Broadcast Monitoring System using MFCC-based Audio Fingerprinting

Myo Thet Htun
Faculty of Computer Systems and Technologies
University of Computer Studies, Yangon
Yangon, Myanmar
myothethtun@ucsy.edu.mm

Twe Ta Oo
Faculty of Computer Systems and Technologies
University of Computer Studies, Yangon
Yangon, Myanmar
twetaoo@ucsy.edu.mm

***Abstract*—An efficient broadcast monitoring system is really needed in Myanmar music industry to solve the issues of copyright infringements and illegal benefit-sharing between artists and broadcasting stations. In this paper, a broadcast monitoring system is proposed for Myanmar FM radio stations by utilizing Mel Frequency Cepstral Coefficient (MFCC) based audio fingerprinting. The proposed system is easy to implement and achieves the correct and speedy music identification even for noisy and distorted broadcast audio streams. In this system, we deploy an audio fingerprint database of 4,379 songs and broadcast audio streams of 3 local FM channels of Myanmar to evaluate the performance of the proposed system. Experimental results show that the system achieves reliable performance.**

Keywords—MFCC, audio fingerprinting, music identification, broadcast monitoring

I. Introduction

In recent years in Myanmar, CD music distribution system has totally been destroyed by piracy problems[1], same as in the global music industry[2]. After changing from physical sales to online sales system in Myanmar since 2011, unauthorized online music distribution has become an un-solving and headaching issue. Major concerning problems are copyright violations and benefit-sharing by the weakness of rules and laws for the protection of intellectual property in Myanmar[3]. It thus demands an efficient broadcast monitoring system to monitor the broadcast media streams and to detect illegal usage of music contents in multiple digital platforms like YouTube, Facebook, etc. Broadcast monitoring is also mainly used to monitor music airplay for radio stations, advertisements for online broadcasting media, copyrighted interview programmes, and background music for TV stations. Such systems should also be reliable and legal for content owners such as artists and composers.

Audio fingerprinting [1-3], which is a well-known music information retrieval technique, is widely used in broadcast monitoring systems. Audio fingerprint is none other than a unique identifier of an audio piece generated by analyzing the acoustic property of the audio itself. It is best known for its ability to identify the correct music information such as artist name, song name, etc., of a short unlabeled audio clip by linking to fingerprint database of known audio clips. This feature makes audio fingerprinting attractive to monitor the usage of music contents in broadcast digital streams. It also helps to solve the copyright infringements and illegal benefit-sharing between artists and broadcasting stations.

In this paper, we develop a broadcast monitoring system for FM radio stations in Myanmar. It uses audio fingerprints to monitor the broadcast media streams and generates a report of specific information such as song name, artist name, broadcasting duration, etc. of the broadcasted songs. That report can be used for legal benefit-sharing between the artists and broadcasting stations.

The rest of the paper is organized as follows. Section 2 briefly explains our previously proposed Mel Frequency Cepstral Coefficient (MFCC) based audio fingerprinting method [4, 5]. It is applied in this paper to extract the audio fingerprints. Section 3 discusses the proposed broadcast monitoring system in detail. The databases used in this system and experimental setup are presented. Section 4 discusses the experimental results. Finally, section 5 concludes the proposed research work.

II. MFCC-based Audio Fingerprinting Method

Various feature extraction methods can be used to extract an audio fingerprint that uniquely identify an audio clip. Among them, MFCC is one of the commonly used methods because of its speaker identification efficiency [6] and the most effective choice of Mel filter bank [7].

In [4], we proposed a space-saving audio fingerprinting method based on MFCC which is closely related to human ear scale. Its general block diagram for extracting an audio fingerprint from a 3-sec audio clip is shown in Fig. 1 and the processes are briefly explained below.

