Journal of Computer Science and Cybernetics, V.36, N.4 (2020), 325–345
DOI 10.15625/1813-9663/36/4/15609

SOME NEW RESULTS ON AUTOMATIC IDENTIFICATION OF VIETNAMESE FOLK SONGS CHEO AND QUANHO

CHU BA THANH[1,2,*], TRINH VAN LOAN[1,2], NGUYEN HONG QUANG[2]

[1] *Faculty of Information Technology, Hung Yen University of Technology and Education*
[2] *School of Information Communication and Technology, Hanoi University of Science and Technology*

Abstract. Vietnamese folk songs are very rich in genre and content. Identifying Vietnamese folk tunes will contribute to the storage and search for information about these tunes automatically. The paper will present an overview of the classification of music genres that have been performed in Vietnam and abroad. For two types of very popular folk songs of Vietnam such as *Cheo* and *Quanho*, the paper describes the dataset and Gaussian Mixture Model (GMM) to perform the experiments on identifying some of these folk songs. The GMM used for experiment with 4 sets of parameters containing Mel Frequency Cepstral Coefficients (MFCC), energy, the first and the second derivatives of MFCC and energy, tempo, intensity, and fundamental frequency. The results showed that the parameters added to the MFCCs contributed significantly to the improvement of the identification accuracy with the appropriate values of Gaussian component number M. Our experiments also showed that, on average, the length of the excerpts was only 29.63% of the whole song for *Cheo* and 38.1% of the whole song for *Quanho*, the identification rate was only 3.1% and 2.33% less than the whole song for Cheo and *Quanho*, respectively. The identification of *Cheo* and *Quanho* was also tested with i-vectors.

Keywords. Identification; Folk songs; Vietnamese, *Cheo*; *Quanho*; GMM; MFCC; Excerpt; Tempo; F0; i-vectors.

1. INTRODUCTION

The researches related to music data mining are very diverse and have been going on for many years in various ways: genre classification, artist/singer identification, emotion/mood detection, instrument recognition, music similarity searching... However, the research on music genre classification is the most complex and very difficult problem to solve.

There were so many researches which have been done in music genre classification with the various approach to problem-solving, such as Naïve Bayes Classifier (NBC) [1,2], Decision Tree Classifier (DTC) [3], K Nearest Neighbor (KNN) [4, 5], Hidden Markov Model (HMM) [6–8], GMM [9–11], Support Vector Machine (SVM) [12, 13] and Artificial Neural Network (ANN) [14–16].

The general architecture diagram of a music genre classification is as shown in Figure 1 [17]. In general, researches in this field can be divided into two steps. The first step is to extract the features from the music signal, and the second step is to use machine learning

[*] Corresponding author.
E-mail addresses: thanhcb.fit@utehy.edu.vn (C.B.Thanh); loantv@soict.hust.edu.vn (T.V.Loan); quangnh@soict.hust.edu.vn (N.H.Quang).

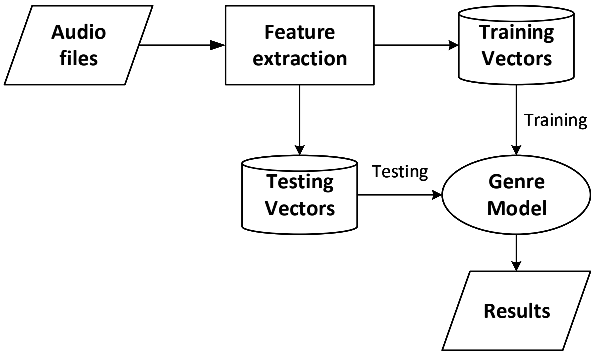


Figure 1. Diagram of the music genre classification system [17]

algorithms for training and testing. The accuracy rate depends on the variants, parameters of the algorithm, and the number of features used.

Vietnam is a multi-ethnic country with a long history of culture, so Vietnamese folk songs are multifarious. The Vietnamese folk songs are available in many regions with different genres: In the North, we have *Quanho* Bac Ninh, Cheo, Xoan singing, Vi singing, Trong Quan singing, Do singing ...; there are singing such as Vi Dam, Ho Hue, Ly Hue, Sac bua in the Middle...; In the South, there are Ly, Ho, poetry speaking...; In the northern mountainous region there are folk songs of Thai, H'Mong, and Muong ethnic groups; In the Highlands, there are folk songs of Gia Rai, Ede, Ba Na, Xo Dang... each has their own identity.

There are many types of folk songs in Vietnam, but *Cheo* and *Quanho* are two very popular types, and their number of songs is richer than others. According to the statistics in [18], there are 213 songs of *Quanho*, and from composer Mai Thien's statistics [19–23] and his notes, there are 190 songs of *Cheo*.

In our published paper [24], we performed the identification experiment with 10 *Quanho* folk songs on the dataset including 100 files using the WEKA toolkit with SMO (Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier), multilayer perceptron and multiclass classifier (A metaclassifier for handling multi-class datasets with 2-class classifiers) algorithms. The results showed that the average accuracy rate for these algorithms is 89%, 86%, and 71% respectively. Also, in this dataset, we have tested on GMM model [25] and the highest average accuracy rate was 79%. In the most recently published paper [26], we used the GMM model for classification and identification on the dataset including 1000 files of two types of folk songs *Cheo* and *Quanho*, each type of 500 files of 25 folk songs. The classification accuracy of two types of *Cheo* and *Quanho* folk songs is 91.0%, and the highest identification rates are 81.6% for *Cheo* folk songs and 85.60% for *Quanho* folk songs. In this paper, we present the results of the *Cheo* and *Quanho* identification using different parameter sets for the full length of audio files and the short excerpts with variable lengths.

This paper is organized as follows: Section 2 is an overview of music genre classification, Section 3 describes the dataset and GMM for our experiments, Section 4 gives the experiment results. Finally, the conclusion is done in Section 5.

2. AN OVERVIEW OF MUSIC GENRE CLASSIFICATION

Although researches on music data mining have been going on for a long time, however, research scale is still limited. Based on ongoing researches from around in the world, the International Symposium on Music Information Retrieval (ISMIR) was officially launched on October 23-25, 2000, in Massachusetts, USA. Since then the symposium has been held annually and is the world's leading research forum on the processing, searching, organization, and extraction of information directly related to music.

One of the most cited papers in music genre classification is that of Tzanetankis et al. [27]. In this paper, the authors conducted a classification experiment on a dataset of six genres (Classical, Country, Disco, Hip Hop, Jazz, Rock) with nine features (Mean-Rolloff, Mean-Flux, Mean-Zero Crossings, Std-Centroid, Std-Rolloff, Std-Flux, Std-Zero Crossings, Low Energy). The results of the experiment have the highest accuracy rate of 58.67%. The same experiment was conducted in 2003 by Tao Li et al., but they have proposed a new extraction feature method called DWCH (Daubechies Wavelet Coefficient Histograms). This method has improved the accuracy rate up to 78.5%.

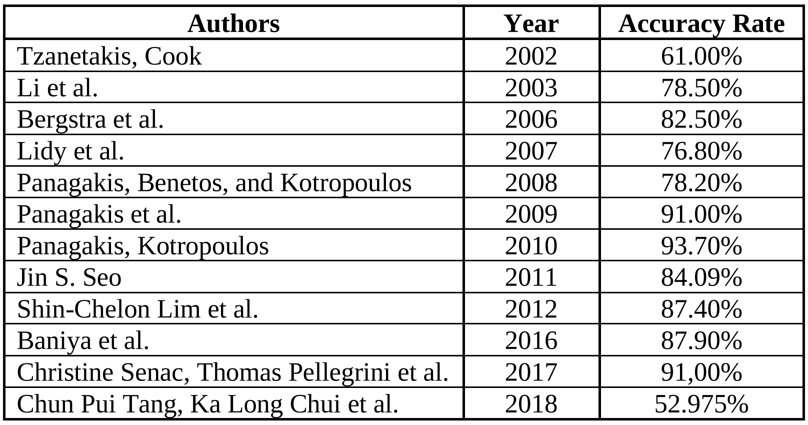
In 2002, George Tzanetankis had completed the dataset named GTZAN [28]. It consists of 10 genres (Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, and Rock), each genre has 100 excerpts, and length of each excerpt is 30 seconds. Tzanetakis & Cook [29] had experimented with classification on this dataset with sets of features including 30-dimensional features vector (19 FFT, 10 MFCC, the rhythmic content features (6 dimensions) and the pitch content features (5 dimensions)). The result has improved the average accuracy rating to 61%. This dataset has been using by many authors for music genre classification experiments with different features.

