Received: August 10, 2019

*International Journal of*
Intelligent Engineering & Systems

http://www.inass.org/

# Hybrid Approach of Structural Lyric and Audio Segments for Detecting Song Emotion

**Fika Hastarita Rachman**[1,2]* **Riyanarto Sarno**[1] **Chastine Fatichah**[1]

[1]*Departement of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia*
[2]*Departement of Informatics, University of Trunojoyo Madura, Indonesia*
* Corresponding author's Email: hastarita.fika@gmail.com

**Abstract:** Detecting song emotion is very important, however many studies have been done based on song lyrics and song audio separately. This research proposes a method for detecting song emotion based on integrated song lyrics and audio. Synchronizing the right structural segment and lyrics of song can be used for hybrid approach to detected right emotion. Song emotion can be classified into Thayer emotion label. The features of a song lyric are extracted using Psycholinguistic and Stylistic; whereas the features of a song audio are extracted using analyze audio signal waveform using Fast Fourier Transform (FFT) method. A song can be divided into 5 structural segments, which are intro, chorus, bridge, verse and outro. A preprocessing method of audio uses Correlation Features Selection (CFS) and preprocessing text for lyrics. Six classification methods are used for classifying emotion based on lyrics and audio of song structural segments separately. The aggregate method is used to analyze the results of classification before to obtain structural segments that represent emotions. Then, the final process Hybrid approach used to combine audio and lyrics features in emotion detection. Sum of matrix and Majority Voting Concept are used for Hybrid approach. The value of F-Measure is 0.823.

**Keywords:** Emotion detection, Song structural segment, Audio features, Lyric features, CFS, Aggregate method, Hybrid approach.

## 1. Introduction

Digital songs are a popular thing among the people. The amount of song data confuses us when we have to choose a song that suitable with emotions. Song emotion detection is preferred and necessary to support other applications such as song emotion recommendation for drivers [1], children [2], teenagers, etc.

There are two types of data in songs that can be processed, audio and lyrics. Many researches on detecting emotion has been done based on song lyrics and song audio separately. The previous research [3-6] uses audio features for song emotion and genre detection. In research [3] mood of song is classified using SVM. It used the audio power and audio harmony features of audio form with discrete wavelet transform (DWT) to reduce noice. Intensity, timbre and rhythm features are used in [4] using 20 second music audio clips to detecting of music mood. The other features, MFCC and Chroma from Echonest analyzer are used to detect the genre of music classification [5]. The mood tracking using the audio features from Marsyas extractor and smooth process method is done for [6] to detect the mood of audio music [6]. Their research shows that audio is an important feature for song emotion detection. Lyrics can also be a feature in song emotion detection. The previous research [7] uses Sentiwordnet to extract sentiment features in lyrics. Mood classification is done using sentiment features, feature selection process and classification models. Research [8] build ANCW from ANEW to detect emotion of chinese song dataset using fuzzy clustering. The emotional classification of songs based on lyrics can also be done using partial syntatic analysis model [9]. The result of [10] show that the use of LSA for lyric features is not optimal. The classification results are still low. Lyrics text mining is used to find the role of lyrics that can improve the accuracy of mood detection [11]. Besides lyrics, the combined testing of audio and lyric features produced the conclusion that not all audio is the dominant feature in determining the emotions of songs, not all mood labels can produce high accuracy [11]. Even though these studies obtained good results but in the process they used audio datasets that the determination of the audio duration is based on the expert. In fact, a song audio has duration of two until five minutes. It is difficult to determine which segment that represent emotion of song.

Generally Music Emotion Recognition (MER) used Bimodal dataset [12]. Bimodal dataset that combine of audio vocals and instruments. Emotion detection usually uses all lyrics of song. Bimodal Dataset [1] is one of the songs emotional datasets that uses 30 seconds audio and all lyrics of song. The 1000 song dataset [13] is also a song emotional dataset that uses 45 seconds of audio data. Both datasets uses short audio that the duration is determined based on the expert. Duration of segment determinated by expert is different of each song, so it is difficult to determine automatically when a song does not have information from the expert. Song emotion is wrongly detected if the system does not have the right segment of audio.

Previous research presents that there are emotional differences in each song segment [14]. The result shows that the best accuracy for music mood detection uses 8 and 16 seconds duration. Other research uses data spoken speech dataset (Emo-DB). The emotion of spoken speech is analyzed after segmentation audio process with 400ms duration. The result of emotion prediction is different of each segment. There is a mechanism to fuse segment emotion in order to make a global emotion in full audio [15].

Emotion detection in audio and lyrics was performed by Ricardo Malheiro, 2016 [12]. They used Bimodal dataset with 133 songs data. Each data has 30 seconds of audio and all lyrics of song. They used 1701 audio subset features from melody, timbre, rhythm, and others [12]. Features of lyrics are obtained from lyric tools (Synesketch, ConceptNet, LIWC, and General Inquirer). The fuse of audio and lyrics features in emotion recognition can increase the F-Measure value to 88.4%. Although the results are very high, but only songs that have right segment audio can be detected emotion of song accurately. Audio data that used in this study is short audio, while the lyric data that used are overall data of the song. There is no synchronization of data between audio and lyrics. If there is a new song, an expert is needed to determine the segment of audio. Ricardo [12] uses short audio to detect song emotions. The determination of short audio duration comes from experts. Supposedly without the short audio from the expert, the system is able to detect emotions. This research can detect song emotions automatically because the system can determine structural segments that represent emotion based on Correlation Feature Selection (CFS) and Aggregate method.

The structure segment of song is part of the song that can be processed to emotion detection beside audio and lyrics. Song structure has a unique form and composed to be a song. In song writing, song is generally composed of five parts: intro, verse, bridge, chorus, outro [16]. Each part can be different position in a song. The intro is the beginning of a song and usually just an instrument. Verse is the verses of songs with lyrics. Chorus is the message or core of the song. The verse and chorus part can be repeated. Bridge is the connecting part between the parts of the song. While outro is the closing part of the song. Position of the intro and outro are often at the beginning and end of the song. Position of the verse, chorus, and bridge located in the middle of the music. It can be different order. The example of song structure position in a song is: Intro – Verse – Chorus – Verse – Bridge – Chorus – Outro.

Chia [16] conducted research on the chorus detection algorithm and emotion detection of songs based on audio data. The value of intensity, frequency band and rhythm regularity in the chorus are used to detect song emotions. Their emotion detection algorithm provides similar results for the same melody in various languages and lyrics. There is an influence on the selection of song structure for song emotion detection.

We propose song emotion detection that used synchronously between audio and lyrics. Song structural segments are also used in this research. The proposed system can automatically recognize the song emotion with the existence of structural segment data.

We have two contributions in this research. First contribution is automatically segment selection that represent the emotions of the whole of song by analyzing structure of song. Second contribution is hybrid approach to combine audio and lyrics features from song structure segment for detection emotion using prediction frequency matrix. With this contribution, to find out the emotions of the song, we need to know the structure of certain songs only. So even though the song doesn't have audio samples from experts, we can still find out the emotions of the song.

This paper is organized as follows. Related work and contribution is shown in introduction in section 1. Section 2 shows a proposed method of this research. Section 3 shows a structural emotion of song dataset that used in this research. Section 4 shows a method of audio and lyric extractions. Section 5 shows a emotional of structural segmen analysis to know what structural segment that represent a whole of song emotion. Section 6 shows a model and method that we used for hybrid features in song emotional detection. Result and discussion are presented in section 7. Finally, section 8 provides conclusions and future work.