MFCC-based Audio Fingerprint Extraction [4]

[1] *Pre-processing*

a) Down sampling: Down-sample the input (3-sec) audio to 5512 Hz to achieve more compact fingerprint and to eliminate the effect of different playback speeds.

b) Pre-emphasis: Apply the pre-emphasis filter shown in Eq. 1 to boost the signal energy in high frequencies.

$$y(t) = x(t) - \alpha x(t-1), \tag{1}$$

where the filter coefficient α is usually between 0.9 and 1.0, and we set it as 0.97 in this system.

c) Framing and overlap: Split the filtered signal into 370 ms frames with 11.6 ms frame shift duration.

d) Windowing: Apply the Hanning window of Eq. 2 on each frame to obtain the smooth frame adjacency.

$$w(n) = 0.5(1 - \cos 2\pi(n / N)), 0 \le n \le N - 1, \tag{2}$$

where N is the window length.

[2] *MFCC Features Extraction*

a) Fast Fourier Transform (FFT): Apply the FFT on each frame of the windowed signal to extract the spectral information.

[1] https://www.mmtimes.com/lifestyle/7248-myanmar-music-set-to-go-online.html

[2] https://www.techdirt.com/articles/20121018/10023520751/30-years-cd-digital-piracy-music-industry-cluelessness.shtml

[3] https://www.tilleke.com/resources/myanmar-enacts-copyright-law

b) Band pass filter: Warp the frequency spectrum to the Mel-scale in order to adapt the frequency resolution to the properties of the human ear, as defined in Eq. 3.

$$F(mel) = 2595 * \log_{10}\left[1 + \frac{f}{700}\right]. \qquad (3)$$

c) Discrete Cosine Transform (DCT): Apply the DCT to convert the log Mel spectrum into time domain. The result is a 13 x 227 MFCC feature vector.

[3] *Bits Difference Computation*

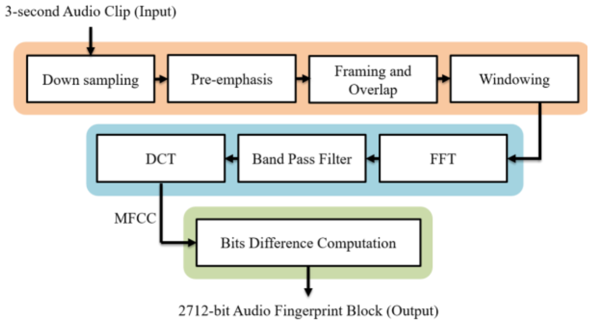
Convert the MFCC features to a binary string (i.e. a 2712-bit fingerprint string in this system) by computing the sign differences between the features of the adjacent rows and columns of the feature vector, as shown in Eq. 4.

$$f = \begin{cases} 1, & (m(r,c) - m(r,c+1)) - \\ & (m(r-1,c) - m(r-1,c+1)) > 0, \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

where $m(r,c)$ is the Mel coefficient value of row r and column c of the feature vector and f is the resulting fingerprint bit.

The above method has significant advantages over Philips Robust Hashing (PRH) [1], one of the most influential works in audio fingerprinting. The main difference is that the above method [4] considers Mel features as fingerprint, whereas the PRH uses the FFT-based spectral information. Mel scales are human ear scales. Thus, it should be more appropriate to extract a compact digital summary of a sound to approximate human perception. As a proof, the resulting fingerprints are found to be more robust to background noise, pitch shifting, and linear speed changes of input audio, which are the most occurred attacks in broadcast monitoring systems [5]. In addition, the method in [4] achieves smaller fingerprint size (2712 bits) for a 3-sec audio clip, whereas the PRH results 8192 bits [1]. It is a big reduction in storage, which is also good for speedy music identification.

Thus in this paper, we apply the above method [4] to design an efficient monitoring system for FM radio stations. We use that method to prepare fingerprint database for known copyrighted registered songs and to extract fingerprint from unlabeled broadcast radio streams. For more information on the fingerprinting method, please refer [4, 5].



3-second Audio Clip (Input)
Down sampling
Pre-emphasis
Framing and Overlap
Windowing
DCT
Band Pass Filter
FFT
MFCC
Bits Difference Computation
2712-bit Audio Fingerprint Block (Output)

Fig. 1. MFCC-based audio fingerprint extraction [4]

III. Proposed Broadcast Monitoring System

A monitoring system that monitors the broadcasting audio streams and automatically generates the playlists of registered songs will be an invaluable tool for copyright enforcement organizations and companies reporting statistics on the music broadcast. Audio fingerprinting is a technology that is able to detect a similar part from an unknown song without using embedded watermarking or any other external metadata. Figure 2 shows how a broadcast monitoring system works.

