Ph.D Dissertation Presentation

3rd , May, 2019

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

Jaehun Bang

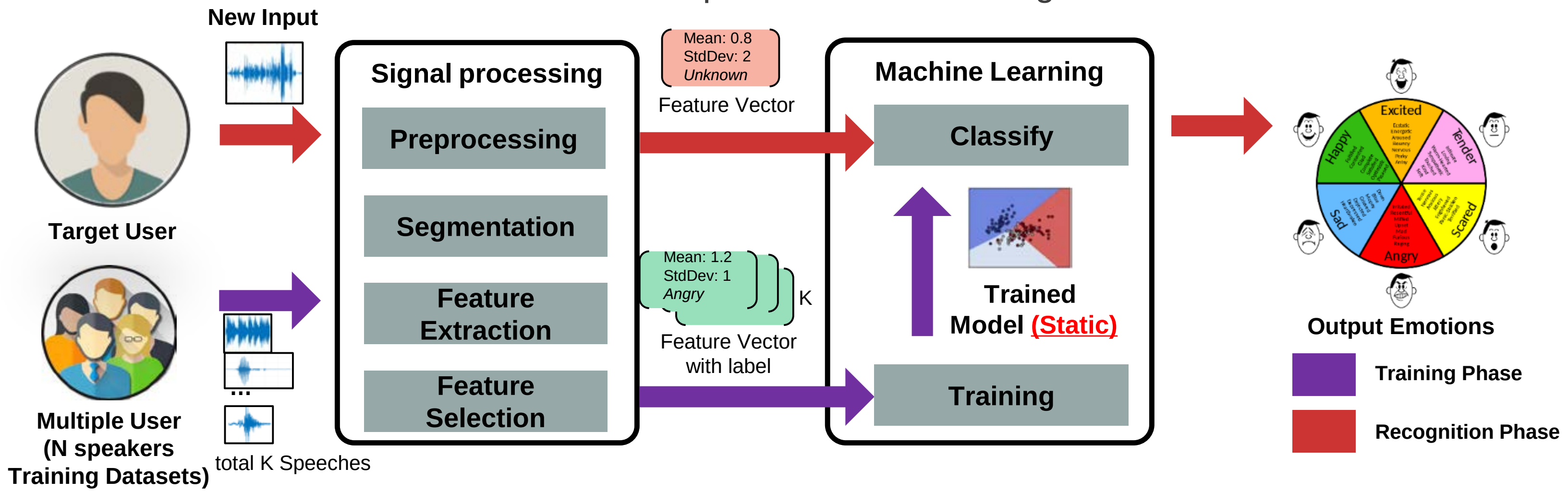
Department of Computer Science and Engineering
Kyung Hee University

Advised by
Prof. Sungyoung Lee, PhD

Table of Contents

Background

Traditional Speech Emotion Recognition

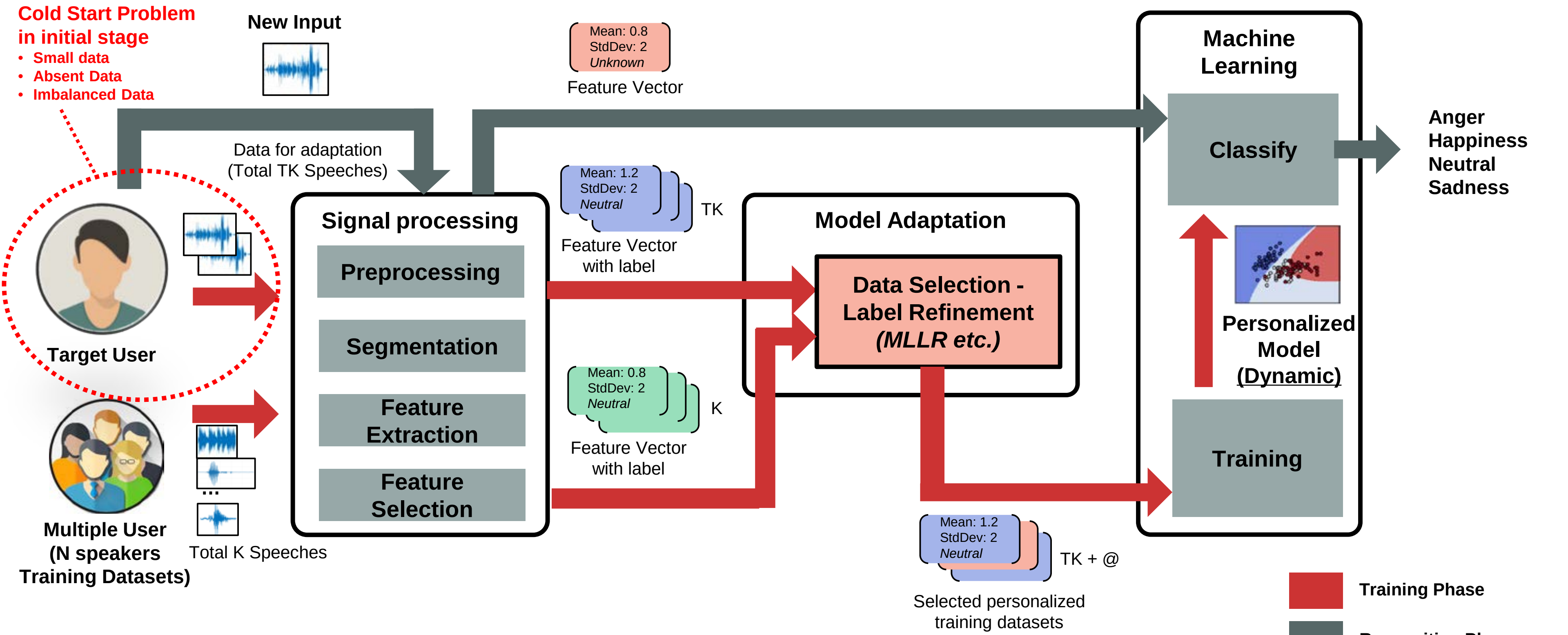


- **Limitation of Traditional Frameworks**
  - Performed **low accuracy in speaker independent evaluations**
  - **Impossible to modify training model** due to implement by static model

- **Recently, the emotion recognition researches are studying on creating a personalized emotion recognition model suitable for target user [1]**

Background

Speaker adaptation for personalized emotion recognition



MLLR: Maximum Likelihood Linear Regression

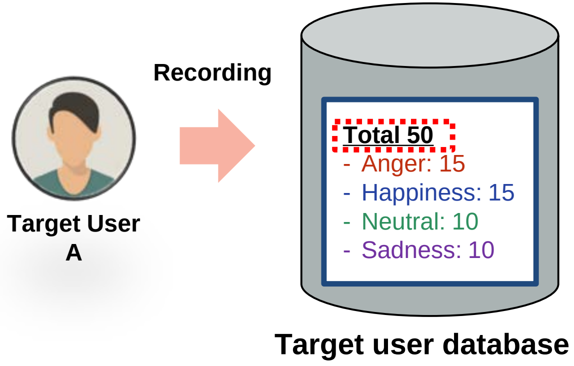
Motivation & Problem Statement

Issues in the personalized emotion recognition

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. **(Cold-Start Problem)**

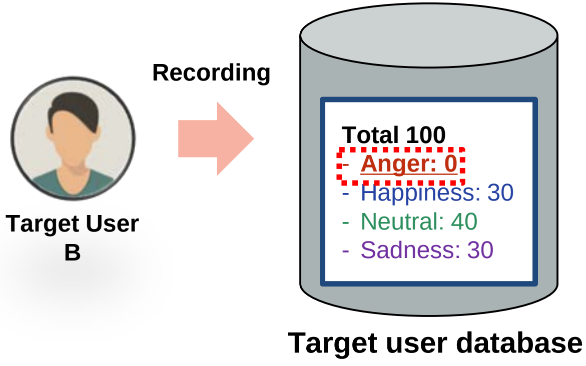
◆ Small data

- Insufficient amount of data to create personalized model



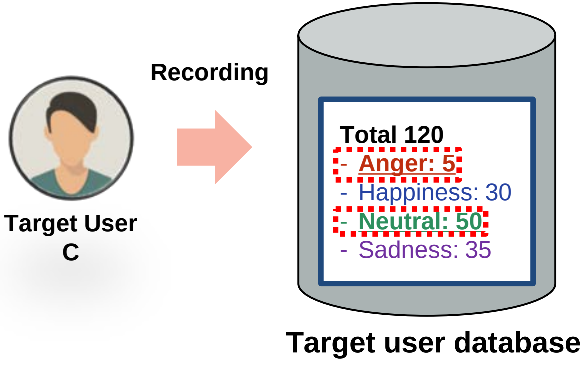
◆ Absent data

- Impossible to reflect personalized model about absent emotion



◆ Imbalanced data

- Uneven accuracy occurs between minor class and major class



Related works

Personalized emotion recognition comparison

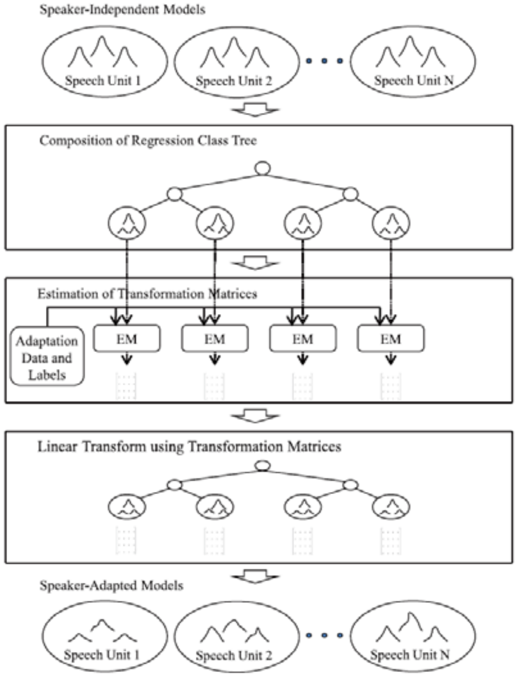
- Proposed Methodology in comparison with other approaches

Categories	Methodologies	Small Data Environment	Absent Data Environment	Imbalanced Data Environment	Emotions
		3 cold-start problems			
Small & Absent Data	conventional MLLR [2]	X (about 700 data required)	△ (Utilize Initial Model)	X	Neutral, Anger, Happiness, Sadness
	MLLR-SLR [3]	X (about 700 data required)	△ (Utilize Initial Model)	X	Neutral, Anger, Happiness, Sadness
	LDM-MDT MLLR [4]	△ (about 360 data required)	△ (Utilize Initial Model)	X	Neutral, Anger, Happiness, Sadness
	Incremental Adaptation [5]	△ (300 data required)	X	X	Neutral, Anger, Happiness, Sadness
	Domain Adaptation [6]	△ (Over 200 data required)	X	X	Arousal, Valance
Small & Imbalanced Data	Iterative Feature Normalization [7]	△ (Over 400 data required)	X	△	Neutral, Emotional
Imbalanced Data	SMOTE [8]	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**

Related works

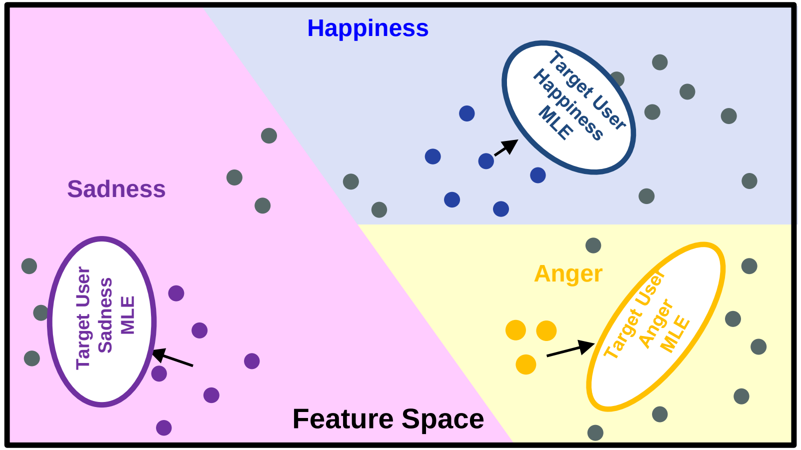
MLLR(Maximum Likelihood Linear Regression) based Model Adaptation [2]

MLLR based Model Adaptation



[Procedure for the conventional MLLR adaptation]

- MLLR Adaptation updates the Linear parameters of existing models based on acquired target user data.
- This approach **requires sufficient target user data** [9] to modify to personalized linear parameter value due to utilization all of the existing model data.