## 2. Proposed method

The research was carried out according to the proposed method in Fig.1. The song data that processed is audio structural data and lyrics. Lyrics data is synchronized with segments in audio data.

Audio segment data experiences extraction features using analyze audio signal waveform using Fast Fourier Transform (FFT) method in MIR Toolbox [17]. The features of an audio segment are sub features from dynamics, rhythm, timbre, pitch, and tonality features. Then the feature selection process is performed to reduce the number of features used in the classification. This research use Correlation Feature Selection (CFS) [18-19] for selection feature method. The audio feature of each structural segment is classified using the Random Forest Method. The results of the classification show that the best structural segment represent the emotions of the song. Each song has several song structural segments. Each of structural segments represented as a matrix, therefore there are several audio and lyric matrices. The row of a matrix represents the index of song data, and the column of a matrix represents count number of emotional label prediction for each song structural segment.

The lyric data is synchronized first with structural segment audio using extension file configuration lrc (short for LyRiCs). Preprocess of lyrics are done in several processes: slang word repair, POS tagging, Porter stemming and stopword removal. Slang word repair is done by using slang word corpus which is often found in the lyrics. Stanford Part of Speech (POS) Tagging [18] is done to find out the position of words in the lyrics. There are several word positions that rarely describe the emotions of the lyrics (shown in Table 1). The word in that position is deleted.

Table 1. Filtering POS tagging

| POS tagging position | Meaning of POS |
|---|---|
| DT | Determiner |
| CD | Cardinal number |
| WRB | Wh – adverb |
| TO | To |
| CC | Coordinating Conjungtion |
| PRPS | Personal pronoun |
| MD | Modal |

After preprocessing the lyrics, next process is the feature extraction process. The extracted features from the lyrics are psycholinguistic and stylistic features. The psycholinguistic feature is a feature based on the emotional psychology dataset. This dataset is a CBE (Corpus Based Emotion) dataset from the results of previous research[19] which is expanded according to the Thayer emotion label.

The stylistic features are words that are often found in lyrics, but are not in the English dictionary (Such as: ‘ah’, ‘ooh’, ‘yeah’). Those unique words, exclamation marks, and question marks in the lyrics are a feature of stylistic. These features are used as input in the emotional classification process from each lyric structural segment. Emotion predicted label of each segment presented using matrix segment.

The hybrid of audio and lyric matrix for each structural segment use operation sum of matrices. This hybrid matrix used for emotion detection model to get a song emotion predicted label. In the emotion detection model, sum of matrix and majority voting are used in several structural segment combinations analyzed.

The emotional label used in this research is the emotional label of Thayer[6]. Thayer emotional label has four (4) quadrants that depicted in a 2-dimensional model. Coordinate axis is Valance and Arousal which have a 'low' and 'high' area. Thayer's emotional label is: Quadrant 1 (Q1), Quadrant 2 (Q2), Quadrant 3 (Q3), and Quadrant 4 (Q4). Q1 is in a high valance and high arousal areas. The example of Q1 are happy and exited emotion. Q2 is in low valance and high arousal. Angry and nervous emotion includes in Q2. Q3 is in low valance and low arousal (The example: sad). Q4 is in low valance and high arousal (The example: calm, relaxed)[20].

## 3. Structural emotion of song dataset

This research requires a dataset of structural segments of songs that have labeled Thayer emotions. Previous datasets are separated between structural song dataset and emotional song dataset. Bimodal dataset is an emotional song dataset which has 133 songs with 30 second audio data, full data text lyrics and thayer emotion label [12]. One of the structural song datasets that has data in file xml is

Ep_groundtruth_excl_Paulus (Ep-dataset) [21]. The duration of each structural segment data is between 20-45 second [21]. This structural song dataset was created with the concept of the Chroma [22] feature on song audio data. This research, used combination of two datasets, Bimodal dataset and Ep-dataset. Bimodal dataset has emotional labels for each song but does not have song structural data. So the audio and lyrics of song synchronization based on the official store song (file wav) and file lrc are needed to complete this dataset. Ep-dataset have structural segment data but do not have emotional label for each song. A music expert is needed to complete this dataset. All data in two datasets are not used as a whole. The availability of full song data (.wav), lyrics data (.lrc) , and the balance of the amount of data between emotional labels is also a consideration.

This research dataset consists of 100 songs, with 25 data for each Thayer emotion label. From 100 songs, there are 875 song structure data along with audio segments (.wav), duration structural segment data, and Thayer emotion label of song. Audio data that has lyrics only 738 data. So in this research we used 738 data. The distribution amount of data for each segment is shown in Table 2.

Table 2. Distribution amount of data structural segment

| Structural Segment | Amount of data |
|---|---|
| Intro | 39 |
| Chorus | 283 |
| Bridge | 76 |
| Verse | 299 |
| Outro | 41 |

The genre of song is becoming one of the considerations as a limit in the scope of research data or not. For this purpose, we analyze the influence parent of genre music with Thayer's emotional label on a bimodal dataset. Fig. 2 shows that in certain genres can emerge certain emotions as well. Examples for song with pop genres tend to have Q4 emotional labels while song with rock genre tends to have Q2 emotional labels. With this analysis, the data on the structural emotion of song datasets do not differentiate between music genres.