West & Cox [30] conducted an assessment of the factors affecting the automatic classification of musical signals. Firstly, they describe and evaluate the classification performance of two different evaluation methods based on spectral shape characteristics, Mel frequency filters, and spectral contrast characteristics. Secondly, their study time models of selected features from music and finally minimize the number of dimensions in the characteristic vector. It was then used to train and evaluate the performance of different classifiers and the results showed an increase of accuracy rate.

Mohd et al. [31] used Marsyas software to extract the same features as Tzanetankis et al. in 2001. Using the J48 classifier in WEKA software (a tool for preprocessing, classifying, clustering, selection of features and modeling) [32] and the dataset is Malaysian music. The results show that factors affecting classification results include features, classifiers, size of the dataset, length of the excerpts, the location of the excerpt in the original and parameters in classifiers.

Bergstra et al. [33] suggested that it would be better to classify the features synthesized from a set of audio signals by using a generic vector for each song or classification into individual features vectors. They focus on synthesizing frame-level features into appropriate segments and classifying and experimenting with the segment itself. Their software won the first prize in the MIREX 2005 competition for the music information retrieval. They used two datasets, with about 1500 original songs. Features are extracted in frames of 1024 samples and m frames are synthesized. With $m = 300$ for each segment of 13.9s, the accuracy rate was 82.3%. They conducted the experiment on the GTZAN dataset with the same parameters set, the accuracy rate reached 82.5%.

Table 1. A summary of the experimental results on the GTZAN dataset



Authors	Year	Accuracy Rate
Tzanetakis, Cook	2002	61.00%
Li et al.	2003	78.50%
Bergstra et al.	2006	82.50%
Lidy et al.	2007	76.80%
Panagakis, Benetos, and Kotropoulos	2008	78.20%
Panagakis et al.	2009	91.00%
Panagakis, Kotropoulos	2010	93.70%
Jin S. Seo	2011	84.09%
Shin-Chelon Lim et al.	2012	87.40%
Baniya et al.	2016	87.90%
Christine Senac, Thomas Pellegrini et al.	2017	91,00%
Chun Pui Tang, Ka Long Chui et al.	2018	52.975%

The results of classification by Panagakis et al in 2009 and 2010 on the GTZAN dataset were 91% and 93.7%. In 2009, they used SRC (Sparse Representation-based Classifier) technique to reduced dimensions of represent information. By 2010, the authors used TNPMF (Topology Preserving Non-Negative Matrix Factorization) to reduce dimensions instead of the SRC. Nowadays, the GTZAN dataset is still widely used by researchers in music classification by genre. Table 1 is a summary of the experimental results on the GTZAN dataset from 2002 to 2018 (sorted by ascending of the year).

The first research with folk songs was conducted by Wei Chai and Barry Vercoe [34] in the Multimedia Laboratory of the Massachusetts Institute of Technology in 2001. The dataset includes 187 Irish folk songs, 200 German folk songs, and 104 Austrian folk songs. This dataset was collected from (1) Helmut Schaffrath's folk collection Essen (Germany) and (2) an Irish music collection by Donncha Ó Maidín. The authors used the HMM tool with the scale of data for training and testing assigned to 70% and 30%. The highest accuracy rate when using binary classification between the union of three music genres including Irish - Germany, Irish - Austrian, and Germany - Austrian was 75%, 77%, and 66%. The result of the classification of the three music genres has the highest accuracy rate of 63.0%.

Until 2015, Nikoletta Bassiou and his colleagues [35] experimented with the classification of Greek folk songs into two genres by using CCA (Canonical Correlation Analysis) technique between the lyrics and the sound. The dataset for experiment includes 98 songs from Pontus and 94 songs from Asia Minor, in which 75% data for training and 25% data for testing. Using the cross-evaluation method, the accuracy result is an average of 5 times testing and achieved 97.02%.

In 2016, Rajesh, Betsy, and DG Bhalke [36] conducted the classification Tamil folk songs (Southern India) on a dataset of 216 songs (103 traditional songs + 113 folk songs) with 30 seconds duration for each song. The data for training in each type is 70 songs and the data for testing is 33 songs and 43 songs for each type. The classifier is KNN, the accuracy rate is 66.23%, and with the SVM classifier, the accuracy rate is 84.21%. For Chinese folk songs, in 2017 Juan Li, Jianhang Ding, and Xinyu Yang proposed the GMM-CRF (Conditional Random Field) [37–39] model and used it for music classification by region on the dataset including 344 Chinese folk songs (109 Shaanxi, 101 Jiangsu and 134 Hunan). On average, the highest accuracy rate reached 83.72% [40].

Most recently in 2018, Juan Li et al. [41] overcame the limitations of the GMM-CRF model (improving calculation accuracy when the number of Gauss components is restricted) by proposing the GMM-RBM (Restricted Boltzmann Machine) model [42, 43] for the classification experiment by region on the Chinese folk songs dataset including 297 folk songs from Northern Shaanxi, 278 from Jiangsu, and 262 from Hunan. The experiment results showed that the GMM-RBM model gives the better results (84.71%) than the GMM-CRF model (83.72%) [40].

In Vietnam, Phan Anh Cang and Phan Thuong Cang [44] conducted a music genre classification experiment on the GTZAN dataset. They used the discrete wavelet transform to extract 19 timbral features, 6 beat features, and 5 pitch features. The experiment used k-NN classifier (with $k = 4$), the highest accuracy result is 83.5%.

The Zalo AI Challenge [45] was first held in Vietnam in 2018. In this competition, the group of Dung Nguyen Ba used CNN model for music genre classification on the Vietnam's music dataset including ten classes with 867 files. As a result, they won the first prize in this competition.

The following sections will describe dataset, GMM for our experiment on *Cheo* and *Quanho* classification and the classification results.

3. DATASET AND GAUSSIAN MIXTURE MODEL

3.1. Vietnamese folk songs dataset

Our dataset has a total of 1000 audio files equally distributed between two types of folk songs *Cheo* and *Quanho*, each file duration is of 45-60 seconds with a sample rate at 16kHz and 16 bits per sample. *Cheo's* dataset is extracted from 25 songs, each song has 20 audio files. *Quanho's* dataset is also extracted from 25 songs with 20 audio files for each song. The average full length of *Cheo* songs is 54 seconds and this value is 43 seconds for *Quanho* songs.

3.2. Gaussian mixture model

In image and speech processing and some other areas, neural networks with deep learning techniques have enabled significant results. However, traditional models and classifiers are still used in the pattern recognition field. The GMM model was used in studies related to music data processing and music genre classification [46, 51]. Over the last few years and so far, GMM has continued to be used for music genre recognition, indexing, and retrieval of music [52–60]. This is because the GMM model is characterized by the parameters related averages and variance of data also allow modeling of data distribution with optional precision. In the same way, GMM proved appropriate for the problem of recognizing information contours such as speaker recognition, dialect recognition, language identification, emotion recognition, and music genre identification [61–67]. On the other hand, in terms of model implementation, GMM allows for training in a much shorter time than ANN, leading to unnecessary use of complex and expensive hardware configurations including GPUs. Therefore, for our research, in addition to our research on the ANN model [68], GMM has been selected as one of the models or classifiers to identify music genres.

The GMM model with mixed Gaussians distribution can be considered as a linear superposition of Gaussian distributions in the form [10]

$$p(\mathbf{x}) = \sum_{k=1}^{M} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k). \tag{1}$$

When using GMM for classification Vietnamese folk songs, $\mathbf{x}$ in (1) is the data vector that contains the set of features vector of each folk song in which each element of the set has D dimensions. $\pi_k (k = 1...M)$ are the weights of the mixture that satisfy the condition $\sum_{k=1}^{M} \pi_k = 1$. Each Gaussian density function is a component of the mixture with mean μ_k and covariance Σ_k

$$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x} - \mu_k)\right\}. \tag{2}$$

The complete GMM model is described by a set of three parameters $\lambda = \{\pi_k, \mu_k, \Sigma_k\},\ k = 1...M$. To identify a folk song that has been modeled by λ, it is necessary to determine the likelihood $p(\mathbf{X}, \lambda)$

$$p(\mathbf{X}, \lambda) = \prod_{n=1}^{N} p(\mathbf{x}_n|\lambda), \tag{3}$$

where N is the number of feature vectors and also the number of segments of the audio file for each folk song. In fact, λ is a statistical model, so we have to use the Expectation-Maximization (EM) algorithm [10] to determine $\log\, p(\mathbf{X}|\lambda)$ such as it get the maximum.

