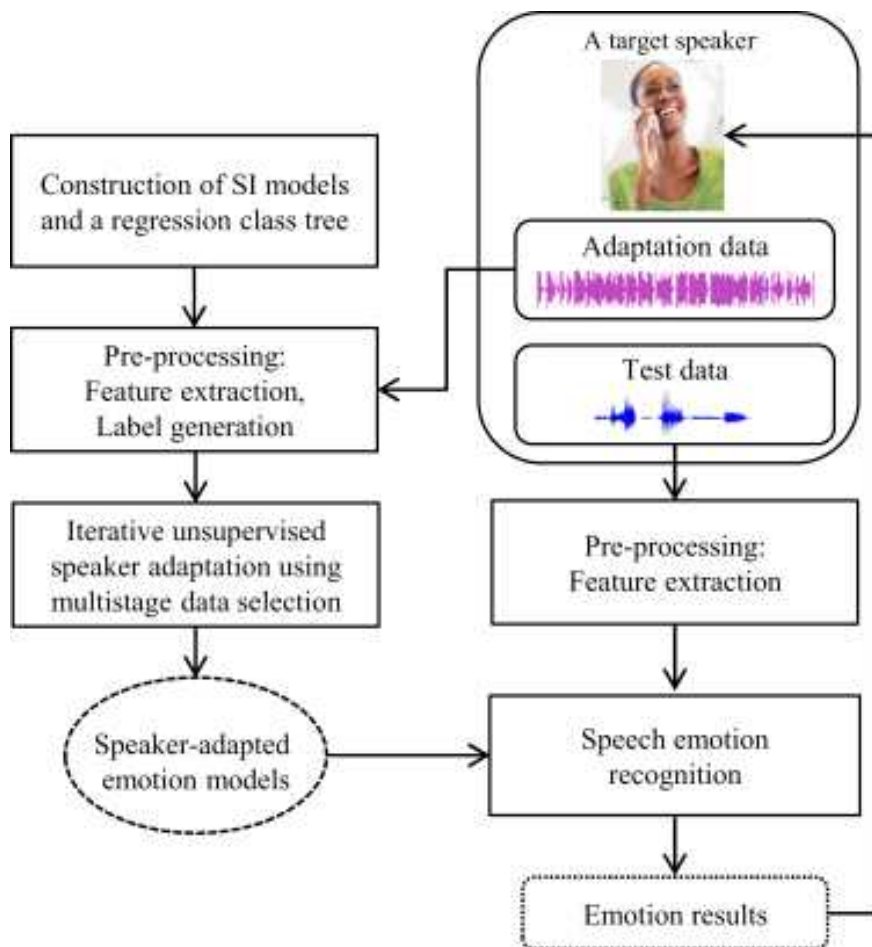


Figure 2-11. Multistage data selection based on LDM-MDT MLLR Algorithm [33]

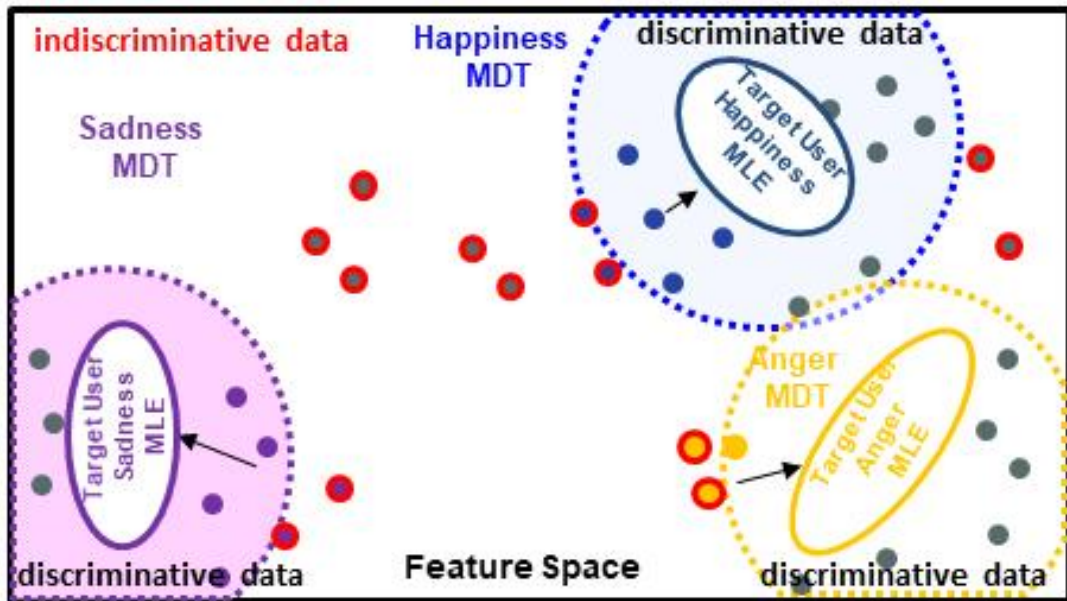


Figure 2-12. The example of data selection based on LDM-MDT MLLR Algorithm

These approaches can generate a personalized model by reflecting the data immediately even in a small data environment, but it has a disadvantage that the model adaptation speed is slow. In addition, there is no process corresponding to the absence data environment, the imbalanced environment problem may occur because the accuracy of the emotion part is not significantly increased by the data of the SI model. Finally, since the imbalanced data environment cannot be solved, there is a disadvantage that the accuracy per each emotion cannot be displayed evenly.

2.2.4 Deep Domain Adaptation Network

Deep Neural Networks have achieved great success in various applications [48]. The deep domain adaptation network (DAN) adjust a model from the source domain knowledge to a different related target domain [49]. Almost deep domain model adaptation techniques are researched in image-based recognition such as face recognition and object recognition to overcome the problem of the different background environment [50]. This method creates virtual data related to target image samples of the newly added class and uses the Deep Learning algorithm to create training models to improve the recognition accuracy of new images in various background environments. This approach is highly focused on image compositing methods because it can not label many real-case images considering the directionality and accuracy of images.

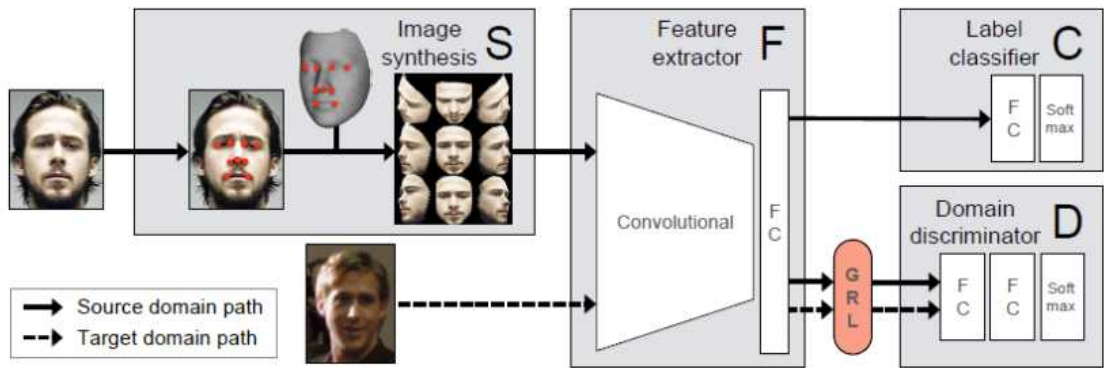


Figure 2-13. Outline of the SSPP-DAN

(Single Sample Per Person - Domain Adaptation Networks) [50]

Speech emotion recognition area also have common problem. The performance of speech emotion recognition is affected by differences in data distribution between training and test datasets used to build and evaluate the models [51]. Many speech emotion recognition domain adaptation researches using Deep Neural Network (DNN)

aim to solve the adversarial example that is that adding some noise to the original input data will degrade DNN performance. This DAN techniques are very powerful to solve the differences in data distributions. Therefore, the DAN researches in speech emotion recognition handle cross-corpus evaluation problem and noise recording environments.

The representative cross-corpus research method is the Domain Adversarial Neural Network (DANN) for emotion recognition. The DANN approach has constructed the network is trained using labeled data from the source domain and unlabeled data from the target domain. The network is learned by two classifiers with tasks classifier and domain classifier based on samples entered into the target sources domain. Both classifiers share the first few layers that determine the representation of the data used for classification. This approach can increase the robustness of the speech emotion recognition system against certain type of noise and readjust weights to find the new representation that satisfies all conditions when change the training data set.

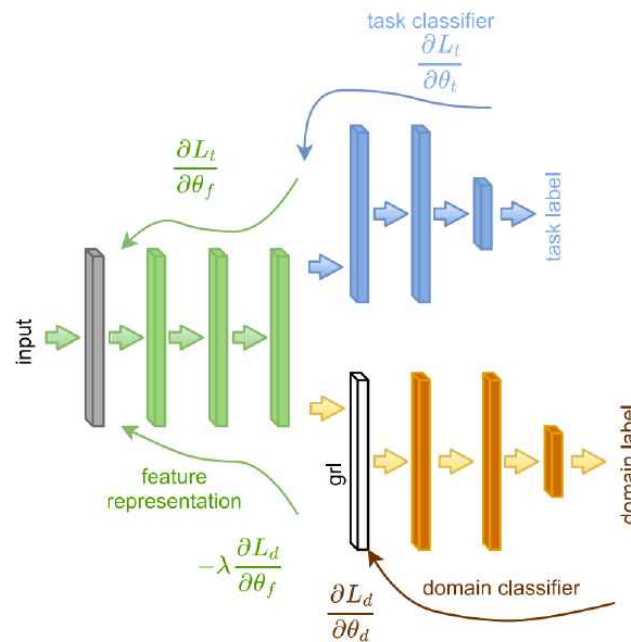


Figure 2-14. Architecture of the domain adversarial neural network (DANN) [51]

The representative robustness for audio noise research using DAN is a Generative Adversarial Networks (GAN) based defense for speech emotion recognition system [52]. This approach aim to investigate the utility of adversarial examples to achieve robustness in speech emotion classification to adversarial attacks. It can better withstand adversarial examples compared to the previous defense solutions [53-54] such as adversarial training and random noise addition.

These Deep Domain Adaptation Networks are a methodology that adapts quickly to the domain of a new recording environment by utilizing existing data sets. In other words, this approach creates new models by comparing and modifying existing data with new target domains considering data distributions. Therefore, this technique is very powerful for changes in the data environment, but it is not suitable for generating a personalization model reflecting the characteristics of the target user data. This method requires virtual speech data generation and synthesis techniques to apply this methodology to personalized speech based emotion recognition that covers various cases in a limited personalized data environment.

Chapter 3

Proposed Robust Speaker Adaptation Methodologies

3.1 Robust Speaker Adaptation Framework

The framework introduced in this section incrementally creates an acceptable training model using a minimal number of target user samples via the proposed Robust Speaker Adaptation methods. This framework is an innovative system that can resolve the cold-start problem present in small and emotionally-imbalanced data environments. The proposed Robust Speaker Adaptation, which is the core methodology of this framework, consists of data reinforcement and data augmentation. The data reinforcement method selects real data by determining the similarity of speech datasets between the acquired target speech and the initial multiple-user training model. The data augmentation method generates virtual data to create more scenarios by utilizing SMOTE. The augmented data extracted via the Robust Speaker Adaptation process constructs the personalized training model using a machine learning algorithm.

This framework can create and update a personalized model incrementally for a target user by implementing a re-training process with only a single target user input. Figure 3-1 shows the system architecture of the proposed method.

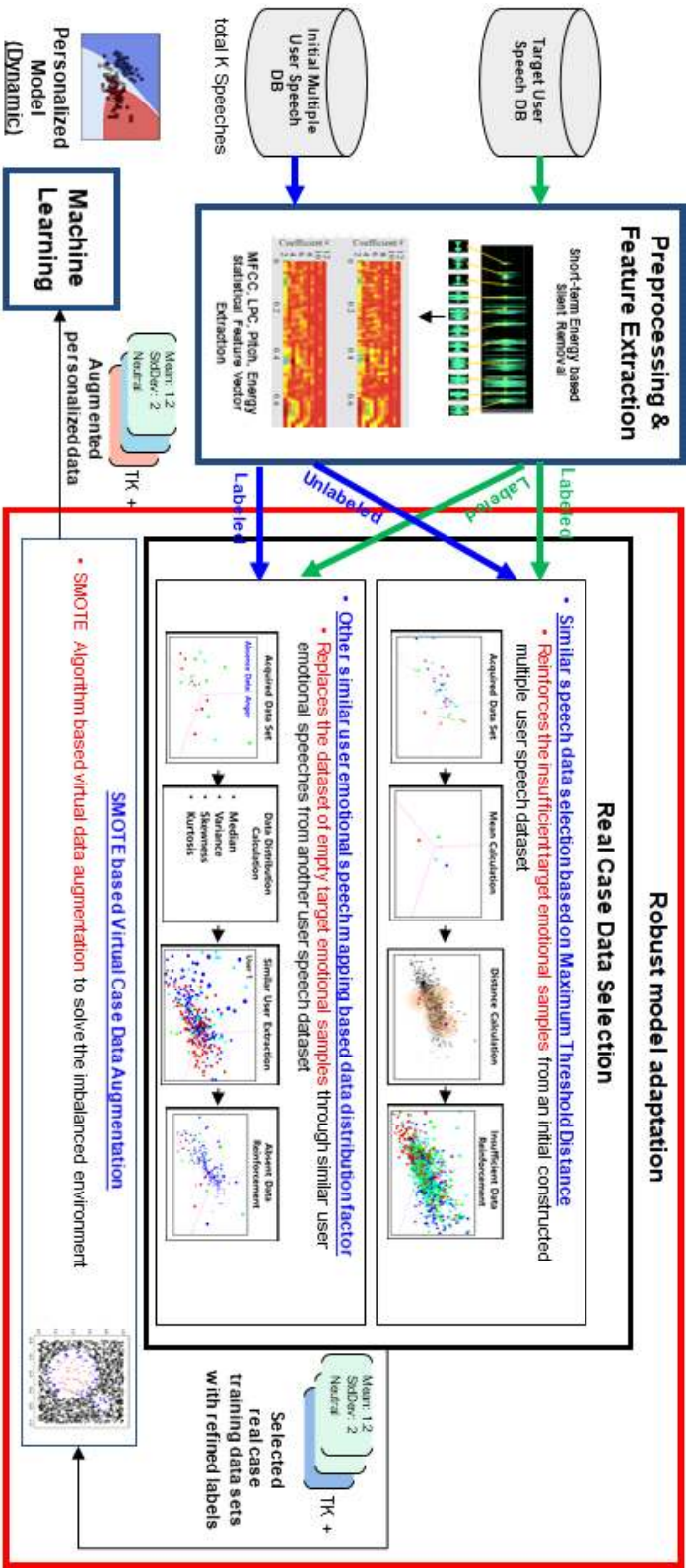


Figure 3-1 Proposed Robust Speaker Adaptation Framework

① Preprocessing

This module performs normalization and the silence removal process. this thesis employed the peak normalization implemented by jAudio [55] from the precedent research [56], which is the default approach to adjusting the data value based on the highest signal level present in the audio. Additionally, also employed the existing silent removal approach based on the zero crossing rate (ZCR) for speaker identification [57] to discard the blank area in the speech. This approach divides audio into frames, where each duration is segmented in 15ms by a hamming window. Then, speech boundaries are estimated based on the short time energy (STE) algorithm [58]. After that, silence areas are removed by the extracted threshold values.

When the converted STE signal information increases from the minimum threshold value to the maximum threshold value, it is judged to be a start point of voice. Thereafter, the maximum threshold value is decreased from the maximum threshold value to the minimum threshold value, and when it falls below the threshold value, it is determined as the end point. Equation 1 represents the Short-term Energy conversion formula, and Equations 2, 3, 4, 5, and 6 represent the threshold value extraction process.

$$E(m) = \sum_{n=1}^N x_m(n)^2 \quad (1)$$

$$Energy_{\max} = \max(E(i)), i = 1, 2, \dots, M \quad (2)$$

$$Energy_{\min} = \min(E(i)), i = 1, 2, \dots, M \quad (3)$$

$$T_{\min} = 1 + 2\log_{10} \frac{Energy_{\max}}{Energy_{\min}} \quad (4)$$

$$SL = \frac{\sum_i E(i)}{\sum_i 1} \quad (5)$$

$$T_{max} = T_{min} - 0.25(SL - T_{min}) \quad (6)$$

Figure 3-2 shows the waveforms of the original speech Signal and Figure 3-3 shows the waveforms of the original speech signal converted into Short-term Energy information.

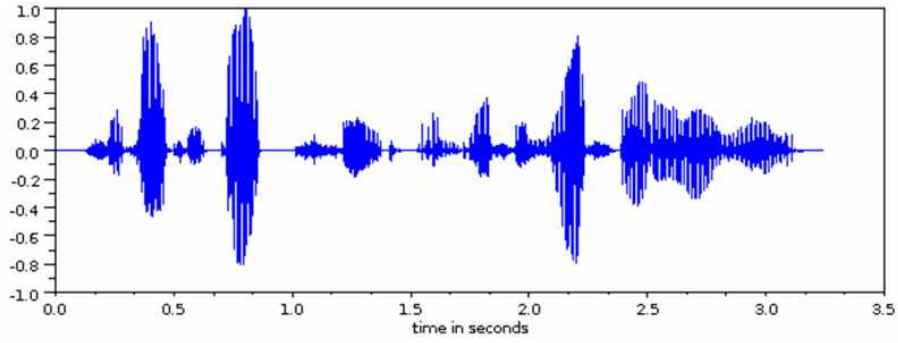


Figure 3-2. Waveform of Original Raw Speech Signal

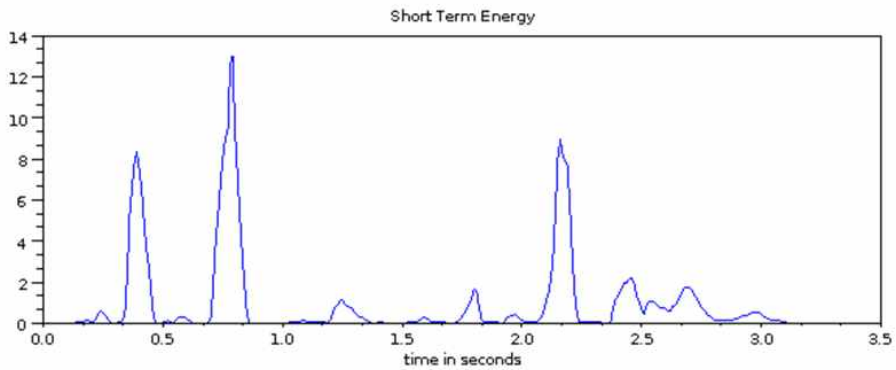


Figure 3-3. Waveform of Converted STE Signal

This method can extract user's speech in consideration of the noise level. Figure 3-4 presents the examples of waves of before and after applying the silent remover.

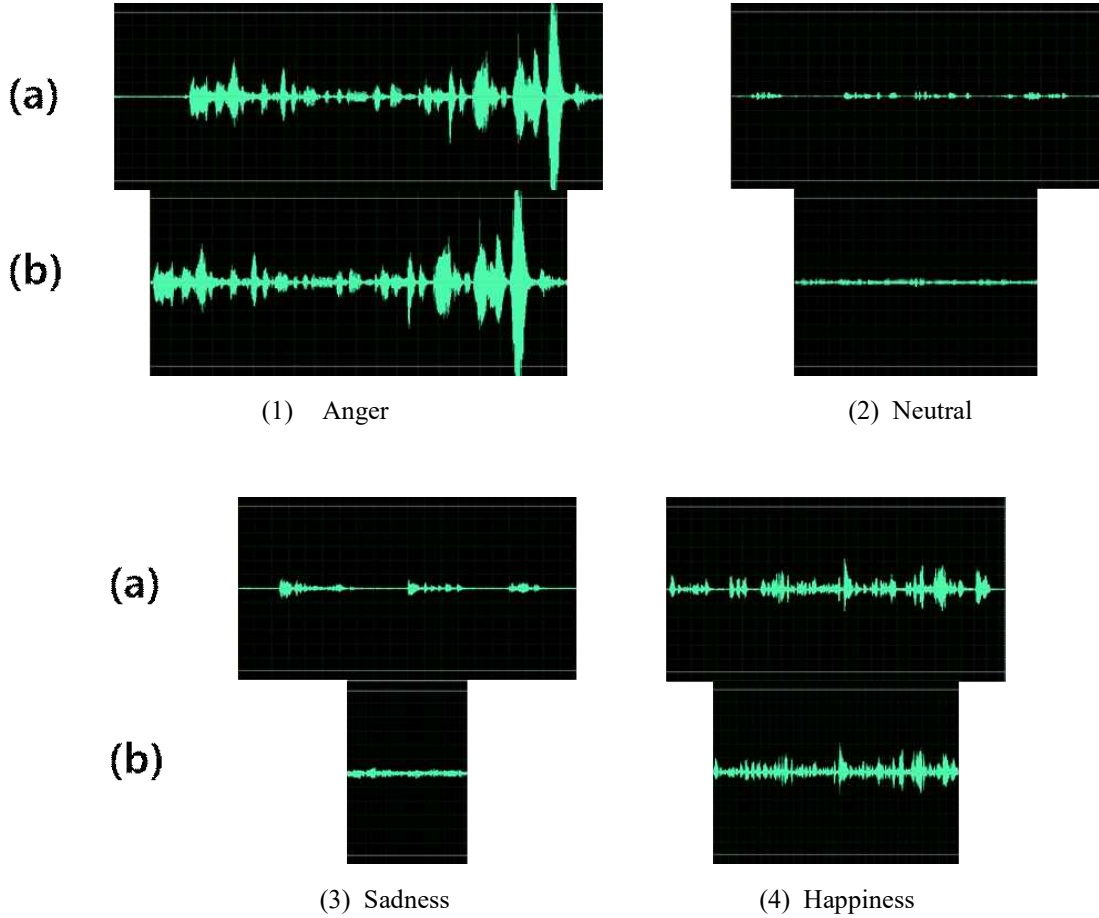


Figure 3-4. Waves of (a) before and (b) after preprocessing module in a sentence

② Feature Extraction & Selection

This module extracts the feature vector from the speech. this thesis employed various basic feature vector in existing methods of speech emotion recognition area [59]. At first, the speech data is split to 16ms and then the filter-bank values are extracted, including 13 MFCC (Mel Frequency Cepstral Coefficient), 10 LPC (Linear

Predictive Coding), Energy, and Pitch in each frame. Then, it calculates the statistical feature vector, which includes the mean, standard deviation, max, and min. Table 3-1 shows the feature vector scheme description.