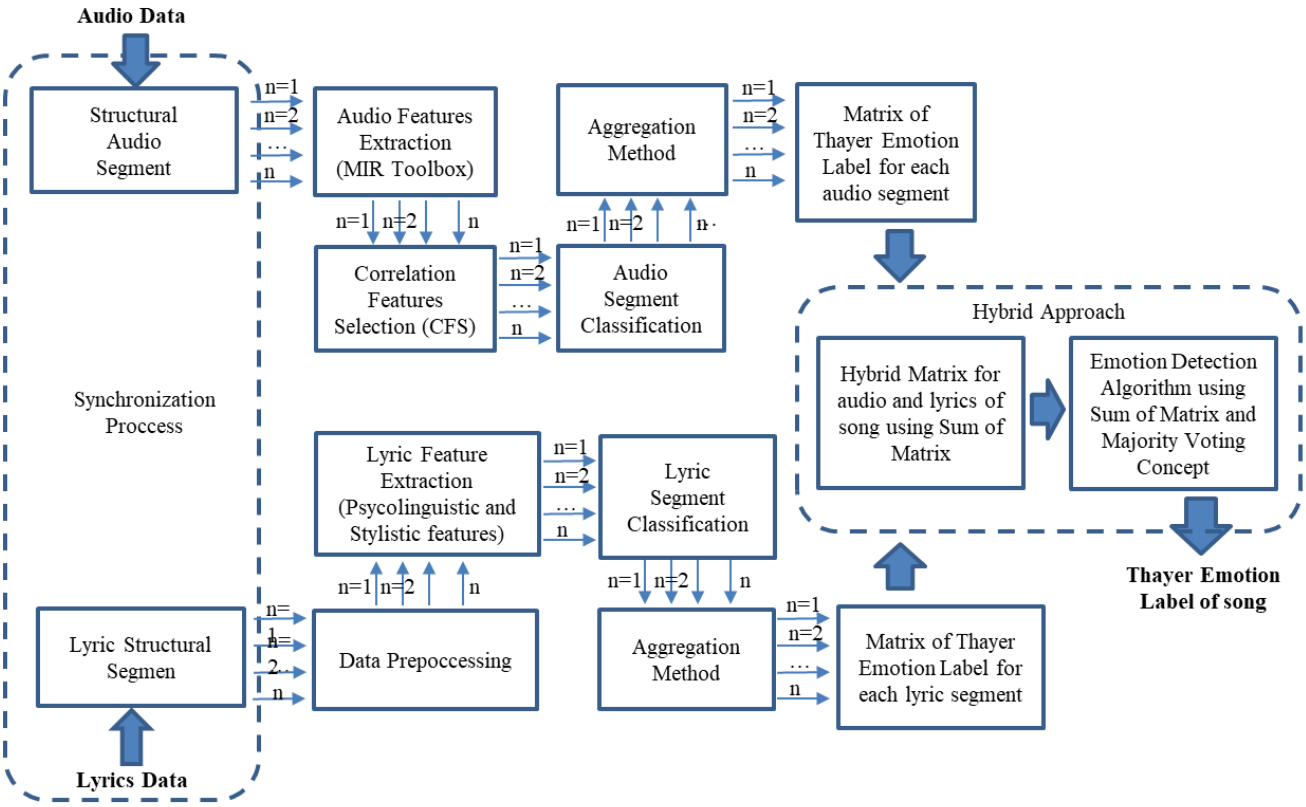

Audio Data
Structural Audio Segment
n=1
n=2
…
n
Audio Features Extraction (MIR Toolbox)
n=1
n=2
n
Correlation Features Selection (CFS)
n=1
n=2
…
n
Audio Segment Classification
n=1
n=2
n
Aggregation Method
n=1
n=2
…
n
Matrix of Thayer Emotion Label for each audio segment
Hybrid Approach
Hybrid Matrix for audio and lyrics of song using Sum of Matrix
Emotion Detection Algorithm using Sum of Matrix and Majority Voting Concept
Thayer Emotion Label of song
Synchronization Proccess
Lyric Feature Extraction (Psycolinguistic and Stylistic features)
n=1
n=2
…
n
Lyric Segment Classification
n=1
n=2
n
n=1
n=2
n
Lyric Structural Segmen
n=1
n=2
…
n
Data Prepoccessing
Aggregation Method
n=1
n=2
…
n
Matrix of Thayer Emotion Label for each lyric segment
Lyrics Data

Figure. 1 Proposed method

Received: August 10, 2019

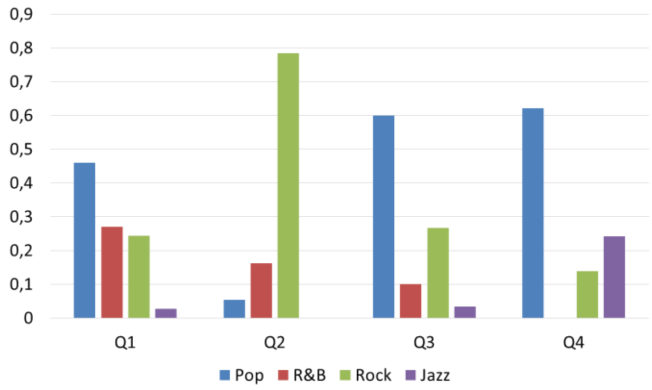


Figure. 2 Analysis graph of bimodal dataset in the song genre

## 4. Feature extraction

The extraction feature is done for two data: audio data and song lyrics. Feature extraction uses different concepts for different data. The feature extraction results are used in structural segment classification process.

### 4.1 Audio feature extraction

The audio features extraction used to analyze audio signal waveform from Fast Fourier Transform (FFT) method [23]. That extraction method has been applied into Music Information Retrieval Matlab tools, namely MIRToolbox version 1.6.1 [17]. Besides MIRToolbox, MPEG-7 is another tool for extracting audio signals [26]. MPEG-7 was not used in this research, because MPEG-7 is a Low Level Description (LLD) including: basic spectral features, basic signal parameters, and timbral description. Previous research [26-27] using this MPEG-7, the feature that used are audio power and audio harmonicity. In Mirtoolbox, audio waveform is extracted into standard level feature. There are dynamic, rhythm, timbre, pitch, and tonality features. From these features there are several sub features that extracted. These subset features are in scalar and signal data. Subset features in signal data form is simplified statistically into statistical parameters. The statistical parameters that used are *average (avg)*, *standard deviation* (*std*), *median* (*med*) of the signal data value vector. Total of parameters that used as sub-features are 54 parameters. The feature extraction scheme using MIRToolbox can be seen in Fig. 3.

From these 54 parameters, a feature selection process is carried out using Correlation Feature Selection (CSF). The CFS configuration using the Best First Search method with the Forward-backward (Bidirectional) search model and 60% threshold.

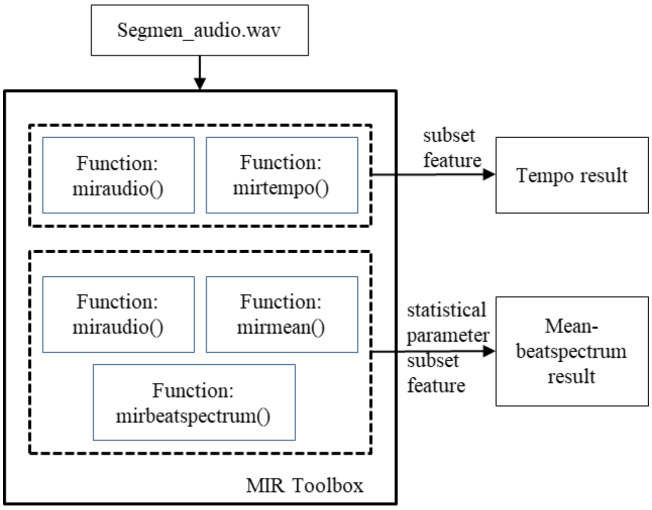


Figure. 3 Feature extraction scheme using MIRToolbox

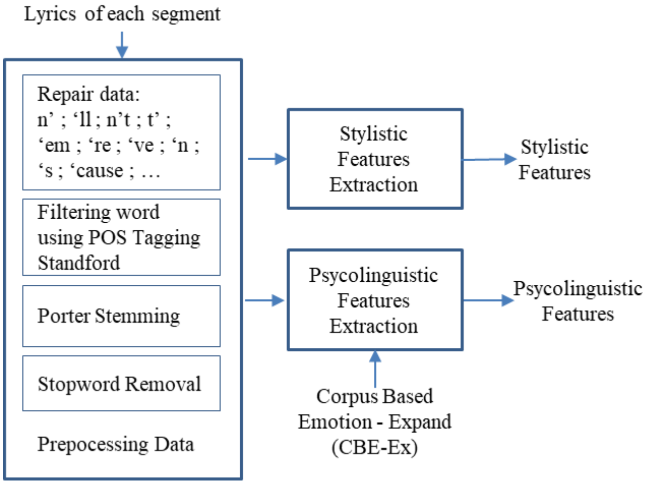


Figure. 4 The process of extracting lyric data

With this method, 16 subset features use in the next process. The sixteen subset features include: *std-beatspectrum, eventdensity, pulseclarity, mean-attacktime, mean-decreaseslope, med-decreaseslope, std-decreaseslope, zerocross, kurtosis, mean-roughness, std-hcdf, mean-mfcc, std-mfcc* , *mean-spectrum, mean-chromagram,* and *envelope-halfwavediff.*

### 4.2 Lyric features extraction

The process of extracting lyric data can be seen in Fig.4. Preprocessing data is done before the data extraction process. The preprocessing data include repairing data, filtering word using POS Tagging Stanford, Porter Stemming and Stopword removal.