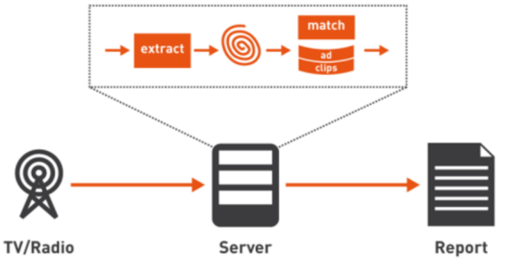


Fig. 2. How a broadcast monitoring system works

A broadcast monitoring system must have an already created fingerprint database of known songs. To monitor the broadcast stream,
Step 1: Capture the broadcasting media streams;
Step 2: Extract fingerprint (for each 3-sec clip in this proposed system) of captured streams;
Step 3: Each of the extracted fingerprints is matched with the ones in the fingerprint database. If the bit error rate (BER) is less than a ***threshold*** (0.35 in this system), it considers as a perfect match and generates report. Otherwise, it is assumed as not match. The BER is calculated, as shown in Eq. 5, by comparing the captured fingerprint bits to known fingerprint bits and counting the number of errors. A number of experiments has proved that when the BER is less than 0.35, matching results can be regarded as effective [1].

$$\text{BER} = \text{Number of errors} / \text{Number of bits}. \qquad (5)$$

Experimental setup of the proposed system is discussed below.

A. Research Aided Tools

1) Software

a) Matlab R2018a: Pre-processing steps, MFCC feature extraction, bits difference computation, and BER calculation are simulated in Matlab R2018a.

b) Audacity 2.4.2: Audacity is used to degrade the audio clips by injecting common signal distortions such as adding background noise, pitch shifting, speed changes, etc.

c) Microsoft Visual Studio 2017: Broadcast monitoring system is implemented in C# by using Microsoft Visual Studio 2017.

d) Microsoft SQL Server 2014: Extracted audio fingerprints and song clips are stored in the database by using Microsoft SQL Server 2014.

2) Runtime Environment

a) PC: Dell Inspiron 5458 Laptop

b) OS: Microsoft Windows 10 Pro 64-bit

c) Processor: Intel(R) Core i3-5005U 2.00 GHz

d) RAM: 4096 MB

e) HDD: 500 GB

3) FM Capturing Device: In this research work, we use the FM Radcap PCIe device shown in Fig. 3, which is an audio signal capture card designed for recording of multiple radio stations at the same time. The Radcap achieves exceptionally low audio distortion through the use of linear phase filtering, mathematically precise FM demodulation, and stereo decoding. The card can be configured to operate in stereo, mono, or paired mono (two mono stations combined on a 2-channel audio stream) modes. Multiple cards can be used in a single PC, subject to available CPU bandwidth.

The card uses a high-speed A/D converter to digitize the entire FM band, with up to 32 individual tuners. The card is factory-configured for PC-FM 6, 12, 18, 24, or 32 stations, which can be expanded in the field for an additional charge. With the purpose of capturing only 10 local FM channels in Myanmar, we use PC-FM12 card in this research work as shown in Fig. 4.

Fig. 3. FM Radcap PCIe device

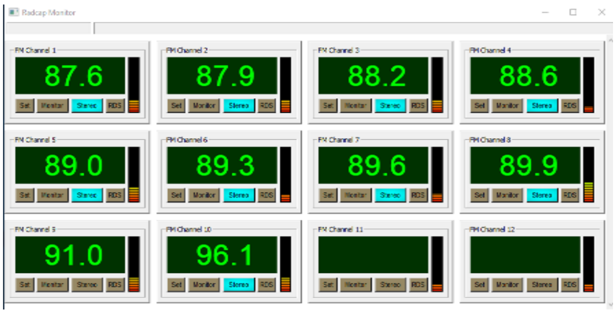


Fig. 4. Capturing local FM channels using FM Radcap PCIe device

B. Databases

In order to be matched with the query fingerprint of the captured broadcast stream, a broadcast monitoring system must have an already created fingerprint database of registered songs. For the purpose of matching fingerprints and linking to the relevant contents of the matched audio clip, the proposed system use three main databases: (1) *Myanmar Music Store (MMS)*, (2) *ChannelRing*, and (3) *FingerprintsDb*.