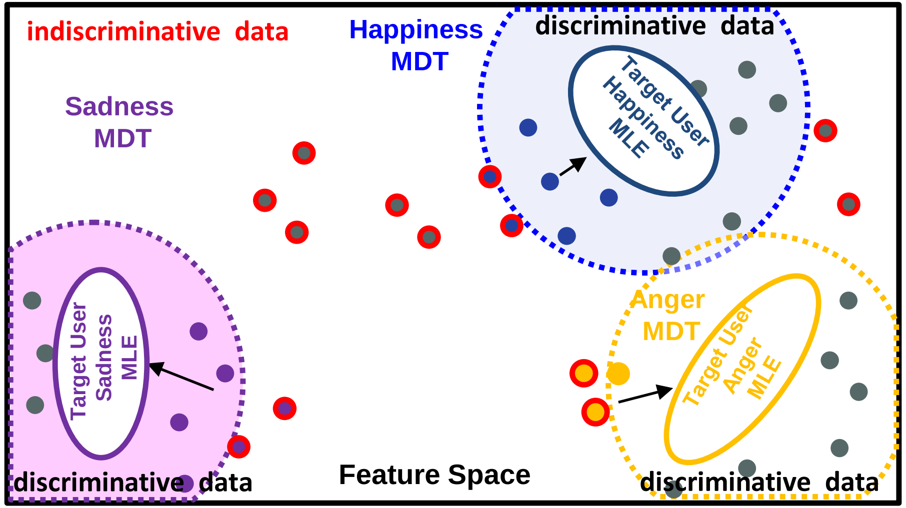
<Example of MLLR Adaptation>

Related works

“Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition”

Engineering Applications of Artificial Intelligence, Volume 52, p.126-134, June 2016

LDM-MDT MLLR based Data Selection [4]



<LDM–MDT MLLR Based Data Selection Example>

- This paper solved conventional MLLR adaptation problem
- Select useful data selection for target user from the initial model by discarding indiscriminative emotion data based on MDT after MLLR based global adaptation process
- Approximately **half of all of user adaptation data are determined to be indiscriminative and are disregarded**.

1. Compute MLE(means) value with new data

$$\hat{\mu_i} = W_i\mu_i = \frac{\sum_x p(i|x,\pi_i,\mu_i,\Sigma_i)x_i}{\sum_x p(i|x,\pi_i,\mu_i,\Sigma_i)}$$

2. LDM based Data Selection

$$LDM(X_i) = \frac{1}{E-1}\sum_{r=1}^{E-1}\{\log P(X_i|\lambda_{R_r(X_i)}) - \log P(X_i|\lambda_{R_{r+1}(X_i)})\}^2$$

3. MDT based Indiscriminative data classification

$$MDT(\lambda_{R_1(X_i)}) = \frac{1}{T_e}\sum_{j=1}^{T_e} LDM(X_j)$$

- **Still requires sufficient target user data (about 360 samples)**
- **If absence data exists, utilize Initial model (Imbalanced Problem)**
- **There is no process to solve imbalanced data problem (Uneven Accuracy)**

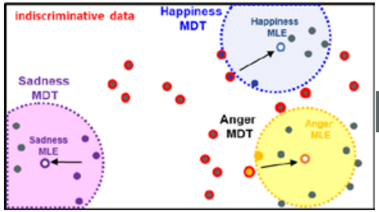
LDM: Log-likelihood Distance based confidence Measure
MDT:Model based Dynamic Threshold, MLE: Maximum Likelihood Estimation

Proposed Idea

Existing Method [4]

Small data

- Due to **global adaptation and small threshold range,** small amount of data is selected

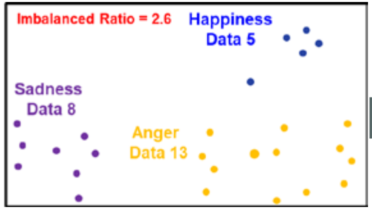


Absent Data

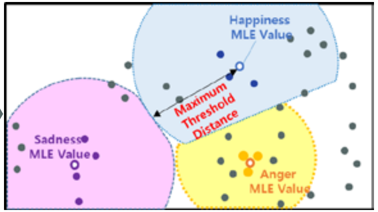
- **Utilize all of the existing emotional data set** for absent data

Imbalanced Data

- **There is no Imbalanced solution process**, it depends on the class of initial data ratio



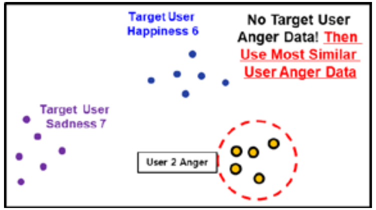
Proposed Solutions



Small data

Similar data selection based on MTD (Solution 1)

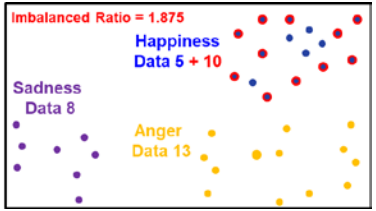
- Select relevant data based on more centroid and large range to target user



Absent Data

Other similar user emotional speech mapping (Solution 2)

- Reinforcement absent data area to extracted similar user emotional data



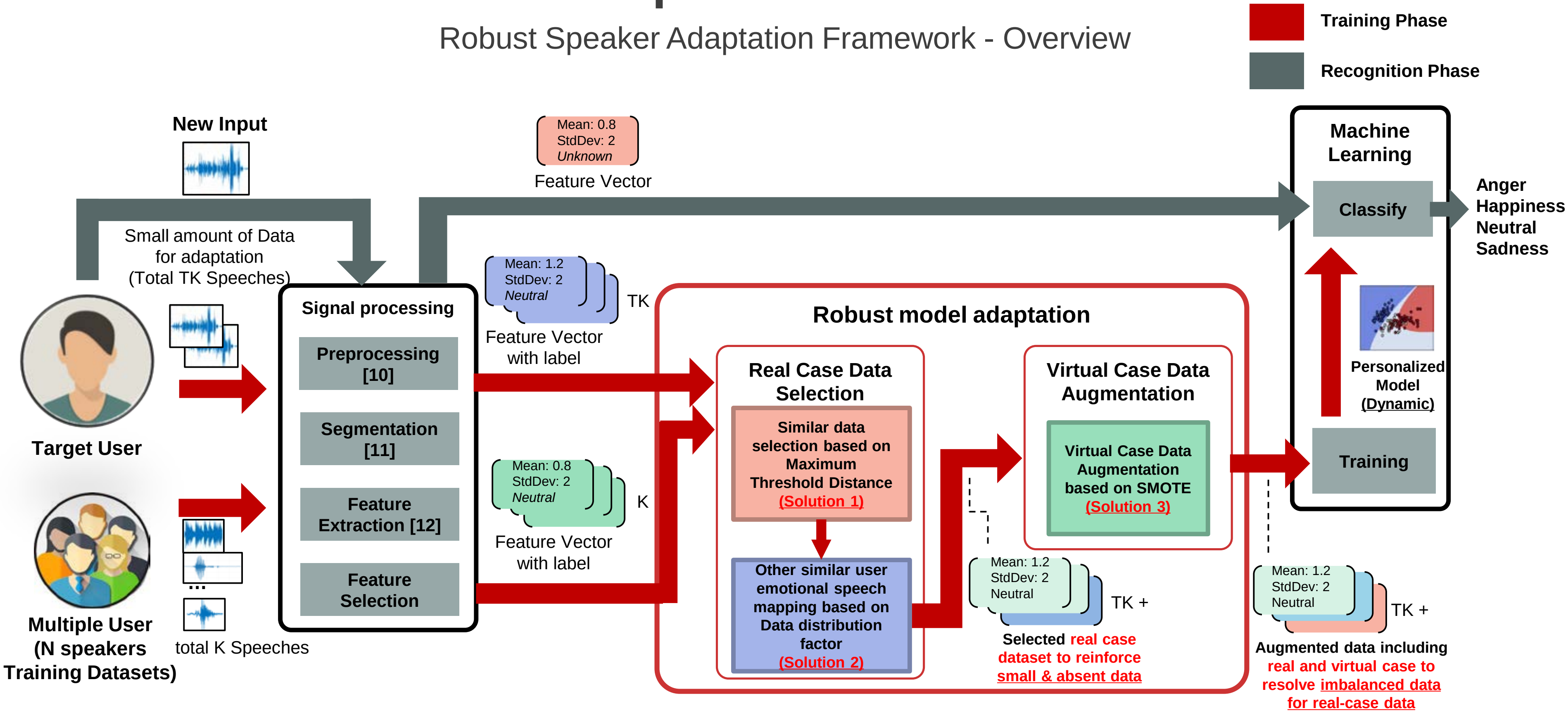
Imbalanced Data

Virtual Case Data Augmentation based on SMOTE (Solution 3)

- Mitigate imbalanced ratio through Iterative SMOTE process

Proposed Idea

Robust Speaker Adaptation Framework - Overview



[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.