This research uses the stylistic feature and psycholinguistic features for lyrics feature of song. The stylistic feature is a feature taken from unique words and special punctuation in the lyrics. Unique and informal words are often found in lyrics, such as: '*oh*', '*ooh* ','*ah* ','*yeah*','*huuu*','*whoo*' and others.

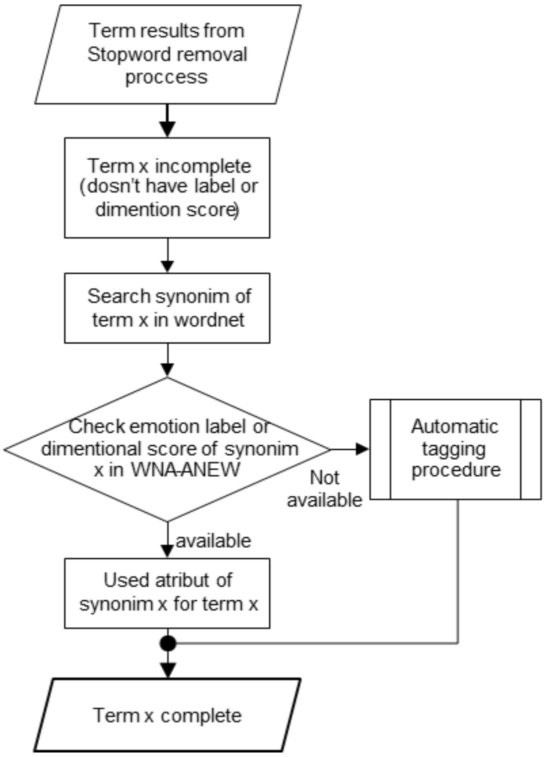

Term results from Stopword removal proccess
Term x incomplete (dosn't have label or dimention score)
Search synonim of term x in wordnet
Check emotion label or dimentional score of synonim x in WNA-ANEW
Not available
Automatic tagging procedure
available
Used atribut of synonim x for term x
Term x complete

Figure. 5 Automatic tagging procedure

Likewise special punctuation found in the lyrics, such as exclamation marks (!) and question marks (?).In previous research [19] the lyrics that were processed are overall lyrics of each song data. For this research, the feature extracted from lyrics of the structural segment data that has been synchronized with the audio structural segment.

Psycholinguistic feature is a feature that was obtained with the help of the existing corpus of emotion. The emotional corpus that we used is the expand of Corpus Based Emotion (CBE). The CBE in the previous research [10] used 5 emotional labels from MIREX [24]. Because the bimodal dataset used previously has the Thayer's emotional label, the emotional label on the CBE is also adjusted. The change in the emotional label on the CBE causes a change in the center of the cluster data. Other developments occur in the number of corpus data due to the Automatic tagging procedure. From all lyrics data in the dataset, data are found to be incomplete. Uncompleted data are the data term does not have the label or emotional value of Arousal-Valance. Not all term of the lyric in the dataset has own emotion label in CBE. With the concept of synonym of term and automatic tagging procedure, CBE is expanding.

Automatic tagging procedure [23] is a procedure of labeling emotions or looking for Valance-Arousal dimension values using Cluster center concept and LESK similarity measure [24]. The Automatic Tagging Procedure shown in Fig.5. The amount of corpus data in the previous CBE was 2122 terms, while Expand CBE (CBE-Ex) obtained an additional 1649 terms to 3771 terms. CBE is emotion corpus that merging of two corpus: WordNet Affect Emotion (WNA) [25] and Affective Norms for English Words (ANEW) [26]. The reliability of the four corpuses (ANEW, WNA, CBE and CBE-Ex) is tested by calculating the F-measure for the emotional classification of the whole of lyrics with 11 psycholinguistic and 19 stylistic features using Random Forest method 5 fold cross-validation. The results are shown in Fig. 6. It is seen that CBE-Ex is better than CBE with the value of F-Measure of 0.399. This research uses CBE-Ex to obtain the psycholinguistic feature. The overall lyrical features used in this research are 30 features: 19 features are stylistic features and 11 features are psycholinguistic features.

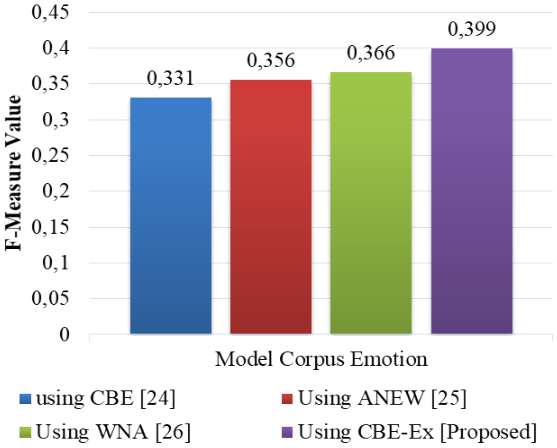


Figure. 6 Comparison of F-measure result using CBE [24], ANEW [26], WNA[25] and CBE-Ex

## 5. Emotional structural segment analysis from audio and lyrics of song

The structural segment of the song consists of the intro, chorus, verse, bridge and outro [27]. In this research, the structure of songs that represent the emotions of the song is analyzed. The analysis is seen from the audio and lyrics features. Purpose of this step is to reduce the duration of audio data to be processed, so that all classification features can be obtained properly.

Emotional prediction of song structural segments is searched using six classification method: Logistic method (ML1), C45 method (ML2), Random Forest method (ML3), Multilayer Perceptron method (ML4), Bayes Net method (ML5), Naive Bayes method (ML6), then find the results using the aggregate method. Six types of classification methods are used because previous research [28] also uses them.

Table 3. The result of structural segment classification from song audio

| Song Structural Segments | F-measure | | | | | |
|---|---|---|---|---|---|---|
| | ML1 | ML2 | ML3 | ML4 | ML5 | ML6 |
| Bridge$^{3}$ | 0.471 | 0.606 | 0.598 | 0.57 | 0.415 | 0.433 |
| Chorus$^{1}$ | 0.539 | 0.574 | 0.707 | 0.683 | 0.5 | 0.557 |
| Intro$^{4}$ | 0.432 | 0.382 | 0.429 | 0.469 | 0.236 | 0.475 |
| Outro$^{5}$ | 0.425 | 0.285 | 0.364 | 0.314 | 0.138 | 0.329 |
| Verse$^{2}$ | 0.534 | 0.598 | 0.658 | 0.615 | 0.528 | 0.49 |

Table 4. The result of structural segment classification from song lyrics

| Song Structural Segments | F-measure | | | | | |
|---|---|---|---|---|---|---|
| | ML1 | ML2 | ML3 | ML4 | ML5 | ML6 |
| Bridge$^{2}$ | 0.546 | 0.474 | 0.46 | 0.577 | 0.186 | 0.373 |
| Chorus$^{1}$ | 0.609 | 0.644 | 0.770 | 0.678 | 0.447 | 0.487 |
| Intro$^{4}$ | 0.283 | 0.327 | 0.287 | 0.254 | 0.167 | 0.323 |
| Outro$^{5}$ | 0.201 | 0.179 | 0.218 | 0.216 | 0.155 | 0.206 |
| Verse$^{3}$ | 0.449 | 0.394 | 0.510 | 0.458 | 0.274 | 0.362 |

Aggregate method [29] is used to determine the best choice, because of various classification methods have different ranking result. If in each classification result of each method there are *n* alternatives, then the ranking of each alternative is given a value. The first rank is given the highest value, *n*, the second rank is given a value of *n-1*, the third rank is given a value of *n-2* and so on. Addition with the highest value is the aggregate ranking.