Table 3-1. Feature Vector Scheme Description

Categories	Statistical Values	Number of Features (100)	Description
13 MFCC	<ul style="list-style-type: none"> - Mean - StdDev - Max - Min 	52 (13 x 4)	This filterbank algorithm takes into account human auditory characteristics and is widely used in speech recognition, having excellent recognition performance [60].
10 LPC		40 (10 x 4)	This filterbank algorithm is also widely used in speech recognition as a kind of parameter speech synthesis method based on a humans vocalization model. [61]
Energy		4	This is a feature that is mainly used in speech-based emotion recognition by measuring the strength of a voice waveform in a speech impulse signal. [62]
Pitch		4	This is a feature which is frequently used in speech recognition and includes the main acoustic correlation of tone and intonation generated by vocal frequency per second. [63]

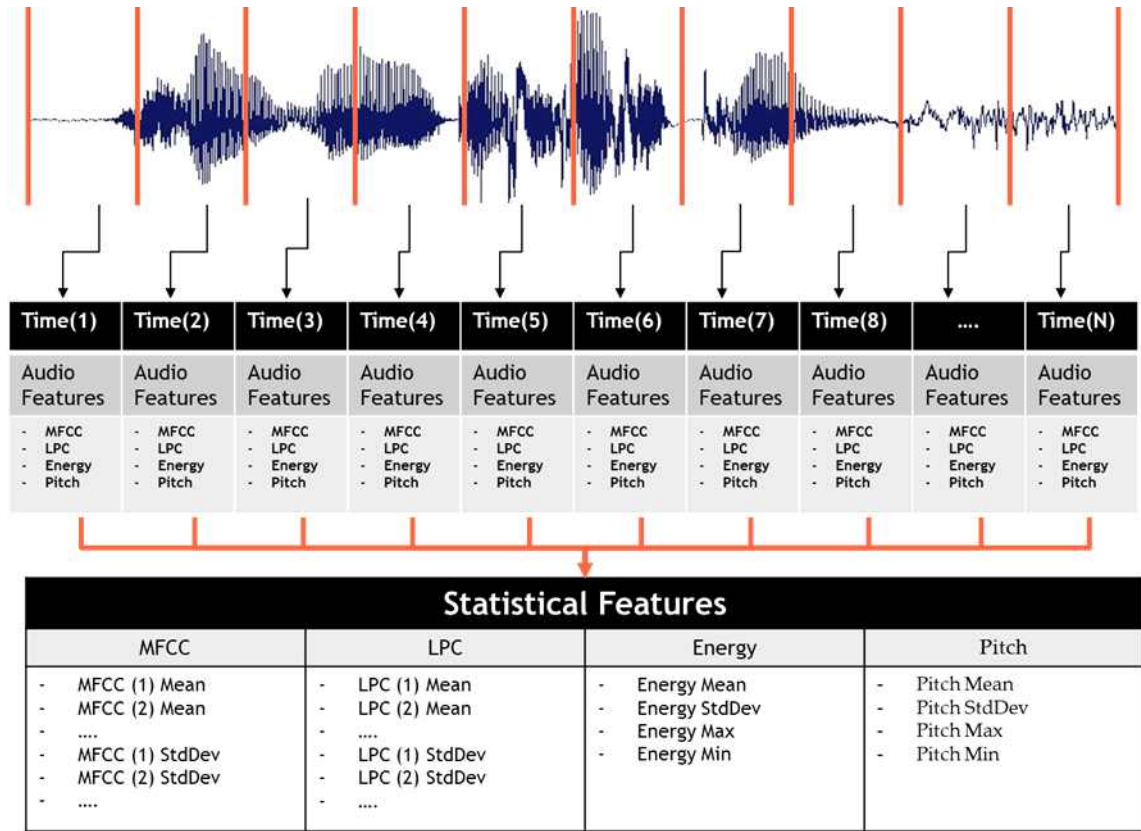


Figure 3-5. Feature Extraction Procedure

In the 100 feature schemes, The Feature Selection process is performed to optimize the feature vector. There are many feature selection methods such as Correlation Feature Selection [64], Information Gain [65], Principal Component Analysis (PCA) [66] etc. These almost feature selection algorithms optimize by reducing features. Therefore, the accuracy level also is decreased from 1% to 2% affected by reducing features. Furthermore, we do not know how deleted features affect the recognition model in personal data. Therefore, we employed the weighted correlation feature selection method [64] that assigns weights to each feature without deleting features. First, global feature extraction is performed through 10-fold cross validation using the existing public data set IEMOCAP (The Interactive Emotional Dyadic Motion Capture) dataset [67], and Merit score points of each feature are measured. IEMOCAP consists of 10 people's emotional speeches by natural conversation, which are useful for selecting more

objective features. Details of IEMOCAP is described in the Experimental section. Table 3-2 shows the correlation merit points measured using the IEMOCAP dataset.

As a result of the correlation merit point for each feature, all of the 10 LPC features have a correlation value of 0, which is a useless feature. Finally, this module uses a total of 96 features excepting four 10 LPC features and reflect weighted values for each feature in the recognition and training process. This feature selection approach can increase the accuracy a little bit. Table 3-3 are show a comparison of accuracy between original emotion recognition and applied correlation feature selection by 10-fold cross validation.

Table 3-2. Correlation Feature Selection Matrix

ID	Feature	Avg. Merit	ID	Feature	Avg. Merit
1	1 MFCC Mean	0.14	51	12 MFCC Min	0.092
2	2 MFCC Mean	0.212	52	13 MFCC Min	0.104
3	3 MFCC Mean	0.019	53	1 LPC Mean	0.081
4	4 MFCC Mean	0.026	54	2 LPC Mean	0.077
5	5 MFCC Mean	0.04	55	3 LPC Mean	0.047
6	6 MFCC Mean	0.182	56	4 LPC Mean	0.106
7	7 MFCC Mean	0.152	57	5 LPC Mean	0.105
8	8 MFCC Mean	0.153	58	6 LPC Mean	0.068
9	9 MFCC Mean	0.156	59	7 LPC Mean	0.058
10	10 MFCC Mean	0.152	60	8 LPC Mean	0.093
11	11 MFCC Mean	0.156	61	9 LPC Mean	0.08
12	12 MFCC Mean	0.162	62	10 LPC Mean	0
13	13 MFCC Mean	0.09	63	1 LPC StdDev	0.107
14	1 MFCC StdDev	0.153	64	2 LPC StdDev	0.101

15	2 MFCC StdDev	0.183	65	3 LPC StdDev	0.118
16	3 MFCC StdDev	0.187	66	4 LPC StdDev	0.169
17	4 MFCC StdDev	0.18	67	5 LPC StdDev	0.129
18	5 MFCC StdDev	0.177	68	6 LPC StdDev	0.09
19	6 MFCC StdDev	0.172	69	7 LPC StdDev	0.144
20	7 MFCC StdDev	0.034	70	8 LPC StdDev	0.068
21	8 MFCC StdDev	0.04	71	9 LPC StdDev	0.089
22	9 MFCC StdDev	0.056	72	10 LPC StdDev	0
23	10 MFCC StdDev	0.027	73	1 LPC Max	0.053
24	11 MFCC StdDev	0.136	74	2 LPC Max	0.092
25	12 MFCC StdDev	0.089	75	3 LPC Max	0.052
26	13 MFCC StdDev	0.073	76	4 LPC Max	0.107
27	1 MFCC Max	0.065	77	5 LPC Max	0.124
28	2 MFCC Max	0.069	78	6 LPC Max	0.128
29	3 MFCC Max	0.128	79	7 LPC Max	0.072
30	4 MFCC Max	0.148	80	8 LPC Max	0.124
31	5 MFCC Max	0.158	81	9 LPC Max	0.041
32	6 MFCC Max	0.163	82	10 LPC Max	0
33	7 MFCC Max	0.153	83	1 LPC Min	0.057
34	8 MFCC Max	0.218	84	2 LPC Min	0.053
35	9 MFCC Max	0.062	85	3 LPC Min	0.049
36	10 MFCC Max	0.185	86	4 LPC Min	0.105
37	11 MFCC Max	0.176	87	5 LPC Min	0.064
38	12 MFCC Max	0.16	88	6 LPC Min	0.091
39	13 MFCC Max	0.132	89	7 LPC Min	0.058
40	1 MFCC Min	0.128	90	8 LPC Min	0.07

41	2 MFCC Min	0.155	91	9 LPC Min	0.079
42	3 MFCC Min	0.141	92	10 LPC Min	0
43	4 MFCC Min	0.149	93	Energy Mean	0.113
44	5 MFCC Min	0.156	94	Energy StdDev	0.045
45	6 MFCC Min	0.145	95	Energy Max	0.092
46	7 MFCC Min	0.079	96	Energy Min	0.132
47	8 MFCC Min	0.14	97	Pitch Mean	0.035
48	9 MFCC Min	0.157	98	Pitch StdDev	0.043
49	10 MFCC Min	0.144	99	Pitch Max	0.099
50	11 MFCC Min	0.07	100	Pitch Min	0.014

Table 3-3. The Comparison of accuracy between original emotion recognition (Left) and applied correlation feature selection (Right) - IEMOCAP (Unit %)

J48 (without feature selection)					J48 (applied feature selection)				
class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
Ang.	44.22	10.82	20.05	24.92	Ang.	44.11	11.16	19.99	24.75
Sad.	14.15	39.22	16.84	29.79	Sad.	14.45	39.75	16.54	29.27
Hap.	23.48	15.97	27.67	32.88	Hap.	23.00	16.24	28.62	32.14
Neu.	20.04	16.89	21.24	41.83	Neu.	19.91	16.76	21.71	41.62
SMO (without feature selection)					SMO (applied feature selection)				
class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
Ang.	59.34	5.49	5.66	29.50	Ang.	59.34	5.49	5.66	29.50
Sad.	8.53	51.12	2.92	37.43	Sad.	8.53	50.97	2.99	37.50
Hap.	21.92	11.10	21.58	45.40	Hap.	21.85	11.10	21.72	45.33
Neu.	12.24	13.82	6.57	67.38	Neu.	12.20	13.82	6.57	67.42
Random Forest (without feature selection)					Random Forest (applied feature selection)				
class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
Ang.	58.15	4.36	3.51	33.98	Ang.	58.61	4.42	2.94	34.03
Sad.	8.31	44.84	3.22	43.64	Sad.	8.16	45.21	3.22	43.41
Hap.	23.61	9.61	15.63	51.15	Hap.	22.67	8.80	16.78	51.76
Neu.	9.81	10.11	4.69	75.39	Neu.	10.15	9.55	5.07	75.22

③ Similar data selection based on Maximum Threshold Distance

This module reinforces the insufficient target emotional samples from initial constructed multiple user speech dataset when the acquired target user speech samples in particular emotional label is not enough to train. Regarding reinforcement of target user training dataset from other users, the overall labeled dataset in multiple user speech dataset is transformed into an unlabeled statement. Then measure the distance from the extracted feature vectors through module 1 and 2 from not only labeled target user speeches but also unlabeled multiple user speeches. The distance between the unlabeled speech data and the mean value of the acquired target user speech is calculated to measure the similarity. Then, the training dataset is reinforced with the speech that has the most similarities.

④ Other similar user emotional speech mapping based on Data Distribution Factor

This module replaces the dataset of empty target emotional samples through similar user emotional speeches from other user speech dataset when some particular emotional label samples were never acquired from target user. Regarding the similar user emotional speech selection from other users, the distance is measured on each emotional category between target user and other user through data distribution factors such as median, variance, skewness, and kurtosis for the target user as well as every user in the initial constructed multiple user dataset. Then, the most similar emotion data among the other users is copied to the empty target user emotional label dataset based on the distance from the distribution factors.

⑤ Virtual Case Data Augmentation based on SMOTE

Using the SMOTE algorithm [68] is an efficient way to reinforce and augment different speech cases. SMOTE is a well-known over-sampling technique that can

resolve the imbalanced data problem where particular class is biased. The SMOTE method reduces the gap in the number of samples compared to the majority and minority classes by augmenting the samples of the minority class. However, the main limitation of this method is the cold-start problem, in which there is no accurate data generated when the initial input data are prime numbers. The reason is that SMOTE generates the random data in the nearest boundary of acquired data [69]. In small amount of data, the boundary area is narrowed. Therefore, it can fall into overfitting problem and show low accuracy with the new input data. To solve this problem, it is important to acquire enough initial samples before oversampling. Therefore, This module builds the final dataset by reinforcing the virtual dataset using the SMOTE algorithm, based on the selected sufficient real-case dataset from 3 and 4 module.

⑥ Model Creation and Classification (Model Creation)

This module creates a training model based on the generated dataset from module 3 to 5 and then classifies emotions from a new speech input from the target user. Accuracy in all recognition research domains affects the selection of machine learning algorithms as well as feature extraction. The selection of machine learning algorithms is also important in the field of emotion recognition because it greatly affects recognition accuracy depending on which machine learning algorithm is selected. Therefore, in this module, utilize the best performance classifier based on the machine learning comparison evaluation results.

3.2. Similar data selection based on Maximum Threshold Distance

The target user speech data is not always acquired in sufficient amount to create the personalized emotion recognition model. Especially, the target user emotional samples are collected in prime number in initial stage of personalized emotional speech acquisition. If the personalized model is trained in prime number of target user emotional speech, we cannot achieve high performance on new input data due to the lack of real case data. The proposed method can solve to overcome the insufficient data problems by adding the similar emotional speech of other users to the training dataset of the personalized model.

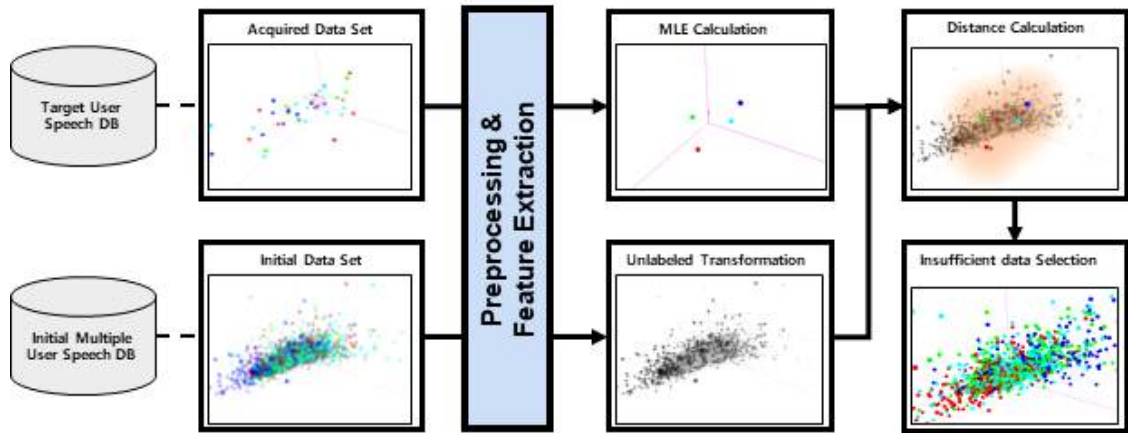


Figure 3-6. Similar data selection based on Maximum Threshold Distance

This section introduces the proposed technique to reinforce insufficient emotional speech of the target users. To increase the amount of insufficient target user emotional speech, the dataset is selected based on the similarity between the target user speech and the multiple-user speech. Figure 3-5 shows the process of Insufficient Data Reinforcement.

For the similarity calculation between the target user speech and the multiple-user speech, preprocessing and a feature extraction process are performed first. Then, the target user dataset is separated into different emotion classes and the MLE value of each feature is obtained for each emotion. The distance between the speech relative to the initial multiple-user speech database is calculated and the target user MLE values are obtained. Among this process, the labeled data in the initial multiple-user speech database are transformed into unlabeled data. This means the label information is ignored in multiple-user speech database.

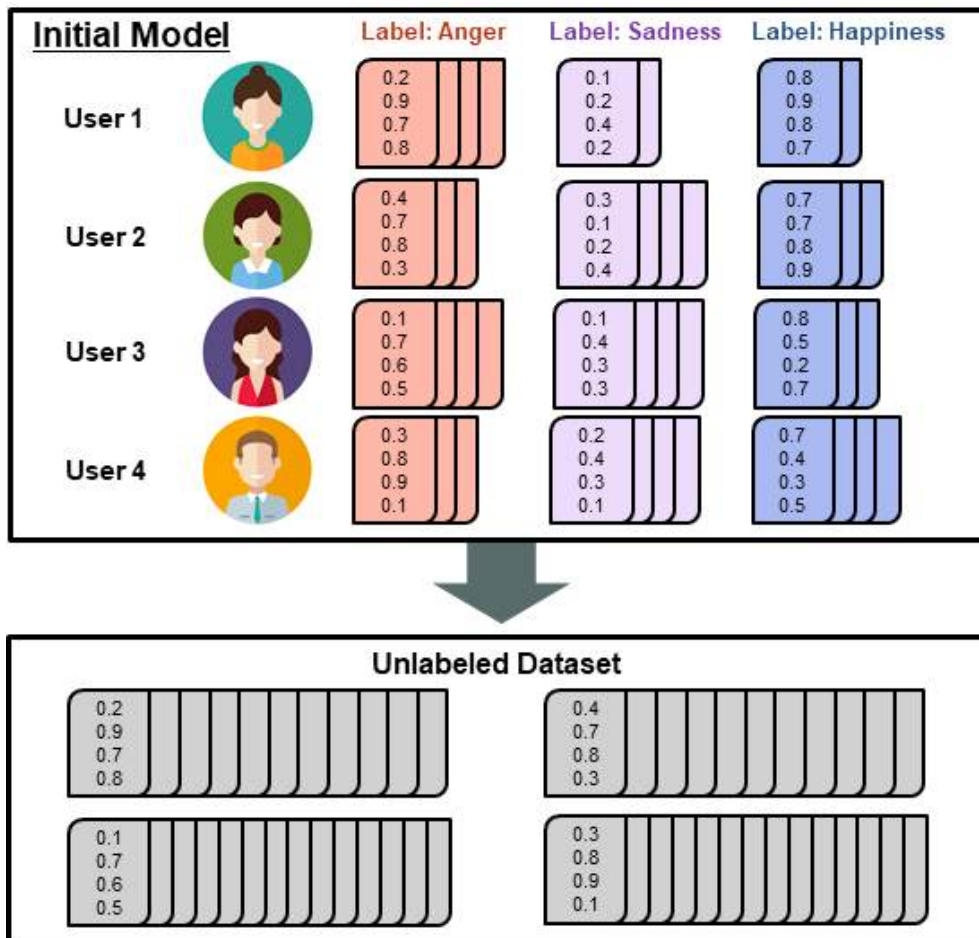


Figure 3-7. Unlabeled transformation

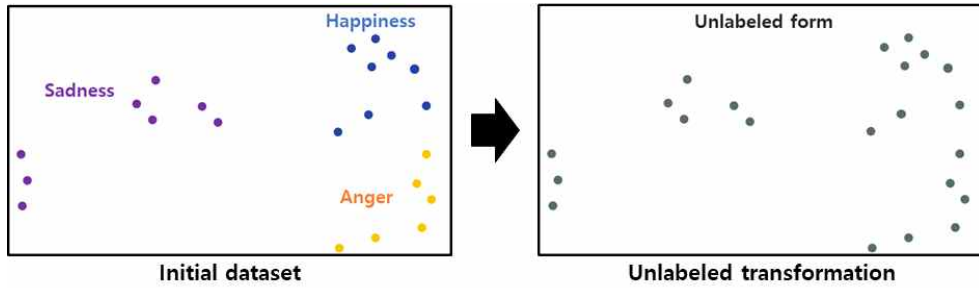


Figure 3-7. Unlabeled transformation in feature space

The reason for using an unlabeled transformation is that emotional expressions are different for each user. For example, if the target user's anger speech pattern is similar to the happiness pattern from the multiple-user speech database, the system classifies the target user's anger as happiness. This means that the target user's particular emotional speech can be similar to different emotional speech in other user emotional speech when the acoustic pattern is almost same. Therefore, this method ignore the labeled information in the multiple-user speech database when reinforcing the target user training dataset with other user similar speech.

