Illegal Audio Copy Detection using Fundamental Frequency Map

Heui-su Son[1,*], Sung-woo Byun[1] and Soek-Pil Lee[2]
[1]*Department of Computer Science, Graduate School, SangMyung University, Seoul, South Korea*
[2]*Department of Electronic Engineering, SangMyung University, Seoul, South Korea*

**https://www.smu.ac.kr*

Keywords: Illegal Use, Copy Detection, Audio Fingerprinting, Fundamental Frequency Map, Pearson's Correlation Score.

Abstract: In this paper, we present a new audio identification system which is robust to various attacks. The types of attacks employed are modification such as changes of tempo, pitch and speed and noise addition. We propose a two-dimensional representation for the audio signal called FFMAP. This consists of pitch components and frame components. We also employ Pearson's correlation score to calculate similarity between original audio data and query. Experimental results show that the proposed algorithm has a high performance.

1 INTRODUCTION

As social network services such as YouTube, Facebook, Instagram, and others have become more common, it is easier for users to access large-sale media content than before, and approaches are becoming more diverse. With the increased accessibility of information, the dissemination of digital content has become more convenient, leading to an increase in concerns related to the illegal use of media content. Since distributing audio files without permission is a representative example of such illegal use, studies searching for identified audio data among unlabeled metadata have been proposed in the past decade(Wu et al., 2000), (Haitsma et al., 2001), (Jiao et al., 2007). These audio identification technologies can be classified into two technologies: audio watermarking (Arnold, 2000), (Hu and Hsu, 2015), (Milaš et al., 2016) and audio fingerprinting (Chen et al., 2013), (Bellettini and Mazzini, 2010), (Seo, 2014). Audio fingerprinting is a method to confirm copyright infringement from audio files by extracting features from audio sources and comparing them with a copyright database. The extracted features are content-based information that contains unique values from the audio data itself.

A number of studies on audio fingerprinting algorithms have been researched. Among these studies, the best-known public algorithm is Shazam (Wang, 2003). It uses the spectrogram peaks as the local key points of an audio stream. It then generates a local descriptor using certain pairs of these key points. The local descriptor is based on the time difference between two adjacent peaks as well as their frequencies. The extracted fingerprints are highly robust to audio compression, foreground voices, and other types of noise. However, because of the nature of the descriptors, the algorithm is vulnerable to pitch or tempo modifications. Most papers emphasizing invariance to tempo/speed and pitch extracted local feature values by converting audio signals into frequency bands to generate fingerprints. Recently, to achieve better robustness to the modifications, a fingerprint method using Chroma vectors obtained from the frequency domain has been proposed (Ewert et al., 2009), (Müller et al., 2005), (Jiang et al., 2011). A recent publication proposing audio fingerprinting algorithm using Chroma, that also meets the demand of robustness against speed, tempo and pitch modification of query audio is proposed by Malekesmaeili and Ward (2014). Also, there are an audio fingerprinting algorithm using representable characteristic feature combinations (Sonnleitner and Widmer, 2016), (Sonnleitner and Widmer, 2014). They propose audio fingerprinting method that is not only robust to noise and audio quality degradation, but also to large amounts of speed or frequency scaling. In addition, it can accurately estimate the scaling factors of applied time/frequency distortions. "lost in space" problem (Lang et al., 2010) is used to make those possible. More precisely, Sonnleitner and Widmer (2014) use a compact four-dimensional, continuous hash representation of quadruples of points called "quad". We will take this as our

Son, H., Byun, S. and Lee, S.
Illegal Audio Copy Detection using Fundamental Frequency Map.
DOI: 10.5220/0008113403500355
In *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019)*, pages 350-355
ISBN: 978-989-758-378-0

reference method in the present paper, because it is the latest publication on this topic, and it reports high precision results for a certain range of speed and noise modification.

Most common fingerprint methods based on frequency components extract local maxima from the frequency band and generate fingerprints with local features (Anguera et al., 2012), (Yang et al., 2014), (Haitsma and Kalker, 2002). However, when the modifications, such as speed, tempo, or pitch change, occur to the original audio signal, it causes irregular changes of the local maxima in the frequency band, which leads to audio matching inaccuracy.

In this paper, we propose a fingerprint extraction method based on fundamental frequency, which improved robustness against the pitch and tempo modifications. To this end, we first extracted the fundamental frequency from audio data and generated a fundamental frequency map (FFMAP) using the extracted fundamental frequency. Then, the generated FFMAP was changed with a specific size through affine transformation. Lastly, the fingerprint was obtained by reconstructing the pitch array from the FFMAP. To compare the similarity between fingerprints, Pearson similarity was used. Tests comparing robustness against the pitch and tempo modifications were conducted for determining the feasibility of the proposed method.

