Audio fingerprint for automatic Balinese rindik music identification using gaussian mixture model

A A R S Widarsa and I D M B A Darmawan

Informatics Department, Faculty of Mathematics and Natural Sciences, Universitas Udayana, Kampus Bukit Jimbaran, Badung, Indonesia

dewabayu@cs.unud.ac.id

Abstract. Rindik is a traditional gamelan instrument that has been known for a long time in which is made from bamboo. Rindik music can be identified by just hearing a little bit snippets from the music that can only be done by professional rindik players. This research assumes doing some training in the system can make the system recognize the music by just using some rindik music snippets. This research focuses on comparing some features extraction method for automatic rindik music identification by using some rindik music snippet to find out which of these features perform the best to represent rindik music. This research will use an audio fingerprint method for song identification. MFCC (Mel Frequency Cepstral Coefficient) and Spectral Subband Centroid will be used and compared as the music’s feature, and Gaussian Mixture Model will be used to model the music’s fingerprint. The result of this research show both methods give excellent results. Both features only need 10-second duration data to get over 90% overall accuracies. Both Feature gives excellent performance with GMM for automatic rindik song identification task. However, Spectral subband centroid show better result with the highest accuracies is 99.3%.

1. Introduction

Rindik is a traditional gamelan instrument that has been known for a long time in which is made from bamboo except for the one that is used for playing it which is called panggul made from rubber. Rindik is played by hitting the bamboo rod or usually called don rindik using the panggul. Rindik has eleven bamboo rods. Each bamboo rode have different notes with laras slendro that start from ndung, ndang, nding, ndong, ndeng, and so on [1]. As time went on, more and more new music emerged from rindik itself, which made rindik musical variations more and more. With so many new rindik music has been made, the existence of classical music from rindik began to be abandoned. Besides, with the number of rindik pieces of music that exist today, a tool or system is needed that can recognize the rindik music to help people to know rindik music better, so there is no rindik music abandoned.

A rindik music can be identified by just hearing a little bit snippets from the music. However, that can only be done by professional rindik players that have been trained and used to listen to those musics. This research assumes that by doing some training to the system can make the system can recognize the rindik music. This research will use audio fingerprint technology to do the song identification task. Audio fingerprint technology extracts relevant acoustic characteristics from audio content and stores them in a database. When there is unknown audio content, the features of the audio chunk are calculated and matched against those stored in the database. Traditional rindik music is different compared to a modern music instrument. Rindik uses laras slendro, and there is no same rindik made because each rindik that made by different people has different frequency structures. Rindik music has a different musical structure where rindik music consists of two rindik where both
rindik playing a different melody one rindik plays a polos melody or melody that goes according to the tempo and one rindik play nyangsih melody or melody that is hit after the polos melody. Because of that, this research needs to find useful musical features that can represent the rindik music well, so that the rindik music snippets can be recognized well by the system. This research will try to use spectral subband centroid and MFCC for the audio features because both features are having good results for audio fingerprint and also audio analysis. For the fingerprint modelling method, this research will use the Gaussian Mixture Model method. This research uses 10 rindik song to identified. GMM are known for modelling data from several groups where each group can be different or similar from another. This research use several rindik music where some of them are similar. Moreover, GMM are used to get really good fingerprints that is robust to audio distortion and signal noise [2] because to get good result, we need audio fingerprint that is robust from audio distortion.

Some research has done audio fingerprinting for recognizing modern songs automatically [2–4]. However, none of them has been applied to traditional Balinese music, especially Balinese rindik music, where Balinese rindik music have a unique frequency structure if compared to modern music. Music identification using audio fingerprint technology that is done in [2–4] shows that audio fingerprint can produce high accuracy for this task, where the accuracy is about 80%. Besides music identification task, audio fingerprint also often used in speech recognition [5–7] and produce good speech recognition accuracy. Research that has been done by [2,8,9] shows that spectral subband centroid can be a useful feature for music identification and speech recognition by giving 90% accuracy for speech recognition and music identification. Moreover, [2,4,10,11] shows that MFCC can be a useful feature for audio analysis and music retrieval because it can characterize the sound approaching the human auditory systems. Gaussian mixture model proved it could give good performance by research [2,11,12], where GMM is being used to model the feature and can produce a good performance.