The "*MMS*" database and "*ChannelRing*" database are developed by Legacy Music Network Co., Ltd which is a leading music company that handles various fields within the Myanmar music business[4]. In the "*MMS*" database, huge amount of copyrighted songs is stored by linking with file directories. As until now, we have generated the fingerprints for 4,379 songs in that database by using the proposed method in [4]. Those fingerprints as binary representation patterns are stored as the registered fingerprints in the "*FingerprintsDb*" database together with specific song id.

The "*ChannelRing*" database stores all the related data of a song such as song title, featuring artists, studio, band, producer, album, audio length, engineer, and genres for most of the songs in the "*MMS*" database (as until now, a total of 65,369 songs).

In this system, we use those databases to generate a report of specific information such as song name, artist name, album name, song ID, and broadcasting duration of the detected registered songs in the monitored 3 local FM broadcast streams.

C. Query Data

In this system, we record the broadcast audio streams from 3 local FM channels mentioned in Table 1 for evaluating the monitoring performance of the proposed system. The system will monitor those broadcast streams and identify the perceptually similar songs by linking with the above mentioned databases. Table 1 also compares the storage requirement of the fingerprint extraction method used in this system and the PRH. After extracting the audio fingerprints for each 3-sec clip of the broadcast stream (e.g. 18 minutes and 16 seconds for Cherry FM), the fingerprint extraction method used in this system needs only 0.11 MB for fingerprint storage. In contrast, the PRH method needs 0.35 MB storage. From space-saving point-of-view, the fingerprinting method used in this system only needs one-third of the PRH's fingerprint size. It is very computationally efficient by reducing the huge amount of memory allocation for large-scale music libraries.

TABLE I. TESTING DATA USED FOR PERFORMANCE EVALUATION

FM Channels	**Tuning Range**	**Length (mm:ss)**	**File Size**	**Fingerprint Size**	
				Proposed System	***PRH***
Cherry FM	89.3 MHz	18:16	25.1 MB	0.11 MB	0.35 MB
City FM	89.0 MHz	26:11	35.9 MB	0.16 MB	0.51 MB
Thazin FM	88.6 MHz	20:30	12.9 MB	0.13 MB	0.40 MB
Total		**64:57**	**73.9 MB**	**0.40 MB**	**1.26 MB**

D. How to Link Databases

Figure 5 shows the tables and attributes of the databases discussed in section *B* and how they are linked when performing the fingerprint matching process in this system.

Step 1: Fingerprint from an unlabeled audio stream is extracted and matched with the fingerprints in the table "*tblFingerprints*" of the "*FingerprintsDb*" database. Then, the "*TrackId*" of the fingerprint with the least BER is retrieved.

Step 2: The "*Id*" in the table "*tblTracks*" of the "*FingerprintsDb*" database that is the same as the "*TrackId*" in step 1 is searched and used to retrieve the corresponding "*ISRC*" (International Standard Recording Code).

Step 3: The "*SongID*" in the "*Song*" table of the "*ChannelRing*" database that is the same as the "*ISRC*" in step 2 is searched. Then, the relative contents of that song are used to generate the report of played list and duration, etc.

[4] https://www.legacy.com.mm

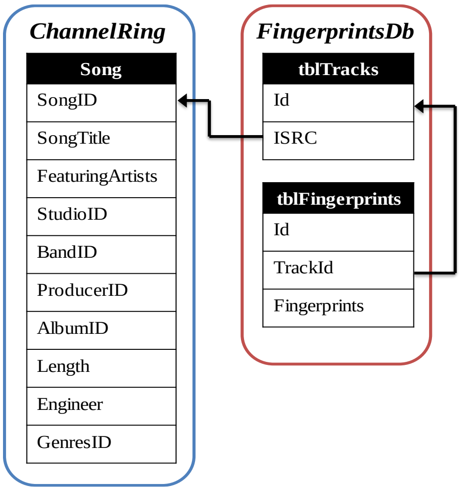


Fig. 5. Link of databases when fingerprint matching

Figure 6 shows an example fingerprint matching process and Fig. 7 depicts the proposed broadcast monitoring system.