4. EXPERIMENT RESULTS

4.1. The test results with GMM

Our experiments used Spro [69], Praat [70], and Matlab [71, 72] tools to extract a set of 63 parameters including 60 parameters related to MFCC and energy (19 MFCCs + energy = 20, the first and the second derivatives of these 20 parameters), tempo, intensity and fundamental frequency (F0). In musical terminology, the tempo is the speed or pace of a given piece. The tempo is often defined in units of beats per minute (BPM). The beat is often defined as the rhythm listeners would tap their toes to when listening to a piece of music [73]. The 63 parameters are divided into 4 sets of parameters in our experiments as in Table 2.

Table 2. Four sets of parameters

Sets of Parameters	Content	Sets of Parameters	Content
S1	60 parameters	S3	S1 + F0 + intensity
S2	S1 + tempo	S4	S3 + tempo

The ALIZE toolkits [74, 75] were used to implement the GMM model for classification with Gaussian component number M varied as a power of 2, from 16 to 4096.

The experiments were conducted in 2 cases: In the first case (Figure 2), a cross-evaluation was performed for two datasets *Cheo* and *Quanho* with the full length of the songs. In this case, 80% dataset is used for training, and 20% dataset for testing. The purpose of this experiment is to consider the effect of tempo, intensity, and F0 parameters on the identification results.

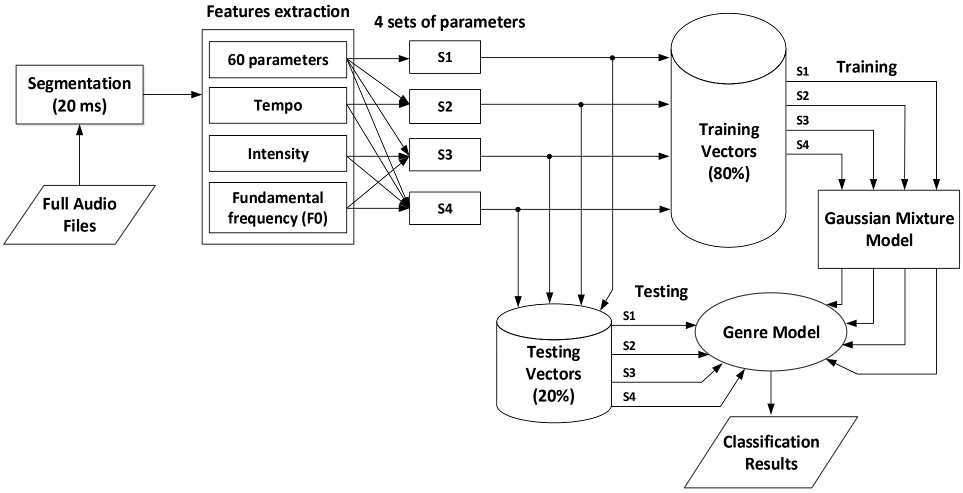


Figure 2. Diagram of the classification of *Cheo* and *Quanho* for the full length of audio files.

Figure 3 is the result of testing on the *Quanho* dataset in the first case with 4 sets of parameters. In general, the addition of parameters increases the identification rate. The average accuracy rate of identification for S1 is 96.62%, meanwhile for S2, S3, and S4 these rates are 96.67%, 96.76%, and 96.69% respectively, and higher than the average identification rate of S1 (Figure 4).

The results of the experiment on the *Cheo* dataset shown in Figure 5 also give a similar conclusion as above. The average accuracy rate of identification for the set of parameters S1 is 93.91%. For S2, S3, and S4 these rates are 94.00%, 94.20%, and 94.18% respectively (Figure 6). The experimental results on the *Quanho* dataset for the full length of the songs are higher (about 2 to 3%) in comparison with the *Cheo* dataset and this is also true for the corresponding excerpt as we can see below.

In the second case, the data for training use the full length of audio files, but the data for testing use only the short excerpts extracted from the dataset (Figure 7). These excerpts vary in length from 4, 6, 8 ... to 16 seconds (the excerpts were extracted randomly from the dataset). The purpose of this experiment is to determine how the accuracy rate will change when changing the length of the excerpts.

For the experiment results in the second case, within the scope of this paper, we present only the results corresponding to the three values of $M = 512$, 1024, and 2048. These values of M show more clearly the effect of parameters such as tempo, intensity, and F0 on the identification results for both *Cheo* and *Quanho* dataset.

Figure 8 is the results of the *Cheo's* excerpts with three values of M mentioned above. It can be seen that when the length of the excerpts is short, the parameters such as tempo, intensity, and F0 have no significant influence on the identification rate. With $M = 512$ (Figure 8a), the effect of these additional parameters is more noticeable when the length of

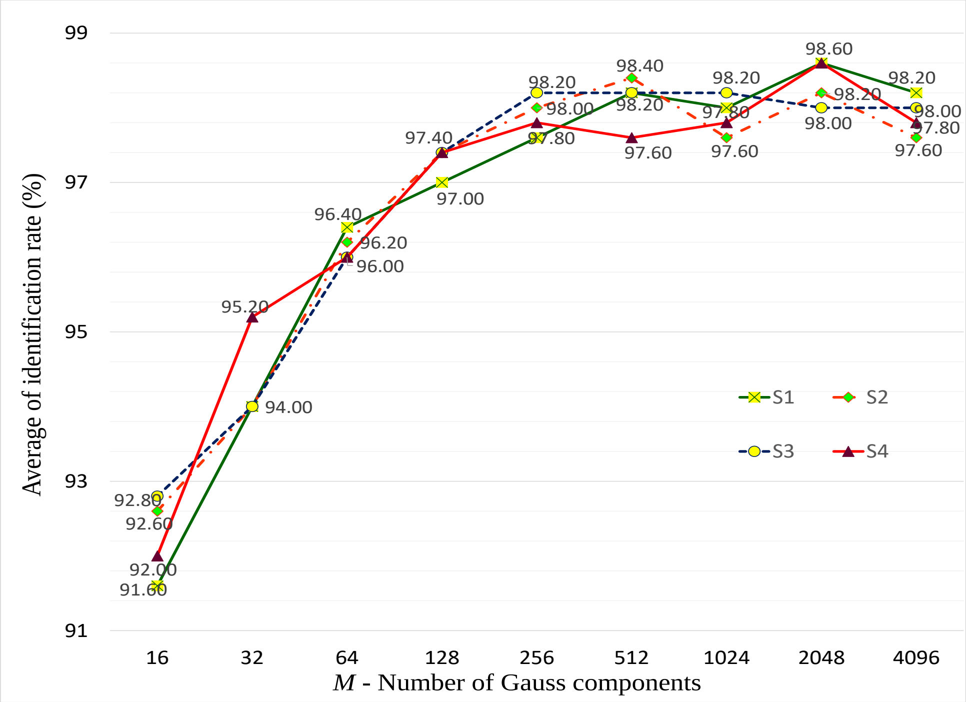


Figure 3. The identification rates correspond to 4 sets of parameters with the different values of M on the *Quanho* dataset

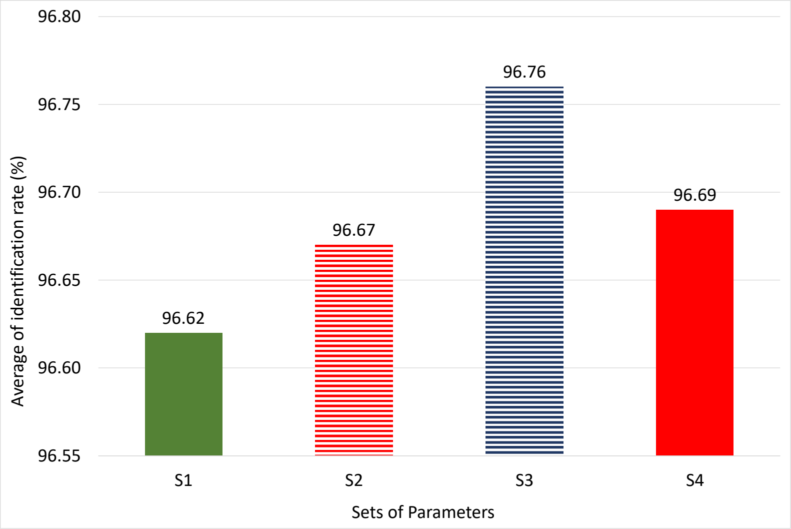


Figure 4. The average identification rates with 4 sets of parameters on the *Quanho* dataset