This research focuses on comparing some features extraction method for automatic rindik music identification by using some rindik music snippet to find out which of these features perform the best to represent rindik music. The rest of this paper will be organized as follows. Section 2 will contain a brief description of the method used. Section 3 will explain the results of the system testing and also a discussion about MFCC and Spectral Subband Centroid as features used in identification of rindik music. Section 4 will explain the conclusion of this research.

2. Method
Figure 1 shows the general process of the system that used in this research, as another audio fingerprint method, this research consists of two processes [13], it is fingerprint modelling or training process and identification process. The training method will take full rindik music as input data and will produce a GMM model that will be used in the identification process. For the identification process, it will take two inputs, it is the rindik music snippets and the GMM model. The final result of the process is the music name that corresponded to the rindik music snippets.

![Diagram](image_url)

**Figure 1.** General process
As shown in Figure 2, the training process will extract the audio features first. In this research, there will be two audio feature extraction method used which is MFCC (Mel Frequency Cepstral Coefficient) with 13 coefficient and Spectral Subband Centroid. The Audio feature will be used to train the system by using GMM. Each feature will be used separately, so there will be two models produced by this process.

![Figure 2. Detailed training process](image)

Figure 2. Detailed training process

Figure 3 shows a more specified identification process. In this process, the audio snippets will processed to get the rindik music corresponding to the audio snippets. Matching process in this research will measure the log-likelihood between audio snippets and the GMM model in the database. Rindik music with the highest score will be chosen as the right rindik music for the audio snippets. Both feature extraction methods will be used separately in this process. If the output from this process matches the corresponding rindik music, then it will be marked as right, and vice versa. This research will compare the accuracy of both the feature extraction method to find the best method combination for GMM between both features for automatic song identification with rindik music.

![Figure 3. Identification process](image)

Figure 3. Identification process

2.1. Audio Fingerprint
Audio fingerprint is a compact content-based signature to encapsulate audio recordings. Audio fingerprint or content-based identification (CBID) technology extracts relevant acoustic characteristics from an audio content and stores them in a database. When there is unknown audio content, the characteristics of the audio chunk calculated and matched with those stored in the database [2]. By using the fingerprint and matching algorithms, any distorted audio from a single recording can be identified as the same music title.

The fingerprint generated from the fingerprint extraction process is a set of relevant perceptual characteristics of the recording in a concise and robust form. During the matching process, a record will be given to the system to match the fingerprint of the model in the database. The distance between the records and the fingerprint models in the database will calculated to determine the suitability of the existing fingerprint models.
2.2. Feature Extraction.
There are two audio features used in this research: MFCC and Spectral Subband Centroid. Both of the features will be extracted separately and used separately. The features with higher accuracy mean the feature is better in characterizing the rindik music. Before MFCC and Spectral Subband Centroid extracted, pre-processing like pre-emphasis, framing and windowing, and FFT are applied to the audio data. The FFT result is used to extract audio features.

2.2.1. MFCC
Mel-Frequency Cepstrum Coefficients (MFCC) are popular acoustic features with high success in music searches, as they approach the human auditory system [12]. MFCC is a suitable feature vector for representing human voices or music signals [7]. MFCC is calculated by extracting the power spectrum of each frame, then calculating it using FFT. Furthermore, Mel-filter bank is used to map the audio power spectrum into the Mel Scale. The Mel-filter bank consists of a set of overlapping triangular band-pass filters, which are uniformly placed on the Mel scale. The Mel-filter bank equation to find the frequencies that is in the mel scale written in equation (1).

\[
F_{mel} = \begin{cases} 
F_{real} \cdot F_{real} < 1000 \\
2595 \log_{10} (1 + \frac{F_{real}}{700}) \cdot F_{real} > 1000 
\end{cases}
\]  

(1)

where \(F_{mel}\) is the frequencies in the mel scale, and \(F_{real}\) are the real frequencies before mapped to the mel scale. Based on the mel scale, for frequencies below 1kHz \((f < 1\text{kHz})\) it will be linear, and frequencies above 1kHz \((f > 1\text{kHz})\) will be logarithmic.