Proposed Idea

Model Adaptation Comparison

Target User Speech DB

Initial Multiple User Speech DB

Existing Method [4]

Signal Processing

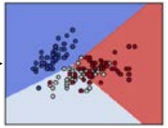
Proposed Method

LDM-MDT MLLR based Model Adaptation [4]

MLLR based Global Adaptation

LDM-MDT based Data Selection

Gaussian Mixture Model



Personalized Model

- Select **only** similar **real-case data** based on **modified model from existing data set**

Robust Model Adaptation

Real-case Data Selection

Similar data selection (Solution 1)

MLE & MTD Estimation

Similar Data Selection

Other similar user emotional speech mapping (Solution 2)

EDS based Similar User Estimation

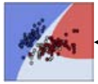
Absence Data Mapping

Virtual-Case Data Augmentation

Virtual Case Data Augmentation based on SMOTE (Solution 3)

SMOTE based Imbalanced Data Reinforcement

Machine Learning



Personalized Model

- Augment sufficient balanced personalized data set consisting of **real-case data & virtual case data** based on **user centric data.**

Proposed Idea

Problem statements / Goal / Challenges

◆ Problem Statements

- Creating personalized emotion recognition model is very difficult in limited data environments such as having **① small data, ② absent data and ③ imbalanced data (Cold Start problems)**

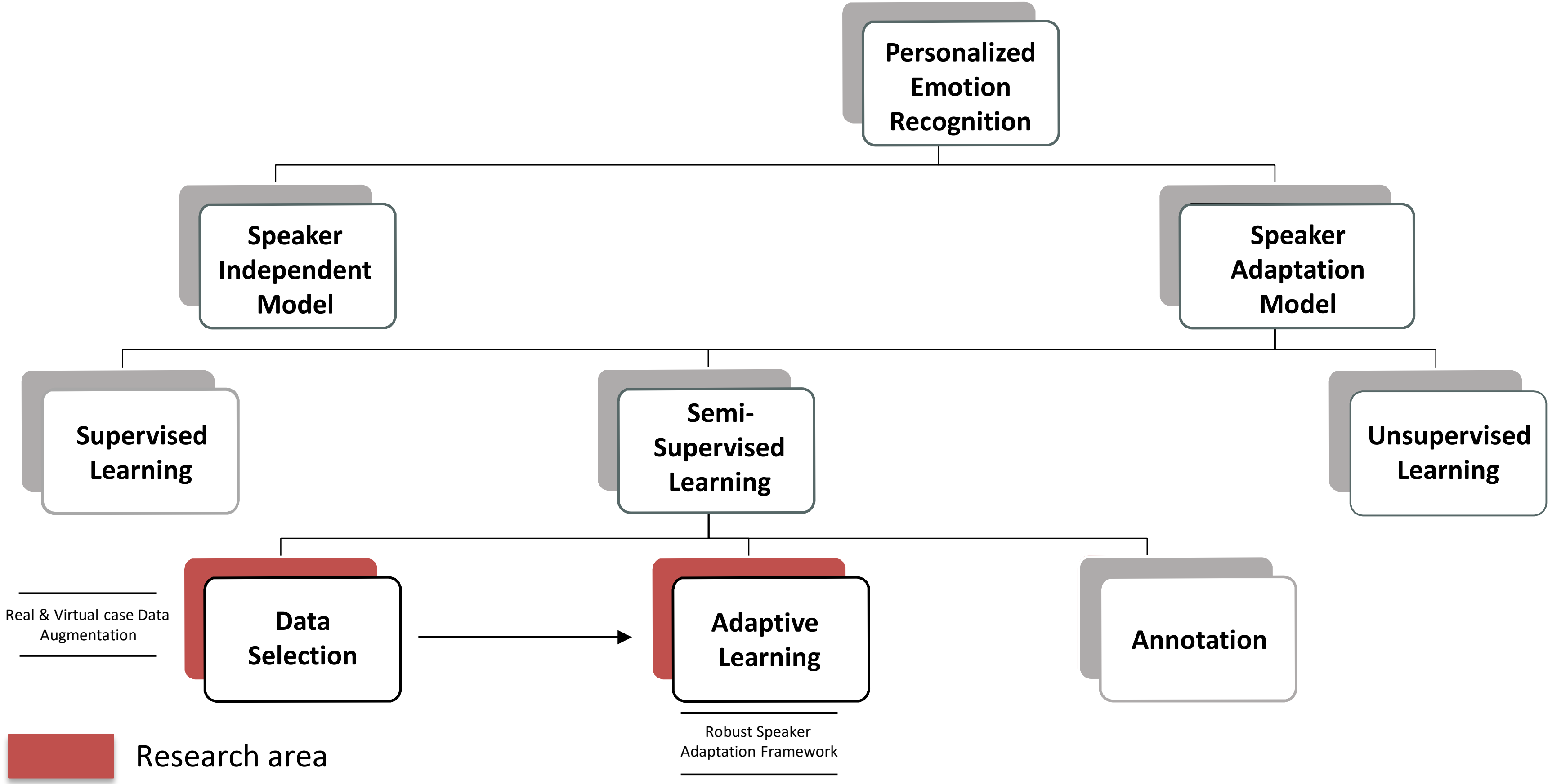
◆ Goal

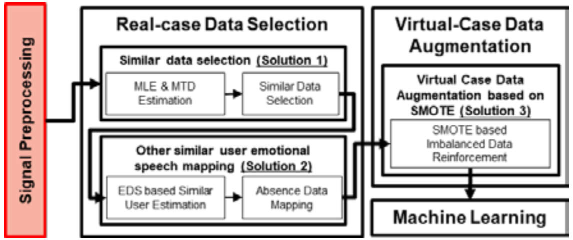
- Research the process and methodologies to create personalized emotion model to solve cold-start problems

◆ Challenges

- **Challenge 1** – Increasing target user oriented training data set for **small data**
- **Challenge 2** – Reinforcing **absent data** to target user relevant data
- **Challenge 3** – Solving **imbalanced data** problem from selected real-case dataset

Research Taxonomy





Methodologies

Preprocessing & Feature Extraction

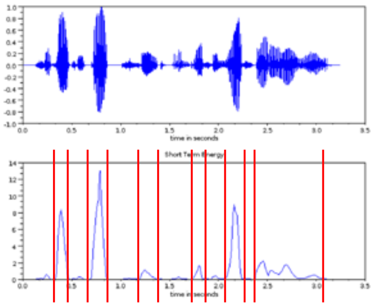
Preprocessing

Peak based Volume Normalization [10]

- The default approach to adjusting the data value based on the highest signal level present in the audio

STE based Silent Removal [11]

- This approach divides audio into frames, where each duration is segmented in 15 ms by a hamming window. Then, speech boundaries are estimated based on the short time energy (STE) algorithm.



Short-term Energy Transformation

$$e(n) = \sum_{m=-\infty}^{\infty} (s(m).w(n-m))^2$$

Threshold based Silent Removal

$$Tmin = 1 + 2\log_{10}\frac{Energy\max}{Energy\min}$$

$$Energy\max = \max(E(i)), i = 1,2, \ldots, M$$

$$Energy\min = \min(E(i)), i = 1,2, \ldots, M$$

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

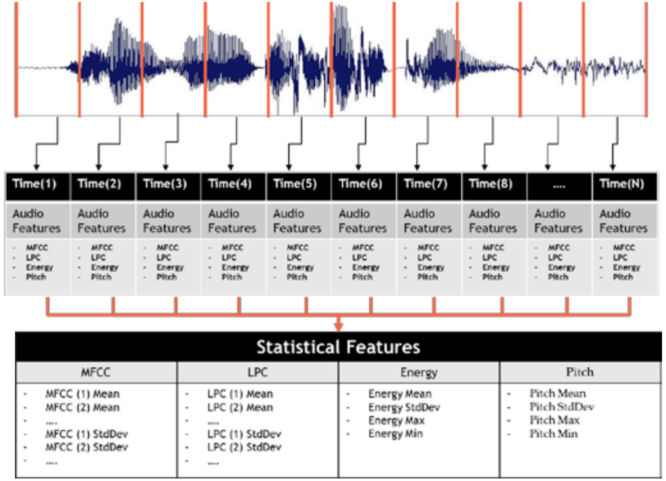
$$Tmax = Tmin - 0.25(SL - Tmin)$$

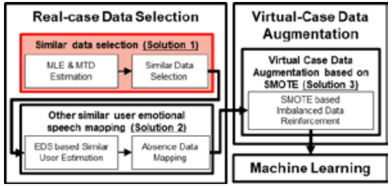
Feature Extraction

Statistical Feature Extraction [12]

- Extract the 100 statistical features with popular feature in SER
  - ✓ 13 MFCC - Mean, StdDev, Min, Max (13 x 4 = 52)
  - ✓ 10 LPC - Mean, StdDev, Min, Max (10 x 4 = 40)
  - ✓ Energy - Mean, StdDev, Min, Max (1 x 4 = 4)
  - ✓ Pitch - Mean, StdDev, Min, Max (1 x 4 = 4)

MFCC - Mel frequency cepstral coefficient
LPC - Linear predictive coding

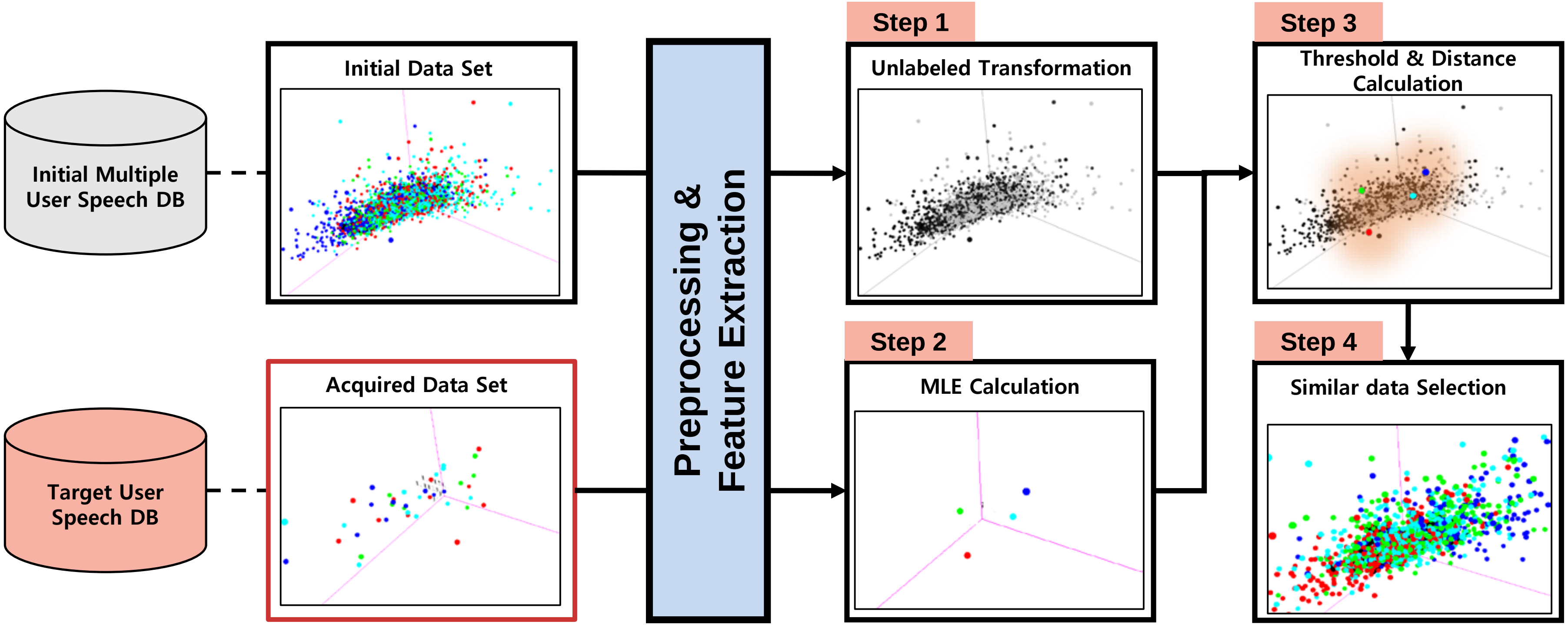


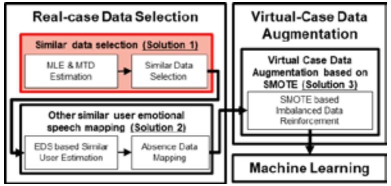


Solution 1 for small data

Similar data selection based on Maximum Threshold Distance

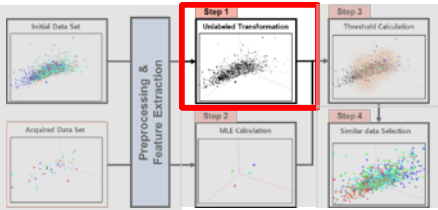
<u>**Reinforces the target user small data environment**</u> utilizing an initial constructed dataset