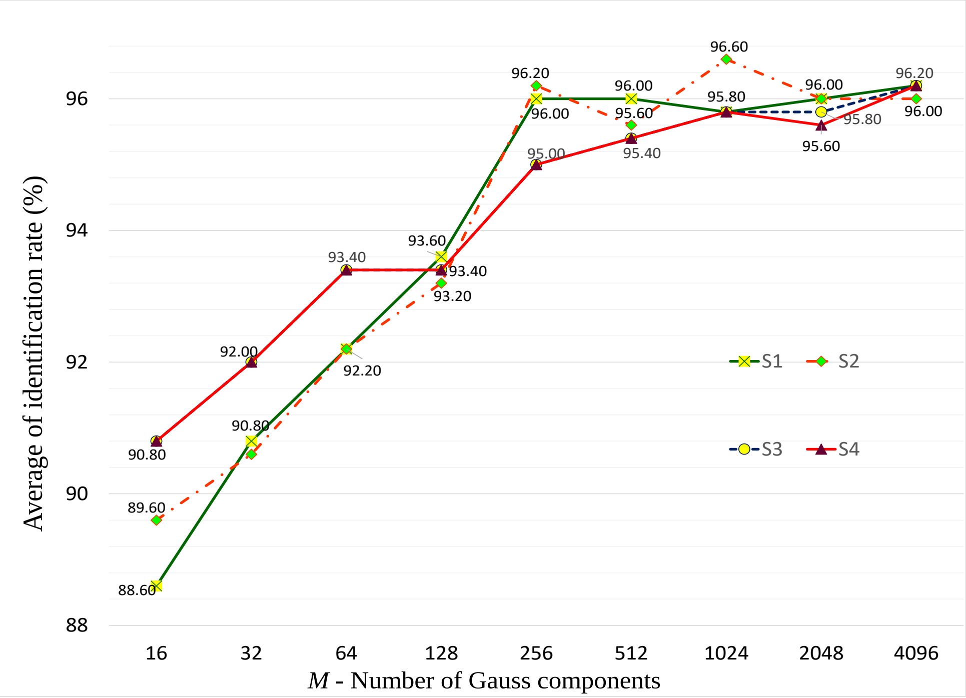


Figure 5. The identification rates correspond to 4 sets of parameters with the different values of M on the *Cheo* dataset

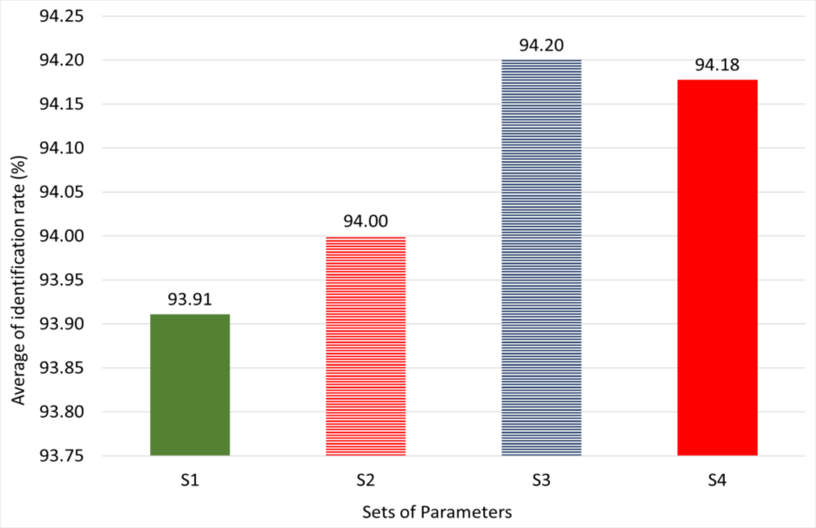


Figure 6. The average identification rates with 4 sets of parameters on *Cheo* dataset

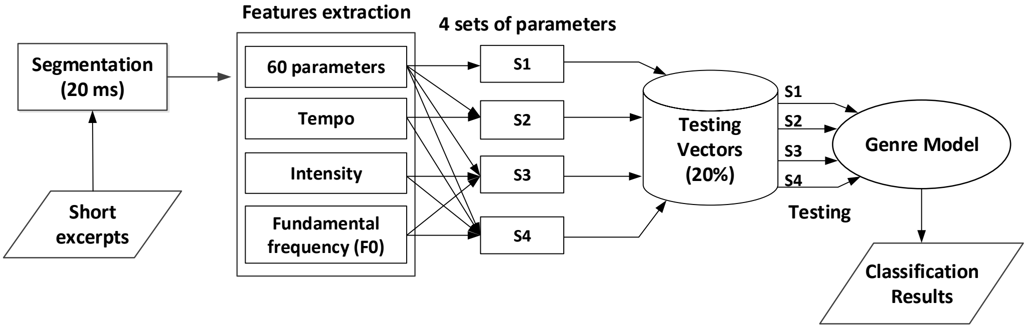
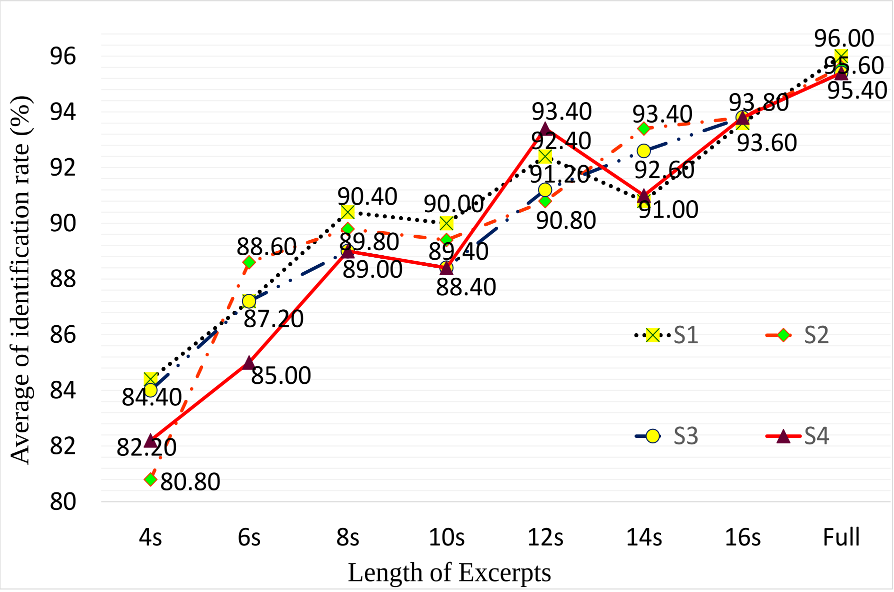


Figure 7. Diagram of the classification of *Cheo* and *Quanho* for the short excerpts of audio files

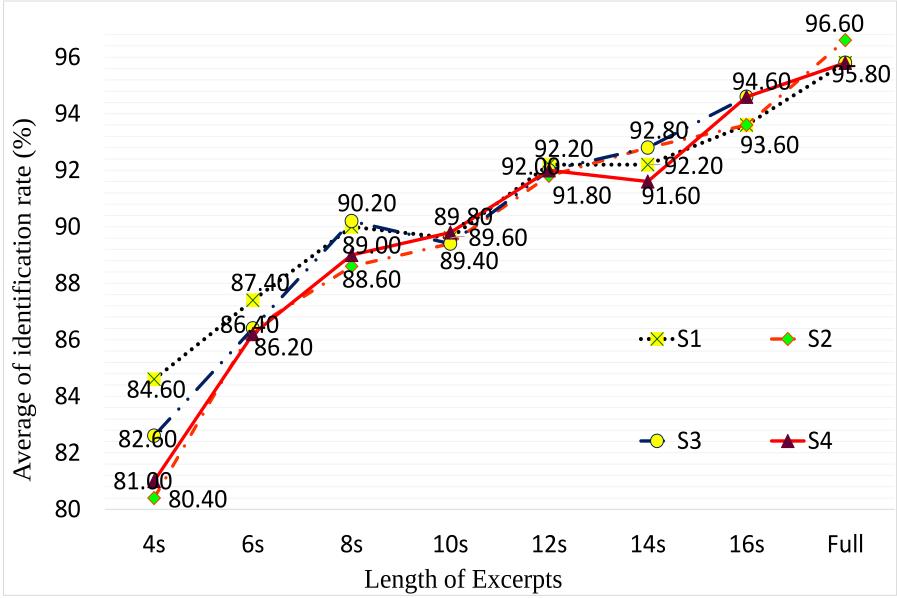
excerpts is 14 seconds or longer. With $M = 1024$ (Figure 8b), this value of length is 10 seconds. With $M = 2048$ (Figure 8c), the impact of additional parameters tends to decrease significantly. It can be seen that, on average, the S2, S3, and S4 parameter sets had a higher identification rate than S1 parameter set, which means that additional parameters had a good influence on the identification results. In particular, as the length of the excerpt increases, the influence of additional parameters becomes even more evident.