Furthermore, the frequencies obtained from the FFT results will be grouped into filters. The grouping process is done by multiplying the FFT value by the corresponding gain filter, and the results will be added. The frequency wrapping equation is written in equation (2).

\[
X_i = \log_{10}(\sum_{k=0}^{N-1} X(k) \cdot H_f(k))
\]

(2)

where \(X\) is the frequency wrapping value in the filter where \(i = 1,2,3,\ldots,n\) (number of filters) , \(X(k)\) is the frequency magnitude value of \(k\) frequencies , and the height value of the filter \(i\) triangle and \(k\) frequency with values \(k = 0,1,\ldots,N-1\).

Then Discrete Cosine Transform (DCT) is applied to reduce the correlation between the coefficients and compress the information into lower-order coefficients. DCT equation written in equation (3). MFCC then represents the amplitude of the resulting spectrum [12]. In this study, the number of MFCC components used is 13. The results of the vector feature of this MFCC will be used as input in the formation stage of the Gaussian Mixture Model.

\[
Y_j = \sum_{i=0}^{F} X(i) \cdot \cos\left(\frac{j(i-1)}{2} \cdot \frac{\pi}{F}\right)
\]

(3)

with \(Y\) is the \(j^{th}\) coefficient, where \(j=1,2,3,\ldots\) (numbers of coefficient used), \(X(i)\) is the \(X\) value on the frequency wrapping with \(i=0,1,2,\ldots,n\) (numbers of wrapping) and \(F\) is the filter numbers.

2.2.2. Spectral Subband Centroid
Spectral Subband Centroid is the centre of gravity of the spectral energy for the frequency band. The spectral centroid defined as the sum of the power frequencies of the normalized power spectrum with an unweighted number. The frequency band is divided into a fixed number of subbands, and the centroid is computed for each subband using the power spectrum of the audio signal [8,9]. Filterbank applied to divide the frequency band into a fixed number of frequency subbands. The centroid of each filterbank will extracted between lower and higher edges of each subband. The Spectral Subband Centroid equation is written in equation (4).
where $\text{SC}(n)$ is the spectral centroid of the $n$-th band, $l_n$ and $h_n$ is the lower and higher edge of the frequency subband, $w_n(f)$ is the filter shape, $P(f)$ is the power spectrum, and $\gamma$ is a constant controlling the dynamic range of the power spectrum where $\gamma < 1$.

2.3. Fingerprint Modelling
The audio feature from the feature extraction process will form the fingerprint of the rindik music. The fingerprint modelling process is done by using the Gaussian mixture model method. This process will produce the rindik songs fingerprint in the probability density function made in the GMM process.

Gaussian Mixture Model
Gaussian Mixture Model (GMM) is a function that is useful for modelling data from several groups, where each group can be different from one another. However, data points that are in a group can be modelled well by the Gaussian Mixture Model. Generally, GMM used as a parametric model of a continuous measurement probability distribution or features for biometric systems, such as audio-channel related spectral features in automatic song identification systems. The Gaussian density function, as shown in [14] written in equation (5).

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\}$$ (5)

where $g(x|\mu_i, \Sigma_i)$ is the gaussian density function, $x$ is the D-dimensional continuous data vector (audio features), $\mu_i$ is the mean vector of the features vector and $\Sigma_i$ is the covariance matrix. The gaussian Mixture model is a weighted sum of $M$ component gaussian density function which is written in equation (6).

$$f(x|\theta) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i)$$ (6)

with $f(x|\theta)$ is the gaussian mixture model, $w_i$ as the weight where $w_i > 0$, $\sum_{i=1}^{M} = 1$ and $g(x|\mu_i, \Sigma_i)$ is the gaussian density function.