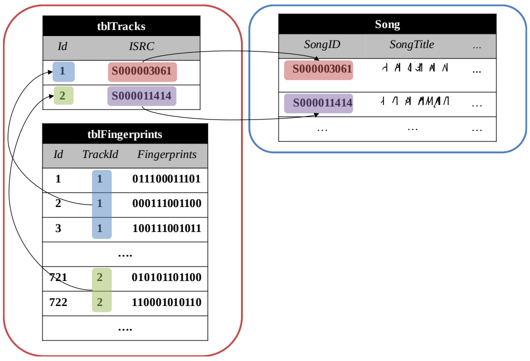


Fig. 6. An example fingerprint matching process

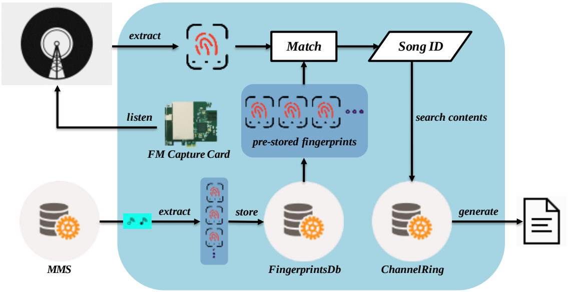
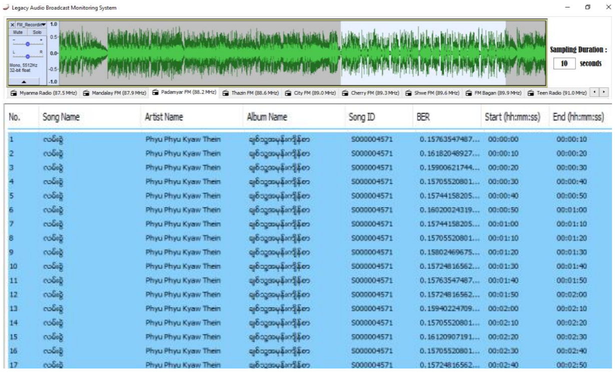


Fig. 7. The proposed broadcast monitoring system

IV. Experiments

A. Demo Version of the Proposed System

As experiments, we develop a demo version of the proposed broadcast monitoring system by using Matlab R2018a and Microsoft Visual Studio 2017. First, we record the broadcast audio streams from the 3 local FM channels using the Radcap PCIe device. Then, the demo is used to monitor the captured audio streams, extract fingerprints from each 3-sec audio clip, and match each fingerprint with the pre-stored audio fingerprints from the "*FingerprintsDb*" database. After the matching process, the relevant information of the least BER fingerprint is retrieved from the "*ChannelRing*" database.



Sampling Duration : 10 seconds

No.	Song Name	Artist Name	Album Name	Song ID	BER	Start (hh:mm:ss)	End (hh:mm:ss)
1	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15763547487...	00:00:00	00:00:10
2	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.16182048927...	00:00:10	00:00:20
3	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15900621744...	00:00:20	00:00:30
4	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15705520801...	00:00:30	00:00:40
5	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15744158205...	00:00:40	00:00:50
6	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.16020024319...	00:00:50	00:01:00
7	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15744158205...	00:01:00	00:01:10
8	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15705520801...	00:01:10	00:01:20
9	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15802469675...	00:01:20	00:01:30
10	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15724816562...	00:01:30	00:01:40
11	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15763547487...	00:01:40	00:01:50
12	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15724816562...	00:01:50	00:02:00
13	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15940224709...	00:02:00	00:02:10
14	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15705520801...	00:02:10	00:02:20
15	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.16120907191...	00:02:20	00:02:30
16	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15705520801...	00:02:30	00:02:40
17	လမ်းခွဲ	Phyu Phyu Kyaw Thein	ချစ်သူ့အမှန်းကျိုနိတာ	S000004571	0.15724816562...	00:02:40	00:02:50

Fig. 8. Demo version of the proposed broadcast monitoring system

Figure 8 shows the example report generated by the demo version of the proposed broadcast monitoring system. It expresses the matching song list with song ID, song name, artist name, album name, BER values, and broadcasting duration in start time and end time. This one is a kind of loyalty report for copyright owners who can analyze information such as the duration of music airplay. By analyzing the reported monitoring list, benefit-sharing and collecting charges for the usage of songs among artists can be effectively determined.