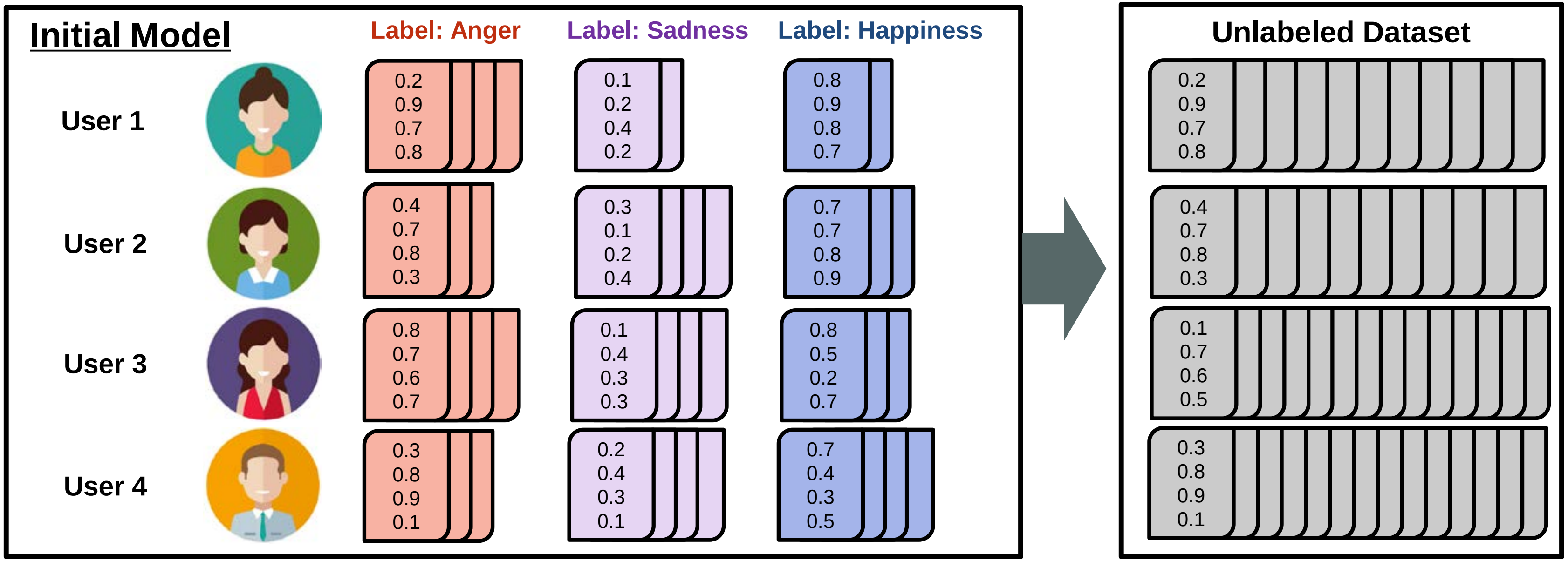
Solution 1 for small data

Step 1. Unlabeled transformation in initial model

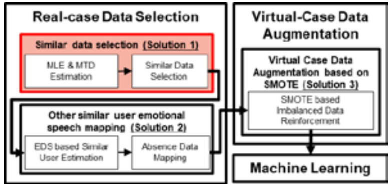


Use the unlabeled data to ignore label information in initial model

- **Unlabeled transformation**
  - The reason for using an unlabeled transformation is that **emotional expressions are different for each user.**
  - **The target user's particular emotional speech can be similar to different emotional speech in other users' emotional speech** when the acoustic pattern is almost the same. **(User 1 Happiness ≒ User 3 Anger)**

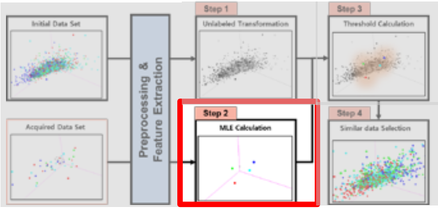


[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.

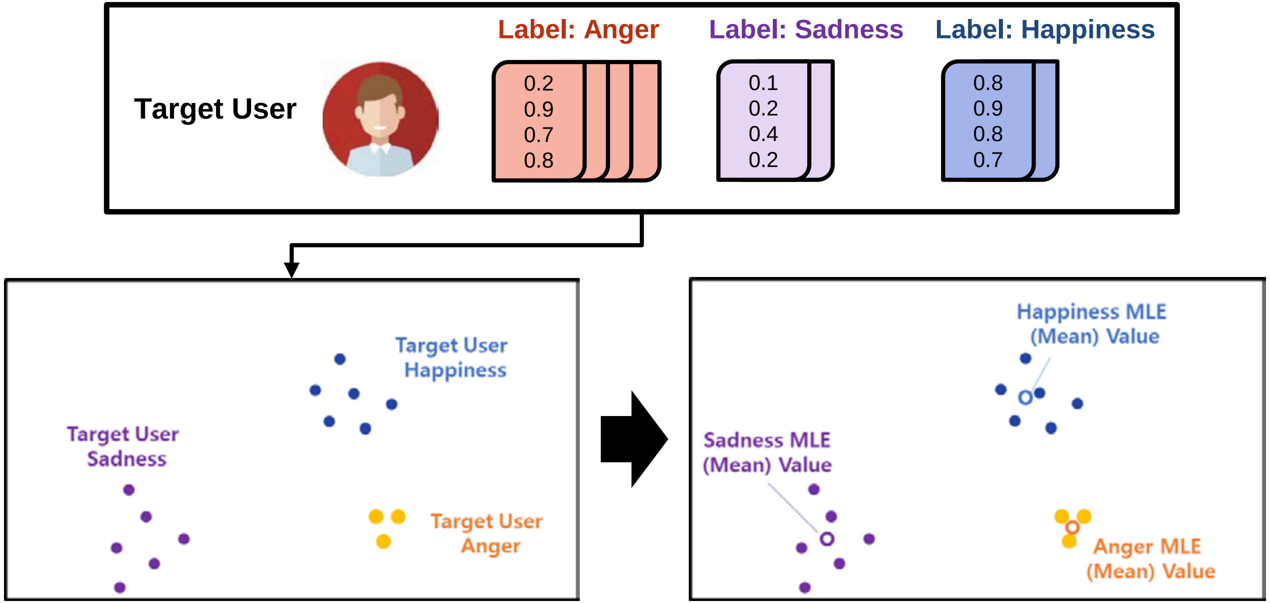


Solution 1 for small data

Step 2. MLE value calculation based on target user data



Calculate target user MLE value from feature vector based on only target user samples



Acquired Target User Speech **MLE Calculation**

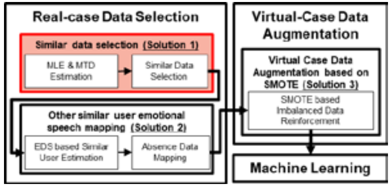
◆ **Target user MLE (Maximum Likelihood Estimation) Calculation**

MLE (means) Calculation

$$TMLE_{ei} = \frac{1}{N}\sum_{j=1}^{N} TfeatureVector_{ji}$$

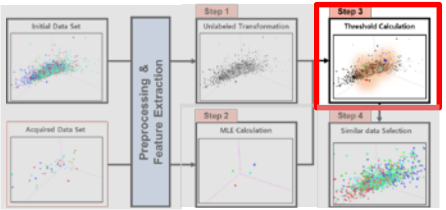
- ***TMLE*** is two-dimensional array that stores the average value of the acquired target user emotion voice feature vectors
- ***e*** is the corresponding emotion index
- ***i*** is the index of the feature vector
- ***N*** is the number of data
- ***j*** is the index of the data
- ***TfeatureVector*** is the extracted statistical feature vector via signal processing.

[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.



Solution 1 for small data

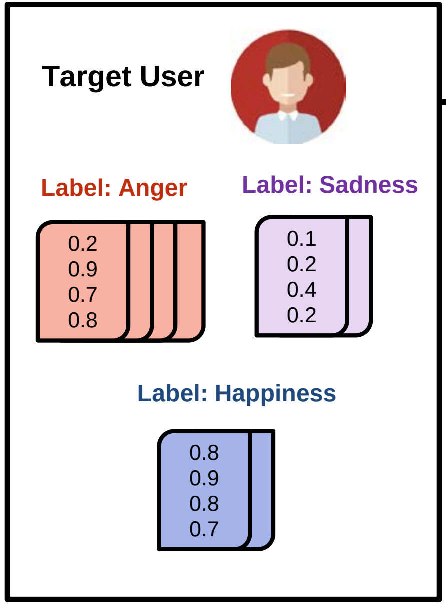
Step 3. Maximum Threshold Distance Calculation

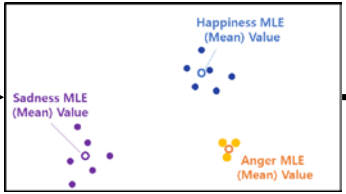


Set thresholds to select as much similar data as possible.

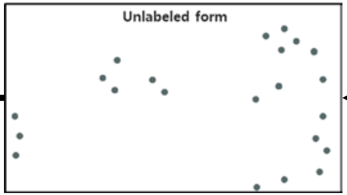
◆ Maximum Threshold Distance Calculation

- The Maximized Threshold value is computed by half of maximum distance of the means values and decide which data is discarded for data selection

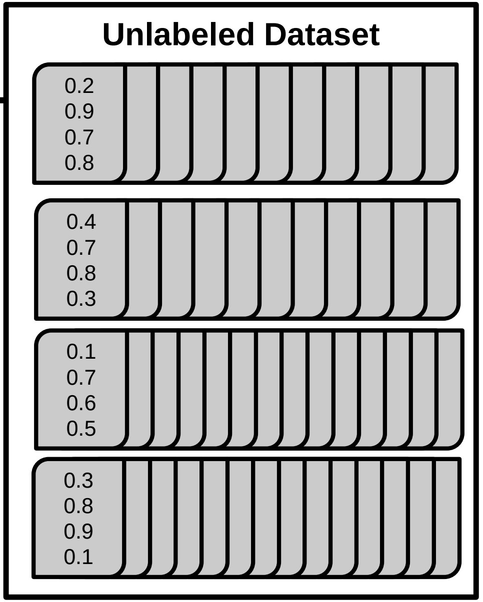


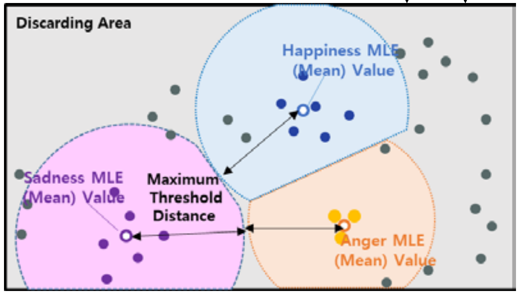


Step2. Extracted Target User MLE values



Step1. Unlabeled Dataset





Step3. Set Threshold value by Maximum Threshold Distance (MTD)

Compute Euclidean Distance

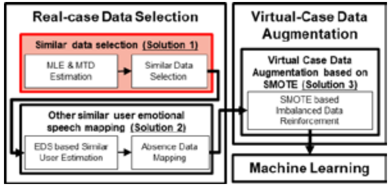
$$d\left(TMLE_{e_i}, TMLE_{e_j}\right) = \sqrt{\sum_{k=1}^{FN}\left(TMLE_{e_ik} - TMLE_{e_jk}\right)^2}$$

Compute MTD Estimation

$$MTD(TMLE_{ei}) = \frac{1}{2}argmax(d\left(TMLE_{e_i}, TMLE_{e_j}, \ldots\right))$$

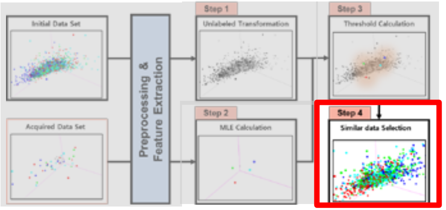
- ***TMLE*** is two-dimensional array
- e_i *and* e_j *are* the corresponding different emotion index
- ***k*** is the index of the feature vector
- ***FN*** *is the number of features*

[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.