Then, the speech samples from the user closest to the target speech MLE value are selected. After that, selected unlabeled data of other users are map to the most similar target user emotional label and add the target user training data set.

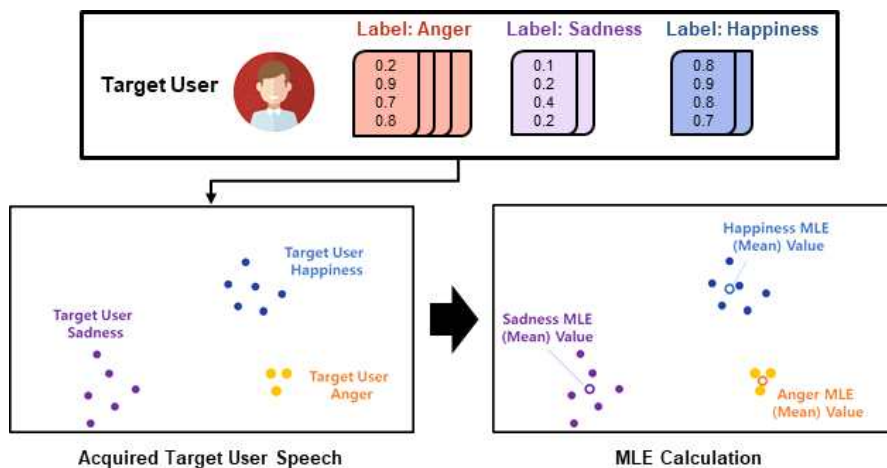


Figure 3-9. MLE value calculation based on target user data

There are various techniques such as Euclidean Distance [70], Log-likelihood Distance [33] and Jaccard Distance [71]. Although the distance measurement method and the calculation result of each technique are different, the distance ranking between the target data and the existing model data does not change. Therefore, in this thesis, the distance measurement method is employed as the method with the lowest complexity and time cost in the calculation of the above three distance measurement methods. Table 3-4 shows the distance measurement process time cost between each data between target data using IEMOCAP dataset and shows the extracted data Index ranking as an example. The time cost was measured using about 6000 data by 10 users for only one target emotion. And the lowest distance ranking index shows the example of user-1 related data.

Table 3-4. Distance measurement performance

Distance Measurement	Description	Time cost	Example of lowest distance ranking (User 1)
Euclidean Distance	The Euclidean distance is the straight-line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space.	53ms	Total 235 3158, 3428, 1064, 733, 6778, 6384, 2460, 5180, 2016, 4630, 1063, 2048, 6870, 2991, 1399, ... 6557, 840, 2359, 3397, 5843
Log-Likelihood Distance	Log-Likelihood Distance is measured the distance by N-best result from log-likelihood. This method can observe large value for a similar small distance	123ms	Total 235 3158, 3428, 1064, 733, 6778, 6384, 2460, 5180, 2016, 4630, 1063, 2048, 6870, 2991, 1399, ... 6557, 840, 2359, 3397, 5843
Jaccard Distance	Jaccard distance considers two objects to be compared as sets of characteristics. Basic concepts or notations are based on set theory.	83ms	Total 235 3158, 3428, 1064, 733, 6778, 6384, 2460, 5180, 2016, 4630, 1063, 2048, 6870, 2991, 1399, ... 6557, 840, 2359, 3397, 5843

Based on the results in Table 3-4, the distance is measured using a Euclidean Distance measurement between the target user's mean feature vector and each of the other user's feature vectors, which is then used to determine the similarity. The following Equations provide distance measurements.

$$TMLE_{ei} = \frac{1}{N} \sum_{j=1}^N TfeatureVector_{ji} \quad (7)$$

$$d(means_{ei}, IDS_m) = \sqrt{\sum_{i=1}^{FN} (means_{ei} - IDS_i)^2} \quad (8)$$

$$MTD(TMLE_{ei}) = \frac{1}{2} \text{argmax}(d(TMLE_{ei}, TMLE_{ej}, \dots)) \quad (9)$$

In Equation 7, is a two-dimensional array that stores the average value of the acquired target user emotion voice feature vectors, where is the corresponding emotion index, is the index of the feature vector, N is the number of data is the index of the data, and is the extracted statistical speech feature vector via the feature extraction module. In Equation 8, represents the distance between two vectors, where is the index of the initial multiple-user speech and is the initial dataset consisting of multiple users. Equation 7 is performed independently for each emotional label of the acquired target user, and Equation 8 is performed based on the results of Equation 7. In the case of the initial dataset in Equation 8, all of the data are retrieved regardless of the label, and then the distance is calculated for each emotion. Then, the maximum threshold value for data that is relevant to the user is specified through Equation 9.

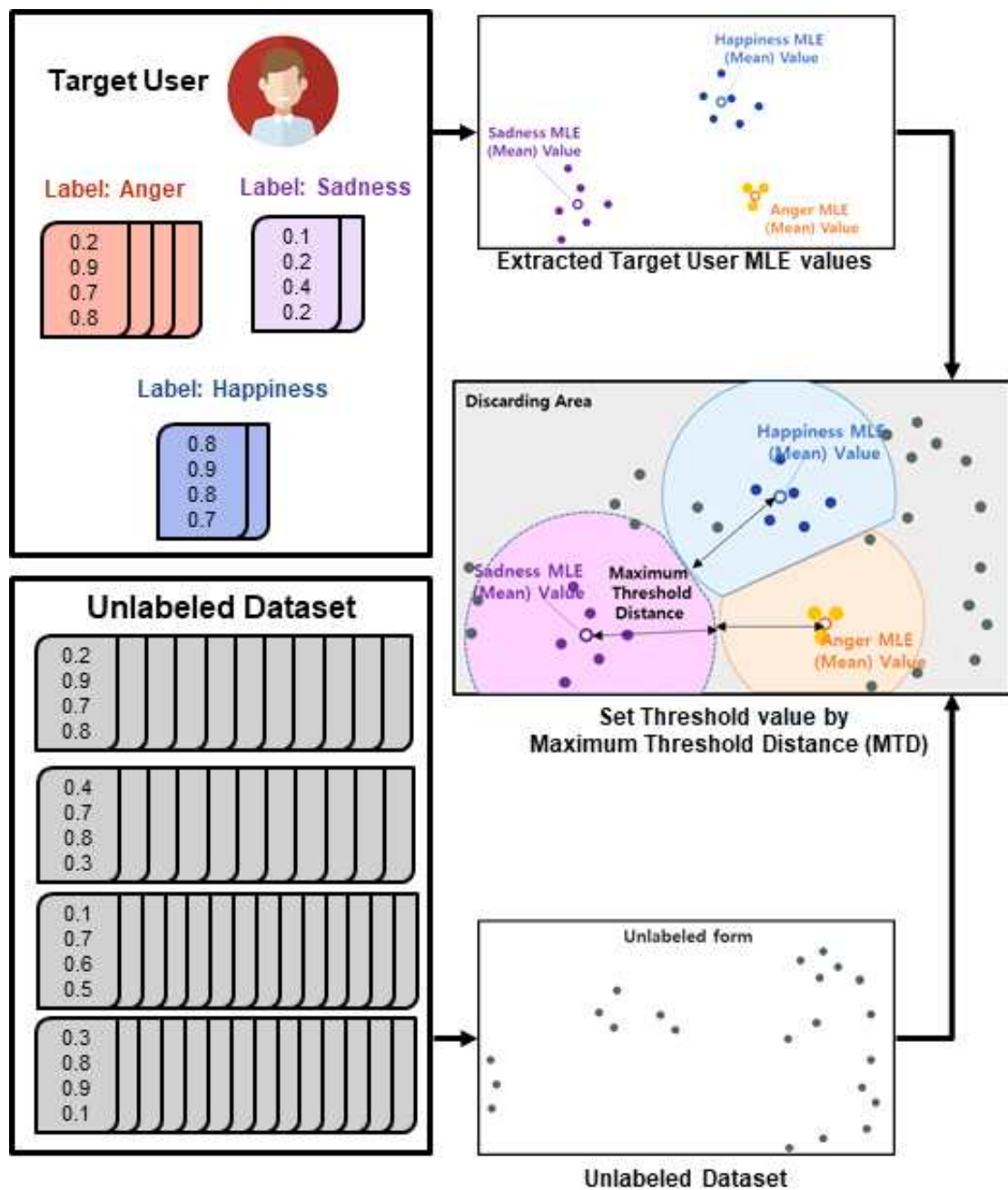


Figure 3-10. MTD based Similar Data Selection Process

Finally, the process of sequentially selecting similar data to reinforce the insufficient data according to distance is performed via the following figure 3-7 and 3-8.

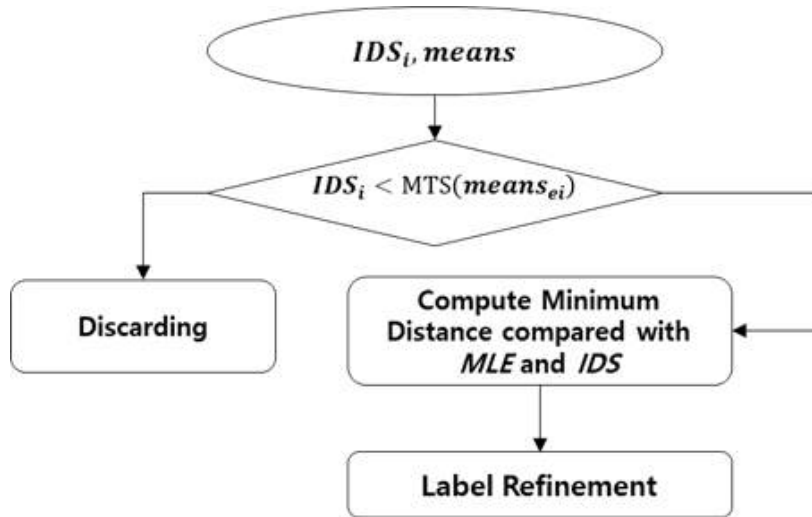


Figure 3-11. Similar Data Selection Algorithm

Proposed Algorithm 1. Similar speech data selection based on Maximum Threshold Distance

Input: $TDS(1 \dots N)$ – Target User Dataset

$IDS(1 \dots M)$ – Initial Multiple User Dataset

Output: $S(1 \dots K)$ – Selected Similar Emotional Speeches Dataset

$TMLE_e = \text{Calculate MLE } (TDS_e)$

$MTD_e = \text{Calculate MTD } (TDS_e, TMLE_e)$

for $i = 1$ to M

$Distance = \text{Calculate Euclidean Distance } (IDS_i)$

 if $Distance \leq MTD_e$ then

$mEmo = \text{Calculate Minimum Distance } (TMLE_e, IDS_i)$

 add $S(IDS_i, mEmo)$

 end

end

end

Return S

Figure 3-12. The Proposed Algorithm of Similar Speech Data Selection based on Maximum Threshold Distance

Figure 3-9 present the output of the this proposed data selection algorithm. Compared with existing method [33], we can see that the proposed technique selects more target user related data.

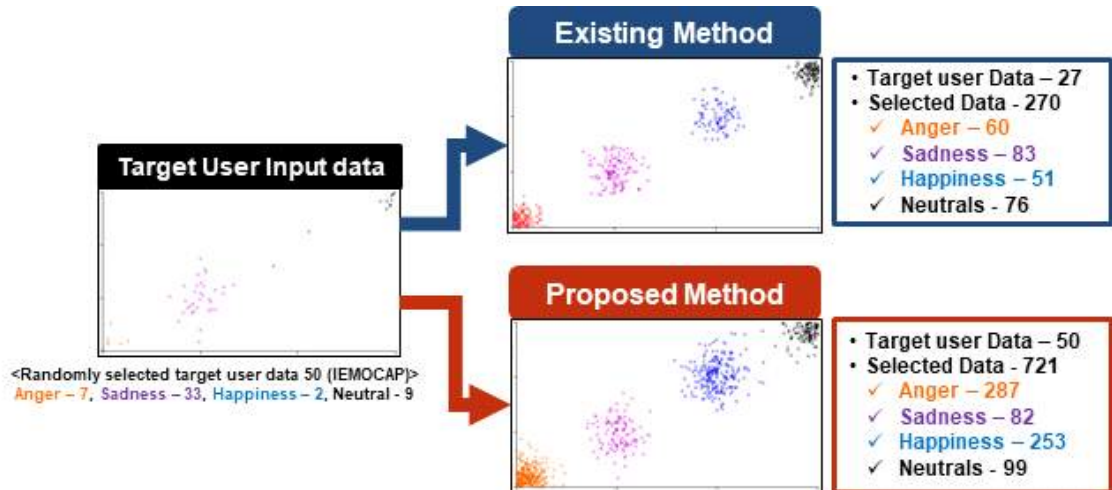


Figure 3-13. Output comparison of the proposed method and existing method (LDT-MDT- MLLR [33])

3.3. Other similar user emotional speech mapping based on Data Distribution Factor

Normally, humans do not express different emotions at the same rates in daily life [72]. If the target user does not express a particular emotion for a long time, the training model will be created without any samples for that particular emotional speech. In this case, this particular emotion is not recognized by the system and the accuracy is 0%. We can assume that the target user's absent emotion data will be similar to that of another user's emotional speech if they have a similar speech pattern. Therefore, I propose the reinforcement method to replace absent target user emotion data to similar user's emotional speech based on this assumption.

This section introduces the proposed method to reinforce data that is not collected from the target user's particular emotional speech. The proposed method selects the user most similar with the target user from among the emotional speech data of multiple users, and then selects the speech from this similar user. Then calculates the distribution similarity based on the speech of each user's training dataset and selects the most similar user relative to the acquired target user. Finally, this particular absent emotion data will be reinforced regarding the target user's training dataset considering its similarity with the other user's emotion speech data.

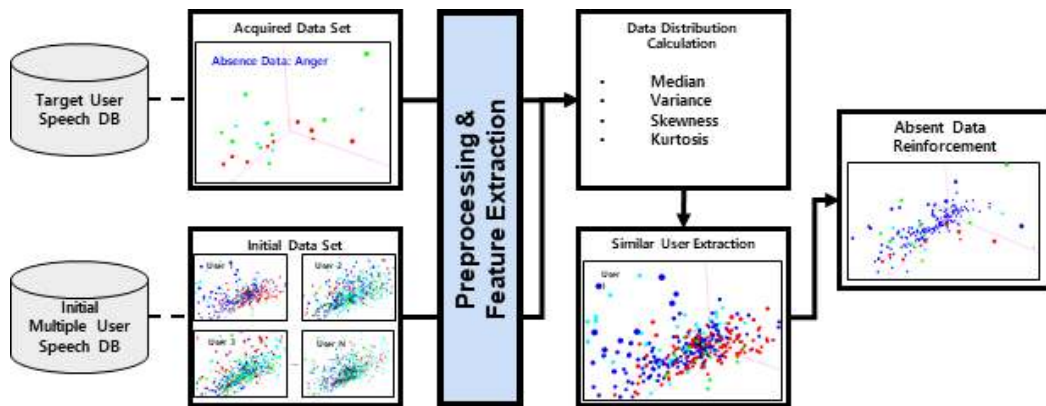


Figure 3-14. Absent Emotion Data Reinforcement Workflow

We compute the statistical distribution, including the median, variance, skewness, and kurtosis, from the speech data of both the target user and the other users considering the speech feature vectors extracted in step 2 of Section 2. Then, the similarity degree between the target user and the other users is calculated. The similarity calculation procedure is the data of the user with the highest the Euclidean Distance Similarity distance value of data distribution factors of each user. The contents of the speech feature vector distribution to be considered are as follows.

- **Average** – The average is a single number taken as representative of a list of numbers [73]. The value is often used as an estimate of a central tendency such as a mean. The calculation procedure is shown in Equation 10.

$$Average(x) = \frac{1}{N} \sum_i^N x_i \quad (10)$$

- **Median** - The median is the value separating the higher half from the lower half of a data sample (a population or a probability distribution). For a data set, it may be thought of as the "middle" value [74]. The calculation procedure is shown in Equation 11.

$$Median (SortedFeatueValues) = SortedFeatueValues_{N/2} \quad (11)$$

- **Variance** - It is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value [75]. The calculation procedure is shown in Equation 12.

$$Variance (X) = \frac{1}{N} \sum_{i=1}^N (x_i - m)^2 \quad (12)$$

- **Standard Deviation** - The standard deviation is a measure of the scattering of the data, defined as the square root of the amount of variance [76]. The smaller the standard deviation, the closer the distance of the variables from the mean value. Statistics and probabilities mainly represent distribution of probability, random variable or measured population or redundancy set. The calculation procedure is shown in Equation 13.

$$\text{Standard Deviation } (x) = \sqrt{\text{Variance } (X)} \quad (13)$$

- **Skewness** - It is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined [77]. The calculation procedure is shown in Equation 14.

$$\text{Skewness } (x) = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - m)^3}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - m)^2 \right)^{\frac{3}{2}}} \quad (14)$$

- **Kurtosis** - It is a measure of the "tailedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population [78]. The calculation procedure is shown in Equation 15.

$$\text{Kurtosis } (x) = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - m)^4}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - m)^2 \right)^2} - 3 \quad (15)$$

- **Maximum** - The maximum is the extreme value in a data set [79]. The maximum value should be the largest value within a given range and is useful to measure the extent of the feature vector data distribution. The calculation procedure is shown in Equation 16.

$$\textit{Maximum}(\textit{sortedFeatureValues}) = \textit{sortedFeatureValues}_1 \quad (15)$$

- **Minimum** - The minimum is the lowest value in a data set [79]. The minimum value should be the smallest value within a given range with maximum value and is useful to measure the extent of the feature vector data distribution. The calculation procedure is shown in Equation 16.

$$\textit{Minimum}(\textit{sortedFeatureValues}) = \textit{sortedFeatureValues}_N \quad (16)$$

Finally, compute the Euclidean Distance Similarity based on data distribution factors for each user to estimate similar user. Then we select the most similar user's real-case data for absent area. The calculation procedure of distance similarity is shown in Equation 17 and 18.

$$d(p_1, p_2) = \sqrt{\sum_{i \in \text{item}} (s_{p_1} - s_{p_2})^2} \quad (17)$$

$$\textit{Similarity} = \frac{1}{1 + d(p_1, p_2)} \quad (18)$$

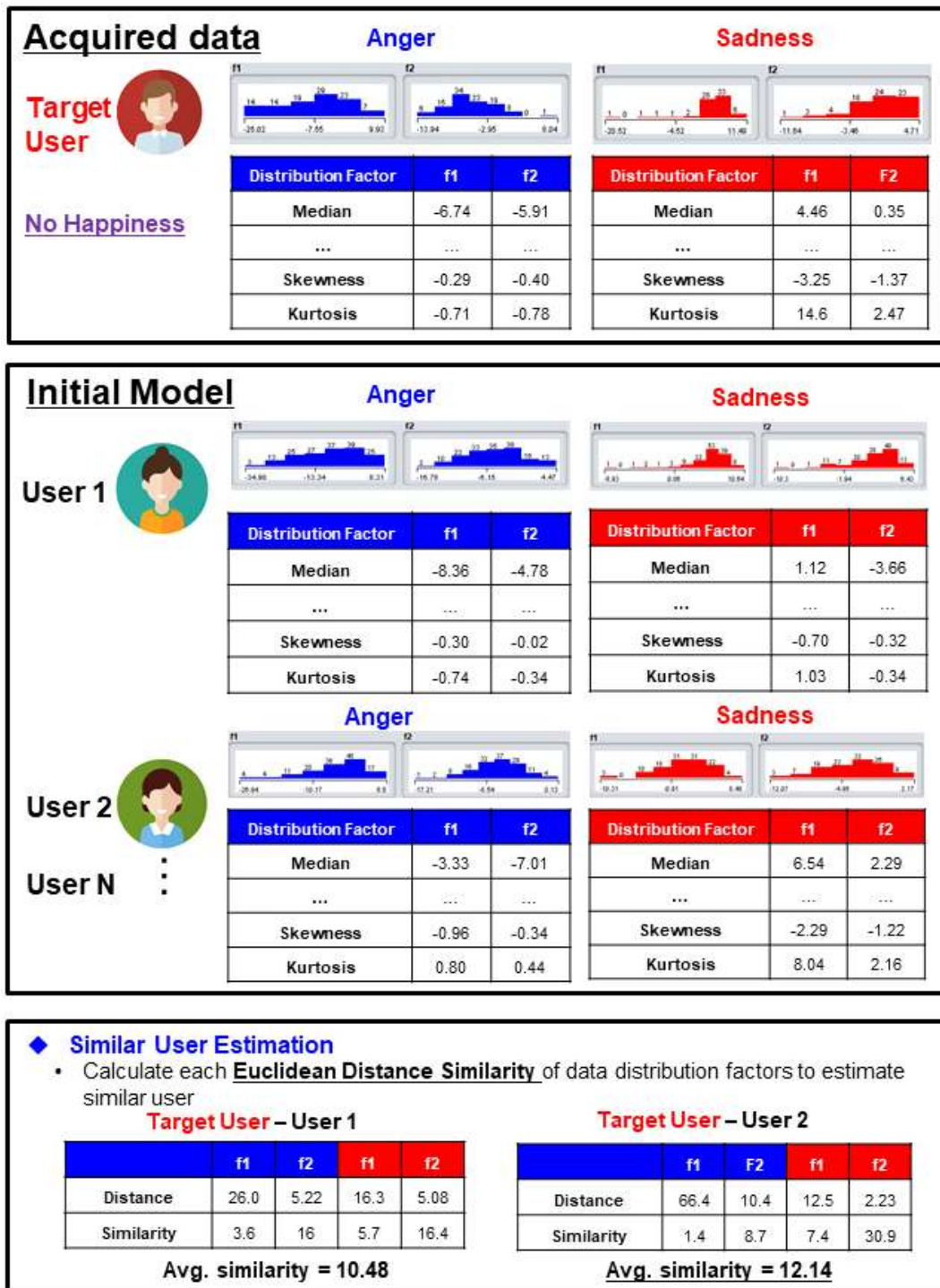


Figure 3-15. Other similar user emotional speech mapping based on Data Distribution Factor

Figure 3-16 present the output of the this proposed data selection algorithm. Compared with existing method [33], By comparison with existing techniques, we can see that the proposed technique selects more target user related data and solves the imbalanced problem to some extent.