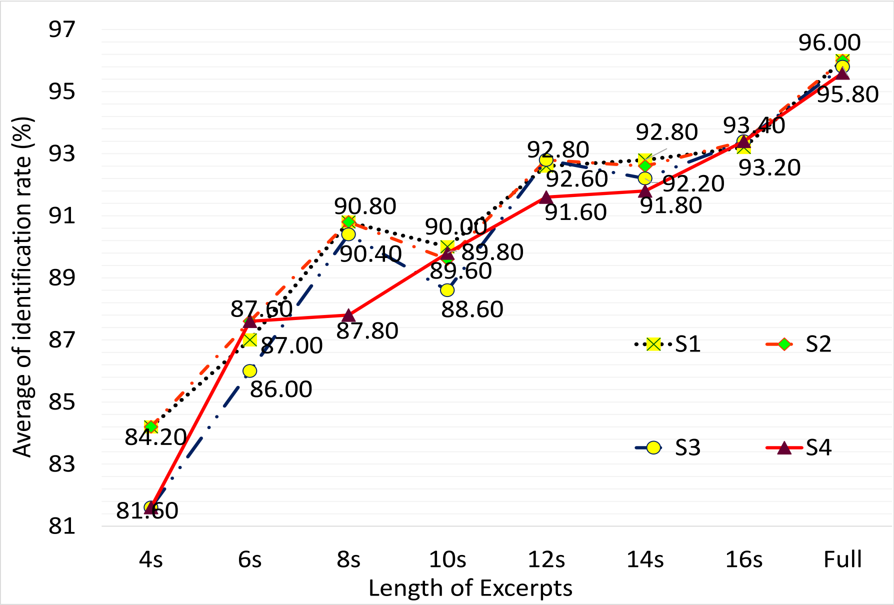
a) $M = 512$

The experimental results on *Quanho's* excerpts with three values of $M = 512$, 1024, and 2048 are shown in Figures 9a), 9b) and 9c) respectively. In this case, the additional parameters have also a positive effect on the identification results as in the experiment on the *Cheo's* excerpts. When $M = 2048$, this effect becomes more obvious (Figure 9c).

The results show that with 16-seconds excerpt length, on average, the identification rate reaches 91.09% compared to 94.18% when using the entire length of *Cheo* songs. With 16-



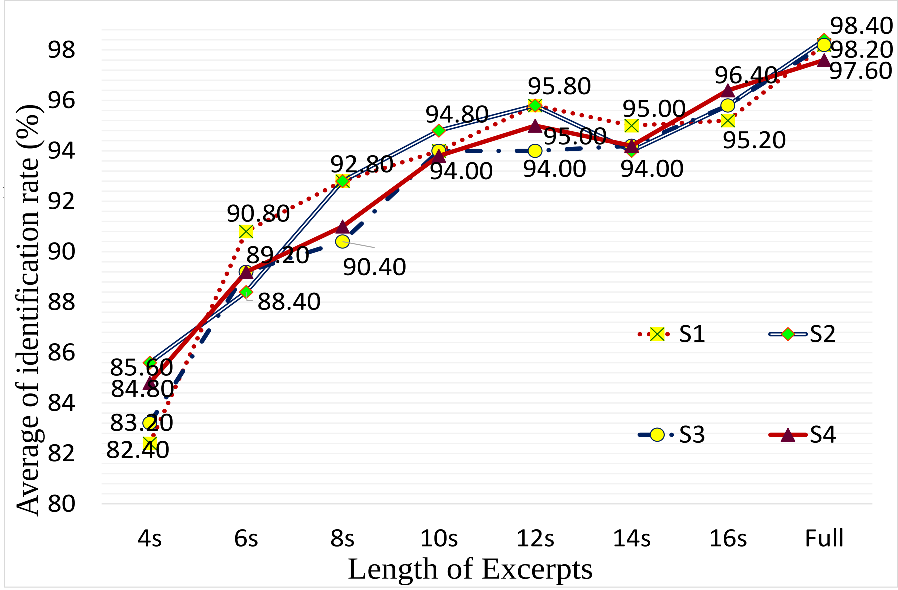
b) $M = 1024$



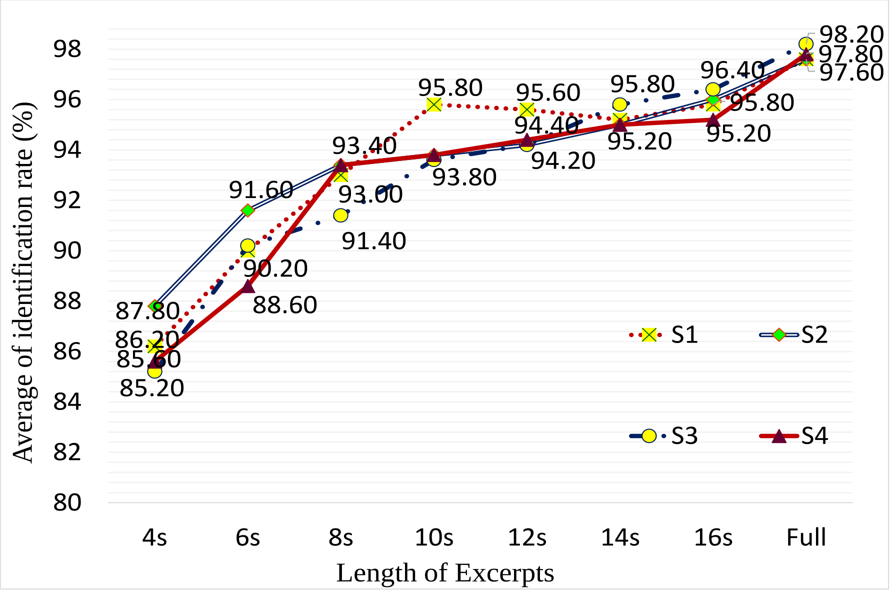
c) $M = 2048$

Figure 8. The identification rate based on lengths of *Cheo's* excerpts with $M = 512$, 1024, and 2048.

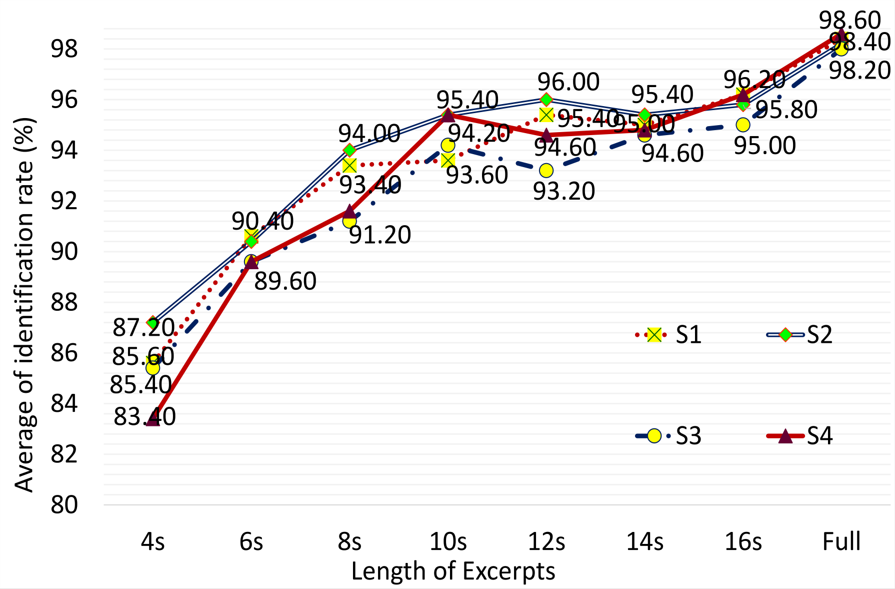
seconds excerpt length for *Quanho* songs, this identification rate reaches 94.44% compared to 96.89% for the full length of audio files.



a) $M = 512$



b) $M = 1024$



c) $M = 2048$

Figure 9. The identification rate based on lengths of *Quanho's* excerpts with $M = 512$, 1024, and 2048.