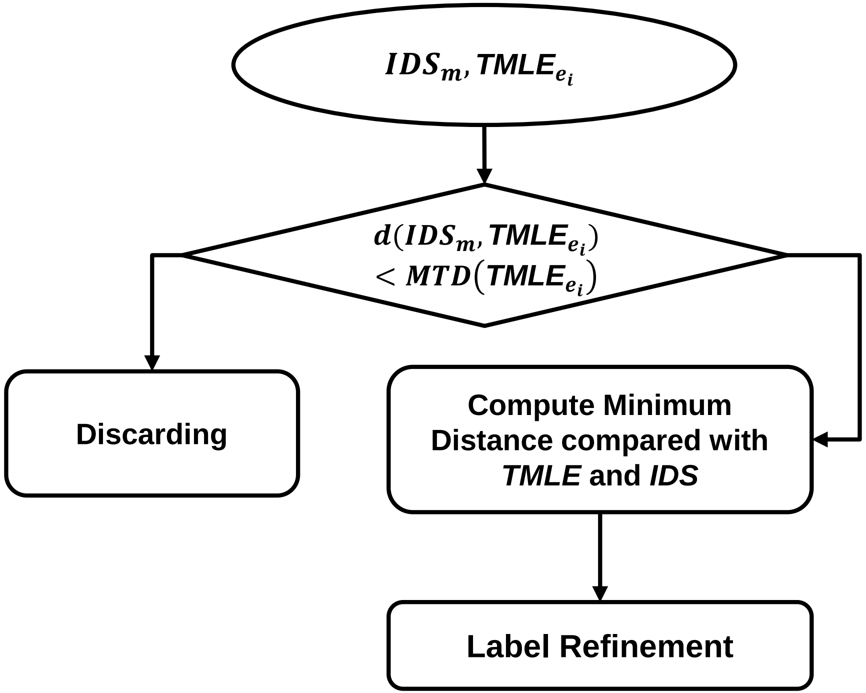


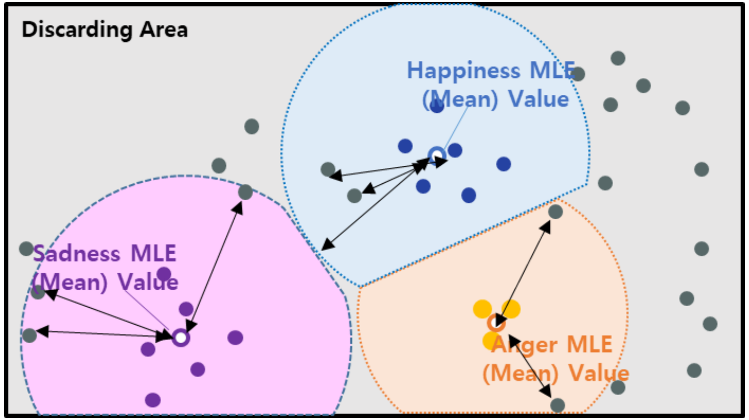
Solution 1 for small data

Step 4. Similar Speech Data Selection



The process of sequentially selecting similar data to reinforce the insufficient data according to distance is performed





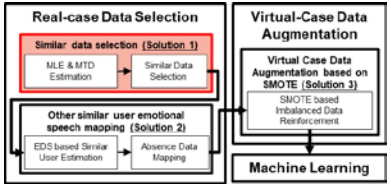
Compute Euclidean Distance
$d(TMLE_{e_i}, IDS_m) = \sqrt{\sum_{k=1}^{FN} (TMLE_{e_i k} - IDS_k)^2}$

- ***TMLE*** is two-dimensional array
- e_i is the corresponding emotion index
- IDS_m is the unlabeled initial dataset
- ***k*** is the index of the feature vector
- ***FN*** *is the number of features*

◆ **Similar Speech Data Selection**

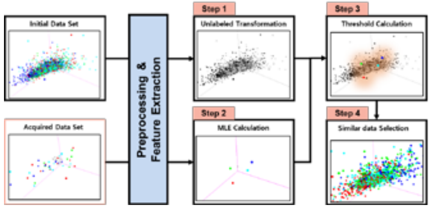
- Discarding area is not useful data to emotional speech model for target user
- The speech samples from the **user closest to the target speech mean value are selected.**

[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.



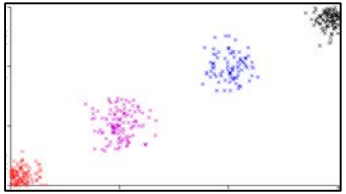
Solution 1 for small data

Detailed Algorithm Comparison

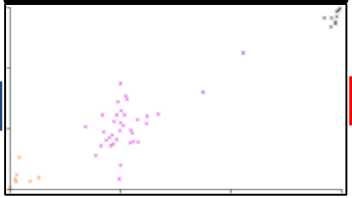


Existing Method [4]

- Target user Data – 27
- Selected Data - 270
  - ✓ Anger – 60
  - ✓ Sadness – 83
  - ✓ Happiness – 51
  - ✓ Neutrals - 76

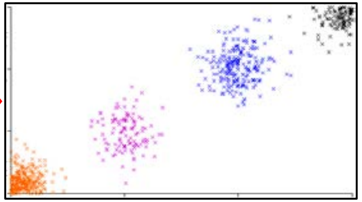


Target User Input data



<Randomly selected target user data 50 (IEMOCAP)>
Anger – 7, Sadness – 33, Happiness – 2, Neutral - 9

Proposed Method



- Target user Data – 50
- Selected Data - 721
  - ✓ Anger – 287
  - ✓ Sadness – 82
  - ✓ Happiness – 253
  - ✓ Neutrals - 99

Exist Algorithm 1. LDM-MDT MLLR based Data Selection

Input: $TDS(1 \dots N)$ – Target User Dataset
$IDS(1 \dots M)$ – Initial Multiple User Dataset

Output : S (1...K) – Selected Similar Emotional Speeches Dataset

MLE_e = *Calculate MLE* (TDS_e , IDS_e)
MDT_e = *Calculate MDT* (TDS_e, MLE_e) // *Average of* TDS_e *log-likelihood distance*
for i = 1 to M
  $Distance$ = *Calculate Log-likelihood Distance* (IDS_i)
  if $Distance$ <= MDT_e **then**
    $mEmo$ = *Calculate Minimum Distance* ($TMLE_e$, IDS_i)
    add S (IDS_i , $mEmo$)
  end
end
Return S

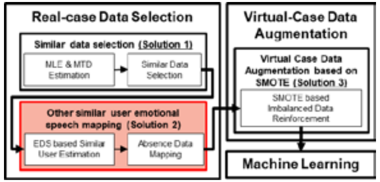
Proposed Algorithm 1. Similar speech data selection based on Maximum Threshold Distance

Input: $TDS(1 \dots N)$ – Target User Dataset
$IDS(1 \dots M)$ – Initial Multiple User Dataset

Output : S (1...K) – Selected Similar Emotional Speeches Dataset

$TMLE_e$ = *Calculate MLE* (TDS_e)
MTD_e = *Calculate MTD* (TDS_e, $TMLE_e$)
for i = 1 to M
  $Distance$ = *Calculate Euclidean Distance* (IDS_i)
  if $Distance$ <= MTD_e **then**
    $mEmo$ = *Calculate Minimum Distance* ($TMLE_e$, IDS_i)
    add S (IDS_i , $mEmo$)
  end
 end
end
Return S

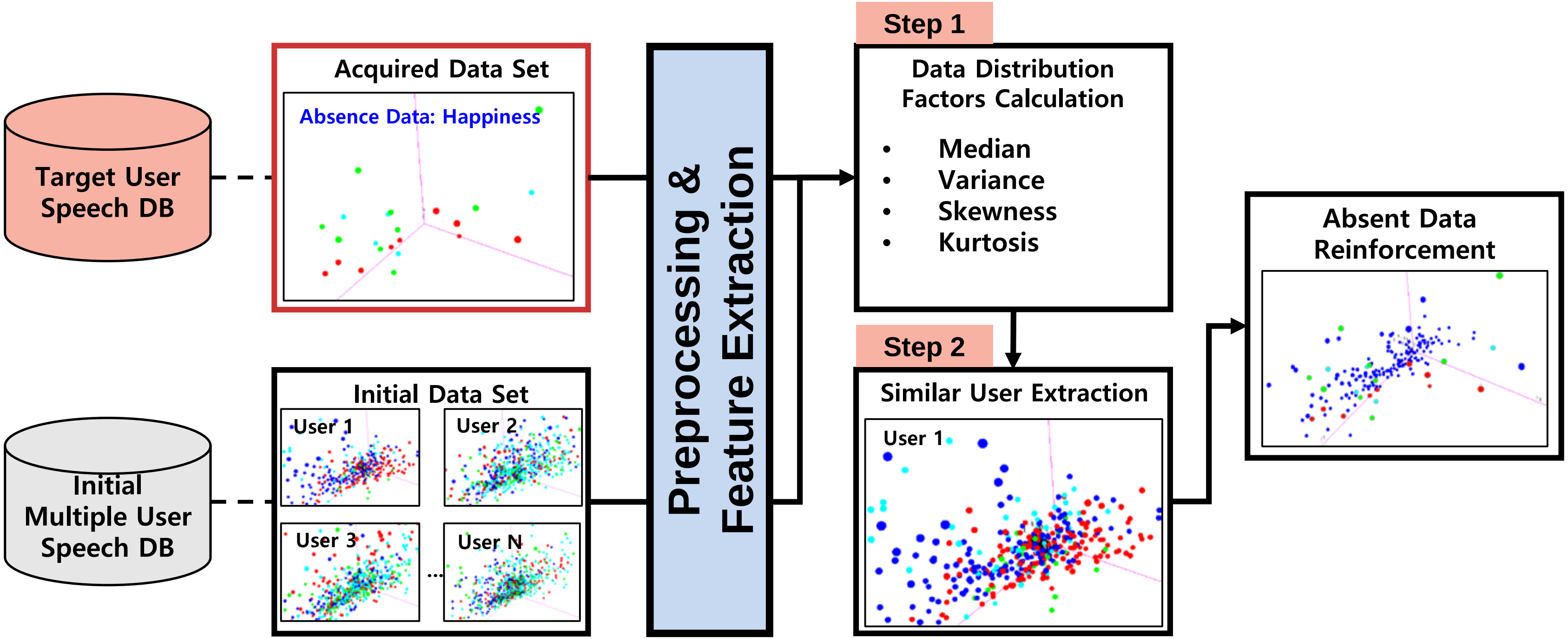
[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.



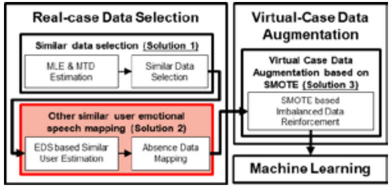
Solution 2 for absent data

Other similar user emotional speech mapping based on data distribution factor

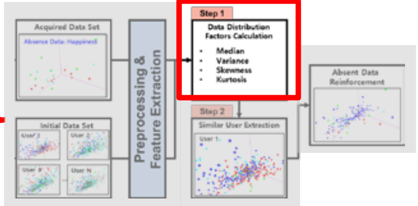
Reinforce absent data environment in target user emotional dataset to add the similar user emotional data



[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.