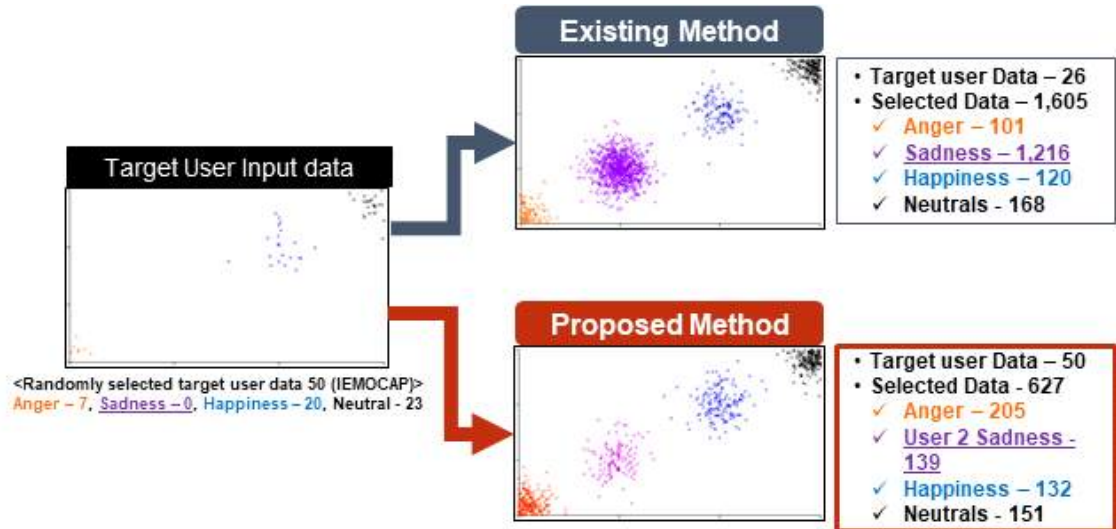


Figure 3-16. Output comparison of the proposed method and existing method (LDT-MDT- MLLR [33])

3.4. Virtual Case Data Augmentation based on SMOTE

SMOTE is the method used to generate the dataset for a minority number of particular class samples in the classification model. At first, SMOTE finds the K nearest neighbors of the minor class samples and finds the difference between the current sample and these K neighbors. This difference is multiplied by a random value between 0 and 1 and is then added to both the training data as well as the original sample. The SMOTE algorithm increases the number of minority classes, which has the smallest number of samples, repeating this several times until the numbers of samples for all classes are balanced. In addition, this algorithm reinforces untrained case data by oversampling this data virtually. This method increases the recognition accuracy of the new input data.

However, the cold-start problem, in which the mis-recognition rate increases during the initial stage, occurs due to the generation of limited ranges of oversampled data, which itself occurs when the number of acquired sample data is too small and thus cannot generate accurate samples for the absent emotion data for SMOTE. The cold-start problem of SMOTE can be solved using the dataset extracted from the proposed real-case data selection process. Then, if the data are amplified using SMOTE, the accuracy can be improved even at the initial stage. Therefore, the final training dataset is constructed by reinforcing the virtual case data using the SMOTE algorithm for the training dataset, which is selected via the data reinforcement technique.

This method performed by following algorithm such as Figure 3-16 and 3-17. And Figure 3-18 show the selected final data compare with existing work and proposed Robust Speaker Adaptation.

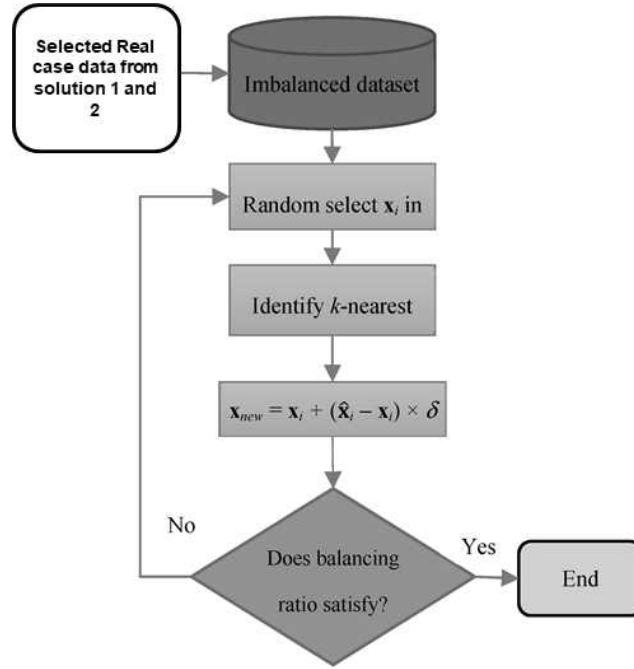


Figure 3-17. The Flowchart of SMOTE Algorithm

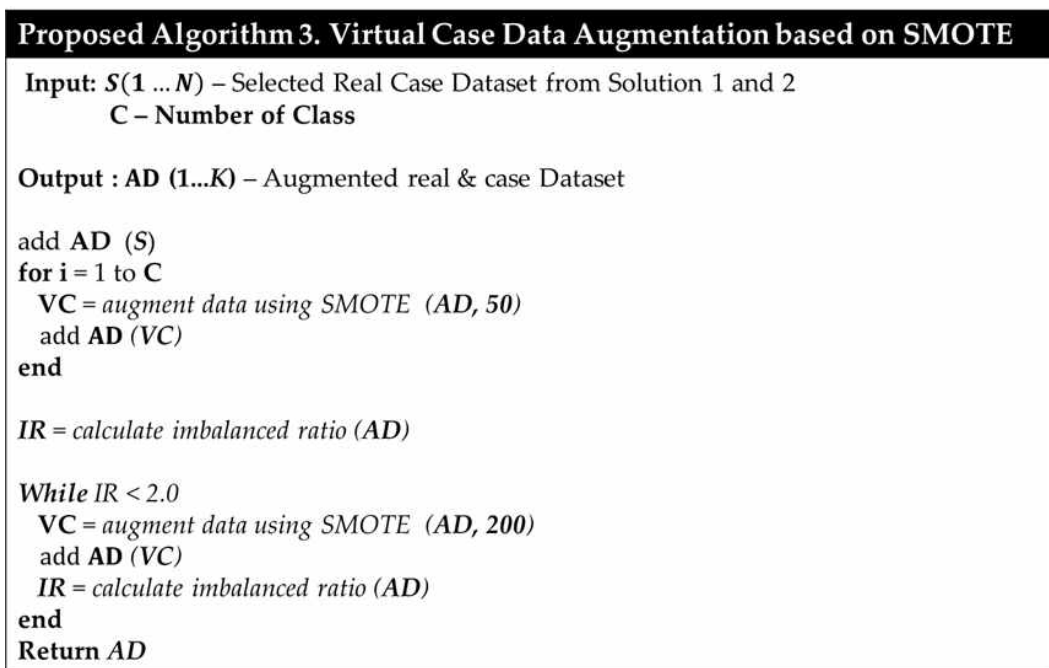


Figure 3-18. The Proposed Algorithm of Virtual Case Data Augmentation based on SMOTE

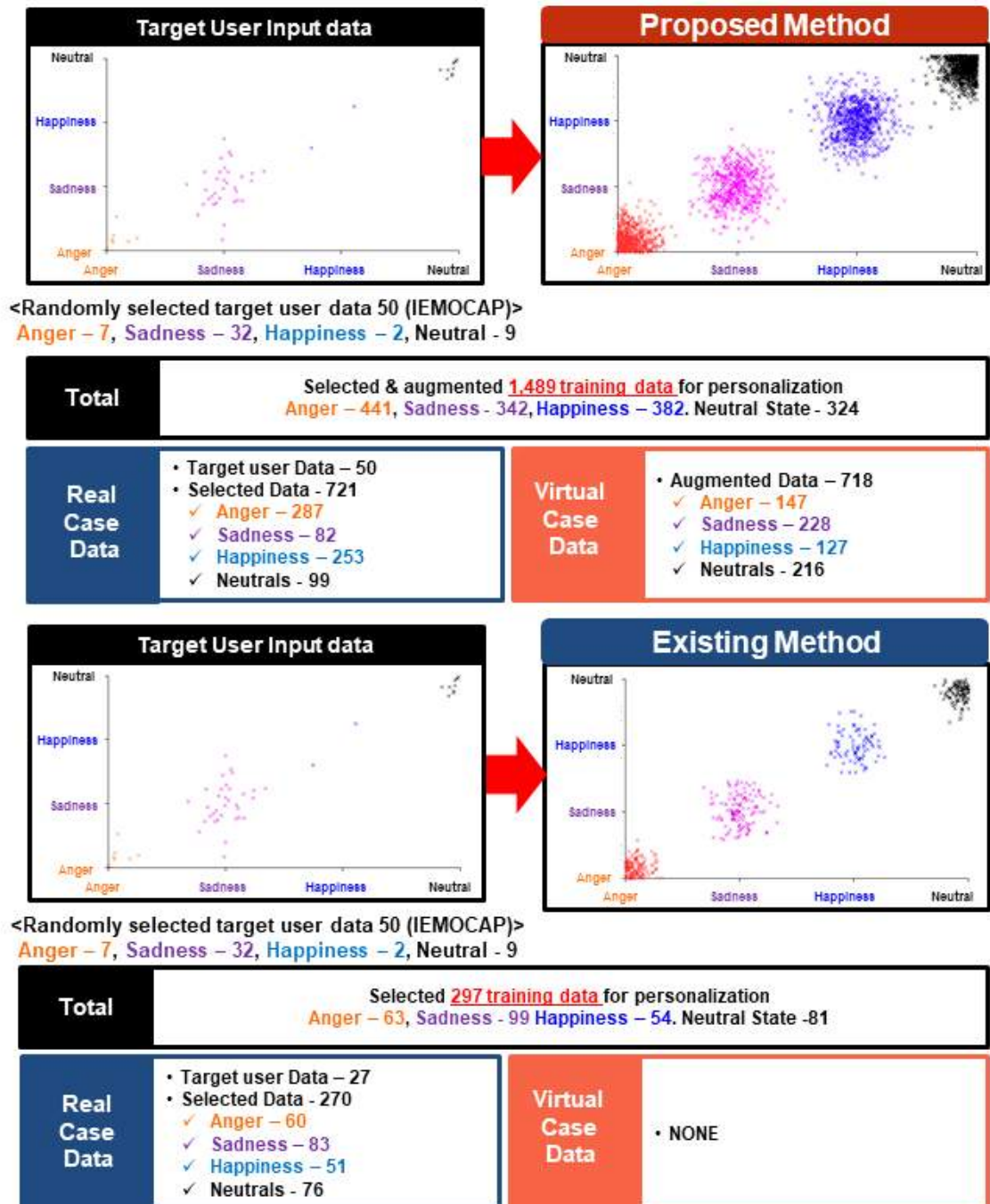


Figure 3-19. Final output comparison of the proposed method and existing method (LDT-MDT- MLLR [33])

3.5. Model Creation and Classification

In this section, we generate the training model using common classification techniques. Choosing an appropriate classifier is important for creating a training model in speech emotion recognition. Machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forest have unique characteristics when generating and recognizing training models.

In this paper, we use a random forest classification algorithm [80] to perform training model generation and recognition. This random forest algorithm was first introduced to mitigate the disadvantages of overfitting and instability common among decision trees. A random forest is a method of creating a single model by combining multiple decision trees. Multiple trees are created by applying randomness to observations and variables. This process generates N bootstrap samples, N trees with arbitrary bootstrap samples and variables, and an ensemble training classifier, which has the advantage of excellent prediction and high stability [81]. Therefore, this classifier is an effective algorithm for speech-based emotion recognition, which can build a reliable training model with few data.

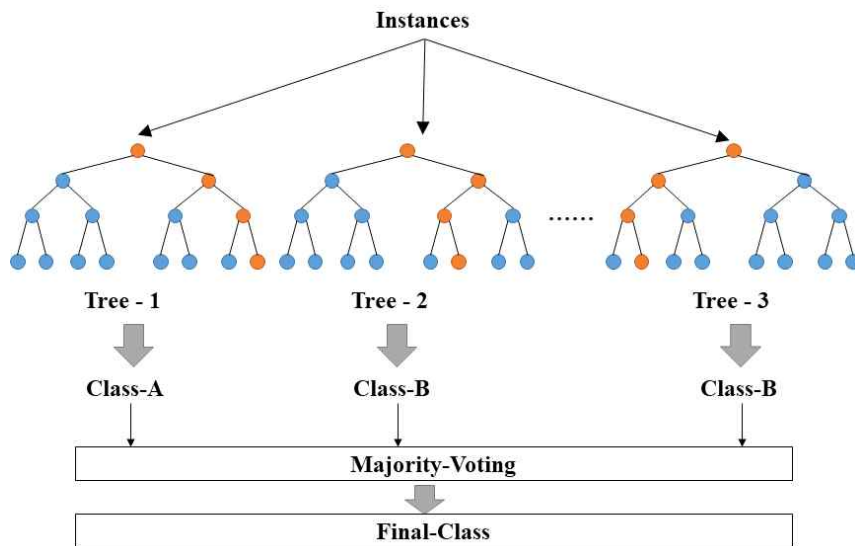


Figure 3-20. Simplified Random Forest

In this section, evaluate the performance of the proposed Robust Speaker Adaptation framework to solve the cold-start problem. Therefore, we proposed new evaluation methodologies to evaluate the cold-start problem. First of all, we selected a dataset with a lot of personal voice data. And make a similar real-world data acquisition environment through randomly training data selection for each user. Finally, this experiment was repeated 10 times. And we also performed cross corpus evaluation for validation of proposed method.

4.1. Experimental Setup

In this work, performed the experiment using Interactive Emotional Dyadic Motion Capture (IEMOCAP) [67], which is a public emotion speech dataset. The IEMOCAP dataset has an extremely large number of data compared to other similar datasets consisting of various speech patterns from real environments. In other recent studies, the 5-fold Cross Validation technique with the four emotions of Anger, Sadness, Happiness, and Neutral has shown a low accuracy of about 60%, which has been challenging to overcome [82-83]. Therefore, the IEMOCAP dataset was selected for our experimental dataset, for which individual datasets are sufficient and clearly exhibit accuracy improvements. In our experimental method, the accuracy of the personalization model generation was calculated by randomly selecting training data and test data from the target user and increasing the number of training data. And we also select the CRowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [84] for cross domain evaluation in this experiment.

4.1.1 Dataset

The purpose of the experiment in this paper is to verify the performance of the personalized emotion recognition model creation method. The proposed method uses the existing SI model when the target user's data is 0. Since the user data is collected more than once, the training model is rapidly changed by retraining process using proposed Robust Speaker Adaptation (Robust Speaker Adaptation). In order to verify the performance of this technique, the number of personalized data must be enough to be able to train and test.

In speech emotion recognition area, there are many well organized open dataset such as eNTEFACE [85], EMO-DB [86], Surrey Audio-Visual Expressed Emotion (SAVEE) Database [87], The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [88] and Crowd-sourced emotional multimodal actors dataset (CREMA-D) [84]. These databases consist of hundreds to thousands of samples. Most of the existing SI studies used k-fold cross validation when evaluating their algorithm. It means they utilize all of data fully to train and test. However, our approach can verify the utilization of individual target user dataset only to train and test. Therefore, for accurate evaluation, we have required a large amount of individual emotional speech data. Table 4-1 shows the representation of existing speech database organization. Existing databases have insufficient amount of individual emotion data such as about 20. These environments have limited choice of user training data and test data, making it difficult to conduct accurate indirect comparison experiment. Finally, we have selected IEMOCAP which has the largest number of total samples about 100 emotional samples of each emotion per person.

Table 4-1. Organization of Existing Emotional Speech Database

Emotional Database	Total Samples	Emotions	Speakers	Avg. Samples per person	Avg. Samples of Each Emotion per Person
Emo-DB	535	7	10	53.5	7.6
eINTERFACE	1,166	6	42	27	4.5
SAVEE	480	8	4	120	15
RAVDESS	1,440	8	24	60	7.5
CREMA-D	7,442	6	91	81.7	13.61
IEMOCAP	10,038	10	10	1003.8	100.3

The IEMOCAP dataset is composed of 10,038 corpus samples with 10 labels (Neutral, Frustration, Anger, Sadness, Happiness, Excited, Other, Surprise, Fear, and Disgust), which are speech data continually collected through a script. Each sample from the IEMOCAP dataset is annotated with multiple labels from many audiences. We chose a representative label through voting. However, the dataset contains ambiguous emotions such as Excited and Frustration. Further, the number of data among Surprise, Fear, Disgust, and Other is too small. Therefore, it is difficult to conduct precise experiments when the data is divided into training and test datasets. Table 4-2 shows the original IEMOCAP dataset structure.

Table 4-2. Original IEMOCAP Dataset Structure

Emotion	Number of Samples	Rate
Anger	1,229	12.24%
Sadness	1,182	11.78%
Happiness	495	4.93%
Neutral	575	5.73%
Excited	2,505	24.96%
Surprise	24	0.24%
Fear	135	1.34%
Disgust	4	0.03%
Frustration	3,830	38.16%
Other	59	0.59%
Total	10,038	100%

Therefore, we transformed the data for the Excited and Frustration emotion labels to other annotated emotion labels that these labels are ambiguous and have high composition ratio in the dataset. We did this by selecting the second most voted label from the IEMOCAP dataset. In addition, We use four emotions based on Valence & Arousal model [89] without using all ten emotions. In most emotion recognition studies using Valence Arousal Model, four emotions as Angry, Happy, Calm(Neutral), Sad are used as cognitive measures by deleting ambiguous emotions such as Afraid, Excited,

Content and Depressed. Therefore, we conducted experiments using data for only four basic emotions which is normally used in speech emotion recognition [90-91]: Neutral, Anger, Sadness, and Happiness.

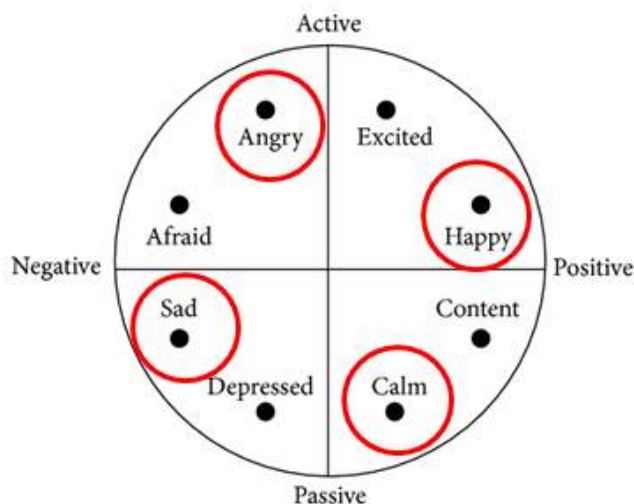


Figure 4-1. Valence-arousal Dimensional Model

Table 4-3 shows the number and ratio of refined data and Fig. 4-1 shows the number of user-specific samples.

Table 4-3. Refined IEMOCAP Dataset Organization

Emotion	Number of Samples	Rate
Anger	1,766	25.51%
Sadness	1,336	19.29%
Happiness	1,478	21.34%
Neutral	2,345	33.86%
Total	6,925	100%

Subject 1		Subject 2		Subject 3		Subject 4		Subject 5	
Emotion	Total	Emotion	Total	Emotion	Total	Emotion	Total	Emotion	Total
Anger	131	Anger	169	Anger	140	Anger	138	Anger	166
Sadness	120	Sadness	139	Sadness	119	Sadness	163	Sadness	143
Happiness	84	Happiness	23	Happiness	106	Happiness	148	Happiness	154
Neutral	96	Neutral	48	Neutral	268	Neutral	256	Neutral	272

Subject 6		Subject 7		Subject 8		Subject 9		Subject 10	
Emotion	Total	Emotion	Total	Emotion	Total	Emotion	Total	Emotion	Total
Anger	229	Anger	263	Anger	201	Anger	159	Anger	170
Sadness	178	Sadness	124	Sadness	95	Sadness	136	Sadness	119
Happiness	183	Happiness	247	Happiness	143	Happiness	233	Happiness	157
Neutral	296	Neutral	227	Neutral	274	Neutral	321	Neutral	287

Figure 4-2. Refined IEMOCAP Dataset Represented by Each User

The CREMA-D selected for the cross-corpus evaluation is constructed with 91 different user speeches and is suitable for generating as an initial model. This dataset was recorded a common method that every user (actors) collect training data speech from a given script. And this dataset has a high diversity including various ages.