The EM algorithm serves to identify each gaussian in the Gaussian Mixture distribution to find maximization. Maximization meant to find and develop each gaussian to become a more reasonable condition. In the identification phase, the fingerprint from the test audio file compared with the previously computed GMM model database, and the model that gives the highest log-likelihood for the fingerprint identified as the correct match for the test audio. With the feature vectors $x_1, x_2, \ldots, x_n$ the maximum likelihood estimation for a likelihood function is shown in equation (7)

$$\theta_{ML} = \arg \max \sum_{i=1}^{n} \log f(x_i|\theta)$$ (7)

where $\theta_{ML}$ is the maximum likelihood and $i$ is the number of feature vectors.

3. Result and Discussion
This research will test 10 rindik musics that obtained by recording it manually from the rindik player. There are three audio recordings per each music used in this research to make more samples used in training to make more probability because each recording is unique to each other. Each recording is using the mono channel and 44100 sample rate. For the gaussian mixture model, this research use the Scikit-learn module to do the Gaussian mixture modelling task and log-likelihood measurement for the identification task [15]. For MFCC and Spectral Subband Centroid feature extraction, this research use the Python Speech Features module [16].

Each rindik music will modelled using three audio recordings. This research uses audio snippets with 5 seconds, 10 seconds, and 15 seconds to test the system. Each audio snippet picked from the random part of each audio recording. Each audio recording will produce ten audio snippets; therefore, there will be 300 audio snippets to test the system. The system will test the audio snippets for each features, so that the system will do 300 testings for each duration and each feature. So the total testing is 1800 times.
3.1. MFCC Result
To test the MFCC feature for the rindik music, this research use 13 MFCC coefficients. The result is shown in Table 1. From the table, MFCC shows a good result. Most of the testing result shows over 80% accuracy. In some music, it shows 70% and 60% accuracy. As expected, MFCC can show excellent results. Between 3 duration, the most effective duration for the identification is 10 seconds. This duration has the highest accuracy between 3 duration, and its average identification time is not far from the 5-second duration. Compared to 15-second data, the 10-second data shows the better result, in 10-second duration, it has 100% accuracy in 6 music. However, in 15-second duration, it only has 10% accuracy. The 15-second duration data have the longest average identification time. There is some rindik music that has under 80% accuracy, that is, Gending 2 and Gending 5. It can be caused by the music’s melody structure caused by some kind of pattern in the music. Some rindik patterns cannot be recognized well by MFCC features.

| Rindik Music | 5-second audio result | 10-second audio result | 15-second audio result |
|--------------|-----------------------|------------------------|------------------------|
|              | Correct | Accuracy (%) | Avg time (s) | Correct | Accuracy (%) | Avg time (s) | Correct | Accuracy (%) | Avg time (s) |
| Gending 1    | 30      | 100         | 1.06        | 30      | 100         | 0.85        | 29       | 96.67        | 1.04        |
| Gending 2    | 20      | 66.67       | 0.9         | 23      | 76.67       | 0.80        | 19       | 63.33        | 0.81        |
| Gending 3    | 28      | 93.33       | 0.48        | 30      | 100         | 0.79        | 29       | 96.67        | 0.93        |
| Gending 4    | 25      | 83.33       | 0.47        | 30      | 100         | 0.62        | 28       | 93.33        | 1.17        |
| Gending 5    | 23      | 76.67       | 0.45        | 23      | 76.67       | 0.61        | 22       | 73.33        | 0.90        |
| Gending 6    | 29      | 96.67       | 0.46        | 29      | 96.67       | 0.57        | 29       | 96.67        | 0.88        |
| Gending 7    | 30      | 100         | 0.47        | 30      | 100         | 0.58        | 30       | 100          | 0.89        |
| Gending 8    | 29      | 96.67       | 0.47        | 30      | 100         | 0.57        | 29       | 96.67        | 0.88        |
| Gending 9    | 29      | 96.67       | 0.48        | 30      | 100         | 0.57        | 29       | 96.67        | 0.89        |
| Gending 10   | 28      | 93.3        | 0.48        | 29      | 96.67       | 0.80        | 29       | 96.67        | 0.86        |
| Total        | 272     | 90.67       | 0.537       | 283     | 94.33       | 0.680       | 274      | 91.33        | 0.93        |