Table 3 shows the results of the emotion classification using song audio. The aggregate ranking for audio show that the chorus is in the first position, then sequentially occupied by verse, bridge, intro and outro. Table 4 shows the result of emotion classification using song lyrics. The superscript *m* of a structural segment in Tables 3 and 4 (*segment*$^{m}$) means ranked. The aggregate ranking for lyrics shows that the chorus is in the first position, then sequentially occupied by bridge, verse, intro and outro. It shows that for structural segments of audio and lyrics, the highest position is in the same section, except bridge and verse. Three highest positions, Chorus-Bridge-Verse is used for next process.

From the six classification methods, the Random Forest method (ML3) in Tables 3 and 4 show that the ranking results are the same as the aggregate ranking, so the prediction results that used for the next process are emotion label predictions from ML3.

In the next process to obtain the best analysis results, three structural segments are used which represent the emotions of the song. There are chorus, bridge, and verse. Emotion predicted labels are presented in the form of a matrix for each structural segment lyric and audio separately.

## 6. Hybrid approach

A hybrid approach is a specific way to combine matrix audio and matrix lyrics. The hybrid matrix is used in emotion detection algorithms to produce one Thayer emotion label for the entire song.

### 6.1 Hybrid matrix thayer emotional label

The song structural segment that analyzed at this step is the chorus, bridge and verse for audio data and lyric data. Of the three structural segments, it is known that verse has the most part in a song. Each verse in the lyrics may also vary. To balance the number between segments, verse to be analyzed is categorized again into 3 parts. The 3 parts of the verse are verses on the beginning of the song (v1), verses in the middle of the song (v2) and verses at the end of the song (v3). The following is an example of the xml segment data from a song titled *Anna (go to him)*, artist *The Beatles*. It can be seen that the intro has 1 section, the chorus has 2 parts, the bridge is 2 parts, the verse has 4 parts, and the outro has 0 parts. The verse is divided into 3 segments: segment v1 there are 2 parts, segment v2 there is 1 part and segment v3 there is 1 part.

```
<segmentation>
  <segment label="Intro" start="00:00:708"
    end="00:09:669" start_sec="0.7084616"
    end_sec="9.6695437"/>
  <segment label="Verse" start="00:09:669"
    end="00:29:410" start_sec="9.6695437"
    end_sec="29.4107308"/>
```

```
<segment label="Chorus" start="00:29:410"
  end="00:36:239" start_sec="29.4107308"
  end_sec="36.2399923"/>
<segment label="Verse" start="00:36:239"
  end="00:57:710" start_sec="36.2399923"
  end_sec="57.7109001"/>
<segment label="Bridge" start="00:57:710"
  end="01:31:965" start_sec="57.7109001"
  end_sec="91.9659766"/>
<segment label="Verse" start="01:31:965"
  end="01:49:562" start_sec="91.9659766"
  end_sec="109.5626313"/>
<segment label="Bridge" start="01:49:562"
  end="02:24:200" start_sec="109.5626313"
  end_sec="144.2006599"/>
<segment label="Verse" start="02:24:200"
  end="02:39:659" start_sec="144.2006599"
  end_sec="159.6600000"/>
<segment label="Chorus" start="02:39:659"
  end="02:54:530" start_sec="159.6600000"
  end_sec="174.5304761">
 <alt_label>outro</alt_label>

```

$CA$ is a prediction frequency matrix with audio chorus vector elements. Matrix $CA$ containing count of Thayer label emotion predict in chorus audio with elements $ca_{i,j}$. Matrix $CA$ shown in Eq.(1). The first subscript $(ca_i)$ will refer to the row position or row of song id in the array. The second subscript $(ca_j)$ will refer to to the column position or column of Thayer label emotions in audio. In this research the number of datasets is 100 songs, so the maximal value is $i$ = 100 and $j$= 4 (representing the data of the Thayer label emotion Q1, Q2, Q3 and Q4). $BA$ is a Prediction frequency matrix with audio bridge vector elements which contain the count of Thayer label emotion predict in bridge audio with elements $ba_{i,j}$. Prediction frequency matrix $BA$ shown in Eq. (2). $V1A$ is a matrix that have a count of Thayer label emotion predict in v1 audio as its elements data. The elements of the matrix $V1A$ are $v1a_{i,j}$. Prediction frequency matrix $V1A$ shown in Eq.(3). Similarly with Matrix $V1A$, Prediction frequency matrix $V2A$ represent the v2 audio matrix and Prediction frequency matrix $V3A$ represent the v3 audio matrix. Beside 5 audio matrices, we also have 5 lyric Prediction frequency matrices namely: matrix $CL$ (chorus lyrics), matrix $BL$ (bridge lyrics), matrix $V1L$ (v1 lyrics), matrix $V2L$ (v2 lyrics), and matrix $V3L$ (v3 lyrics).

$$CA = \begin{bmatrix} ca_{11}, & ca_{1,2} & ca_{1,3} & ca_{1,4} \\ ca_{2,1} & ca_{2,2} & ca_{2,3} & ca_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ ca_{i,1} & ca_{i,2} & ca_{i,3} & ca_{i,4} \end{bmatrix} \quad (1)$$

$$BA = \begin{bmatrix} ba_{1,1} & ba_{1,2} & ba_{1,3} & ba_{1,4} \\ ba_{2,1} & ba_{2,2} & ba_{2,3} & ba_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ ba_{i,1} & ba_{i,2} & ba_{i,3} & ba_{i,4} \end{bmatrix} \quad (2)$$

$$V1A = \begin{bmatrix} v1a_{1,1} & v1a_{1,2} & v1a_{1,3} & v1a_{1,4} \\ v1a_{2,1} & v1a_{2,2} & v1a_{2,3} & v1a_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ v1a_{i,1} & v1a_{i,2} & v1a_{i,3} & v1a_{i,4} \end{bmatrix} \quad (3)$$