Table 4-4. CREMA-D Actors' Age Distribution

Age	# actors
20-29 YRS	25.51%
30-39 YRS	19.29%
40-49 YRS	21.34%
50-59 YRS	33.86%
60-69 YRS	5
over 70 YRS	1
Total	100%

This dataset is targeted the 6 kinds of emotions such as happy, sad, anger, fear, disgust and neutral. This corpus consist of 12 sentences for each rendered in all of the emotional states. This dataset was recorded by directed to express the first sentence in three levels of intensity: low, medium, and high. For the other remaining 11 sentences, the intensity level was unspecified.

The semantic content of all 12 sentences was rated as emotionally neutral in a prior study [92]. The 12 sentences is presented in Table 4-5.

Table 4-5. CREMA-D sematic contents

Description
<ul style="list-style-type: none"> ● It's 11 o'clock. ● That is exactly what happened. ● I'm on my way to the meeting. ● I wonder what this is about. ● The airplane is almost full. ● Maybe tomorrow it will be cold. ● I would like a new alarm clock. ● I think I have a doctor's appointment. ● Don't forget a jacket. ● I think I've seen this before. ● The surface is slick. ● We'll stop in a couple of minutes.

The CREAM-D is organized by emotionally balanced manner. Table 4-6 present the organization of CREMA-D of selected only 4 kinds basic emotions such as anger, sadness, happiness, neutral.

Table 4-6. Organization of CREMA-D

Emotion	Number of Samples	Rate
Anger	1271	25.94%
Sadness	1270	25.92%
Happiness	1271	25.94%
Neutral	1087	22.2%
Total	4899	100%

4.1.2 Experimental Methodologies

The traditional emotion recognition experiments were conducted using the 5-fold cross validation method usually. This evaluation method yields a high accuracy and includes the target user data in the training dataset, where the number of training data is relatively large. However, this method is not suitable for measuring the performance in personalized emotion recognition experiments, as there is only a small amount of target user training data. Therefore, we aimed to verify the individual accuracy performance using a minimal target user training dataset combined with a new experimental method.

In this new experiment, the training dataset and test dataset were randomly divided without considering emotion label balance in order to create an environment similar to real speech acquisition with a limited dataset. At first, we decided the number of maximum training data samples. We allocated the training data and test data to half and half, and we also constructed the sufficient test data samples for evaluation. As a result, we set maximum training data to 300 as considering the total number of data is 6,925 and the minimum number of data is 379 in subject 2. The remained data not included in the training dataset were used as the test dataset. Second, we incrementally increased the size of the training dataset for each target user starting from a minimum of 50 to a maximum of 300.

This is done to progressively measure the accuracy according to the number of target user training data when creating the personalized training model. And the average accuracy and precision were measured by repeating the experiment 10 times for fairness. In other words, test data is randomly fixed in each experiment and the training data changes from 50 to 300 incrementally. (e.g. Subject 1 had 431 utterances; total dataset:431, Training dataset: 50-300, Test dataset: 131)

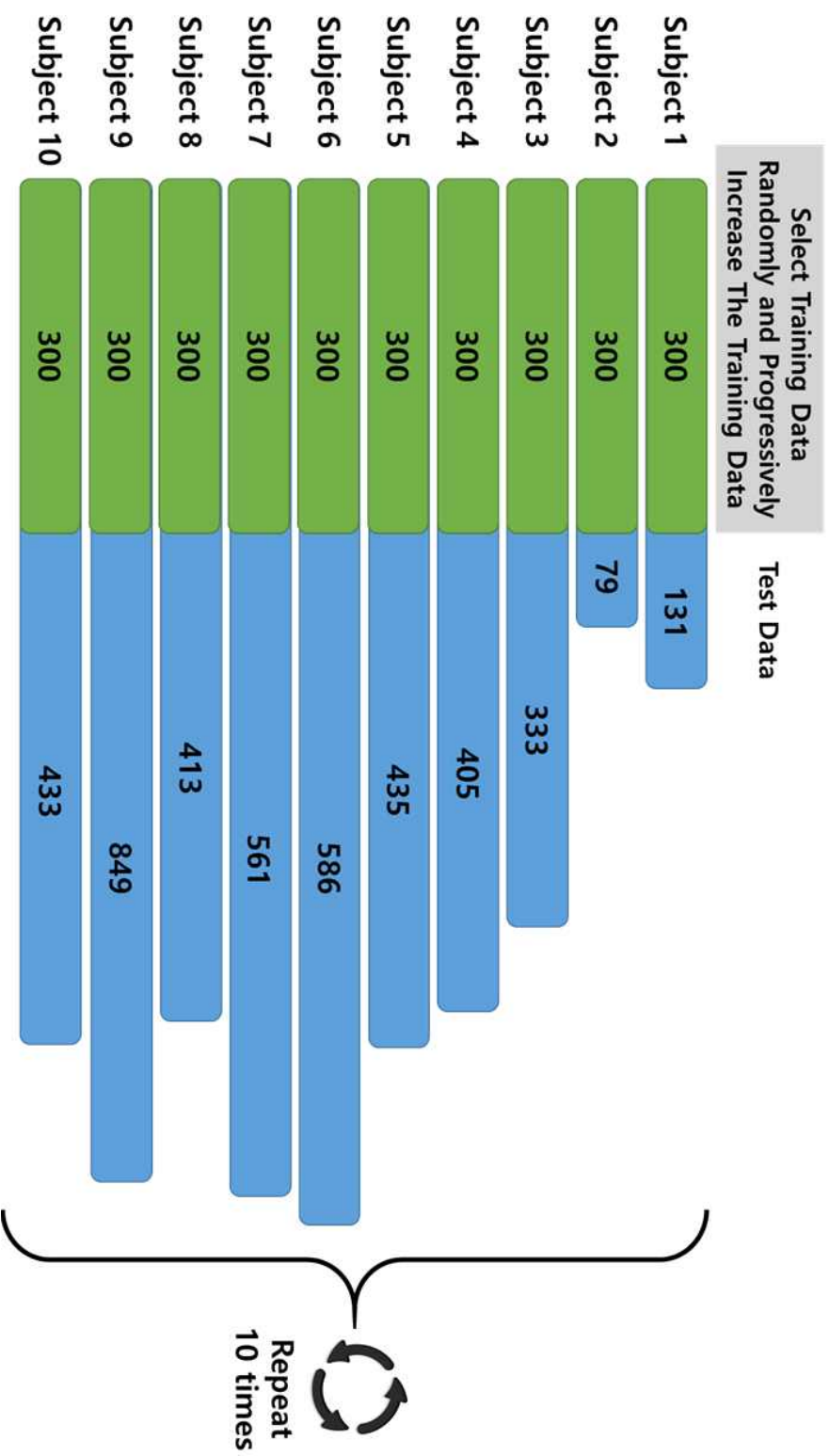


Figure 4-3 The concept of the experimental methodologies (The example of IEMOCAP)

We performed 4 kinds of comparison evaluation to validate if the proposed method is really efficient in emotionally imbalanced small samples environment. And we also employed the Imbalance Ratio (IR) [93] to understand how much emotional data is unbalanced and improved. And we also evaluate the cross corpus environment with CREMA-D. There are 51 to 53 personal data in CREMA-D. Therefore, when CREMA-D is used as evaluation dataset, only 20 data are used as training data sets and the remained data are used as test data. The experiment consists of four criteria as follows.

1) SI (Speaker Independent – baseline) :

The experiment using target user speech data as the test data and creating a training model with the remaining nine users' datasets

2) PM (Personal Model – self learning) :

The experiment conducted by constructing a training model only with personal user speech data.

3) Conventional MLLR with HMM :

The experiment using conventional MLLR based model adaptation

4) LDM-MDT MLLR with GMM :

The experiment using LDM-MDT MLLR based data selection

5) Proposed method :

The experiment using the proposed Robust Speaker Adaptation Framework

4.2. Experimental Results

4.2.1. Recognition Accuracy

In this section, we describe the results of the recognition accuracy of the four experiments introduced in Section 5.2. The experiments were performed using implemented SMO, J48 and Random Forest in WEKA Library [94] to estimate which classifier shows best performance. The WEKA Library is a well known machine learning open source library.

In SMO case, we select the RBF kernel which is normally used in speech emotion recognition area. The advantage of using RBF kernel is that it restricts training data to lie in specified boundaries. The RBF kernel nonlinearly maps samples into a higher dimensional space which means it can handle the case when the relation between class labels and attributes is nonlinear unlike the linear kernel. The RBF kernel has less numerical difficulties than polynomial kernel [95]. Therefore, we used the RBF Kernel for SVM classifier. And the parameter of the Gamma and C is set to default value as in the Weka Library (Gamma Value = 0.01, C value = 1). And we also use the standardization process in RBF Kernel.

Figure 4-4 and Figure 4-5. shows the average accuracy for experiments using various classifiers and how many target user data we use to train. In Figure 4-4. and Figure 4-5, the accuracy in every classifier in all experiments is incrementally increased while target user's training data is increased. The performance of the Random Forest classifier used in the proposed framework is the highest.

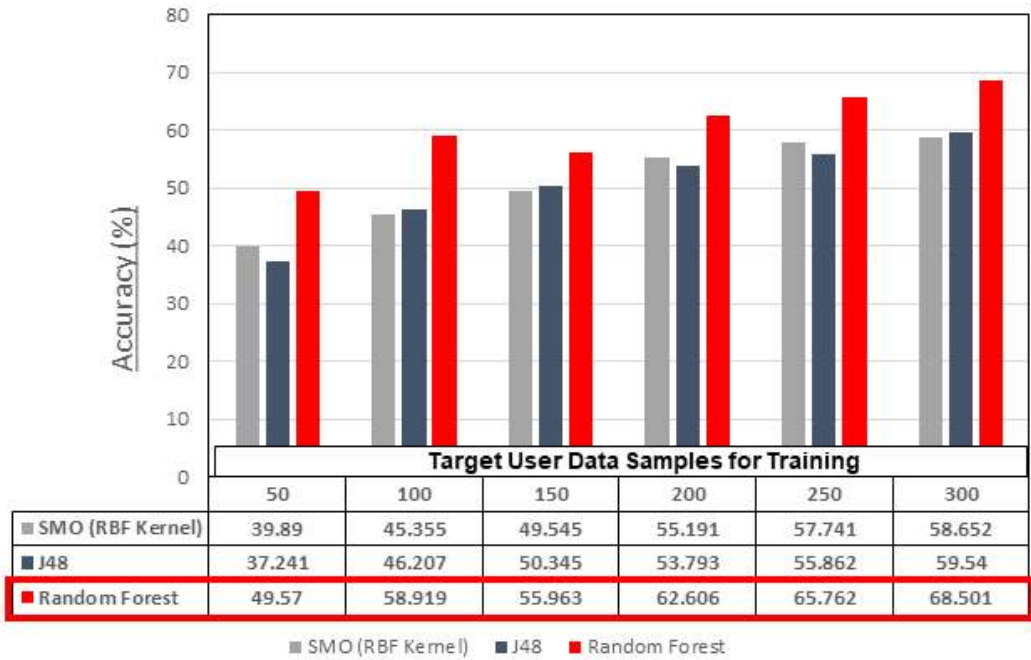


Figure 4-4. Experimental Results for Each Classifier (Unit %)
(Initial Model: CREMA-D, Evaluation Dataset: IEMOCAP)

In SMO experiment, there is a large difference between small amount of training data and large amount of training data. As a result. We can see that the SMO classifier requires lots of target user training data to create personalized model. This means, SMO classifier is more suitable to create general model than personalized model.

In J48 experiment, the result shows that the accuracy improves continuously as the target user data increases (37.241% to 59.54%). However, the accuracy is poor in small data environments. This means that the J48 classifier is hard to create personalized model when acquired amount of data is small.

In Random Forest experiment, the result shows the best accuracy compared with all other experiments (49.57% to 68.501%). Therefore, we can know that the Random Forest classifier is suitable to create personalized model with our proposed method.

To validate the robustness of diversity environment, we performed evaluate again that CREMA-D is set as an evaluation data and IEMOCAP dataset is set as an initial model. Figure 4-5 shows the results of a machine learning algorithm comparison experiment in a reversed cross-corpus environment. This result does not make much difference from the experimental result compared with previous experiment. The reason is the proposed algorithm selects the data of the specific area in feature space. In other words, stable accuracy can be obtained when a large number of data in an initial model is obtained. Finally, we can know that the classifier suitable for the personalized emotion recognition model is the Random Forest through two cross-corpus experiments. And also we can see CREMA-D is higher accuracy than IEMOCAP dataset evaluation because it was recorded by the defined rule and limited dataset.

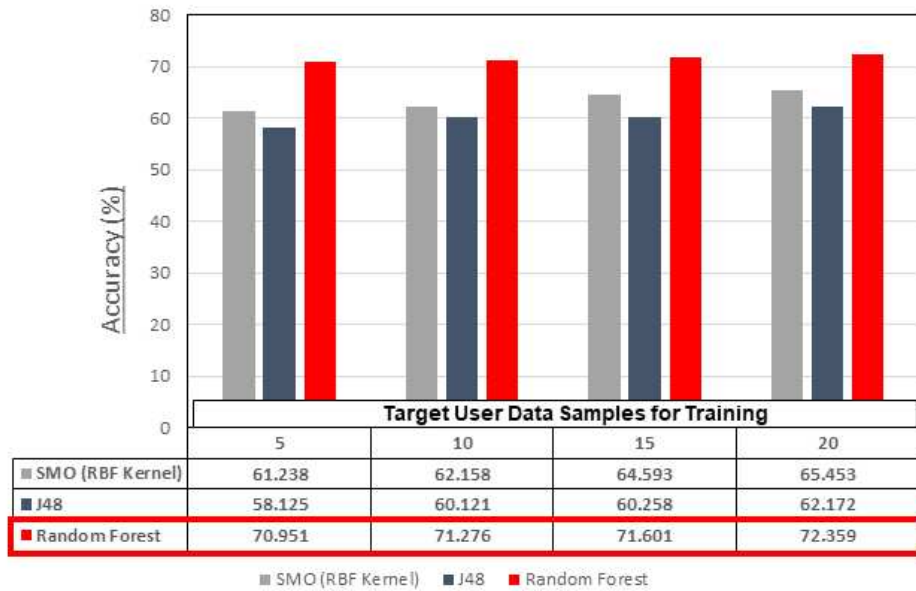


Figure 4-5. Experimental Results for Each Classifier (Unit %)
(Initial Model: CREMA-D, Evaluation Dataset: IEMOCAP)

Figure 4-6 and Figure 4-7 shows the detailed results using the Random Forest classifier with caparison evaluation. We can see that the proposed method always shows the highest accuracy.

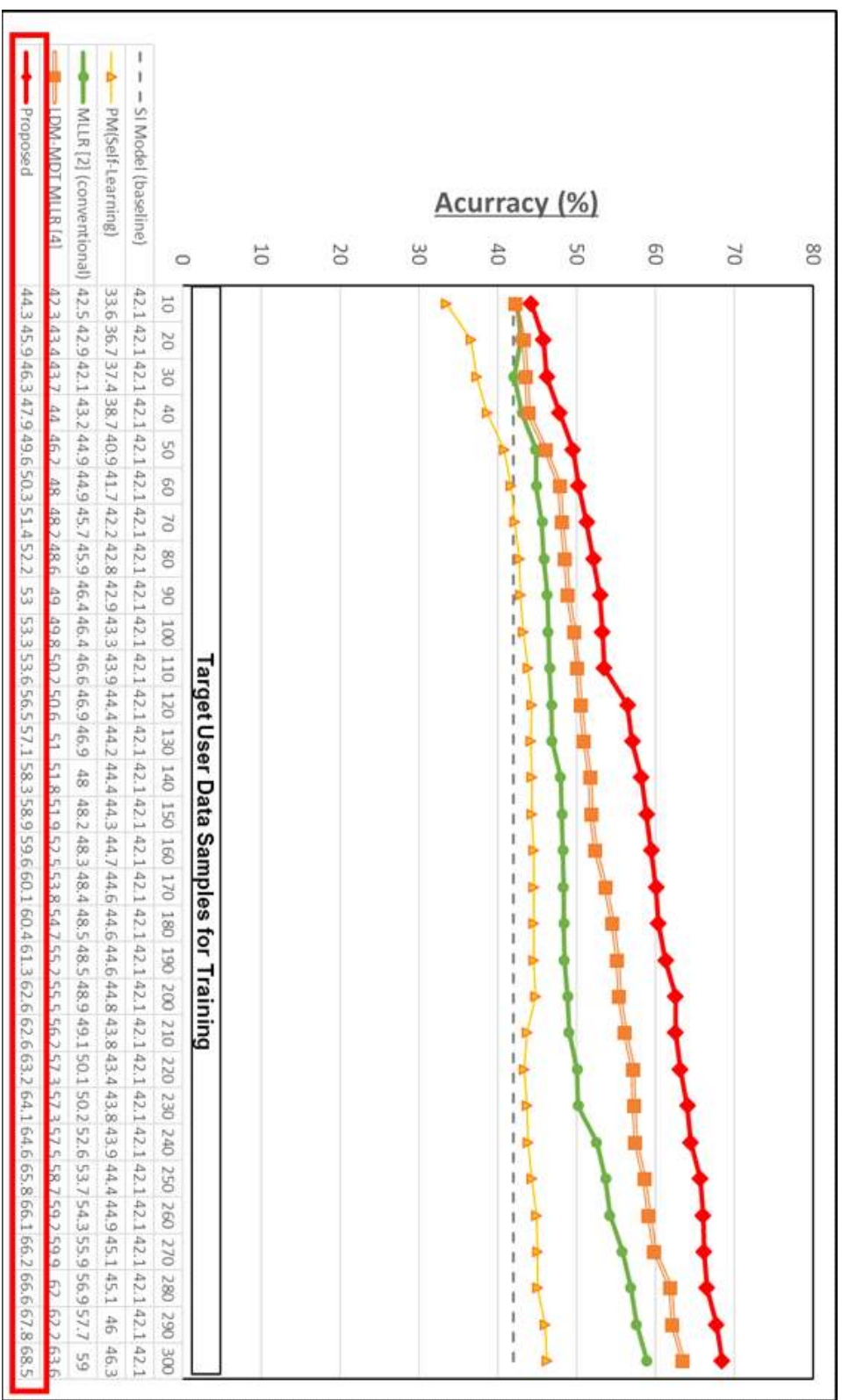


Figure 4-6 Detailed experimental results (Evaluation Data: IEMOCAP ‘ Initial Model: CREMA-D)

The experimental results of SI show an average of 42.1%. Before the target user speech exceeds 70, the performance is higher than both PM and SMOTE. After that value however, PM and SMOTE show a higher accuracy. The SMOTE shows a lower accuracy than PM when the number of target user samples is less than 70. After that, when the target user data is sufficiently acquired, we can see that the accuracy is rapidly increased. The experimental results of conventional MLLR method is increase the accuracy very slowly due to utilize target user data with all of existing data. And the LDM-MDT-MLLR method is performed better than Conventional MLLR method due to utilize useful data in existing data for target user. However, it still adaptation speed is very slow due to select insufficient real case data for personalization in small samples environment.

Proposed method exhibits high performance across all of the experiments over the whole period due to the construction of sufficient number of data with proposed Robust Speaker Adaptation method from other user even small amount of data environment. The result in large amount of data environment of LDM-MDT-MLLR and proposed method are becoming similar. However, in the small amount of the target user data environment, the result of proposed method clearly shows higher accuracy than LDM-MDT-MLLR where the accuracy difference is about 6%. In other words, we can see that the proposed Robust Speaker Adaptation method solves the cold-start problem of existing methods efficiently.

The reversed cross-corpus evaluation also is shown the proposed methods is highest accuracy compared with existing methods. The CREAM-D dataset is very diverse and limited data set. Therefore, we can not see the rapid increment of accuracy level in Figure 4-7. Table 4-7 and 4-8 is show the confusion matrix of comparison of accuracy for each cross-corpus evaluations.