4.2. The test results with i-vectors

The i-vectors and x-vectors are both capable of representing feature parameters of a speech signal in a compact form (as a vector of fixed size, regardless of the length of the utterance). These vectors have been suggested to be suitable for speaker recognition [76,77]. The x-vector concept is newer and these vectors are used in the neural network to recognize a speaker.

The i-vectors have been used for the GMM model for speaker recognition and the following is a brief description of the i-vector and the experimental result using i-vector together with the GMM model to classify the Vietnamese folk-music *Cheo* and *Quanho*.

An important problem posed to the speaker recognition system is how to model the inter-speaker variability and to compensate for channel/session variability in the context of GMM. In Joint Factor Analysis (JFA) [78–80], the speaker's utterance is represented by the $\boldsymbol{M}$ supervector that includes additional components in the speaker and the channel/session subspace. In particular, the speaker-dependent supervector $\boldsymbol{M}$ is defined as follows [76]

$$\boldsymbol{M} = m + V_y + U_x + D_z. \tag{4}$$

Here, m is the session and speaker independent supervector (generally from the Universal Background Model (UBM)), V and D define the subspace of speaker (which are the eigenvoice matrix and diagonal residual, respectively), U defines the session subspace and is the eigenchannel matrix. The vectors x, y, and z are session and speaker dependent factors in their respective subspaces, and each vector is assumed to be a random variable with normal distribution $N(0,\ I)$.

In [81] the author proposed a new space and this new space is only a single space, instead of two separate spaces. This new space is called the “total variability space”, and contains the channel and speaker variabilities simultaneously. With this new space, equation (4) is rewritten as follows

$$\boldsymbol{M} = m + T_w, \tag{5}$$

where, m is the session and speaker independent supervector, which can be taken from UBM, T is a rectangular matrix of low rank, w is a random vector with standard distribution $N(0, I)$. The components of the vector w are the total factors and these vectors are identity vectors or i-vectors. Alize has provided the tool to define i-vectors [74, 75] from the set of parameter S1, and these vectors were used to classify the Vietnamese folk-music *Cheo* and *Quanho* in our case.

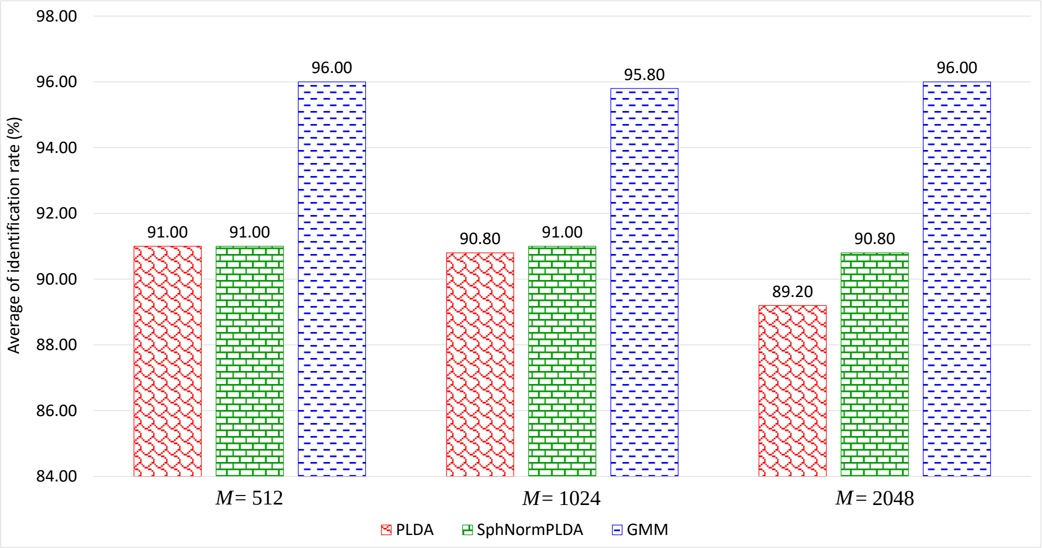


Figure 10. The average of identification rates for PLDA, SphNormPLDA using i-vectors, and for GMM using the set of parameters S1 on *Cheo* dataset

Figures 10, 11 are the identification rates for PLDA (Probabilistic Linear Discriminant Analysis) [76, 82–86] and SphNormPLDA (Spherical Normalization PLDA) [82, 87–89] using i-vectors and for GMM using the set of parameters S1 on the *Cheo* and *Quanho* dataset.

The following is a comment on the above result. In general, the accuracy obtained with i-vectors is lower. These results can be interpreted as follows. As mentioned above, the i-vectors are in a compact form with a fixed size, regardless of the length of the utterance, and feature very well for the speaker. However, for music genre classification, the rhythmic factors that change over time are very important. Because the compact nature of the i-vector does not take into account time-varying factors such as frame-by-frame processing, the result is of lower accuracy.

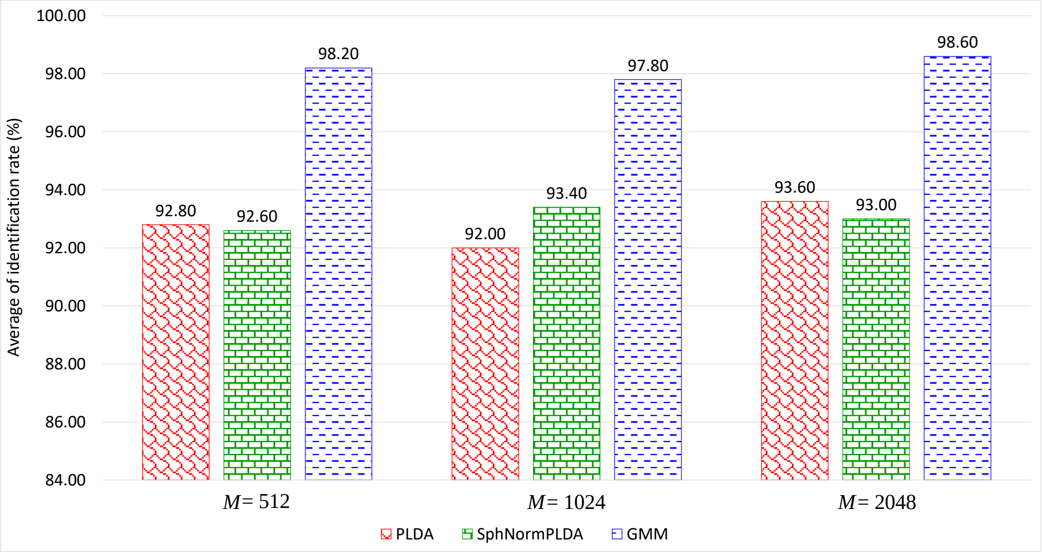


Figure 11. The average of identification rates for PLDA, SphNormPLDA using i-vectors and for GMM using the set of parameters S1 on *Quanho* dataset

5. CONCLUSIONS

The paper presents the experimental results of identification for some of the Vietnamese folk songs *Cheo* and *Quanho* using GMM for which the length of excerpts used for identification is multiples of 2s, from 4s to 16s compared to the full length of audio files and the number of Gaussian components M changes as powers of 2 from 16 to 4096. With the appropriate M values, the identification results showed the important effects of tempo, intensity, and fundamental frequency on the increase of identification accuracy rate. On average, when the excerpt length is 16s (29.63% compared to full length for *Cheo*, 37.2% compared to full length for *Quanho*), the identification rate was only 3.1% and 2.33% less than the whole song for *Cheo* and *Quanho* respectively. In the case of music genre identification, rhythm feature is an important parameter, so the use of i-vectors proved to be not real efficiency versus speaker recognition.

Our forthcoming research direction is to identify Vietnamese folk songs using other models or classifiers including various artificial neural network models.

ACKNOWLEDGMENT

We would like to thank School of Information and Communication Technology - Hanoi University of Science and Technology, and Faculty of Information Technology - Hung Yen University of Technical Education for helping us complete this paper.

REFERENCES

[1] X. Hu, J.S. Downie, K. West, and A. Ehmann, "Mining music reviews: Promising preliminary results," *In Proceedings of the International Conference on Music Information Retrieval,*

pages 536–539, 2005.

[2] C. DeCoro, Z. Barutcuoglu, and R. Fiebrink, "Bayesian aggregation for hierarchical genre classification," *In Proceedings of the International Conference on Music Information Retrieval*, pages 77–80, 2007.