3.2. Spectral Subband Centroid Result
The spectral subband centroid testing result is shown in Table 2. Table 2 shows Spectral Subband Centroid has better performance than the MFCC in all three duration, especially the 10-second duration data almost get 100% accuracy. That because the centroid extracted from each frequency band that helps the system to recognize the complicated rindik music structure. That means the Spectral Subband Centroid is good enough to characterize the rindik music by recognizing more audio snippets than MFCC. Between all three duration, the 10-second duration has the best result overall. The 10-second duration data has the highest accuracy (99.3%), and the average identification time is 0.59 second. If compared to other duration, the 10-second duration has the most 100% identification time, and the average identification time is not much different. None of the test results has under 80% accuracies on Spectral Subband Centroid, which mean this feature can characterize the rindik music well and not be affected by the melody pattern of rindik music.
The best result from GMM to do the automatic rindik music identification task is 94.33\% accuracy and average identification time 0.680 seconds. However, all of the rindik music can be recognized well by the Spectral Subband Centroid, the average time difference is 0.17 second. These results mean that both of these feature extraction methods work well with GMM to do the automatic rindik music identification task, and both features can characterize the rindik music well even the Spectral Subband Centroid has better results in all three duration.

Unlike MFCC, Spectral Subband Centroid does not have identification results that has under 80\% accuracies. Whereas MFCC has two identification results that have under 80\% accuracy. Those results may be affected by the rindik music melody patterns. Some melody pattern cannot be recognized by the rindik music well. However, all of the rindik music can be recognized well by the Spectral Subband Centroid. The Spectral Subband Centroid identification result was not affected by the rindik music melody pattern.

### Table 2. Spectral subband centroid result

| Rindik Music | 5-second audio result | 10-second audio result | 15-second audio result |
|-------------|-----------------------|------------------------|-----------------------|
|             | Correct | Accuracy(\%) | Avg time(s) | Correct | Accuracy(\%) | Avg time(s) | Correct | Accuracy(\%) | Avg time(s) |
| Gending 1   | 30      | 100         | 0.88       | 30      | 100         | 0.59       | 29       | 96.67       | 0.85       |
| Gending 2   | 30      | 100         | 0.37       | 30      | 100         | 0.59       | 29       | 96.67       | 0.79       |
| Gending 3   | 29      | 96.67       | 0.39       | 30      | 100         | 0.60       | 29       | 96.67       | 0.75       |
| Gending 4   | 27      | 90          | 0.39       | 29      | 96.67       | 0.59       | 25       | 83.33       | 0.78       |
| Gending 5   | 28      | 93.33       | 0.36       | 30      | 100         | 0.59       | 29       | 96.67       | 0.77       |
| Gending 6   | 29      | 96.67       | 0.38       | 29      | 96.67       | 0.58       | 28       | 93.33       | 0.79       |
| Gending 7   | 30      | 100         | 0.35       | 30      | 100         | 0.58       | 30       | 100         | 0.77       |
| Gending 8   | 29      | 96.67       | 0.59       | 30      | 100         | 0.58       | 29       | 96.67       | 0.76       |
| Gending 9   | 28      | 93.33       | 0.35       | 30      | 100         | 0.58       | 29       | 96.67       | 0.77       |
| Gending 10  | 30      | 100         | 0.61       | 30      | 100         | 0.59       | 29       | 96.67       | 0.75       |
| Total       | 290     | 96.67       | 0.42       | 298     | 99.33       | 0.59       | 287      | 96.67       | 0.78       |