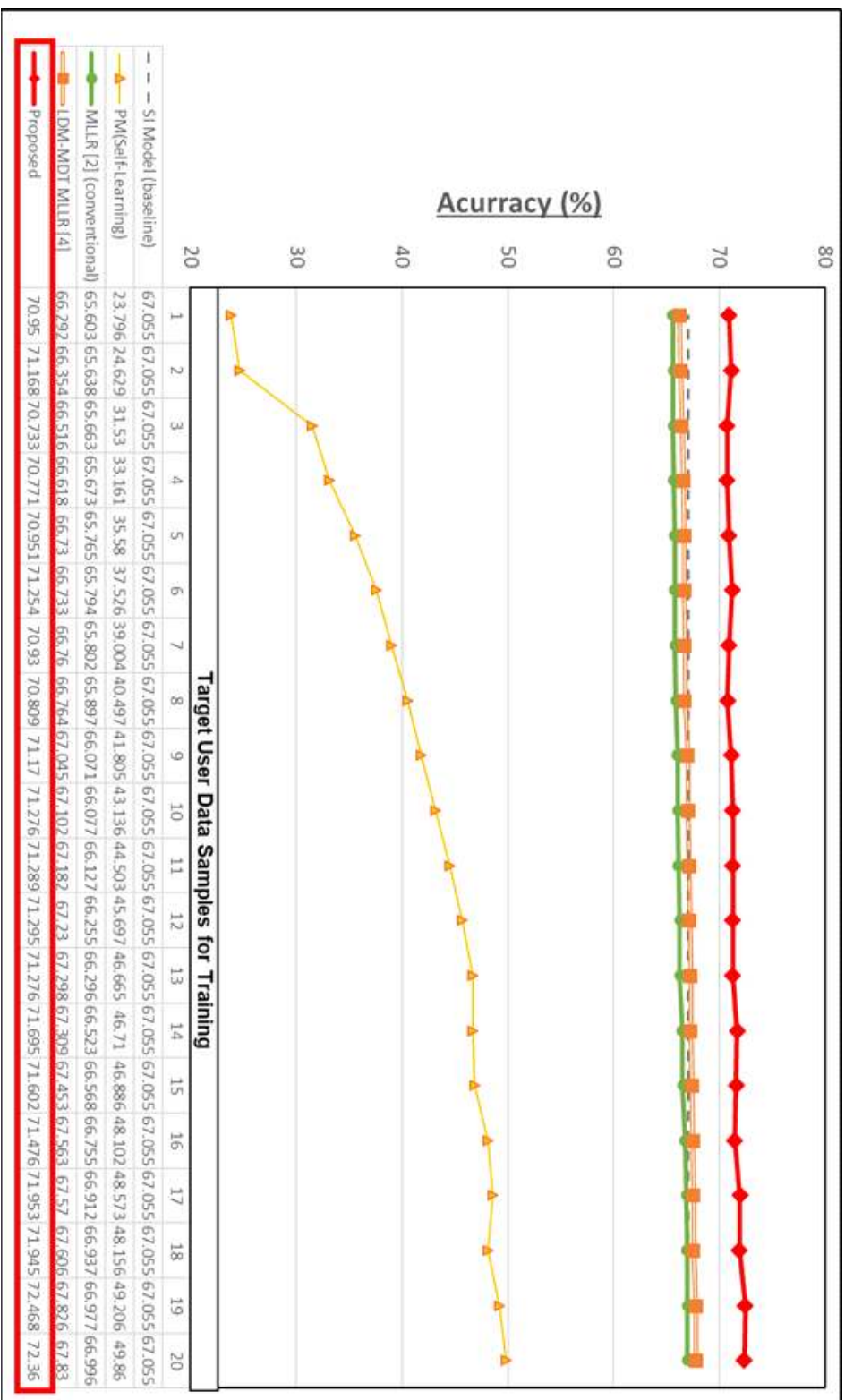


Figure 4-7 Detailed experimental results (Evaluation Data: IEMOCAP ‘ Initial Model: CREMA-D)

Table 4-7. The confusion matrix of comparison of accuracy (Unit %)
(Evaluation Data: IEMOCAP, Initial Model: CREMA-D)

Target user Data 300	PM Model (46.3%)					Conventional MLLR (59%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	35.39	4.42	26.16	34.03	Ang.	64.87	9.08	11.18	14.87
	Sad.	6.66	33.23	15.94	44.16	Sad.	9.00	64.56	7.31	19.13
	Hap.	23.30	9.04	22.81	44.85	Hap.	20.09	12.22	49.71	17.98
	Neu.	8.87	9.55	5.07	76.50	Neu.	13.81	13.30	12.93	59.96
	LDM-MDT-MLLR (63.6%)					Proposed Method (68.5%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	63.47	8.98	11.18	16.37	Ang.	75.50	4.21	7.92	12.38
	Sad.	9.14	64.42	8.72	17.72	Sad.	6.09	72.75	7.54	13.62
	Hap.	20.45	9.87	50.76	18.92	Hap.	14.71	12.42	53.59	19.28
	Neu.	12.34	12.64	11.76	63.26	Neu.	6.38	12.59	8.87	72.16
Target user Data 250	PM Model (44.4%)					Conventional MLLR (52.6%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	51.89	6.47	11.75	29.88	Ang.	53.98	13.77	17.58	14.67
	Sad.	7.21	35.92	11.56	45.31	Sad.	9.86	63.91	10.97	15.26
	Hap.	24.56	9.98	26.21	39.25	Hap.	19.72	18.01	45.67	16.60
	Neu.	12.98	11.79	15.01	60.21	Neu.	14.10	20.54	18.19	47.17
	LDM-MDT-MLLR (58.7%)					Proposed Method (65.8%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	62.48	8.58	12.08	16.87	Ang.	69.06	5.69	9.65	15.59
	Sad.	8.58	62.45	7.74	21.24	Sad.	6.38	71.30	9.57	12.75
	Hap.	22.56	10.34	47.83	19.27	Hap.	13.73	12.75	54.25	19.28
	Neu.	14.62	13.15	12.86	59.37	Neu.	7.45	13.83	10.28	68.44

Target user Data 200	PM Model (44.8%)					Conventional MLLR (48.9%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	42.63	5.58	19.12	32.67	Ang.	54.76	16.91	13.89	14.45
	Sad.	7.35	37.69	12.93	42.04	Sad.	9.22	66.30	11.92	12.56
	Hap.	19.30	11.62	30.26	38.82	Hap.	21.70	20.99	40.71	16.60
	Neu.	7.44	12.03	15.72	64.80	Neu.	18.36	22.72	19.84	39.08
	LDM-MDT-MLLR (55.5%)					Proposed Method (62.6%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	56.33	12.43	16.46	14.78	Ang.	64.85	8.17	9.90	17.08
	Sad.	10.81	63.75	10.65	14.79	Sad.	7.54	71.59	10.14	10.72
	Hap.	19.57	14.75	47.52	18.16	Hap.	16.67	14.38	50.33	18.63
	Neu.	12.53	17.41	16.97	53.09	Neu.	7.98	15.96	12.41	63.65
Target user Data 150	PM Model (44.3%)					Conventional MLLR (48.2%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	41.63	6.18	21.41	30.78	Ang.	53.98	18.03	16.80	11.20
	Sad.	7.48	41.90	10.61	40.00	Sad.	12.24	62.00	12.72	13.04
	Hap.	20.50	11.95	29.50	38.05	Hap.	25.53	22.55	40.14	11.77
	Neu.	10.48	13.28	16.02	60.21	Neu.	18.36	24.72	20.19	36.73
	LDM-MDT-MLLR (52.5%)					Proposed Method (58.9%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	53.98	13.77	17.58	14.67	Ang.	59.90	12.38	11.63	16.09
	Sad.	9.86	63.91	10.97	15.26	Sad.	5.22	66.67	15.94	12.17
	Hap.	19.72	18.01	45.67	16.60	Hap.	14.38	15.03	51.31	19.28
	Neu.	14.10	20.54	18.19	47.17	Neu.	8.16	16.84	17.20	57.80

Target user Data 100	PM Model (43.3%)					Conventional MLLR (46.4%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	44.82	7.57	12.85	34.76	Ang.	47.26	21.61	20.38	10.75
	Sad.	6.80	43.67	9.52	40.00	Sad.	7.95	66.45	13.67	11.92
	Hap.	22.59	13.49	20.07	43.86	Hap.	23.83	24.82	39.72	11.63
	Neu.	9.95	13.40	13.82	62.84	Neu.	16.71	31.94	22.72	28.63
	LDM-MDT-MLLR (49.8%)					Proposed Method (53.3%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	54.76	16.91	13.89	14.45	Ang.	49.50	10.64	22.28	17.57
	Sad.	9.22	66.30	11.92	12.56	Sad.	3.48	60.00	25.80	10.72
	Hap.	21.70	20.99	40.71	16.60	Hap.	15.03	15.03	52.29	17.65
	Neu.	18.36	22.72	19.84	39.08	Neu.	5.85	21.81	20.74	51.60
Target user Data 50	PM Model (40.9%)					Conventional MLLR (44.9%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	29.88	5.08	15.64	49.40	Ang.	45.46	21.50	21.72	11.31
	Sad.	3.27	36.60	5.85	54.29	Sad.	10.17	64.86	17.01	7.95
	Hap.	12.83	10.42	17.87	58.88	Hap.	19.29	27.94	40.28	12.48
	Neu.	3.45	10.07	9.17	77.31	Neu.	17.23	31.51	22.80	28.46
	LDM-MDT-MLLR (46.2%)					Proposed Method (49.6%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	48.49	18.92	21.16	11.42	Ang.	57.92	23.02	4.95	14.11
	Sad.	10.17	64.86	12.72	12.24	Sad.	3.48	51.01	40.00	5.51
	Hap.	20.71	23.83	43.55	11.91	Hap.	7.84	30.39	50.33	11.44
	Neu.	15.06	29.85	21.67	33.42	Neu.	6.56	28.55	25.89	39.01

Table 4-8. The confusion matrix of comparison of accuracy (Unit %)
(Evaluation Data: CREMA-D, Initial Model: IEMOCAP)

Target user Data 20	PM Model (49.86%)					Conventional MLLR (67%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	64.71	4.99	23.57	6.73	Ang.	73.85	0.81	22.81	2.53
	Sad.	4.17	63.50	11.82	20.51	Sad.	5.07	76.18	6.84	11.91
	Hap.	30.50	13.65	39.61	16.24	Hap.	16.93	6.68	67.71	8.69
	Neu.	14.09	29.27	27.77	28.86	Neu.	8.47	22.46	18.36	50.71
	LDM-MDT-MLLR (67.8%)					Proposed Method (72.36%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	73.18	1.50	23.12	2.20	Ang.	70.93	3.28	18.64	7.15
	Sad.	5.66	75.00	6.13	13.21	Sad.	0.85	82.22	4.11	12.82
	Hap.	19.80	8.38	63.87	7.94	Hap.	12.77	9.12	56.78	21.32
	Neu.	11.05	18.64	15.17	55.14	Neu.	1.61	9.37	9.66	79.36
Target user Data 15	PM Model (46.8%)					Conventional MLLR (66.5%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	65.18	3.94	23.39	7.49	Ang.	72.00	1.73	22.93	3.34
	Sad.	6.79	58.79	14.67	19.76	Sad.	4.94	75.18	6.12	13.76
	Hap.	34.86	13.74	34.73	16.67	Hap.	18.56	7.89	65.00	8.56
	Neu.	17.55	25.91	27.99	28.55	Neu.	9.37	18.23	18.87	53.53
	LDM-MDT-MLLR (67.4%)					Proposed Method (71.6%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	69.46	1.96	25.11	3.48	Ang.	72.03	3.58	17.62	6.77
	Sad.	3.57	75.95	6.21	14.27	Sad.	1.81	80.05	4.47	13.66
	Hap.	18.46	8.55	64.44	8.55	Hap.	16.26	9.23	55.59	18.92
	Neu.	6.60	16.84	17.06	59.50	Neu.	2.49	10.10	8.64	78.77

Target user Data 10	PM Model (43.1%)					Conventional MLLR (66.1%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	60.58	4.10	28.44	6.88	Ang.	66.94	1.96	26.73	4.38
	Sad.	7.41	53.58	19.08	19.93	Sad.	3.54	75.71	6.49	14.27
	Hap.	35.46	14.54	37.63	12.37	Hap.	17.33	9.11	63.89	9.67
	Neu.	17.81	28.33	31.84	22.02	Neu.	6.51	17.58	18.26	57.65
	LDM-MDT-MLLR (67.1%)					Proposed Method (71.2%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	71.21	2.58	23.32	2.89	Ang.	74.56	3.05	17.58	4.81
	Sad.	4.56	77.42	7.34	10.68	Sad.	1.09	80.70	3.38	14.84
	Hap.	17.62	9.59	64.44	8.34	Hap.	16.50	8.76	54.72	20.02
	Neu.	8.74	18.99	17.04	55.23	Neu.	2.72	11.30	11.16	74.82
Target user Data 5	PM Model (35.5%)					Conventional MLLR (65.8%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	49.60	8.22	34.35	7.82	Ang.	65.15	2.82	28.88	3.15
	Sad.	9.68	38.97	32.72	18.63	Sad.	3.56	76.33	6.22	13.89
	Hap.	34.39	15.80	36.94	12.87	Hap.	16.03	9.60	65.61	8.76
	Neu.	18.80	27.72	36.63	16.85	Neu.	8.08	18.53	18.00	55.39
	LDM-MDT-MLLR (66.7%)					Proposed Method (70.9%)				
	class	Ang.	Sad.	Hap.	Neu.	class	Ang.	Sad.	Hap.	Neu.
	Ang.	70.41	2.06	23.13	4.40	Ang.	71.00	4.20	19.29	5.51
	Sad.	5.38	74.89	6.33	13.40	Sad.	0.87	81.59	4.10	13.43
	Hap.	22.44	11.22	53.24	13.09	Hap.	17.00	8.92	55.89	18.19
	Neu.	6.54	12.31	12.99	68.17	Neu.	2.20	10.98	11.71	75.11

In the confusion matrix of cross-corpus evaluation, it can be seen that the proposed algorithm shows balanced accuracy in each emotion class over the whole interval. In the case of the PM model, the accuracy varies depending on the number of selected training data for each emotion, and the most imbalanced accuracy is shown. On the other hand, the model adaptation techniques of MLLR and LDM-MDT-MLLR show somewhat balanced accuracy. However, existing model adaptation techniques show uneven accuracy that the number of selected target user training data is a limited environment such as 5 or 10. The proposed method is more accurate than the conventional methods over the whole interval. And, in a limited environment with very few data, it showed more accurate accuracy than the existing techniques.

4.2.2. Imbalanced Ratio

The Imbalanced Ratio (IR) evaluation is an validation of the imbalance in the final selected training data sets. If the IR is high, the difference between the majority class and the minority class is not so accurate. Therefore, lowering this IR value has a good effect on improving the perception accuracy. The Figure 4-8 shows the status of imbalanced level represented by IR between the majority class and minority class. The IR measurement is calculated by the equation 16.

$$\text{Imbalanced Ratio} = \text{Major Class} / \text{Minor Class} \quad (16)$$

The result of PM experiment means standard IR value in the IEMOCAP dataset. the PM does not solve imbalanced environment over the whole periods, and conventional SMOTE solves a little bit in small amount data environment. The conventional MLLR method basically cannot solves the imbalanced problem due to utilize all of data for adaptation. Because it is affected by imbalanced ratio of initial data set. The LDM-MDT MLLR method looks like solve the imbalanced ratio due to select similar data of small range. However, in absence data, the imbalanced ratio is increased significantly due to utilize exist model for absence data. Proposed method solves the imbalanced data in not only small data environment but also large data environment.

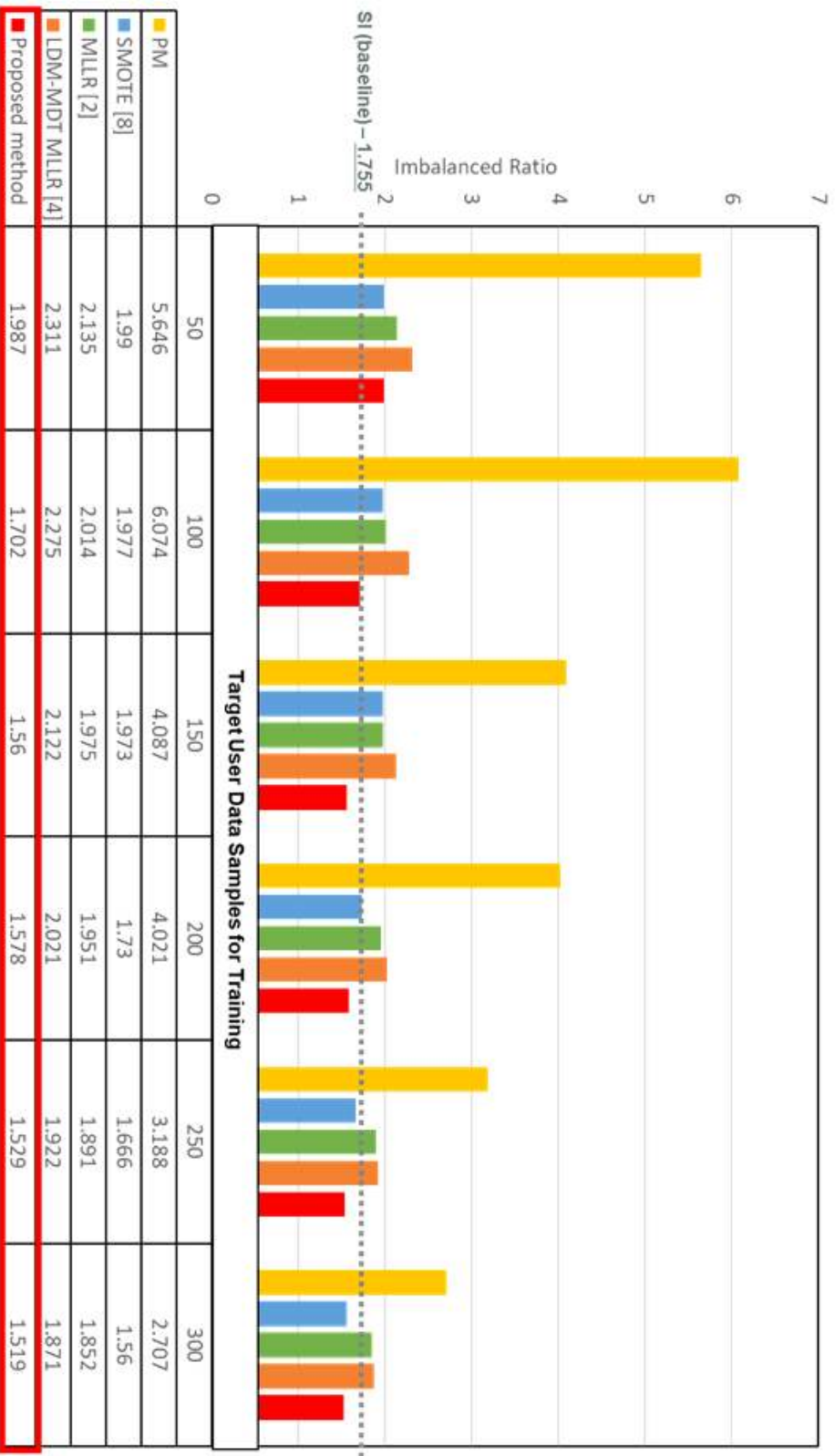


Figure 4-8 Detailed Imbalanced Ratio experimental results

5.1 Conclusion

In this paper, we proposed a robust personalized emotion recognition framework considering the small and imbalanced data environment problem in adaptive speech-based emotion recognition. The proposed Robust Speaker Adaptation Framework provides a personalized training model for the target user utilizing 3 core solutions by selecting the actual case data useful for the target user from collected target user emotional speech with initial training model and incrementally augmenting virtual data based on the SMOTE.

In this thesis, a well-known data set such as IEMOCAP and CREMA-D which is a public emotional speech database is used for comparison experiment evaluation. In experimental result, the proposed method showed about 2.2%~6% accuracy improvement compared to existing methods in small data environment when the number of target user training data is from 10 to 150. And the proposed method reduced imbalanced difference about 178% to 356% from original target user training dataset. The proposed method proved that it is able to provide a stable personalization model which is faster than existing techniques in the limited data environment in the whole section by providing a sufficient and balanced personalized training model even in the initial stage.

5.2 Future Direction

This section presents the direction of future research and the research extension. 4 kinds of future plan are set up which are ① Data acquisition mechanism to apply real environment, ② Enhancement of proposed framework using deep learning method, ③ Apply the proposed framework to other domain, and ④ Research on emotion fusion techniques based on multi-modal sensors.

① Data acquisition mechanism to apply real environment

The first future work is to develop a personalized data collection mechanism to utilize the proposed framework in a real environment. The proposed framework in this research has only considered model adaptation part. Therefore, an effective personalized data acquisition mechanism must be developed for fully personalized speech based emotion recognition in supervised manner. In other words, the direction of effective personalized data acquisition mechanism must be possible to create a personalized emotion recognition model while minimizing target user intervention. And this research have to extend to conduct additional experiments using state of the art classification methods. Currently, we cannot conduct direct comparison with other studies as the data environment, research goal, and methodologies are quite different. However, we will figure out a solution for this later. Also, I will further conduct research integrating emotional speech databases, such as Emo-DB, CREMA-D, eNTERFACE, SAVEE, and IEMOCAP etc. to validate the generalization of our framework.

② Enhancement of proposed framework using deep learning method

The second future direction is to develop enhanced model adaptation techniques using Deep Domain Adaptation Network which is the state of art technology to adapt to enable personalized Speech based emotion recognition. The Deep Domain Adaptation is designed to improve the accuracy level for various data environments. This approach is very powerful method to solve the adversarial problem. For resolving the adversarial problem in the speeches, it should make various data through synthetic speeches. The basic idea of emotional speech synthetic techniques is creating combined speeches randomly on long-term speech. After then, learn through Deep Neural Network Algorithm based on created various real-cases. Therefore, the proposed framework in this research will extend to Deep Domain Adaptation Network research field for personalized emotion recognition through emotional speech data augmentation and synthetic techniques.