Solution 2 for absent data



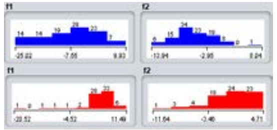
Step 1. Compute Statistical Data Distribution Factors

Calculate Data Distribution Factors without absent data part from target user

Acquired data

Target User

No Happiness



Anger

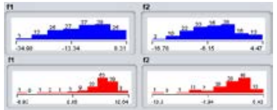
Distribution Factor	f1	f2
Median	-6.74	-5.91
Variance	72.0	13.6
Skewness	-0.29	-0.40
Kurtosis	-0.71	-0.78

Sadness

Distribution Factor	f1	F2
Median	4.46	0.35
Variance	19.7	9.76
Skewness	-3.25	-1.37
Kurtosis	14.6	2.47

Initial Model

User 1



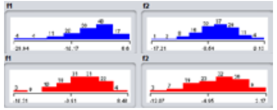
Anger

Distribution Factor	f1	f2
Median	-8.36	-4.78
Variance	99.9	18.4
Skewness	-0.30	-0.02
Kurtosis	-0.74	-0.34

Sadness

Distribution Factor	f1	f2
Median	6.54	2.29
Variance	7.64	7.57
Skewness	-2.29	-1.22
Kurtosis	8.04	2.16

User 2



Anger

Distribution Factor	f1	f2
Median	-3.33	-7.01
Variance	46.2	8.58
Skewness	-0.96	-0.34
Kurtosis	0.80	0.44

Sadness

Distribution Factor	f1	f2
Median	1.12	-3.66
Variance	11.5	8.95
Skewness	-0.70	-0.32
Kurtosis	1.03	-0.34

User N

⋮

◆ Data Distribution Calculation

- 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 data distribution.
- Utilize 4 most commonly used values for data distribution Factors including **median, variance, skewness, and kurtosis**

Data Distribution Factor Extraction

$$Median\left(SortedFeatureVector_{fi}\right) = SortedFeatureVector_{fi_{N/2}}$$

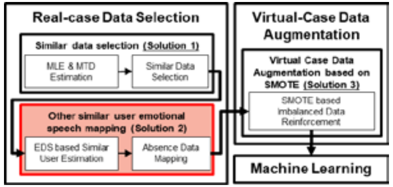
$$Variance\left(FeatureVector_{fi}\right) = \frac{1}{N}\sum_{k=1}^{N} \left(FeatureVector_{fi_k} - means_i\right)^2$$

$$Kurtosis\left(FeatureVector_{fi}\right) = \frac{\frac{1}{N}\sum_{k=1}^{N}\left(FeatureVector_{fi_k} - means_{fi}\right)^4}{\left(\frac{1}{N}\sum_{k=1}^{N}\left(FeatureVector_{fi_k} - means_{fi}\right)^2\right)^2} - 3$$

$$Skewness\left(FeatureVector_{fi}\right) = \frac{\frac{1}{N}\sum_{k=1}^{N}\left(FeatureVector_{fi_k} - means_{fi}\right)^3}{\left(\frac{1}{N}\sum_{k=1}^{N}\left(FeatureVector_{fi_k} - means_{fi}\right)^2\right)^{\frac{3}{2}}}$$

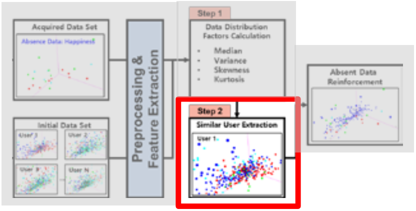
- fi is the index of the feature vector
- $FeatureVector_{fi}$ is particular feature values
- $mean$ is the average value in $FeatureVector_{fi}$
- N is the number of data

[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.



Solution 2 for absent data

Step 2. Similar user estimation and reinforcement



Replace the target user absent emotional data area to similar user's speeches

Acquired data

Target User

No Happiness

Anger

Distribution Factor	f1	f2
Median	-6.74	-5.91
Variance	72.0	13.6
Skewness	-0.29	-0.40
Kurtosis	-0.71	-0.78

Sadness

Distribution Factor	f1	F2
Median	4.46	0.35
Variance	19.7	9.76
Skewness	-3.25	-1.37
Kurtosis	14.6	2.47

Initial Model

User 1

Anger

Distribution Factor	f1	f2
Median	-8.36	-4.78
Variance	99.9	18.4
Skewness	-0.30	-0.02
Kurtosis	-0.74	-0.34

Sadness

Distribution Factor	f1	f2
Median	6.54	2.29
Variance	7.64	7.57
Skewness	-2.29	-1.22
Kurtosis	8.04	2.16

User 2

Anger

Distribution Factor	f1	f2
Median	-3.33	-7.01
Variance	46.2	8.58
Skewness	-0.96	-0.34
Kurtosis	0.80	0.44

Sadness

Distribution Factor	f1	f2
Median	1.12	-3.66
Variance	11.5	8.95
Skewness	-0.70	-0.32
Kurtosis	1.03	-0.34

User N ⋮

◆ Similar User Estimation

- Calculate each **Euclidean Distance Similarity** of data distribution factors to estimate similar user

Target User – User 1

Anger	f1	f2
Distance	26.0	5.22
Similarity	3.6	16
Sadness	**f1**	**f2**
Distance	16.3	5.08
Similarity	5.7	16.4

Avg. similarity = 10.48

Target User – User 2

Anger	f1	F2
Distance	66.4	10.4
Similarity	1.4	8.7
Sadness	**f1**	**f2**
Distance	12.5	2.23
Similarity	7.4	30.9

Avg. similarity = 12.14

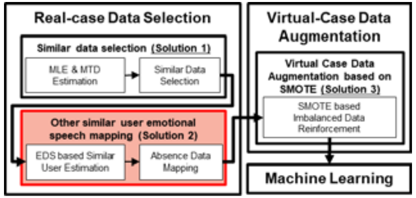
Euclidean Distance Similarity

$$d\left(TDDF_{e_i}, IDSDDF_{u_{e_i}}\right) = \sqrt{\sum_{k=1}^{DFN}\left(TDDF_{e_i k} - IDSDDF_{u_{e_i} k}\right)^2}$$

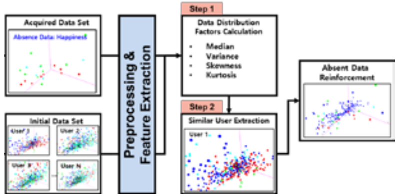
$$Similarity\left(TDDF_{e_i}, IDSDDF_{u_{e_i}}\right) = \frac{1}{1+d(TDF_{e_i}, IDSDDF_{u_{e_i}})} * 100$$

- $TDDF_{e_i}$ is the target user data distribution factors
- $IDSDDF_{u_{e_i}}$ is the initial model data distribution factors
- DFN is the number of data distribution factors

[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.



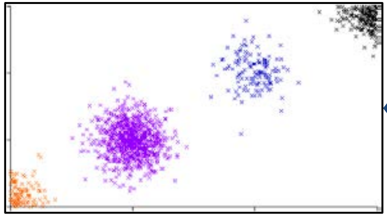
Solution 2 for absent data



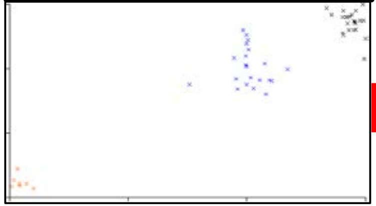
Detailed Algorithm Comparison

Existing Method [4]

- Target user Data – 26
- Selected Data – 1,605
  - ✓ Anger – 101
  - ✓ Sadness – 1,216
  - ✓ Happiness – 120
  - ✓ Neutrals - 168

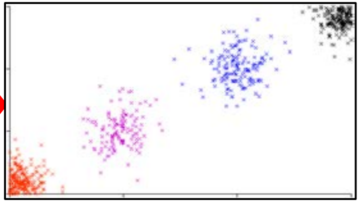


Target User Input data



<Randomly selected target user data 50 (IEMOCAP)>
Anger – 7, Sadness – 0, Happiness – 20, Neutral - 23

Proposed Method



- Target user Data – 50
- Selected Data - 627
  - ✓ Anger – 205
  - ✓ User 2 Sadness - 139
  - ✓ Happiness – 132
  - ✓ Neutrals - 151

Exist Algorithm 2. LDM-MDT MLLR based Data Selection in absent data case

Input: $\boldsymbol{TDS}(\mathbf{1}\ldots\boldsymbol{N})$ – Target User Dataset
$\boldsymbol{IDS}(\mathbf{1}\ldots\boldsymbol{M})$ – Initial Multiple User Dataset

Output : *S* **(1…*K*)** – Selected Similar Emotional Speeches Dataset

Execute Exist Algorithm 1 (TDS, IDS)
if $\boldsymbol{TDS_e}$ *= 0 then*
add **S** (IDS_e)
end

Return *S*

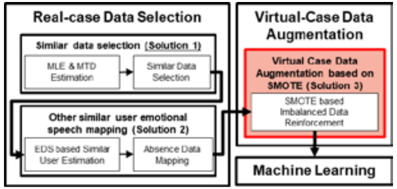
Proposed Algorithm 2. Data distribution factor based other similar user emotional speech mapping

Input: $\boldsymbol{TDS}(\mathbf{1}\ldots\boldsymbol{N})$ – Target User Dataset
$\boldsymbol{IDS}(\mathbf{1}\ldots\boldsymbol{M})$ – Initial Multiple User Dataset
NU - Number of Users

Output : *S* **(1…*K*)** – Selected Similar Emotional Speeches Dataset

Execute Algorithm 1 (TDS, IDS)
if $\boldsymbol{TDS_e}$ *= 0 then*
$\boldsymbol{TDDF_{oe}}$ = *Calculate Data Distribution Factors* ***(TDS)*** *// oe is other emotions*
for i = 1 to *NU*
$\boldsymbol{IDDF_i}$ = *Calculate Data Distribution Factors* $\boldsymbol{(IDS_{ie})}$
Similarity = *Calculate Euclidean Distance Similarity* $\boldsymbol{(TDDF_{oe}, IDS_{ie})}$
end
Muser = Get Maximum Similarity User's Absent Emotional Data (***Similarity***)
add S $\boldsymbol{(IDS_{Muser})}$
end
Return *S*

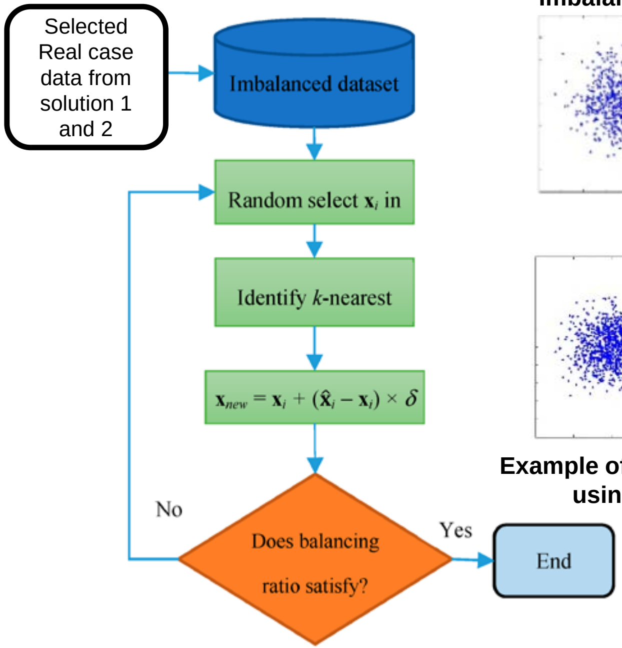
[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.