B. Robustness

The robustness and reliability of the audio fingerprinting method used in this system have already been tested for noise-free high-quality audio clips in [4, 5]. The system is said to be robust if it can identify the correct song from the monitored noisy and distorted audio clips. The results showed that the method works well for those data.

In this paper, we evaluate the robustness of that method for captured broadcast radio streams by means of the BER. The BER less than 0.35 is assumed as effective [1]. No need to doubt, the audio quality of the broadcast streams is really degrading.

Firstly, the robustness to various kinds of signal distortions is tested by adding the distortion types of Hard Clip, Hard Overdrive, Medium Overdrive, Soft Clip, and Soft Overdrive to the broadcast streams shown in Table 1. These distortions are implemented with the factory presets values of Audacity. The resulting BERs are illustrated in Fig. 9. The results show that the fingerprinting method preserves its robustness very well for broadcast streams as well: all the BER values are under the threshold value of 0.35.

The robustness is also tested by adding the background noise to the broadcast streams. The results are illustrated in Fig. 10. As we can see, the fingerprinting method works perfectly for white noise addition. It is more robust to white noise addition than signal distortions.

Robustness of the fingerprinting method to 'pitch shifting' of the broadcast streams is also evaluated by changing the pitch of the streams. Those pitch shiftings affect both the up and down of time-stretching in the original broadcast streams. The resulting BERs are visualized in Fig. 11. All the BERs are under threshold for 'pitch shifting' from -4% to +4%. It shows that the fingerprinting method well preserves its robustness under 'pitch shifting' as well.