### 4. Conclusion

Conclusions can be drawn from the research result for automatic rindik music identification with a gaussian mixture model in section 3. Both of the features show excellent results for the automatic identification task, but spectral subband centroid show better result than MFCC. Both Features only need 10-second duration data to produce the best result compared to other duration data, that means the system just need 10-second rindik music snippets to recognize the rindik music well. The 10-second duration data has average identification time not much different than the 5-second duration data. For MFCC, the average time difference is 0.15 second, and for the Spectral Subband Centroid, the average time difference is 0.17 second. These results mean that both of these feature extraction methods work well with GMM to do the automatic rindik music identification task, and both features can characterize the rindik music well even the Spectral Subband Centroid has better results in all three duration.
References

[1] Dibia I W 1977 Pengantar karawitan Bali (Denpasar: Proyek Peningkatan/Pengembangan ASTI)

[2] Ramalingam A and Krishnan S 2006 Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting IEEE Trans. Inf. Forensics Secur. 1 457–63

[3] Fan Y and Feng S 2016 A music identification system based on audio fingerprinting 2016 4th Intl Conf Appl. Comput. Inf. Technol. Intl Conf Comput. Sci. Appl. Informatics/1st Intl Conf Big Data, Cloud Comput. Data Sci. Eng. 363–7

[4] Negi Y S 2009 Evaluation of audio based searching for Indian traditional music ACT 2009 - Int. Conf. Adv. Comput. Control Telecommun. Technol. 255–9

[5] Abu M, Amir A, Zahri N A H and Ngadiran R 2020 Voice-based Malay commands recognition by using audio fingerprint method for smart house applications IOP Conf. Ser. Mater. Sci. Eng. 767

[6] Leu F Y and Lin G L 2017 An MFCC-based speaker identification system 2017 IEEE 31st Int. Conf. Adv. Inf. Netw. Appl. An 1055–62

[7] Winursito A, Hidayat R and Bejo A 2018 Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition 2018 Int. Conf. Inf. Commun. Technol. ICOIAC T 2018 379–83

[8] Seo J S, Jin M, Lee S, Jang D, Lee S and Yoo C D 2005 Audio fingerprinting based on normalized spectral subband centroids ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. III 213–6

[9] Paliwal K K 1998 Spectral subband centroid features for speech recognition ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2 617–20

[10] Das P P, Allayear S M, Amin R and Rahman Z 2016 Bangladeshi dialect recognition using Mel frequency cepstral coefficient, delta, delta-delta and Gaussian mixture model Proc. 8th Int. Conf. Adv. Comput. Intell. ICACI 2016 359–64

[11] Panagiotou V and Mitianoudis N 2013 PCA summarization for audio song identification using Gaussian mixture models 2013 18th Int. Conf. Digit. Signal Process. DSP 2013

[12] Kaur C and Kumar R 2018 Study and analysis of feature based automatic music genre classification using Gaussian mixture model Proc. Int. Conf. Inven. Comput. Informatics, ICICI 2017 465–8

[13] Cano P, Batle E, Kalker T and Haitsma J 2002 A review of algorithms for audio fingerprinting Proc. 2002 IEEE Work. Multimed. Signal Process. MMSP 2002 169–73

[14] Reynolds D 2009 Gaussian mixture models Encycl. Biometrics 659–63

[15] Pedregosa F, Varoquaux G, Buitinck L, Louppe G, Grisel O and Mueller A 2015 Scikit-learn GetMobile Mob. Comput. Commun. 19 29–33

[16] Lyons J 2020 Python_Speech_Features jameslyons/python_speech_features: release v0.6.1 (Version 0.6.1). Zenodo. http://doi.org/10.5281/zenodo.3607820