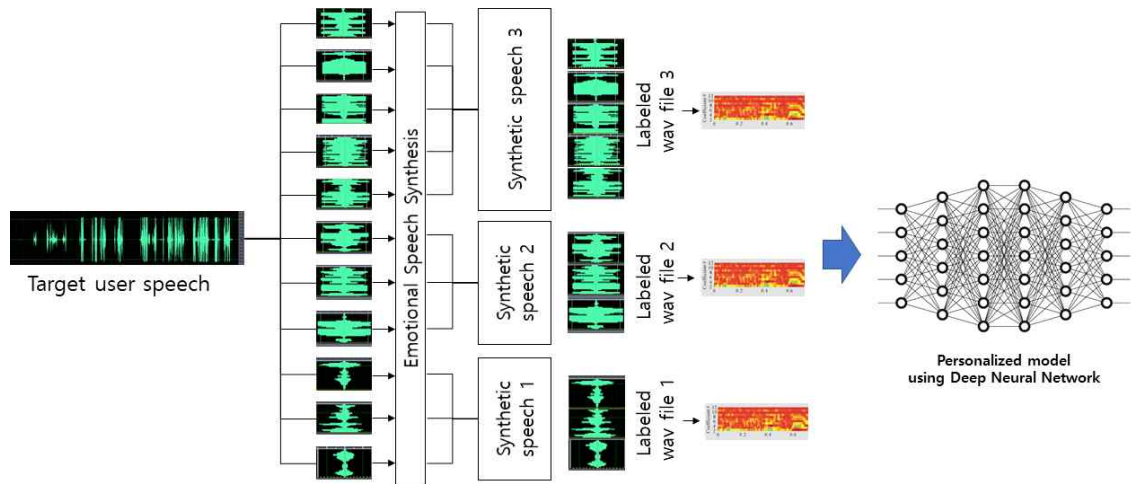


Figure 5-1. Concept of the proposed personalized modeling using DNN based on target user speech synthesis

③ Apply the proposed framework to other domain

The third future work is to apply the proposed framework to other domain. The proposed framework is a very general algorithm for personalization. Therefore, if there is knowledge of feature extraction in other fields, customized learning model can be created regardless of domain. Especially, it is a plan to apply image based face recognition in a lot of machine learning field to prove the generalization of the algorithm. The basic idea of applying the proposed framework to personalized face recognition is to generate real-case data and virtual case data based on the face photograph of the user tagged in the SNS and apply it to the Deep Neural Network. Figure 5-2. show application of proposed framework in face recognition.

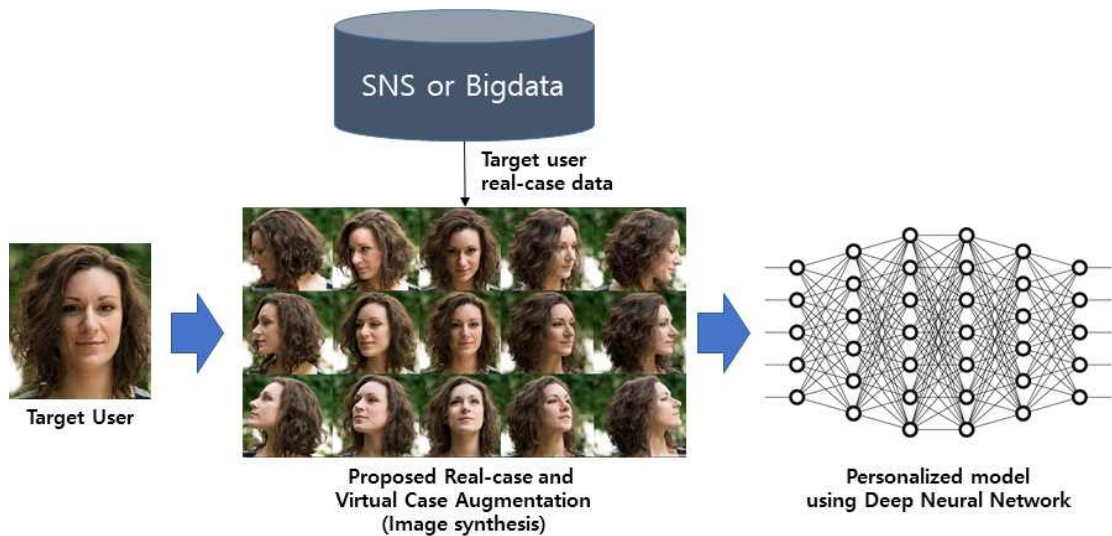


Figure 5-2. Application of proposed framework in face recognition

④ Research on emotion fusion techniques based on multi-modal sensors

The final future work is to extend the work using multi-modal sensors. There are many kinds of data which can be acquired from sensors such as camera, word etc. for

emotion recognition. At first, we consider three different sources such as image, audio, and word from audio. The reason for considering these three data is that they are the most affected representative data sources to recognize emotions. We have a plan to develop feature level fusion technique using DNN algorithm. The DNN algorithm is very powerful to find the optimized feature automatically. Therefore, we will develop the feature aggregation methodologies for feature level emotion fusion. After that, we will try to apply the personalization by proposed framework in this thesis. Figure 5-3 shows the concept of the proposed multi-modal fusion.

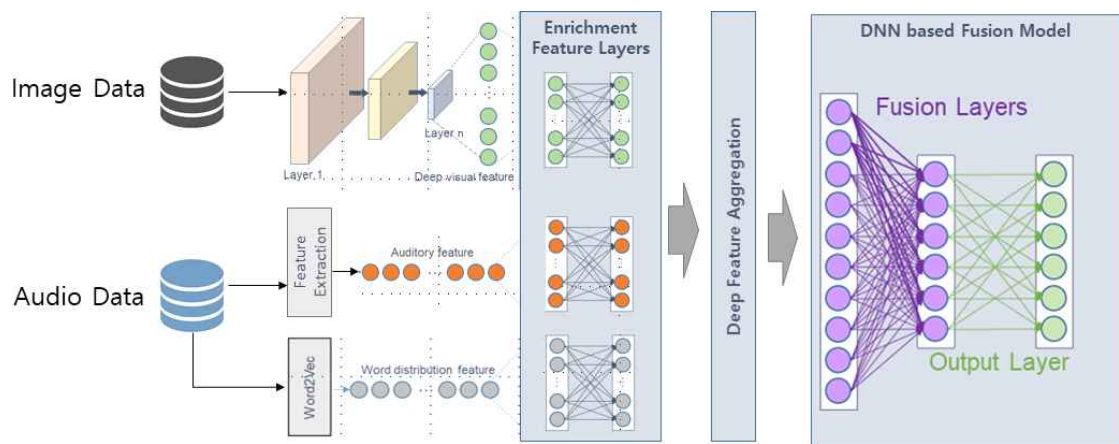


Figure 5-3. Concept of the proposed multi-modal fusion.

- [1] Schuller, Björn W., “Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends”, *Communications of the ACM* Vol. 61, No. 5, pp.90-99, 2018
- [2] Leila Kerkeni¹, Youssef Serrestou¹, Mohamed Mbarki, Kosai Raoof¹, Mohamed Ali Mahjoub, “Speech Emotion Recognition: Methods and Cases Study”, 10th International Conference on Agents and Artificial Intelligence, Jan, 2018
- [3] Kandali, A. B., Routray, A., Basu, T. K., “Emotion recognition from Assamese speeches using MFCC features and GMM classifier”, *TENCON, IEEE Region 10 International Conference*, Hyderabad, India, 19-21, Nov, 2008
- [4] Wang, K. C. “Time-frequency feature representation using multi-resolution texture analysis and acoustic activity detector for real-life speech emotion recognition”, *Sensors*, Vol. 15. Issue 1, pp.1458-1478, 2015
- [5] Zhu, L., Chen, L., Zhao, D., Zhou, J., Zhang, W. “Emotion recognition from chinese speech for smart affective services using a combination of SVM and DBN”, *Sensors*, Vol. 17, Issue 7, 2017.
- [6] Xiao, Z., Dellandréa, E., Chen, L., Dou, W. “Recognition of emotions in speech by a hierarchical approach”. *Affective Computing and Intelligent Interaction 2009. 3rd International Conference*, Amsterdam, Netherlands pp.401-408, 10-12, Spet., 2009
- [7] Cho, Y. H., Park, K. S. “A Study on The Improvement of Emotion Recognition by Gender Discrimination” *Journal of the Institute of Electronics Engineers of Korea* SP, Vol. 45, Issue 4, pp.107-114, 2008
- [8] Poria, S., Cambria, E., Bajpai, R., Hussain, A. “A review of affective

- computing: From unimodal analysis to multimodal fusion” Information Fusion, Vol. 37, pp.98-125, 2017
- [9] Leggetter C. J.; Woodland P. C. “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.” Computer Speech and Language, Vol. 9, No. 2, pp.171–185, 1995
 - [10] Trigeorgis G., Ringeval F., Brueckner R., Marchi E., Nicolaou M.A., Schuller B., Zafeiriou S., “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network”, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5200-5204, 20-25, March, 2016
 - [11] Zhang S., Zhang S., Huang T., Gao W. “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching”, IEEE Transactions on Multimedia, 2018, Vol. 20 Issue 6, pp.1576-1590, 2018
 - [12] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, Guan-Zheng Tan, “Speech emotion recognition based on feature selection and extreme learning machine decision tree”, Neurocomputing, Vol. 273, pp.271-280, 2018
 - [13] Lim, W., Jang, D., Lee, T., “Speech emotion recognition using convolutional and Recurrent Neural Networks”, 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, South Korea, 13-16, Dec., 2016
 - [14] Guo, L., Wang, L., Dang, J., Zhang, L., Guan, H., “A feature fusion method based on extreme learning machine for speech emotion recognition” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2666-2670, Calgary, AB, Canada, 15-20, April, 2018
 - [15] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." Fifteenth Annual Conference of the International Speech Communication Association. 2014.
 - [16] Guo, L., Wang, L., Dang, J., Liu, Z., Guan, H. “Exploration of

- Complementary Features for Speech Emotion Recognition Based on Kernel Extreme Learning Machine” IEEE Access, Vol. 7, pp.75798-75809, 2019
- [17] Kaya, H., Karpov, A. A., “Efficient and effective strategies for cross-corpus acoustic emotion recognition”. Neurocomputing,, Vol. 275, pp.1028-1034, 2018
- [18] Noroozi F., Kaminska D. T. Sapinski and G. Anbarjafari, “Supervised Vocal-Based Emotion Recognition Using Multiclass Support Vector Machine, Random Forests and Adaboost”, Journal of Audio Engineering Society, Vol. 65, pp.562-572, 2017
- [19] Fu, L., Wang, C., Zhang, Y., “Classifier fusion for speech emotion recognition” 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, Vol. 3, pp. 407-410, Xiamen, China, 29-31, Oct. 2010
- [20] Gosztolya, G., Busa-Fekete, R., Toth, L. “Detecting Autism, Emotions and Social Signals Using AdaBoost” INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, pp.220-224, Lyon, France, 25-29, Aug, 2013
- [21] Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., Rigoll, G., Speaker independent speech emotion recognition by ensemble classification. 2005 IEEE International Conference on Multimedia and Expo, pp.864-867, 6, July, 2005
- [22] Shinoda, K. “Speaker adaptation techniques for speech recognition using probabilistic models” Electronics and Communications in Japan, Vol. 88, No. 12, pp.25-42, 2005.
- [23] Busso, C., Mariooryad, S., Metallinou, A., Narayanan, S., “Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech” IEEE Transactions on Affective Computing, Vol. 4, Issue 4, pp.386-397, 2013
- [24] Busso, C., Metallinou, A., Narayanan, S., “Iterative feature normalization for emotional speech detection. In Acoustics”, Speech and Signal Processing (ICASSP), pp.5692-5695, Prague, Czech Republic, 22-27, May, 2011

- [25] Zhao, Y., Li, J., Zhang, S., Chen, L., & Gong, Y., "Domain and speaker adaptation for Cortana speech recognition" 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5984-5988, Calgary, AB, Canada, 15-20, April, 2018
- [26] Wang, Y., Du, S., Zhan, Y., "Adaptive and optimal classification of speech emotion recognition" Natural Computation (ICNC), pp.407-411, Jinan, China, 18-20, Oct, 2008
- [27] Abdelwahab, M., Busso, C., "Supervised domain adaptation for emotion recognition from speech" 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5058-5062, Brisbane, QLD, Australia, 19-24, April 2015
- [28] Mao, Q., Xue, W., Rao, Q., Zhang, F., Zhan, Y., "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes" 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2608-2612, 20-25, March, 2016
- [29] Mishra, T., Dimitriadis, D. "Incremental emotion recognition" In INTERSPEECH, 2013, pp.2876-2880, 2013
- [30] Kuhn, R., Junqua, J. C., Nguyen, P., Niedzielski, N., "Rapid speaker adaptation in eigenvoice space" IEEE Transactions on Speech and Audio Processing, Vol. 8, Issue 6, pp.695-707, 2000
- [31] Abdelwahab, M., Busso, C., "Incremental adaptation using active learning for acoustic emotion recognition" 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5160-5164, New Orleans, LA, USA, 5-9, March, 2017
- [32] Huang, Z., Xue, W., Mao, Q., & Zhan, Y., "Unsupervised domain adaptation for speech emotion recognition using PCANet", Multimedia Tools and Applications, Vol. 76 Issue. 5, pp.6785-6799, 2017
- [33] Kim, J. B., Park, J. S., "Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition" Engineering Applications of Artificial Intelligence, Vol. 52, pp.126-134, 2018

- [34] S. Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh, "Speech emotion recognition", 2014 International Conference on Advances in Electronics Computers and Communications, Bangalore, India, 10-11, Oct., 2014
- [35] Lee, C. C., Mower, E., Busso, C., Lee, S., Narayanan, S., "Emotion recognition using a hierarchical binary decision tree approach" Speech Communication, Vol.53, Issue 9-10, pp.1162-1171., 2011
- [36] Lee, C. C., Mower, E., Busso, C., Lee, S., Narayanan, S., "Emotion recognition using a hierarchical binary decision tree approach" Speech Communication, Vol.53, Issue 9-10, pp.1162-1171., 2011
- [37] Shi, K., Liu, X., & Qian, Y.. "Speech Emotion Recognition Based on SVM and GMM-HMM Hybrid System", NCMMSC2017, LianYunGang, China, Oct, 2017
- [38] Garg, V., Kumar, H., Sinha, R. "Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers" 2013 IEEE National Conference on Communications (NCC). pp.1-5, New Delhi, India, 15-17, Feb., 2013
- [39] Huang, C. L., Tsao, Y., Hori, C., Kashioka, H. "Feature normalization and selection for robust speaker state recognition" 2011 International Conference on Speech Database and Assessments, pp. 102-105, Hsinchu, Taiwan, 26-28, Oct., 2011
- [40] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," NIPS'00 Proceedings of the 13th International Conference on Neural Information Processing Systems, pp.388-394, Cambridge, MA, USA, Oct, 2000
- [41] N.A. Syed, S. Huan, L. Kah, and K. Sung, "Incremental learning with support vector machines" Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence (IJCAI 1999), Stockholm, Sweden, Aug, 1999
- [42] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal

- estimated sub-gradient solver for SVM,” *Mathematical programming*, Vol. 127, No. 1, pp. 3–30, 2011.
- [43] Lawal I.A. “Incremental SVM Learning: Review” *Learning from Data Streams in Evolving Environments*. Vol 41. pp.279-296, 2019
 - [44] J. Yang, R. Yan, and A.G. Hauptmann, “Cross-domain video concept detection using adaptive SVMs,” in *ACM international conference on Multimedia (MM 2007)*, pp. 188–197, Augsburg, Germany, September 2007
 - [45] Choi, D. J., Park, J. S., Oh, Y. H. “Unsupervised rapid speaker adaptation based on selective eigenvoice merging for user-specific voice interaction”, *Engineering Applications of Artificial Intelligence*, Vol. 40, pp.95-102, 2015
 - [46] Leggetter, C.J., Woodland, P.C., “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models” *Comput. Speech Language*, Vol. 9 Issue 2, pp.171–185, 1995
 - [47] Goronzy, Silke. “Robust adaptation to non-native accents in automatic speech recognition” *Lecture Notes in Computer Science* 2560, 2002
 - [48] Bucci, S., Loghmani, M. R., Caputo, B., “Multimodal Deep Domain Adaptation” *arXiv preprint arXiv:1807.11697.*, 2018
 - [49] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Lempitsky, V. “Domain-adversarial training of neural networks“ *The Journal of Machine Learning Research*, Vol. 17, Issue 1, pp.2096-2030, 2016
 - [50] Hong, S., Im, W., Ryu, J., & Yang, H. S. “Ssp-dan: Deep domain adaptation network for face recognition with single sample per person” 2017 *IEEE International Conference on Image Processing (ICIP)*, pp. 825-829. IEEE, 2017
 - [51] Abdelwahab, M., & Busso, C. “Domain adversarial for acoustic emotion recognition” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, Issue 12, pp.2423-2435, 2018
 - [52] Latif, S., Rana, R., & Qadir, J., “Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness”, *arXiv preprint arXiv:1811.11402.*, 2018

- [53] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, Julien Epps. "Transfer learning for improving speech emotion classification accuracy", INTERSPEECH 2018, pages 257–261, 2018.
- [54] Siddique Latif, Rajib Rana, Junaid Qadir, Julien Epps. "Variational autoencoders for learning latent representations of speech emotion: A preliminary study" INTERSPEECH 2018, pages 3107–3111, 2018.
- [55] McKay, C., Fujinaga, I., Depalle, P., "jAudio: A feature extraction library" Proceedings of the International Conference on Music Information Retrieval, London, UK, 11–15, Sept., 2005
- [56] Bang, J.; Lee, S. "Call Speech Emotion Recognition for Emotion based Services" Journal of Korean Institute of Information Scientists and Engineers: Software and Applications, Vol. 41, No. 3, pp.208–213, 2014
- [57] Sahoo, T.R.; Patra, S. "Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification" International Journal of Image, Graphics & Signal Processing. Vol. 6, pp.27–35, 2014
- [58] Nandhini, S., Shenbagavalli, A., "Voiced/unvoiced detection using short term processing", International Journal of Computer Applications, Vol. 975, pp.39–43, 2014
- [59] Anagnostopoulos, C. N., Iliou, T. "Towards emotion recognition from speech: definition, problems and the materials of research" In Semantics in Adaptive and Personalized Services, pp.127–143, Berlin, Heidelberg, 2010
- [60] Lalitha, S., Geyasruti, D., Narayanan, R., "Shravani, M. Emotion detection using MFCC and Cepstrum features" Procedia Computer Science, Vol. 70, pp.29–35, 2015
- [61] El Ayadi, M., Kamel, M. S., Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases" Pattern Recognition, Vol. 44, Issue 3, pp.572–587, 2011
- [62] Schuller, B., Rigoll, G., Lang, M., "Hidden Markov model-based speech emotion recognition" 2003 International Conference on Multimedia and Expo. ICME '03, pp.401–404, Baltimore, MD, USA, 6–9, July, 2003

- [63] Ververidis, D., Kotropoulos, C., "Emotional speech recognition: resources, features, and methods" *Speech Communication*, Vol. 48, Issue 9, pp.1162-1181, 2006
- [64] HALL, Mark Andrew. "Correlation-based feature selection for machine learning" 1999.
- [65] Azhagusundari, B., Antony Selvadoss Thanamani. "Feature selection based on information gain" *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* Vol. 2, Issue 2, pp.18-21, 2013
- [66] Fewzee, P., Karray, F. "Dimensionality reduction for emotional speech recognition" 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 532-537, Feb, 2012
- [67] Busso, C., Bulut, M., Lee, C.C. Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S., "IEMOCAP: Interactive emotional dyadic motion capture database", *Language Resources and Evaluation*, vol. 42, pp.335–359, 2008
- [68] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., "SMOTE: synthetic minority over-sampling technique" *Journal of artificial intelligence research*, Vol. 16, pp.321-357, 2002
- [69] Han, H., Wang, W. Y., Mao, B. H., "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning" In *International Conference on Intelligent Computing*, pp. 878-887, 2005
- [70] Lele, S., Richtsmeier, J. T., "Euclidean distance matrix analysis: A coordinate-free approach for comparing biological shapes using landmark data" *American Journal of Physical Anthropology*, Vol. 86, Issue.3, pp.415-427, 1991
- [71] Yuan, Y., Chao, M., Lo, Y. C., "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance" *IEEE transactions on medical imaging*, Vol. 36 Issue 9, pp.1876-1886, 2017
- [72] Trampe, D., Quoidbach, J., Taquet, M., "Emotions in everyday life" *PloS*

one, Vol. 10, Issue 12, 2015

- [73] <https://en.wikipedia.org/wiki/Average>
- [74] <https://en.wikipedia.org/wiki/Median>
- [75] <https://en.wikipedia.org/wiki/Variance>
- [76] https://en.wikipedia.org/wiki/Standard_deviation
- [77] <https://en.wikipedia.org/wiki/Skewness>
- [78] <https://en.wikipedia.org/wiki/Kurtosis>
- [79] https://en.wikipedia.org/wiki/Maxima_and_minima
- [80] Rong, J., Li, G., Chen, Y.P.P., “Acoustic feature selection for automatic emotion recognition from speech” *Information Processing & Management*, Vol. 45, Issue 3, pp.315–328, 2009
- [81] Noroozi, F., Sapiński, T., Kamińska, D., & Anbarjafari, G., “Vocal-based emotion recognition using random forests and decision tree” *International Journal of Speech Technology*, Vol. 20, Issue 2, pp.239-246, 2017
- [82] Chernykh, V., Sterling, G., Prihodko, P., “Emotion recognition from speech with recurrent neural networks” *arXiv:1701.08071*, 2017
- [83] Tripathi, S., & Beigi, H., “Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning” *arXiv preprint, arXiv:1804.05788.*, 2018
- [84] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., Verma, R., “CREMA-D: Crowd-sourced emotional multimodal actors dataset” *IEEE transactions on affective computing*, Vol. 5, Issue 4, pp.377-390, 2017
- [85] Martin, O., Kotsia, I., Macq, B., Pitas, I., “The enterface’05 audio-visual emotion database”, the 22nd International Conference on IEEE Data Engineering Workshops, p.8, Atlanta, GA, USA, 3–7, April, 2006
- [86] Schuller, B., Steidl, S., Batliner, A., “The interspeech 2009 emotion challenge” In *Proceedings of the Tenth Annual Conference of the International Speech Communication Association*, Brighton, UK, 6–10 September 2009.
- [87] Jackson, P., Haq, S., “Surrey Audio-Visual Expressed Emotion (Savee) Database” *University of Surrey: Guildford, UK*, 2014.