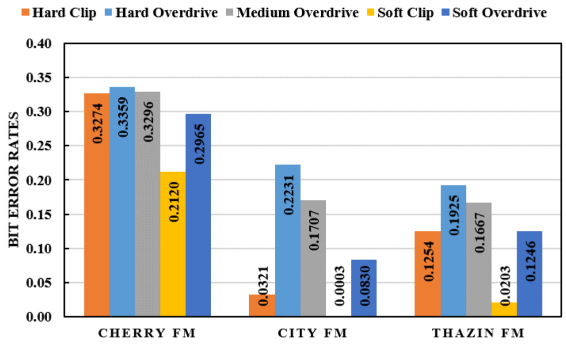


Fig. 9. Illustration for signal distortions

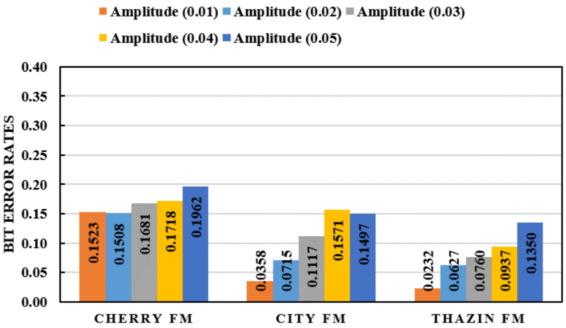


Fig. 10. Illustration for white noise addition

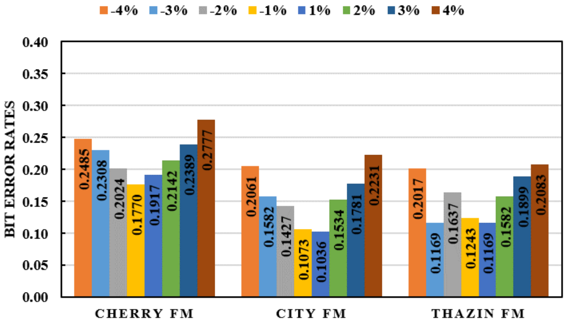


Fig. 11. Illustration for pitch shiftings

From the results in Fig. 9-11, we can also see that the performance of the robustness differs from each channel. The recordings of City FM and Thazin FM perform better than the Cherry FM. It is because of the consequences of extracted MFCC features from various music genres. Our former research in [5] showed that the robustness level of audio clips is different based on the music genres such as Pop, Rock, Jazz, Classical, Hard Rock, Hip Hop, Acoustic, and Traditional. Hard Rock was the most robust while input audio signals were distorted by various signal distortions such as Hard Clip, Hard Overdrive, etc. Traditional music achieved satisfying robustness at pitch shiftings and Classical music performed very well under white noise addition.

In our experiments, the recorded music of the City FM are mostly Hard Rock, whereas the recorded music are Classical for the Thazin FM, and short advertisements and background speech for the Cherry FM. From the BER results in Fig. 9, we can see that the recordings of City FM are mostly more robust to signal distortions than other two channels. It is because the recordings of City FM are mostly Hard Rock music. As for white noise addition, the Thazin FM recordings are more robust because they are classical music. As for the Cherry FM, the background speech in music is perceptually related to the kinds of music genres like Hip Hop and Rap. Based on the research findings in [8], speech and Hip Hop are less rhythmically diverse and more similar in the contents of Rap music. As presented in our previous research [5], Hip Hop music were not robust enough while comparing with other music genres. Therefore, the BER values of the Cherry FM are higher for all attacks compared to other channels.

To summarize, robustness of the MFCC-based fingerprint method depends on the broadcasting music genre as well. However, all of the experimental BER results are under threshold for signal distortion, white noise addition, and pitch shifting attacks, which are the major challenges for broadcast audio streams. Hence, we can conclude that the fingerprinting method works well for broadcast audio streams and can perfectly detect the perceptually similar audio clips through the degraded signals while broadcasting.

V. Conclusion

In this paper, we propose a broadcast monitoring system for FM radio stations in Myanmar. The experimental results show that the proposed system can perfectly retrieve the perceptually similar songs from broadcasting audio streams even under noisy conditions. It can also generate a kind of loyalty report that might be helpful for solving copyright infringements and benefit-sharing problems. Besides, the space-saving approach of the MFCC-based audio fingerprinting method reduces the fingerprint size which is the important one of theoretical considerations for a broadcast monitoring system.

Future research is planned to capture broadcasting streams from more local FM channels and to combine the audio fingerprinting method with hashing algorithm with the aim of achieving more efficient search speed from large-scale fingerprint databases.

References

[1] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” Intl. Symposium for Music Information Retrieval, 2002.

[2] M. Park, H. Kim, and S. H. Yang, “Frequency-temporal filtering for a robust audio fingerprinting scheme in realnoise environments,” J. Electronics and Telecommunications Research Institute, vol. 28, no. 4, pp. 509–512, 2006.

[3] S. Yao, B. Niu, and J. Liu, “A sampling and counting method for big audio retrieval,” IEEE Second Intl. Conf. on Multimedia Big Data, 2016.

[4] Myo Thet Htun and Twe Ta Oo, “Compact and robust audio fingerprinting for speedy music identification,” 11th Intl. Conf. on Future Computer and Communications, February 2019.

[5] Myo Thet Htun, “Analytical approach to MFCC based space-saving audio fingerprinting system,” 17th Intl. Conf. on Computer Applications, February 2019.

[6] F. Y. Leu and G. L. Lin, “An MFCC-based speaker identification system,” IEEE 31st Intl. Conf. on Advanced Information Networking and Applications, 2017.

[7] S. K. Kopparapu and M. Laxminarayana, “Choice of Mel filter bank in computing MFCC of a resampled speech,” 10th Intl. Conf. on Information Sciences, Signal Processing and their Applications, May 2010.

[8] S. Gilbers, N. Hoeksema, K. de Bot, and W. Lowie, “Regional variation in West and East Coast African-American English prosody and rap flows,” Language and Speech Journal, vol. 63, pp. 713–745, 2020.