[3] A. Anglade, Q. Mary, R. Ramirez, and S. Dixon, "Genre classification using harmony rules induced from automatic chord transcriptions," *In Proceedings of the International Conference on Music Information Retrieval*, pages 669–674, 2009.

[4] Cunningham, Padraig, and Sarah Jane Delany, "k-Nearest neighbor classifiers," *arXiv preprint arXiv:2004.04523 (2020).*

[5] Sazaki, Yoppy, "Rock genre classification using k-nearest neighbor", *ICON-CSE,* vol.1, no. 1, pp. 81–84, 2014.

[6] Ghahramani, Zoubin, "An introduction to hidden Markov models and Bayesian networks", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 1, pp. 9–42, 2001. https://doi.org/10.1142/S0218001401000836

[7] Xi Shao, Changsheng Xu and M. S. Kankanhalli, "Unsupervised classification of music genre using hidden Markov model," *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763),* Taipei, 2004, pp. 2023–2026 Vol.3. Doi: 10.1109/ICME.2004.1394661

[8] J. Reed and C.H. Lee, "A study on music genre classification based on universal acoustic models," *In Proceedings of the International Conference on Music Information Retrieval*, pages 89–94, 2006.

[9] Bağcı, Ulaş, Engin Erzin. "Boosting classifiers for music genre classification," *International Symposium on Computer and Information Sciences.* Springer Berlin Heidelberg, 2005.

[10] Christopher M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2013.

[11] Markov, Konstantin, Tomoko Matsui, "Music genre and emotion recognition using Gaussian processes," *IEEE Access 2*, pp.688–697, 2014.

[12] A. Meng, J. Shawe-Taylor, "An investigation of feature models for music genre classification using the support vector classifier," *In Proceedings of the International Conference on Music Information Retrieval*, pages 604–609, 2005.

[13] M. Li and R. Sleep, "Genre classification via an LZ78-based string kernel," *In Proceedings of the International Conference on Music Information Retrieval*, pages 252–259, 2005.

[14] A.S. Lampropoulos, P.S. Lampropoulou, and G.A. Tsihrintzis, "Musical genre classification enhanced by improved source separation techniques," *In Proceedings of the International Conference on Music Information Retrieval*, pages 576–581, 2005.

[15] C. McKay and I. Fujinaga, "Automatic genre classification using large high-level musical feature sets," *In Proceedings of the International Conference on Music Information Retrieval*, pages 525–530, 2004.

[16] A. Meng, P. Ahrendt, and J. Larsen, "Improving music genre classification by short-time feature integration," *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 497–500, 2005.

[17] Jinsong Zheng, M. Oussalah, *Automatic System for Music Genre Classification*, 2006, ISBN 1-9025-6013-9, PGNet.

[18] Le Danh Khiem, Hoac Cong Huynh, Le Thi Chung, *Quanho's cultural space.* Publisher of Bac Ninh Provincial Culture and Sports Center, 2006 (Vietnamese).

[19] Hoang Kieu, *Learn the ancient Cheo folk songs.* Publisher of the stage - Vietnam Cheo theatre, 2001.

[20] Bui Duc Hanh, *50 ancient Cheo folk songs.* Publishing House of National Culture, 2006 (Vietnamese).

[21] Hoang Kieu, Ha Hoa, *Selected ancient Cheo folk songs.* Publishing House of Information Culture, 2007 (Vietnamese).

[22] Nguyen Thi Tuyet, *Cheo singing syllabus.* Publishing House of Hanoi Academy of Theatre and Cinema, 2000 (Vietnamese).

[23] Nguyen Thi Tuyet, *Ancient Cheo melodies.* Publishing House of Hanoi Academy of Theatre and Cinema, 2007 (Vietnamese).

[24] Chu Ba Thanh, Trinh Van Loan, Nguyen Hong Quang, "Automatic identification of some Vietnamese folk songs," *In Proceedings of the 19th National Symposium of Selected ICT Problems*, Ha Noi, 2016. pages 92–97, ISBN: 978-604-67-0781-3.

[25] Chu Ba Thanh, Trinh Van Loan, Nguyen Hong Quang, "GMM for automatic identification of some Quanho Bac Ninh folk songs," *In Proceedings of Fundamental and Applied IT Research (FAIR)*, Da Nang, 2017. pages 416–421, ISBN: 978-604-913-165-3.

[26] Chu Ba Thanh, Trinh Van Loan, Nguyen Hong Quang, "Classification and identification of Cheo and Quanho Bac Ninh folk songs," *In Proceedings of Fundamental and Applied IT Research (FAIR)*, Ha Noi, 2018. pages 395–403, ISBN: 978-604-913-165-3.

[27] George, Tzanetakis, Essl Georg, and Cook Perry, "Automatic musical genre classification of audio signals," *Proceedings of the 2nd International Symposium on Music Information Retrieval*, Indiana, 2001.

[28] Available:http://marsyas.info/downloads/datasets.html

[29] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[30] K. West, S. Cox, "Features and classifiers for the automatic classification of musical audio signals," *In Proceedings of the Fifth International Conference on Music Information Retrieval (ISMIR)*, 2004. CUNG CAP LINK?????????? Check lai

[30] West, Kristopher, and Stephen Cox, "Features and classifiers for the automatic classification of musical audio signals," *ISMIR 2004, 5th International Conference on Music Information Retrieval,* Barcelona, Spain, October 10-14, 2004. https://www.researchgate.net/publication/220723821.

[31] N. Mohd, S. Doraisamy, R. Wirza, "Factors affecting automatic genre classification: An investigation incorporating non-western musical forms," *In Proceedings of the Fifth International Conference on Music Information Retrieval*, 2005.

[32] Witten, Ian H., and Eibe Frank. "Data Mining: Practical machine learning tools and techniques". *Morgan Kaufmann*, 2005.

[33] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, B. Kégl, "Aggregate features and adaboost for music classification," *Machine Learning*, vol. 65, no. (2-3), pp. 473–484, 2006.

[34] Chai, Wei, and Barry Vercoe, "Folk music classification using hidden Markov models," *Proceedings of International Conference on Artificial Intelligence*, vol. 6, no. 6.4, 2001. http://www.haralick.org/ML/Folk_Music_Classification_Using_Hidden_Markov_Mode.pdf.

[35] N. Bassiou, C. Kotropoulos and A. Papazoglou-Chalikias, "Greek folk music classification into two genres using lyrics and audio via canonical correlation analysis," *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA),* Zagreb, 2015, pp. 238–243. Doi: 10.1109/ISPA.2015.7306065.

[36] Rajesh, Betsy, and D. G. Bhalke, "Automatic genre classification of Indian Tamil and western music using fractional MFCC," *International Journal of Speech Technology*, vol. 19, no.3, pp. 551-563, 2016.

[37] J. Lafferty, A. Mccallum, FC. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning,* June 2001. Pages 282–289. http://portal.acm.org/citation.cfm?id=655813

[38] A. Mohamed, D. Yu and L. Deng, "Investigation of fullsequence training of deep belief networks for speech recognition," *INTERSPEECH 2010 11th Annual Conference of the International Speech Communication Association,* Makuhari, Chiba, Japan, September 26-30, 2010.Pages 2846–2849.

[39] I. Heintz, E. Fosler-Lussier and C. Brew, "Discriminative input stream combination for conditional random field phone recognition," *in IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, no. 8, pp. 1533–1546, Nov. 2009. Doi: 10.1109/TASL.2009.2022204

[40] Li, Juan, Jianhang Ding, and Xinyu Yang, "The regional style classification of Chinese folk songs based on GMM-CRF model," *ICCAE '17: Proceedings of the 9th International Conference on Computer and Automation Engineering,* February 2017. Pages 66–72. https://doi.org/10.1145/3057039.3057069

[41] Li, Juan, et al., "Regional classification of Chinese folk songs based on CRF model," *Multimedia Tools and Applications*, vol. 78, pp. 11563–11584, 2019. https://doi.org/10.1007/s11042-018-6637-6

[42] G.E. Hinton, "A practical guide to training restricted Boltzmann machines," In: Montavon G., Orr G.B., Müller KR. (eds), *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, vol 7700. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_32

[43] J. Martel, T. Nakashika, C. Garcia, K. Idrissi, "A combination of hand-crafted and hierarchical highlevel learnt feature extraction for music genre classification," In: Mladenov V., Koprinkova-Hristova P., Palm G., Villa A.E.P., Appollini B., Kasabov N. (eds), *Artificial Neural Networks and Machine Learning – ICANN 2013. ICANN 2013. Lecture Notes in Computer Science,* vol 8131. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40728-4_50

[44] Phan Anh Cang, Phan Thuong Cang, "Music classification by genre using discrete wavelet transform," *In Proceedings of Fundamental and Applied IT Research (FAIR)*, ISBN: 978-604-913-165-3, Can Tho, 2016. Pages 395–403.