- [88] Livingstone, S. R., Russo, F. A., “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English” PloS one, Vol. 13, Issue 5, 2018
- [89] Wanare, M. A. P., Dandare, S. N., “Human Emotion recognition from speech. system”, Journal of Engineering Research and Applications, Vol. 4, Issue 7, pp.74-78, 2014
- [90] Posner, J., Russell, J. A., Peterson, B. S., “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology” Development and psychopathology, Vol. 17, Issue 3, pp.715-734, 2005
- [91] Joshi, A., Kaur, R., “A Study of speech emotion recognition methods” International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 2, Issue 4, pp.28-31, 2013
- [92] J. B. Russ, R. C. Gur, W. B. Bilker, “Validation of affective and neutral sentence content for prosodic testing.” Behavior Research Methods, Vol. 40, No. 4, pp.935-939, 2008
- [93] Hoens, T.R., Chawla, N.V., “Imbalanced Datasets: From Sampling to Classifiers” Imbalanced Learning: Foundations, Algorithms, and Applications; Wiley: Hoboken, NJ, USA, 2013
- [94] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., “The WEKA data mining software: An update”, ACM SIGKDD Explorations Newsletter, Vol. 11, Issue 1, pp.10-18, 2009
- [95] Chavhan, Y.D., Yelure, B.S., Tayade, K.N., “Speech emotion recognition using RBF kernel of LIBSVM” In Proceedings of the 2015 2nd International Conference on IEEE Electronics and Communication Systems (ICECS), pp. 1132-1135, Coimbatore, India, 26-27, February, 2015

Appendix: List of Publication

A-1. Journal Papers

SCI Journal Papers

- [1] Taeho Hur, **Jaehun Bang**, Thien Huynh-The, Jongwon Lee, Jee-In Kim and Sungyoung Lee, "Iss2Image: A Novel Signal-Encoding Technique for CNN-Based Human Activity Recognition", *Sensors* (SCIE, IF: 2.475), Vol.18, Issue 11, pp.1-19, 2018
- [2] **Jaehun Bang**, Taeho Hur, Dohyeong Kim, Thien Huynh-The, Jongwon Lee, Yongkoo Han, Oresti Banos, Jee-In Kim and Sungyoung Lee, "Robust Speaker Adaptation Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments", *Sensors* (SCIE, IF: 2.475), Vol.18, Issue 11, pp.1-21, 2018
- [3] Maqbool Ali, Syed Imran Ali, Dohyeong Kim, Taeho Hur, **Jaehun Bang**, Sungyoung Lee, Byeong Ho Kang, and Maqbool Hussain, "uEFS: An efficient and comprehensive ensemble-based feature selection methodology to select informative features", *PLoS ONE* (SCIE, 2.766), Vol.13 No.8, DOI: <https://doi.org/10.1371/journal.pone.0202705>, 2018
- [4] Jamil Hussain, Wajahat Ali Khan, Taeho Hur, Hafiz Syed Muhammad Bilal, **Jaehun Bang**, Anees Ul Hassan, Muhammad Afzal and Sungyoung Lee, "A Multimodal Deep Log-Based User Experience (UX) Platform for UX Evaluation", *Sensors* (SCIE, IF:2.677), Vol.18, Issue 5, pp.1-31, 2018
- [5] Maqbool Ali, Rahman Ali, Wajahat Ali Khan, Soyeon Caren Han, **Jaehun Bang**, Taeho Hur, Dohyeong Kim, Sungyoung Lee, and Byeong Ho Kang, "A data-driven knowledge acquisition system: An end-to-end knowledge engineering process for generating production rules", *IEEE Access* (SCIE, 3.244), Vol.6, pp.15587-15607,

2018

- [6] Thien Huynh-The, Cam-Hao Hua, Anh Tu Nguyen, Taeho Hur, **Jaehun Bang**, Dohyeong Kim, Muhammad B. Amin, Byeong Ho Kang, Hyonwoo Seung, Soo-Yong Shin, Eun-Soo Kim, Sungyoung Lee, "Hierarchical Topic Modeling With Pose-Transition Feature For Action Recognition Using 3D Skeleton Data", Information Sciences (SCI, IF:4.832), Vol.444, pp.20-35, 2018
- [7] Jamil Hussain, Anees Ul Hassan, Hafiz Syed Muhammad Bilal, Muhammad Afzal, Shujaat Hussain, **Jaehun Bang**, Oresti Banos, and Sungyoung Lee, "Model-based adaptive user interface based on context and user experience evaluation", Journal on Multimodal User Interfaces (SCIE, IF: 1.031), Vol.12, Issue 1, pp.1-16, 2018
- [8] Thien Huynh-The, Cam-Hao Hua, Anh Tu Nguyen, Taeho Hur, **Jaehun Bang**, Dohyeong Kim, Muhammad Bilal Amin, Byeong Ho Kang, Hyonwoo Seung and Sungyoung Lee, "Selective Bit Embedding Scheme For Robust Blind Color Image Watermarking", Information Science (SCI, IF:4.832), Vol. 426, pp.1-18, 2018
- [9] Muhammad Asif Razzaq, Claudia Villalonga, Sungyoung Lee, Usman Akhtar, Maqbool Ali, Eun-Soo Kim, Asad Masood Khattak, Hyonwoo Seung, Taeho Hur, **Jaehun Bang**, Dohyeong Kim and Wajahat Ali Khan, "mlCAF: Multi-Level Cross-Domain Semantic Context Fusioning for Behavior Identification", Sensors (SCIE, IF:2.677), Doi: 10.3390/s17102433., 2017
- [10] Taeho Hur, **Jaehun Bang**, Dohyeong Kim, Oresti Banos and Sungyoung Lee, "Smartphone Location-Independent Physical Activity Recognition Based on Transportation Natural Vibration Analysis", Sensors (SCIE, IF:2.033), Vol.17, Issue 4, 2017
- [11] Oresti Banos, Claudia Villalonga, **Jaehun Bang**, Taeho Hur, Donguk Kang, Sangbeom Park, Thien Huynh-The, Vui Le-Ba, Muhammad Bilal Amin, Muhammad Asif Razzaq, Wajahat Ali Khan, Choong Seon Hong and Sungyoung Lee, "Human Behavior Analysis by Means of Multimodal Context Mining", Sensors (SCIE, IF: 2.033), Vol.16, Issue 8, doi:10.3390/s16081264, 2016
- [12] Shujaat Hussain, **Jae Hun Bang**, Manhyung Han, Muhammad Idris Ahmed, Muhammad Bilal Amin, Chris Nugent, Sally McClean, Bryan Scotney, Gerard Parr

and Sungyoung Lee, "Behavior Life Style analysis for mobile sensory data in cloud computing through MapReduce", Sensors (SCIE, IF:2.048), Vol.14, Issue 11, pp.22001-22020 , 2014

- [13] Manhyung Han, **Jae Hun Bang**, Chris Nugent, Sally McClean, Sungyoung Lee , "A Lightweight Hierarchical Activity Recognition Framework using Smartphone Sensors", Sensors (SCIE, IF:2.048), Vol. 14, Issue 9, pp.16181-16195, 2014

Non-SCI Journal Papers

- [1] Muhammad Bilal Amin, Muhammad Sadiq, Maqbool Ali, **Jaehun Bang**, “Curating Big Data for Health and Wellness in Cloudcentric IoT”, The Journal of The Korean Institute of Communication Sciences, Vol.35, No.2, pp.42-57, 2018
- [2] Muhammad Hameed Siddiqi, Madallah Alruwaili, **JaeHun Bang** and Sungyoung Lee, “Real Time Human Facial Expression Recognition System using Smartphone”, International Journal of Computer Science AND Network Security, Vol.17, No.10, pp.223-230, 2017
- [3] Muhammad Bilal Amin, Wajahat Ali Khan, Bilal Ali, **Jaehun Bang**, Taqdir Ali, Taeho Hur, Shujaat Hussain, Imran Ali, Dohyeong Kim “Health and Wellness platforms: A Survey on Services and Enabling Technologies”, Communications of the Korean Institute of Information Scientists and Engineers, Vol.35 No.7 (Wn.338), pp. 9-25, 2017
- [4] Oresti Baonos, Wajahat Ali Khan, Muhammad Bilal Amin, Tae Ho Hur, **Jaehun Bang**, Donguk Kang, Maqbool Hussain, Muhammad Afzal, Taqdir Ali and Sungyoung Lee, “Big Data based Mining Minds Healthcare Framework”, Journal of Korea Institute Of Communication Sciences, Vol.32, No.11 , pp.12 -pp.20, 2015
- [5] **Jae Hun Bang** and Sungyoung Lee, “Adaptive Speech Emotion Recognition Framework Using Prompted Labeling Technique”, Journal of KIISE : Computing Practices and Letters Vol.21, Issue 2, pp.160-165, 2015
- [6] **Jae Hun Bang**, Sungyoung Lee, “Call Speech Emotion Recognition for Emotion based Services”, Journal of KIISE: Software and Applications, Vol.1, Issue 3, 2014

A-2. Conference Papers

International Conference Papers

- [1] Taqdir Ali, Sungyoung Lee, **Jaehun Bang**, Sun Moo Kang, Muhammad Bilal Amin, "Intelligent Medical Platform for clinical decision making", 15th APAN Research Workshop 2018, Auckland, New Zealand, Aug 6, 2018
- [2] **Jaehun Bang**, Dong Uk Kang, Tae Ho Hur, Thien Hyun The, Muhammad Asif Razzaq, Wajahat Ali Khan, Oresti Banos and Sungyoung Lee, "IoT based Human Centric Context Awareness Framework for Healthcare and Wellness Platform", 2016 International Symposium on Perception, Action and Cognitive Systems (PACS), Seoul, Korea, Oct 27-28, 2016
- [3] Wajahat Ali Khan, Muhammad Bilal Amin, Oresti Banos, Taqdir Ali, Maqbool Hussain, Muhammad Afzal, Shujaat Hussain, Jamil Hussain, Rahman Ali, Maqbool Ali, Dongwook Kang, **Jaehun Bang**, Tae Ho Hur, Bilal Ali, Muhammad Idris, Asif Razzaq, Sungyoung Lee and Byeong Ho Kang, "Mining Minds: Journey of Evolutionary Platform for Ubiquitous Wellness", 12th International Conference on Ubiquitous Healthcare (u-Healthcare 2015), Osaka, Japan, Nov 30- Dec 02, 2015
- [4] Oresti Banos, Jose Antonio Moral-Munoz, Manuel Arroyo-Morales, Hector Pomares, Ignacio Rojas, Claudia Villalonga, **Jae Hun Bang**, Dong Uk Kang, Choong Seon Hong and Sungyong Lee, "Facilitating Trunk Endurance Assessment by means of Mobile Health Technologies", Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015), Osaka, Japan, September 7-11, 2015
- [5] Oresti Banos, **Jaehun Bang**, Taeho Hur, Muhammad Hameed Siddiqi, Huynh-The Thien, Le-Ba Vui, Wajahat Ali Khan, Taqdir Ali, Claudia Villalonga and Sungyoung Lee, "Mining Human Behavior for Health Promotion", International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS2015), Milano, Italy, Augst 25-29, 2015

- [6] Banos, O., Bilal Amin, M., Ali Khan, W., Afzel, M., Ahmad, M., Ali, M., Ali, T., Ali, R., Bilal, M., Han, M., Hussain, J., Hussain, M., Hussain, S., Hur, T. H., **Bang, J. H.**, Huynh-The, T., Idris, M., Kang, D. W., Park, S. B., Siddiqui, M., Vui, L. B., Fahim, M., Khattak, A. M., Kang, B. H. and Lee, S, "An Innovative Platform for Person-Centric Health and Wellness Support", Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2015), Granada, Spain, April 15-17, 2015
- [7] Ba-Vui Le, **Jae Hun Bang** and Sungyoung Lee, "Hierarchical Emotion Classification using Genetic Algorithms", The 4th Symposium on Information and Communication Technology(SoICT '13), Danang, Vietnam, December 5-6, 2013
- [8] Manhyung Han, **Jae Hun Bang**, Chris Nugent, Sally McClean and Sungyoung Lee, "HARF: A Hierarchical Activity Recognition Framework using Smartphone Sensors", 7th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2013), Guanacaste, Costa Rica, December 2-6, 2013 (Best Paper Award)
- [9] Shujaat Hussain, Manhyung Han, **Jae Hun Bang**, Chris Nugent, Sally McClean, Bryan Scotney , Gerard Parr and Sungyoung Lee, "Activity recognition and resource optimization in mobile cloud through MapReduce", 15th International Conference on E-Health Networking, application & service, Lisbon, Portugal, October 9-12, 2013

Domestic Conference Papers

- [1] Hojun Lim, Sukkyun Jung, **Jaehun Bang** and Sungyoung Lee, "User's Location Context Inference System for Advanced LBS in smart phone environment", KIISE Korea Computer Congress 2017 (KCC 2017), pp.1756-1758, Jeju, Korea, 18-20, June, 2017
- [2] Hyuntae Kang, **Jaehun Bang**, Yongkoo Han and Sungyoung Lee "Guide Service System of Personalized Walking Course", KIISE Korea Computer Congress 2017 (KCC 2017), pp.1756-1758, Jeju, Korea, 18-20, June, 2017
- [3] **Jaehun Bang** and Sungyoung Lee, "Cosine Similarity based Similar Speech

- Integration for Efficient Annotation of Emotional Speech from Target User”, KISSE Winter Conference, Pyeong Chang, Korea, 21-23, Dec, 2016
- [4] **Jae Hun Bang** and Sungyoung Lee, “Image based Real-time Emotion Recognition Framework on Smartphone Environment”, KIISE Korea Computer Congress 2015 (KCC 2015), pp.443-pp.445, Jeju, Korea, 25-27, June, 2015
 - [5] **Jae Hun Bang** and Sungyoung Lee, “Personalized Speech Emotion Recognition Frameworks using Prompted Labeling Technique”, KIISE Korea Computer Congress 2014 (KCC 2014), pp.423-pp.425, Busan, Korea, 25-27, June, 2014 (Best Presentation Award)
 - [6] **Jae Hun Bang**, Sungyoung Lee and Taechung Jung “Speech Emotion Recognition Framework on Smartphone Environment” Korea Information Processing Society (KIPS 2013), Busan, Korea, 10-11, May 2013
 - [7] **Jae Hun Bang**, Chan Min Jung and Sungyoung Lee, “Speech Emotion Recognition using Tilted-Time Window in the Smartphone”, Korea Information and Communications Society Conference 2012 (KICS 2012), Yong Pyung, Korea , 8-10, Feb, 2012
 - [8] Chan Min Jung, Hyun Woo Kim, Chang Hyun Lee, **Jae Hun Bang** and Sungyoung Lee, “Personalized Application Recommendation System using Emotion Recognition on Smartphone” , Korea Information and Communications Society Conference 2012 (KICS 2012), Yong Pyung, Korea , 8-10, Feb 2012
 - [9] Sung Ho Lee, **Jae Hun Bang** and Sungyoung Lee, “User Emotion Extraction Engine(E3) based on Fuzzy Inference and Bayesian Networks in Smart Phone Environment”, KIISE Korea Computer Congress 2011, Kyung Ju, Korea, June, 30-July.2, 2012 (Best Paper Award)

A-3. Patents Registration

- [1] Sungyoung Lee, Hosung Lee, **Jaehun Bang**, “Apparatus and method for real-time of activity and posture based on combined sensors data”, Patent Registration: 10-1584458, January, 5th, 2016, Korea
- [2] **Jae Hun Bang**, Sungyoung Lee, Taechung Jung, Jung Hoon Cha “Method for estimating user emotion from inputted string” Patent Registration: 10-1536051, July, 6th, 2015, Korea
- [3] **Jae Hun Bang**, Sungyoung Lee, Taechung Jung, Jung Hoon Cha “Method for estimating user emotion based on call speech” Patent Registration: 10-1449856, Oct, 2nd, 2014, Korea

Korean Abstract (국문 초록)

전통적인 음성기반 감정인식 연구는 여러 사용자로부터 감성음성을 수집하여 범용적인 훈련 모델을 생성하고 이를 인지 프로세스에 적용한다. 전통적인 감정인식 방법들은 사용자마다 감정인식 정확도의 편차가 크다는 문제가 있다. 이러한 정확도 편차를 해결하기 위해 각 대상 사용자에게 대한 개인화된 모델을 생성하여 제공하는 개인화된 감정인식 연구가 활발히 진행되고 있다. 기존의 개인화된 음성 감정인식 연구는 정확한 개인화된 훈련 모델을 생성하기 위해서는 대상 사용자에게 충분하고 균형잡힌 개인화 감성음성 데이터를 요구하는 콜드 스타트 문제가 있다. 콜드 스타트 문제가 발생하는 환경으로는 데이터가 개인화 모델을 생성하기에 충분히 수집되지 않은 ① 적은 데이터 환경, 특정 감정이 수집되지 않은 ② 부재 데이터 환경, 수집된 감정 데이터 개수의 편차가 큰 ③ 불균형한 데이터 환경의 3가지 환경이 있다. 이러한 콜드스타트 문제를 해결하기 위한 다양한 기존 적응형 모델 기법은 기존에 다수의 사용자로부터 구축되어진 초기 모델을 기반으로 수집되는 타겟 사용자의 감성 음성을 반영하여 점진적으로 모델을 변경하는 방식으로 동작한다. 이러한 연구들은 빠른 개인화 모델 생성에 초점을 맞춘 기법으로 적은 데이터 환경에서는 어느 정도 해결하는 반면, 부재 데이터 및 불균형한 데이터 환경에서는 제대로 대응하지 못해 개인화된 감성음성 수집의 초기 단계에서 안정적인 개인화된 훈련 모델 생성이 어렵다는 단점이 있다.

따라서 본 논문에서는 이러한 3가지 데이터 환경에서 발생하는 콜드 스타트 문제를 해결하는 Robust Speaker Adaptation Framework를 제안한다. 제안하는 기술은 적은 데이터 환경에서 기존 데이터 셋에서 더 많은 실제 사례 데이터를 선택하는 ① 최대임계거리 기반의 유사 데이터 선택 기법, 수집된 감성음성 특징벡터 데이터 분포를 타 사용자간의 특징벡터 데이터 분포를 비교하여 가장 감성음성이 비

슷한 사용자를 추출하여 타겟 사용자의 부재 데이터 부분을 유사 사용자의 감성음성으로 대체하는 ② 데이터 분포요소 기반 유사 사용자 음성 매핑 기법, 불균형한 데이터 환경을 개선하기 위해서 Oversampling 알고리즘인 SMOTE (Synthetic Minority Over-Sampling Technique)을 반복적으로 활용하여 가상 데이터를 생성하는 ③ SMOTE 기반 가상 데이터 생성 기법으로 구성된다.

제안된 프레임워크는 제안하는 솔루션들을 결합하여 수집된 대상 사용자 음성과 다른 사용자의 음성을 결합하여 대상 사용자에게 유용한 실제 사례 데이터를 선택하고 SMOTE 기반으로 가상 데이터를 생성하여 개인화된 훈련 데이터 셋을 강화함으로써 대상 사용자에 대한 개인화된 훈련 모델을 점진적으로 제공한다. 제안된 방법은 IEMOCAP (Interactive Emotional Dyadic Motion Capture)와 CREMA-D (CRowd-sourced Emotional Multimodal Actors Dataset) 공공 감성 음성 데이터베이스를 사용하여 기존 기법과 비교실험을 통해 제한된 데이터 환경에서 기존 기법보다 빠르게 개인화된 모델 생성하고 감성음성 초기 단계에서도 충분하고 균형 잡힌 개인화된 훈련 모델을 제공함으로써 전체 구간에서 안정적인 개인화 모델 제공이 가능함을 입증하였다.

핵심어: 화자적응; 기계학습; 적응형 모델; 데이터 선택; 데이터 생성; 음성신호분석

방 재 훈
(컴퓨터공학 전공) 공학박사
경희대학교 대학원
지도교수: 이 승 룡