Solution 3 for imbalanced data

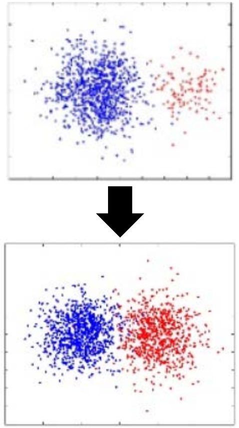
Virtual Case Data Augmentation based on SMOTE

Reinforce imbalanced environment to augment virtual data through iterative process of SMOTE



<The flowchart of SMOTE algorithm [8]>

Imbalanced dataset



Example of Oversampling using SMOTE

- **SMOTE (Synthetic Minority Oversampling Technique) based oversampling to solve imbalanced data environment**
  - SMOTE is the method used to **generate the dataset for a minority number of particular class samples** in the classification model.
  - **Iterative augmentation** using conventional SMOTE algorithm based on the selected real case data from Solution 1 and 2
  - The imbalance ratio is satisfied (IR < 2.0) stop the virtual data augmentation

Proposed Algorithm 3. Virtual Case Data Augmentation based on SMOTE

Input: $S(1 \dots N)$ – Selected Real Case Dataset from Solution 1 and 2
C – Number of Class

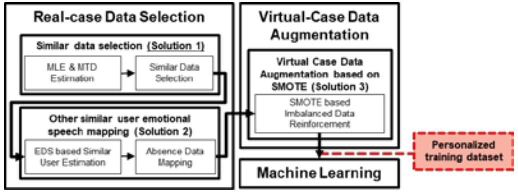
Output : AD (1…*K*) – Augmented real & case Dataset

```
add AD  (S)
for i = 1 to C
  VC = augment data using SMOTE  (AD, 50)
  add AD (VC)
end

IR = calculate imbalanced ratio (AD)

While IR < 2.0
  VC = augment data using SMOTE  (AD, 200)
  add AD (VC)
  IR = calculate imbalanced ratio (AD)
end
Return AD
```

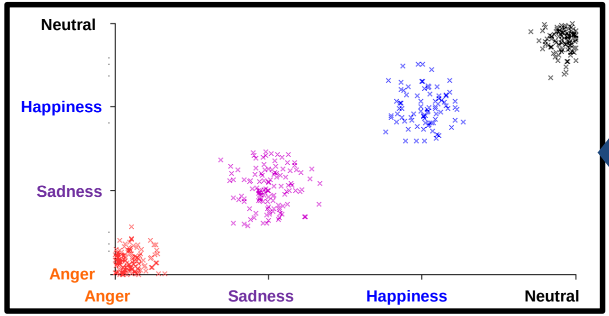
Imbalanced Ratio [18]
= Major Class/Minor Class



Personalized Model

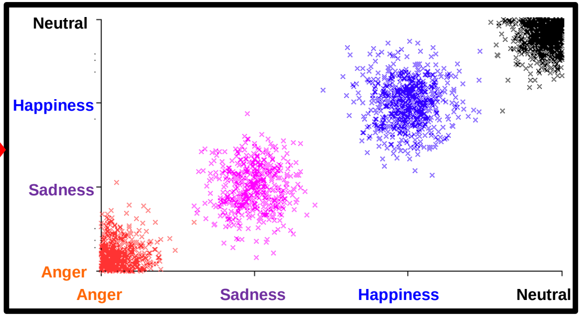
Model Comparison

Existing Method [4]



Target User Input data

Proposed Method



<Randomly selected target user data 50 (IEMOCAP)>

Anger – 7, Sadness – 32, Happiness – 2, Neutral - 9

Existing Method

Total	Selected 297 training data for personalization Anger – 63, Sadness - 99 Happiness – 54. Neutral State -81

Real Case Data		Virtual Case Data	
	• Target user Data – 27 • Selected Data - 270 ✓ Anger – 60 ✓ Sadness – 83 ✓ Happiness – 51 ✓ Neutrals - 76		• NONE

Proposed Method

Total	Selected & augmented 1,489 training data for personalization Anger – 441, Sadness - 342, Happiness – 382. Neutral State - 324

Real Case Data		Virtual Case Data	
	• Target user Data – 50 • Selected Data - 721 ✓ Anger – 287 ✓ Sadness – 82 ✓ Happiness – 253 ✓ Neutrals - 99		• Augmented Data – 718 ✓ Anger – 147 ✓ Sadness – 228 ✓ Happiness – 127 ✓ Neutrals - 216

[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. *Sensors*, *18*(11), 3744.

Experimental Environments

Evaluation Dataset

◆ Evaluation dataset selection

- **Evaluation Dataset: IEMOCAP (Interactive Emotional Dyadic Motion Capture)**
  - 10 actors (5 male, 5 female), 10 Emotions (**Anger, Happiness, Sadness, Neutral**, Frustrated, Excited, Fear, Disgust, Surprise, Others)
- **Initial Model Dataset: CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)**
  - 91 actors (48 male, 43 female), 6 Emotions (**Anger, Happiness, Sadness, Neutral**, Fear, Disgust)

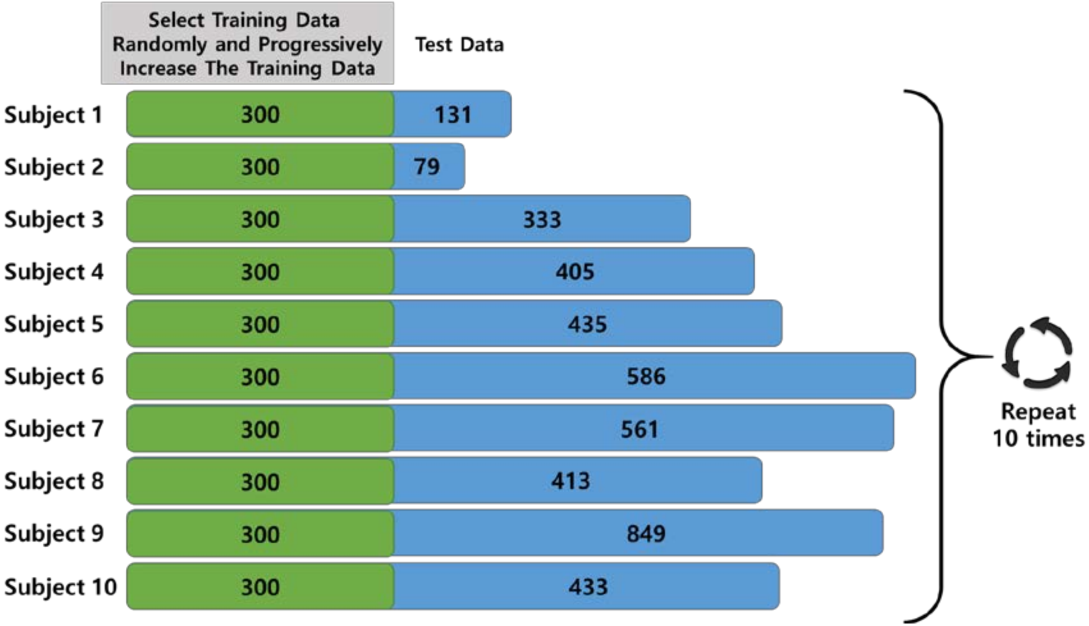
◆ Why IEMOCAP and CREMA-D dataset?

Emotional Database	Total Samples	Emotions	Speakers	Avg. Samples per Person	Avg. Samples of Each Emotion per Person
Emo-DB [13]	535	7	10	53.5	7.6
eNTERFACE [14]	1166	6	42	27	4.5
SAVEE [15]	480	8	4	120	15
RAVDESS [16]	1,440	8	24	60	7.5
CREMA-D [17]	7,442	6	91	81.7	13.61
IEMOCAP [18]	10,038	10	10	1003.8	100.3

Experimental Environments

Evaluation Methodologies

IEMOCAP Dataset



[Evaluation Criteria]

1. **Machine learning accuracy comparison evaluation using proposed method (SMO, J48, Random Forest)**

2. **Avg. accuracy with comparison evaluation of existing method (Evaluation Data: IEMOCAP, Initial Model: CREMA-D)**

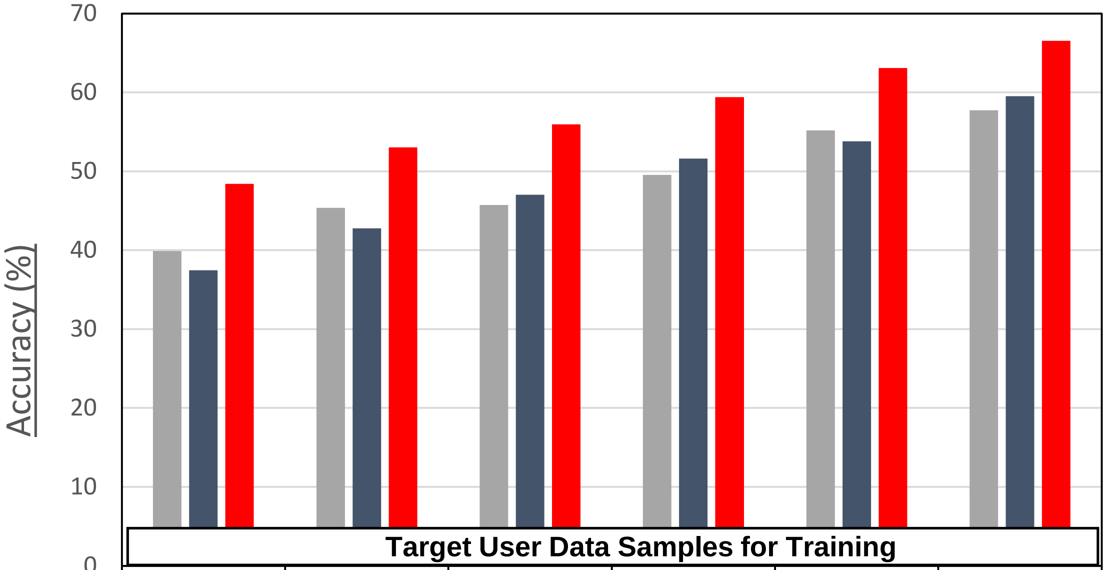
3. **Avg. Imbalanced Ratio with comparison evaluation of existing method**

[Comparison Evaluation]

1) SI (Speaker Independent – baseline)
2) PM (Personal Model – self learning)
3) SMOTE with RF [8]
4) Conventional MLLR with HMM [2]
5) LDM-MDT MLLR with GMM [4]
6) Proposed method

Experimental Results

1. Machine learning comparison evaluation using proposed speaker adaptation framework

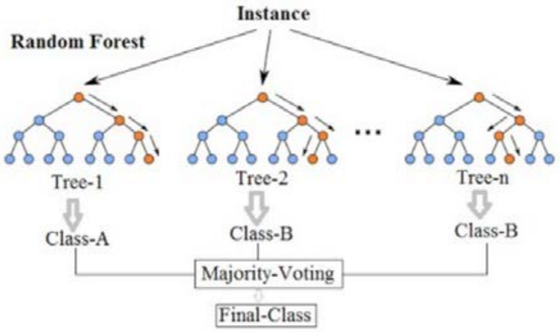


	50	100	150	200	250	300
SMO (RBF Kernel)	39.891	45.355	45.719	49.545	55.191	57.741
J48	37.471	42.759	47.018	51.61839864	53.79310345	59.506
Random Forest	48.389	53.038	55.963	59.39	63.089	66.521

SMO (RBF Kernel) J48 Random Forest

[Experimental Analysis]