[45] [Online]. Available: https://challenge.zalo.ai/

[46] K. Markov and T. Matsui, "Music genre and emotion recognition using Gaussian processes," *IEEE Access,* vol. 2, pp. 688–697, 2014. Doi: 10.1109/ACCESS.2014.2333095

[47] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03),* Hong Kong, 2003, pp. V-553. Doi: 10.1109/ICASSP.2003.1200029

[48] T. Heittola, A. Klapuri, "Locating segments with drums in music signals," *In Proceeding of the 3rd International Conference on Music Information Retrieval*, 2002, pp. 271–272.

[49] M. Marolt, "Gaussian mixture models for extraction of melodic lines from audio recordings", *In Proceedings of the International Conference on Music Information Retrieval,* 2004.

[50] G. Fuchs, "A robust speech/music discriminator for switched audio coding," *2015 23rd European Signal Processing Conference (EUSIPCO),* Nice, 2015. pp. 569–573. Doi: 10.1109/EUSIPCO.2015.7362447

[51] G. Sell and P. Clark, "Music tonality features for speech/music discrimination, *In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Florence, 2014, pp. 2489–2493. Doi: 10.1109/ICASSP.2014.6854048

[52] R. Thiruvengatanadhan, and P. Dhanalakshmi, "Indexing and retrieval of music using Gaussian mixture model techniques," *International Journal of Computer Applications*, vol. 148, no.3, 2016. https://www.ijcaonline.org/archives/volume148/number3/thiruvengatanadhan-2016-ijca-911095.pdf.

[53] Jakubec, Maros, and Michal Chmulik, "Automatic music genre recognition for in-car infotainment," *Transportation Research Procedia*, vol. 40, pp. 1364–1371, 2019.

[54] Evstifeev, Stepan, and Ivan Shanin, "Music genre classification based on signal processing," *In DAMDID/RCDL*, pp. 157–161, 2018.

[55] Bhattacharjee, Mrinmoy, S. R. M. Prasanna, and Prithwijit Guha, "Time-frequency audio features for speech-music classification," *arXiv.org > eess > arXiv:1811.01222.*

[56] Baelde, Maxime, Christophe Biernacki, and Raphaël Greff, "Real-time monophonic and polyphonic audio classification from power spectra," *Pattern Recognition*, vol. 92,pp. 82–92, 2019.

[57] R. Thiruvengatanadhan, "Music genre classification using GMM," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 10, Oct 2018.

[58] C. Kaur and R. Kumar, "Study and analysis of feature based automatic music genre classification using Gaussian mixture model," *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Coimbatore, 2017, pp. 465–468. Doi: 10.1109/ICICI.2017.8365395.

[59] B.K. Khonglah, S.M. Prasanna, "Speech/music classification using speech-specific features," *Digit Signal Process,* vol. 48, pp. 71–83, 2016. https://doi.org/10.1016/j.dsp.2015.09.005

[60] H. Zhang, X.K. Yang, W.Q. Zhang, W.L. Zhang, J. Liu, "Application of i-vector in speech and music classification," *2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),* Limassol, 2016, pp. 1–5. Doi: 10.1109/ISSPIT.2016.7885999.

[61] Stuttle, Matthew Nicholas, "A Gaussian mixture model spectral representation for speech recognition," Diss. University of Cambridge, 2003.

[62] S. G. Bagul and R. K. Shastri, "Text independent speaker recognition system using GMM," *2013 International Conference on Human Computer Interactions (ICHCI)*, Chennai, 2013, pp. 1–5. Doi: 10.1109/ICHCI-IEEE.2013.6887781.

[63] G. Suvarna Kumar et. al., "Speaker recognition using GMM," *International Journal of Engineering Science and Technology*, vol. 2, no. 6, pp. 2428–2436, 2010.

[64] D. Reynolds, R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[65] A. Dustor and P. Szwarc, “Application of GMM models to spoken language recognition,” *2009 MIXDES-16th International Conference Mixed Design of Integrated Circuits & Systems*, Lodz, 2009, pp. 603–606.

[66] Sarmah, Kshirod, and Utpal Bhattacharjee, “GMM based language identification using MFCC and SDC features,” *International Journal of Computer Applications*, vol. 85, no. 5, 2014.

[67] Pham Ngoc Hung, “Automatic recognition of continuous speech for Vietnamese main dialects through pronunciation modality,” Doctoral Thesis - Hanoi University of Science and Technology, 2017.

[68] Quang H. Nguyen, Trang T. T. Do, Thanh B. Chu, Loan V. Trinh, Dung H. Nguyen, Cuong V. Phan, Tuan A. Phan, Dung V. Doan, Hung N. Pham, Binh P. Nguyen and Matthew C. H. Chua, “Music genre classification using residual attention network,” *2019 International Conference on System Science and Engineering (ICSSE)*, Dong Hoi, Viet Nam, 2019, pp. 115–119. Doi: 10.1109/ICSSE.2019.8823100.

[69] [Online]. Available: http://www.irisa.fr/metiss/guig/spro/download.html

[70] [Online]. Available: https://www.fon.hum.uva.nl/praat/download_win.html

[71] [Online]. Available: https://www.tutorialspoint.com/matlab/index.htm

[72] [Online]. Available: https://www.mathworks.com/products.html

[73] [Online]. Available: https://en.wikipedia.org/wiki/Tempo

[74] J. Bonastre, F. Wils and S. Meignier, “ALIZE, a free toolkit for speaker recognition,” *Proceedings. (ICASSP ’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Philadelphia, PA, 2005, pp. I/737–I/740, vol. 1. Doi: 10.1109/ICASSP.2005.1415219.

[75] Tommie Gannert, “A Speaker verification system under the scope: Alize,” Stockholm, Sweden School of Computer Science and Engineering, 2007.

[76] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *in IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788–798, May 2011. Doi: 10.1109/TASL.2010.2064307.

[77] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Calgary, AB, 2018, pp. 5329–5333. Doi: 10.1109/ICASSP.2018.8461375.

[78] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Transaction on Audio Speech and Language Processing*, vol. 15, no. 4, pp.1435–1447, May 2007.

[79] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Speaker and session variability in GMM-based speaker verification,” *IEEE Transaction on Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[80] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of interspeaker variability in speaker verification,” *IEEE Transaction on Audio, Speech and Language*, vol. 16, no. 5, pp. 980–988, July 2008.

[81] N. Dehak, “Discriminative and Generative Approches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification,” Ph.D. thèsis, École de Technologie Supérieure, Montréal, 2009.

[82] Bousquet, Pierre-Michel, et al., "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis,“ Odyssey 2012-The Speaker and Language Recognition Workshop, 2012.

[83] T. Stafylakis, et al., "I-Vector/PLDA variants for text-dependent speaker recognition," *Preprint submitted to Computer, Speech and Language*, pp (2013).

[84] Kanagasundaram, Ahilan, "Speaker verification using I-vector features," Diss. Queensland University of Technology, 2014.

[85] E. Khoury, L. El Shafey, M. Ferras, and S. Marcel, "Hierarchical speaker clustering methods for the NIST i-vector challenge," in Odyssey: The Speaker and Language Recognition Workshop, no. EPFL-CONF-198439, 2014.

[86] S. Novoselov, T. Pekhovsky, and K. Simonchik, "Stc speaker recognition system for the NIST i-vector challenge," *in Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 231–240.

[87] Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Levy, H. Li, J. S. Mason, and J.-Y. Parfait, "ALIZE 3.0 - open source tool-kit for state-of-the-art speaker recognition," in Interspeech, Lyon, France, 2013.

[88] I. Salmun, I. Opher and I. Lapidot, "Improvements to PLDA i-vector scoring for short segments clustering," *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, 2016, pp. 1–4. Doi: 10.1109/ICSEE.2016.7806108.

[89] Fredouille, Corinne, and Delphine Charlet, "Analysis of i-vector framework for speaker identification in TV-shows," *INTERSPEECH 2014 15th Annual Conference of the International Speech Communication Association,* Singapore, September 14-18, 2014.

Received on September 20, 2019
Revised on October 26, 2020