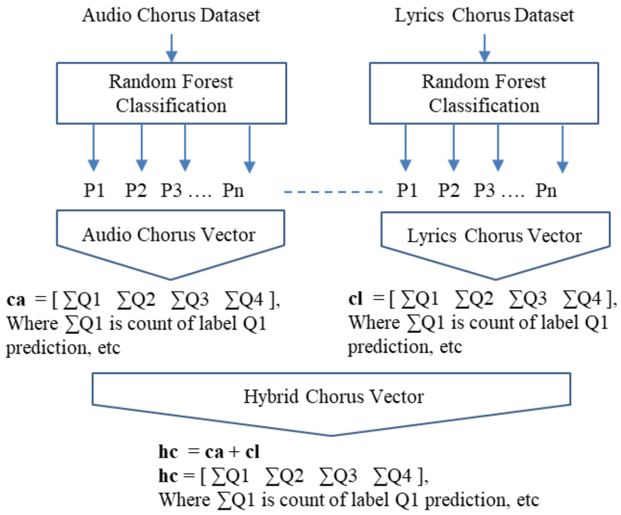
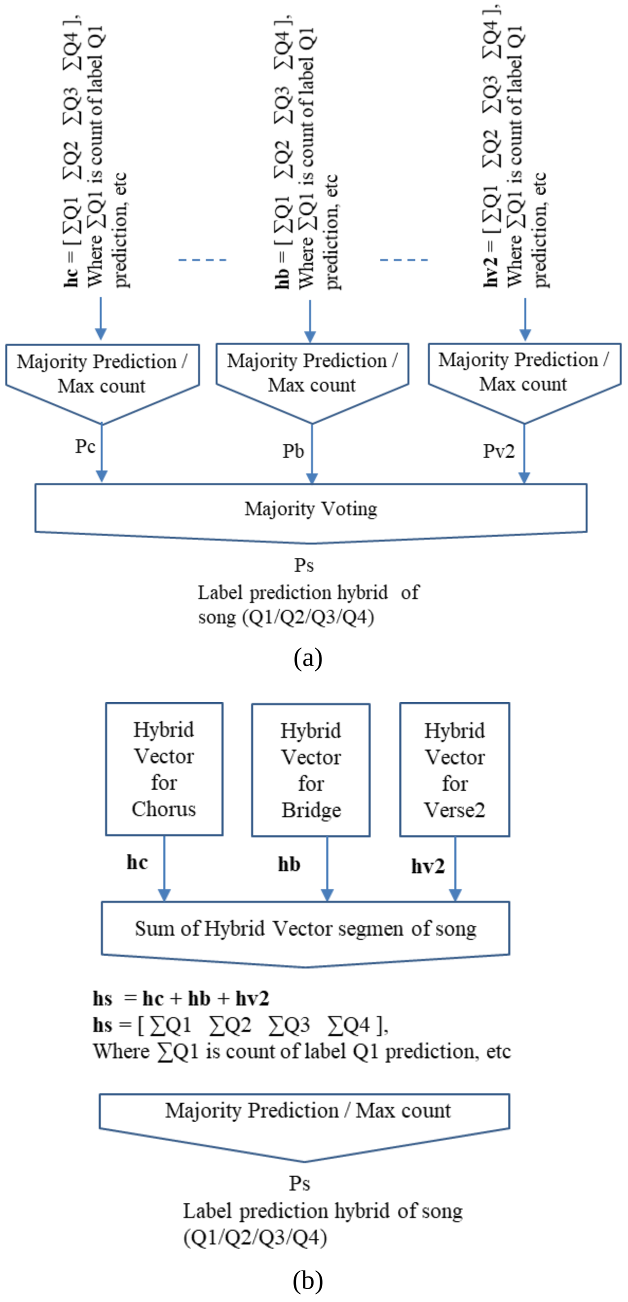

Audio Chorus Dataset
Lyrics Chorus Dataset
Random Forest Classification
Random Forest Classification
P1 P2 P3 .... Pn
P1 P2 P3 .... Pn
Audio Chorus Vector
Lyrics Chorus Vector
ca = [ ∑Q1 ∑Q2 ∑Q3 ∑Q4 ], Where ∑Q1 is count of label Q1 prediction, etc
cl = [ ∑Q1 ∑Q2 ∑Q3 ∑Q4 ], Where ∑Q1 is count of label Q1 prediction, etc
Hybrid Chorus Vector
hc = ca + cl
hc = [ ∑Q1 ∑Q2 ∑Q3 ∑Q4 ],
Where ∑Q1 is count of label Q1 prediction, etc

Figure. 7 Scheme of vector hybrid chorus formation

Furthermore, matrix prediction of lyrics is formed with the same process as forming an audio prediction matrix. Both matrix predictions of the audio and lyrics structural segment is hybridized with the concept of majority voting and the sum of hybrid matrices, to detect the emotions of the whole song. Schematic of forming a hybrid vector, especially for chorus shows in Fig.7. Variable *P1*, *P2*, *P3*, *Pn* are label prediction for each chorus of song (Q1/Q2/Q3/Q4).

### 6.2 Emotion detection algorithm

The next step is detecting emotions by using emotion detection algorithm. There are 12 alternative combination from three (3) models to detect song emotions. Model 1 is the first emotional detection model by using the concept of majority voting. In majority voting, each structural segment is decided its emotional predicted label. It based on maximum value element of a hybrid vector. Emotion label prediction is taken based on the maximum value of vector elements. If each structural segment already has a prediction label, the major prediction label is taken through the concept of majority voting. That major prediction label is the label of the song's emotional prediction. Model 1 shown in Fig. 8(a). Model 2 is a detection emotion model using the Sum of hybrid vector concept from the analyzed structural segments and majority voting. It can be seen in Fig. 8(b). *Pc*, *Pb* and *Pv2* are the abbreviation of predicted chorus label, predicted bridge label and predicted v2 label. Hybrid vector for chorus, bridge and v2 is named *hc*, *hb* and *hv2*. The hybrid matrix has been synchronized between audio and lyrics and *Ps* is label prediction from a hybrid matrix of song.


hc = [ ∑Q1 ∑Q2 ∑Q3 ∑Q4 ],
Where ∑Q1 is count of label Q1
prediction, etc
hb = [ ∑Q1 ∑Q2 ∑Q3 ∑Q4 ],
Where ∑Q1 is count of label Q1
prediction, etc
hv2 = [ ∑Q1 ∑Q2 ∑Q3 ∑Q4 ],
Where ∑Q1 is count of label Q1
prediction, etc
Majority Prediction / Max count
Majority Prediction / Max count
Majority Prediction / Max count
Pc
Pb
Pv2
Majority Voting
Ps
Label prediction hybrid of song (Q1/Q2/Q3/Q4)
Hybrid Vector for Chorus
Hybrid Vector for Bridge
Hybrid Vector for Verse2
hc
hb
hv2
Sum of Hybrid Vector segmen of song
hs = hc + hb + hv2
hs = [ ∑Q1 ∑Q2 ∑Q3 ∑Q4 ],
Where ∑Q1 is count of label Q1 prediction, etc
Majority Prediction / Max count
Ps
Label prediction hybrid of song
(Q1/Q2/Q3/Q4)

(a)

(b)

Figure. 8 Emotion detection model using 3 structural segments (Chorus-Bridge-Verse2): (a) model 1 and (b) model 2

Table 5. Confusion Matrix for Multi Class

| | | Prediction | | | |
|---|---|---|---|---|---|
| | | A | B | C | D |
| Actual | A | TP_A | E_AB | E_AC | E_AD |
| | B | E_BA | TP_B | E_BC | E_BD |
| | C | E_CA | E_CB | TP_C | E_CD |
| | D | E_DA | E_DB | E_DC | TP_D |

## 7. Result and discussion

This study was tested using 12 case studies. Each case study uses 2 to 7 structural segments that represent songs. The 12 case studies were tested using 3 emotion detection models: model 1, model 2 and model 3. Confusion matrix for multiclass show in Table 5. For that confusion matrix, we get True Positive, False Negative and False Positive values. F-Measure calculations for each case study using F-measure formula in Eq. (6). In this formula, $recall$ and $precision$ calculated according to Eq. (4) and (5).

True Positives ($TP$) are on the diagonal position. False Positives ($FP$) are column-wise sums (E_BA, E_CA, E_DA), without the diagonal. False Negatives ($FN$) are row-wise sums (E_AB, E_AC, E_AD), without the diagonal.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F - Measure = 2 \, . \frac{Precision.Recall}{Precision + Recall} \tag{6}$$

In Tables 6 - 8, it can be seen that case studies with the best accuracy are in the structural combination of the chorus-bridge-v3 using emotion detection model 2. The average $F - Measure$ value obtained is 0.823. This is possible because in model 2, hybrid vector are combined first with the sum of vector concept then prediction labels are determined. So there must be a prediction label result.

This study also compared the result of song emotion detection with previous research[12]. But we do not find any features used in [12], so what we compare in this case is the data used as input for emotional detection system.