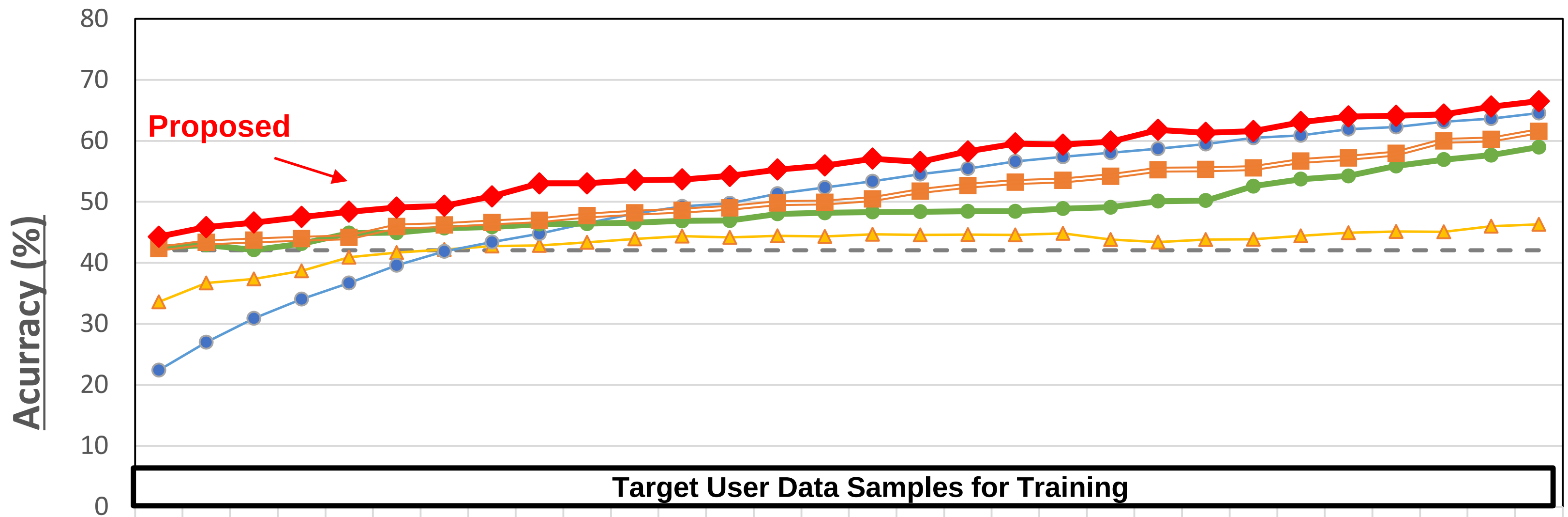
- Augmented personalized dataset is composed of target user oriented data. (Overfitting Problem)
- **Random Forest has good generalization performance through randomization** using the bootstrap method. **(High Prediction)**



< Simplified Random Forest>

Experimental Results

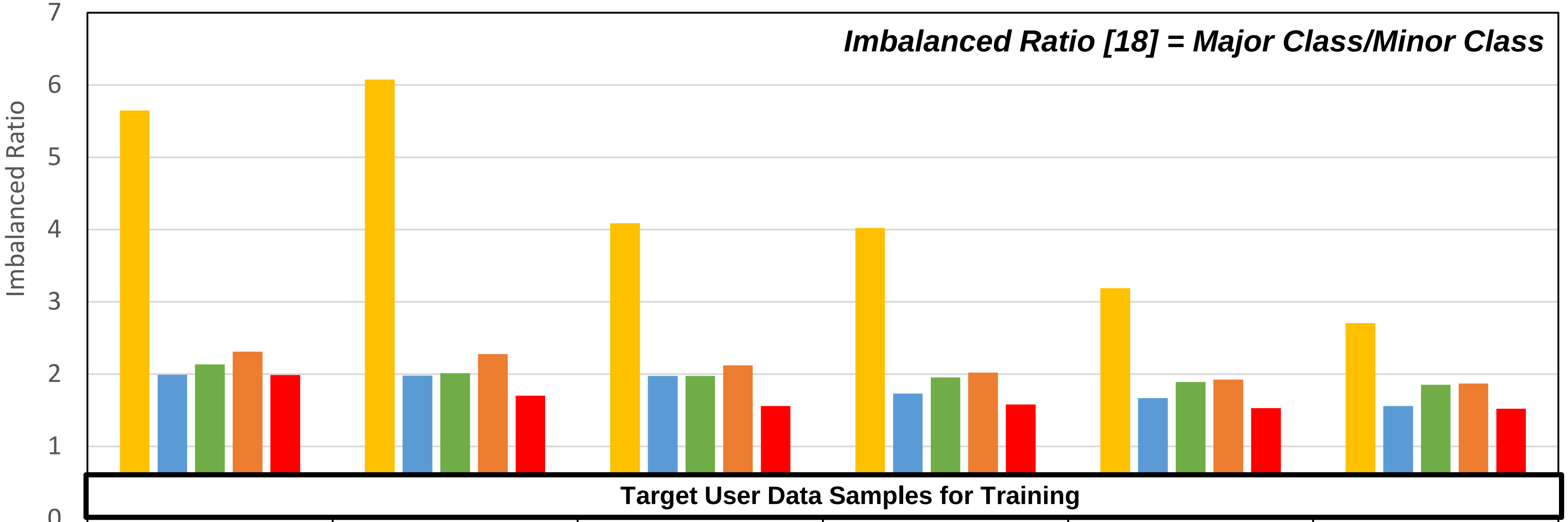
2. Comparison evaluation of average accuracy (**Evaluation Data: IEMOCAP**, **Initial Model: CREMA-D**)



	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
SI Model (baseline)	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1	42.1
PM(Self-Learning)	33.6	36.7	37.4	38.7	40.9	41.7	42.2	42.8	42.9	43.3	43.9	44.4	44.2	44.4	44.3	44.7	44.6	44.6	44.6	44.8	43.8	43.4	43.8	43.9	44.4	44.9	45.1	45.1	46	46.3
SMOTE [8] (only)	22.4	27	30.9	34	36.7	39.6	41.9	43.4	44.8	46.5	48.1	49.3	49.8	51.4	52.4	53.4	54.5	55.5	56.6	57.4	58	58.7	59.5	60.5	60.9	61.9	62.3	63.2	63.7	64.6
MLLR [2] (conventional)	42.5	42.9	42.1	43.2	44.9	44.9	45.7	45.9	46.4	46.4	46.6	46.9	46.9	48	48.2	48.3	48.4	48.5	48.5	48.9	49.1	50.1	50.2	52.6	53.7	54.3	55.9	56.9	57.7	59
LDM-MDT MLLR [4]	42.3	43.4	43.7	44	44.2	46	46.2	46.6	47	47.8	48.2	48.6	49	49.8	49.9	50.5	51.8	52.7	53.2	53.5	54.2	55.3	55.3	55.5	56.7	57.2	57.9	60	60.2	61.6
Proposed	44.3	45.9	46.6	47.5	48.4	49.1	49.4	50.9	53	53	53.6	53.7	54.3	55.3	56	57.1	56.5	58.3	59.6	59.4	59.9	61.8	61.3	61.6	63.1	64	64.1	64.3	65.6	66.5

Experimental Results

3. Comparison evaluation of imbalanced ratio



	50	100	150	200	250	300
PM	5.646	6.074	4.087	4.021	3.188	2.707
SMOTE [8]	1.99	1.977	1.973	1.73	1.666	1.56
MLLR [2]	2.135	2.014	1.975	1.951	1.891	1.852
LDM-MDT MLLR [4]	2.311	2.275	2.122	2.021	1.922	1.871
Proposed method	1.987	1.702	1.56	1.578	1.529	1.519

Conclusion & Future work

This thesis contributes to research the robust speaker adaptation framework that can resolve the cold-start problem

- **Improve the accuracy and imbalanced ratio in limited data environment**
  - ✓ Higher accuracy than existing methods in small samples environment as 10 to 150 **(2.2% ~ 6%)**
  - ✓ Reduce imbalanced difference from original target user training dataset **(178% ~ 356%)**
  - ✓ **The proposed method can fastly create personalized model speaker adaptation in limited data environment** such as small samples and absent data environment.

Future Work

- **Research on effective personalized data acquisition mechanism.**
- **Research on suitable re-training time to create personalized model.**

Publications

Journal : 16

SCI/E	**First author 1 (SCIE)**	**Co-author 12 : 2(SCI) / 10 (SCIE)**
Non SCI/E	**First author 2**	**Co-author 1**

First author

- **SCIE** : Sensors (IF: 2.475 , published, 2018)

Conference : 9

International	**First author : 1**	**Co-author : 5**
Domestic	**First author : 3**	

Patents : 3

Domestic	**First author : 2**	**Co-author : 1**

Total publications : 28

First author : 9

References

[1] Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. Inf. Fusion 2017, 37, 98–125.

[2] Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Language 9 (2), 171–185.

[3] Kim, J. B., Park, J. S., & Oh, Y. H. (2011, May). On-line speaker adaptation based emotion recognition using incremental emotional information. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4948-4951). IEEE.

[4] Kim, J.B.; Park, J.S. Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition. Eng. Appl. Artif. Intell. 2016, 52, 126–134.

[5] Abdelwahab, M.; Busso, C. Incremental adaptation using active learning for acoustic emotion recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5160–5164.

[6] Abdelwahab, M., & Busso, C. (2015, April). Supervised domain adaptation for emotion recognition from speech. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5058-5062). IEEE.

[7] Busso, C.; Mariooryad, S.; Metallinou, A.; Narayanan, S. Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech. IEEE Trans. Affect. Comput. 2013, 4, 386–397

[8] Deng, Jun. Feature Transfer Learning for Speech Emotion Recognition. Diss. Technische Universität München, 2016.

[9] Goronzy, Silke. Robust adaptation to non-native accents in automatic speech recognition. Vol. 2560. Springer, 2003.

[10] McKay, C.; Fujinaga, I.; Depalle, P. jAudio: A feature extraction library. In Proceedings of the International Conference on Music Information Retrieval, London, UK, 11–15 September 2005.

[11] Sahoo, T.R.; Patra, S. Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification. Int. J. Image, Graph. Signal Process. 2014, 6, 27–35.

[12] Anagnostopoulos, C.N.; Iliou, T. Towards emotion recognition from speech: Definition, problems and the materials of research. In Semantics in Adaptive and Personalized Services; Springer: Berlin/Heidelberg, Germany, 2010; 127–143.

[13] 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 Resour. Eval. 2008, 42, 335–359.

[14] Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The enterface'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on IEEE Data Engineering Workshops, Atlanta, GA, USA, 3–7 April 2006; 8.

[15] 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.

[16] Jackson, P.; Haq, S. Surrey Audio-Visual Expressed Emotion (Savee) Database; University of Surrey: Guildford, UK, 2014.

[17] Livingstone, S. R., & Russo, F. A. (2018). 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, 13(5), e0196391.

[18] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

[19] Lele, S.; Richtsmeier, J.T. Euclidean distance matrix analysis: A coordinate-free approach for comparing biological shapes using landmark data. Am. J. Phys. Anthropol. 1991, 86, 415–427

[20] Hoens, T. R.; Chawla, N. V. Imbalanced datasets: from sampling to classifiers. Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley, 2013

[21] Bang, J., Hur, T., Kim, D., Lee, J., Han, Y., Banos, O., ... & Lee, S.. Adaptive Data Boosting Technique for Robust Personalized Speech Emotion in Emotionally-Imbalanced Small-Sample Environments. Sensors, 18(11), 3744.

Thank you

for your attending

Q & A ?