System 1 in previous research used 30 seconds audio and all lyric [12]. System 2 use hybrid matrix for chorus structural_segment audio [16] and all lyrics. System 3 use hybrid matrix for all structural_segment audio and all lyrics. System 4, use [20] to detect emotion using valence and arousal dimension. System 5, our proposed, uses chorus-

Table 6. Recall test result

| Combination of Structural Segments | $Recall$ | | | | |
|---|---|---|---|---|---|
| | **Audio** | **Lyrics** | **Model 1** | **Model 2** | **Model 3** |
| Chorus-Bridge-V1-V2-V3 | 0.670 | 0.720 | 0.550 | 0.770 | 0.650 |
| Chorus-V1 | 0.661 | 0.670 | 0.420 | 0.800 | 0.800 |
| Chorus-V2 | 0.720 | 0.680 | 0.410 | 0.800 | 0.800 |
| Chorus-V3 | 0.690 | 0.730 | 0.580 | 0.807 | 0.790 |
| Bridge-V1 | 0.490 | 0.350 | 0.505 | 0.525 | 0.499 |
| Bridge-V2 | 0.480 | 0.360 | 0.370 | 0.464 | 0.500 |
| Bridge-V3 | 0.406 | 0.260 | 0.328 | 0.408 | 0.440 |
| Chorus-bridge | 0.680 | 0.700 | 0.590 | 0.810 | 0.810 |
| Chorus-bridge-V1 | 0.700 | 0.700 | 0.510 | 0.790 | 0.730 |
| Chorus-Bridge-V2 | 0.710 | 0.710 | 0.470 | 0.810 | 0.760 |
| Chorus-bridge-V3 | 0.700 | 0.710 | 0.590 | 0.830 | 0.770 |
| Chorus-bridge-V1-V2-V3-intro-outro | 0.720 | 0.650 | 0.388 | 0.730 | 0.660 |

Table 7. Precision test result

| Combination of Structural Segments | $Precision$ | | | | |
|---|---|---|---|---|---|
| | **Audio** | **Lyrics** | **Model 1** | **Model 2** | **Model 3** |
| Chorus-Bridge-V1-V2-V3 | 0.736 | 0.664 | 0.70 | 0.773 | 0.68 |
| Chorus-V1 | 0.658 | 0.678 | 0.788 | 0.807 | 0.807 |
| Chorus-V2 | 0.718 | 0.730 | 0.765 | 0.801 | 0.801 |
| Chorus-V3 | 0.713 | 0.776 | 0.861 | 0.841 | 0.818 |
| Bridge-V1 | 0.490 | 0.350 | 0.622 | 0.571 | 0.551 |
| Bridge-V2 | 0.641 | 0.400 | 0.616 | 0.595 | 0.601 |
| Bridge-V3 | 0.573 | 0.368 | 0.613 | 0.612 | 0.643 |
| Chorus-bridge | 0.698 | 0.745 | 0.817 | 0.833 | 0.833 |
| Chorus-bridge-V1 | 0.703 | 0.711 | 0.82 | 0.796 | 0.75 |
| Chorus-Bridge-V2 | 0.721 | 0.732 | 0.81 | 0.819 | 0.778 |
| Chorus-bridge-V3 | 0.703 | 0.734 | 0.825 | 0.825 | 0.778 |
| Chorus-bridge-V1-V2-V3-intro-outro | 0.737 | 0.645 | 0.474 | 0.728 | 0.653 |

Table 8. $F - Measure$ test result

| Combination of Structural Segments | $F - Measure$ | | | | |
|---|---|---|---|---|---|
| | **Audio** | **Lyrics** | **Model 1** | **Model 2** | **Model 3** |
| Chorus-Bridge-V1-V2-V3 | 0.720 | 0.663 | 0.586 | 0.764 | 0.639 |
| Chorus-V1 | 0.651 | 0.653 | 0.533 | 0.795 | 0.795 |
| Chorus-V2 | 0.717 | 0.676 | 0.515 | 0.795 | 0.795 |
| Chorus-V3 | 0.694 | 0.731 | 0.673 | 0.813 | 0.796 |
| Bridge-V1 | 0.487 | 0.351 | 0.409 | 0.519 | 0.499 |
| Bridge-V2 | 0.524 | 0.367 | 0.424 | 0.484 | 0.513 |
| Bridge-V3 | 0.441 | 0.287 | 0.394 | 0.440 | 0.483 |
| Chorus-bridge | 0.681 | 0.699 | 0.671 | 0.811 | 0.811 |
| Chorus-bridge-V1 | 0.692 | 0.681 | 0.578 | 0.785 | 0.722 |
| Chorus-Bridge-V2 | 0.705 | 0.700 | 0.550 | 0.807 | 0.757 |
| Chorus-bridge-V3 | 0.694 | 0.701 | 0.661 | 0.823 | 0.762 |
| Chorus-bridge-V1-V2-V3-intro-outro | 0.719 | 0.645 | 0.369 | 0.728 | 0.641 |

bridge-v3 that are synchronized between audio and lyrics. By using the emotional detection method of model 2 and Bimodal dataset for testing, the result of $F - Measure$ shown in Fig.9 shows that data input model that we proposed have better results, which is 0.695.

From the results of this research, it can be analyzed by using the second detection emotion model, detection emotion based on lyrics is able to reach the F-measure value of 0.731 and detection emotion based on audio reaches a value of 0.720. Emotion detection based on lyrics and audio can reach Recall value 0.830 (shown in Table 6), Precision value 0.861 (shown in Table 7) and F-Measure value 0.823 (shown in Table 8). The highest Recall and F-Measure are obtained for the combined chorus-bridge-V3 on M2 but the highest precision

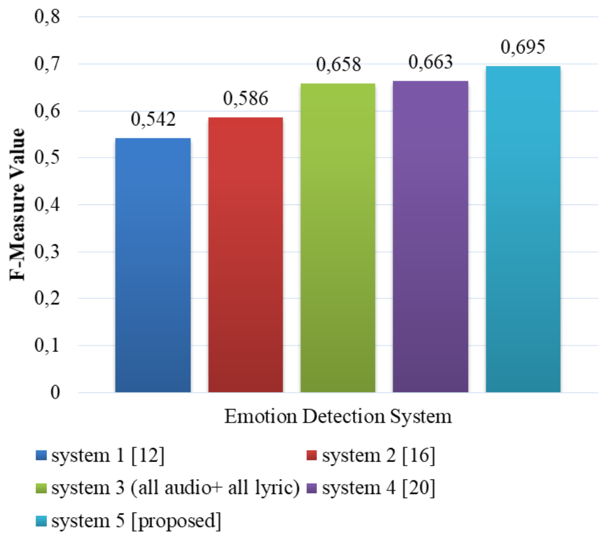


Figure. 9 Comparison song emotion detection result based on input data audio and lyrics

value is obtained for the combined chorus-V3 on M1. F-measure is a combination of precision and recall values, so the reference we use to see the system reliability is the highest F-measure value. Comparison of detection emotion based on audio and lyric features with the previous study also looks better. The F-measure value of proposed system is 0.695.

## 8. Conclusion

The final of this research is to detect the emotion of the whole a song by using synchronized audio and lyrics features. The difference with previous research is the use of input data and the integration of audio and lyrics features. Generally song emotion detection used all lyric and audio data (30-45 seconds) that the duration comes from the expert. In this research song structural segment, audio and lyrics are used as input data. The lyric data are lyrics that are synchronized with the song structural segment audio.

Analyzing of song structural segment that represent a whole emotion of the song is done by 6 classification methods and aggregate method. The result shows that chorus, bridge, and verse are the best segments. The audio features used are extracted using MIRToolbox and the lyrics features are extracted psychologically and stylistically. The results of audio extraction and feature selection process with CFS obtained 16 audio features. In the lyric feature, psycholinguistically extracted 11 features. The lyric extraction process uses CBE-Ex, because system that use it is better than the previous CBE. The F-Measure value is 0.399.

A hybrid approach is used to integrate structural segment audio and lyrics that already synchronized. Data for each structural segment are presented in prediction frequency matrices. Then with a sum of matrices and majority voting, a hybrid matrix is obtained to detect the emotions of the whole song. From 3 emotion detection models for 12 combinations of structural segments, the best F-measure for chorus-bridge-v3 was 0.823.

The development of this research can be done by adding other features to the audio and lyrics features. Word sense disambiguation [30] can be used for additive features of lyrics. Dependency parser [31] is possible to preprocessing data before extracting lyrics. It used to know the relationship between words. When the psycholinguistic feature extraction process is done, the related words can be processed together, so that the results of extraction and detection can be better.

## References

[1] E. Cano, “Mood-Based On-Car Music Recommendation”, *Lecture Notes of the Institute for Computer Science*, No. November, 2016.

[2] J. Jones, “The role of music in your classroom”, in *Music Curriculum Exchange*, Lincoln, pp. 90–92, 2010.

[3] J. A. Ridoean and R. Sarno, “Music Mood Classification Using Audio Power and Audio Harmonicity Based on MPEG-7 Audio Features and Support Vector Machine”, In: *Proc. of International Conference on Science in Information Technology (ICSITech)*, pp. 72–77, 2017.

[4] L. Lu, D. Liu, and H. Zhang, “Automatic Mood Detection and Tracking of Music Audio Signals”, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1, pp. 5–18, 2006.

[5] A. Schindler and A. Rauber, “Capturing the Temporal Domain in Echonest Features for Improved Classification Effectiveness”, In: *Proc. of International Workshop on Adaptive Multimedia Retrieval*, pp. 1–15, 2014.

[6] R. Panda and R. P. Paiva, *Automatic Mood Tracking in Audio Music*. Universidade de Coimbra, 2010.

[7] V. Kumar, “Mood Classifiaction of Lyrics using SentiWordNet”, In: *Proc. of International Conference on Computer Communication and Informatics (ICCCI -2013)*, pp. 1–5, 2013.

[8] Y. Hu, X. Chen, and D. Yang, “Lyric-Based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method”, In: *Proc. of International Society for Music Information Retrieval Conference*, pp. 123–128, 2009.

[9] M. Kim and H. Kwon, “Lyrics-based Emotion Classification using Feature Selection by Partial Syntactic Analysis”, In: *Proc. of International Conference on Tools with Artificial Intelligence*, 2011.

[10] C. Laurier, *Automatic Classification of Musical Mood by Content Based Analysis*. Barcelona, Spain: Universitat Pompeu Fabra, 2011.

[11] J. S. Downie and A. F. Ehmann, “Lyric text mining in music mood classification”, In: *Proc. of International Society for Music Information Retrieval Conference*, pp. 411–416, 2009.

[12] R. Malheiro, R. Panda, P. Gomes, and R. Paiva, “Bi-modal music emotion recognition: Novel lyrical features and dataset”, In: *Proc. of International Workshop on Music and Machine Learning*, pp. 1–5, 2016.

[13] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Sha, and Y. Yang, “1000 Songs Database”, In: *Proc. of ACM International Workshop on Crowdsourcing for Multimedia*, pp. 4–7, 2014.

[14] D. Liu, L. Lu, and H.-J. Zhang, “Automatic mood detection from acoustic music data”, In: *Proc. of the International Conference on Music Information Retrieval*, pp. 13–17, 2003.

[15] M. A. Pandharipande and S. K. Kopparapu, “Audio segmentation based approach for improved emotion recognition”, In: *Proc. of TENCON 2015 - 2015 IEEE Region 10 Conference*, Macao, No. I, pp. 1–4, 2015.

[16] C. Yeh, C.Tseng, W.Chen, C.Lin, Y.Tsai, Y.Bi, H.Lin, Y.Lin, and Ho-yi, “Popular music representation : chorus detection & emotion recognition”, *Multimedia Tools Application*, Vol. 73, pp. 2103–2128, 2014.

[17] O. Lartillot, “MIRtoolbox 1.6.1”, Denmark, 2014.

[18] C. D. Manning, “Part-of-Speech Tagging from 97 % to 100 %: Is It Time for Some Linguistics ?”, In: *Proc. of Computational Linguistics and Intelligent Text Processing*, pp. 171–189, 2011.

[19] F. H. Rachman, R. Sarno, and C. Fatichah, “Music emotion classification based on lyrics-audio using corpus based emotion”, *International Journal of Electrical and Computer Engineering*, Vol. 8, No. 3, pp. 1720–1730, 2018.

[20] V. L. Nguyen, D. Kim, V. P. Ho, and Y. Lim, “A New Recognition Method for Visualizing Music Emotion”, *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 7, No. 3, pp. 1246–1254, 2017.

[21] E. Peiszer, T. Lidy, and A. Rauber, “Automatic Audio Segmentation : Segment Boundary and Structure Detection in Popular Music”, In: *Proc. of LSAS*, Vol. 106, No. August, pp. 45–59, 2008.

[22] S. Ewert, “Chroma Toolbox: Matlab Implementations for Extracting Varians of Chroma-Based Audio Features”, *International Society for Music Information Retrieval*, pp. 1–6, 2011.

[23] O. Lartillot, O. Lartillot, P. Toiviainen, and P. Toiviainen, “A matlab toolbox for musical feature extraction from audio”, In: *Proc. of International Conference on Digital Audio Effects*, No. Ii, pp. 1–8, 2007.

[24] F. H. Rachman, R. Sarno, and C. Fatichah, “CBE : Corpus-Based of Emotion for Emotion Detection in Text Document”, In: *Proc. of ICITACEE*, pp. 331–335, 2016.

[25] C. Strapparava and A. Valitutti, “WordNet-Affect : an Affective Extension of WordNet”, In: Proc. of *LREC*, pp. 1083–1086, 2004.

[26] M. M. Bradley, P. J. Lang, M. M. Bradley, and P. J. Lang, “*Affective Norms for English Words ( ANEW ): Instruction Manual and Affective Ratings*”, The Center for Research in Psychophysiology, University of Florida, 1999.

[27] F. Kaiser, *Music Structure Segmentation*. Berlin: Universit¨at Berlin, 2012.

[28] D. Unal, “Comparison of Data Mining Classification Algorithms Determining the Default Risk”, *Scientific Programming*, pp. 1–8, 2019.

[29] H. Lee and C. Chang, “Comparative analysis of MCDM methods for ranking renewable energy sources in Taiwan”, *Renewable and Sustainable Energy Reviews*, Vol. 92, No. April 2017, pp. 883–896, 2018.

[30] B. S. Rintyarna and R. Sarno, “Adapted Weighted Graph for Word Sense Disambiguation”, In: Proc. of *IcoICT*, 2016.

[31] Z. Jie and W. Lu, “Dependency-based Hybrid Trees for Semantic Parsing”, In: *Proc. of Conference on Empirical Methods in Natural Language Processing*, pp. 2431–2441, 2018.