The remainder of this paper is organized as follows: Section 2 explains the proposed FFMAP-based audio fingerprint method. Section 3 presents the experimental results compared to existing method (Sonnleitner and Widmer, 2014), and Section 4 concludes this work.

2 FFMAP-BASED AUDIO FINGERPRINTING METHOD

2.1 Fingerprint Extract

Audio fingerprinting is defined as confirming copyright infringement from audio files by extracting features from audio sources. In general, an ideal fingerprinting system should fulfill several requirements (Cano et al., 2002):

- It should be able to accurately identify an item, regardless of the level of compression and distortion or interference in the transmission channel;
- Depending in the application, it should be able to identify the titles from excerpts of only a few seconds;
- The fingerprinting system should also be computationally efficient. This computational cost is related to the size of the fingerprints, the complexity of the search algorithm and the complexity of the fingerprint extraction.

In this research, we focused on accurately identifying audio sources. We did not consider computationally efficient methods.

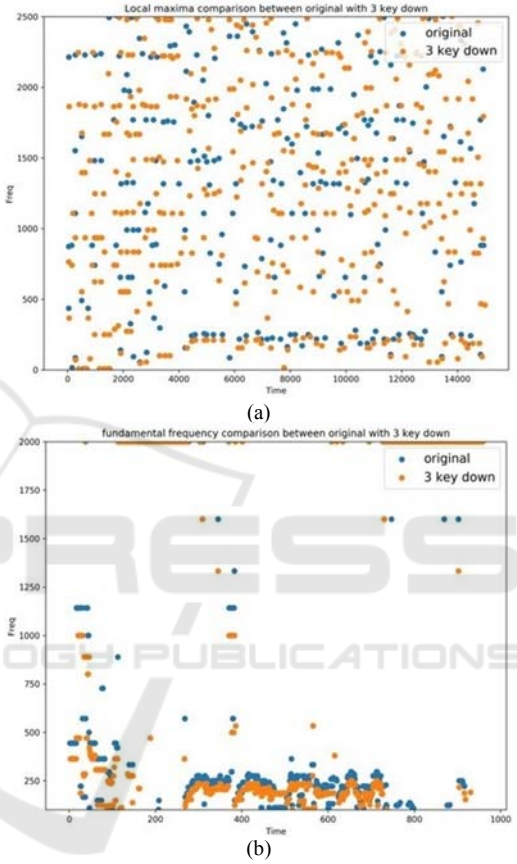


Figure 1: (a) The local maxima of the frequency selected through the existing research method (Wang, 2003) (b) the results of extracting fundamental frequency.

Most studies related to audio fingerprint generate the spectrogram after performing Fourier transform and extract fingerprints by extracting local maxima and analysing the certain structure of the local maxima (Malekesmaeili and Ward, 2012). However, when the pitch change occurs, the local maxima is irregularly changed as shown in Figure 1 (a). This causes audio matching inaccuracy. On the other hand, when using the fundamental frequency, all the frequencies of the audio are scaled by the pitch shift parameter as shown in Figure 1 (b). Using this characteristic, we can design an audio fingerprinting method that resists pitch and tempo changes.

Considering the range of the scales from the music, we extract the fundamental frequency from 100Hz to 2000Hz, which are the frequency of pitches from C3 to C7. Voiceless sounds such as the pronunciation of a person and the hitting sound of a musical instrument have to be removed from original signal in order to extract accurate pitch. In general, voiced sounds are signal with a period characteristic, and voiceless sounds are signal with a non-period characteristic. Therefore, if the normalized autocorrelation value of signal in a certain interval is less than the threshold value, it means that the signal has weak periodicity. As a result, it can be classified as voiceless sounds. The equation of normalized autocorrelation is as follows:

$$R_s(l) = \sum_i s[i] \times s[i-l] \tag{1}$$

$$E_s = \sum_{n=-\infty}^{\infty} |x[n]|^2 \tag{2}$$

$$Nomalized\ autocorrelation = \frac{R_s(l)}{\sqrt{E_{s[n]} \times E_{s[n-l]}}} \tag{3}$$

Where R is autocorrelation function, and S and E are time series signal and the energy of the signal, respectively.

Based on preliminary experiments, we set the threshold to 0.55. If the interval is classified voiceless sound, we mapped it to zero. On the other hand, if the interval is classified voice sound, we find the frequency which get the largest autocorrelation value by increasing frequency from 100Hz to 2000Hz. Then, the frequency which have the largest correlation values become fundamental frequency of the interval. The value of autocorrelation is calculated as the equation (1).

$$\text{FFMAP}(frame_i, frequency_i) = 1, \quad otherwise = 0 \tag{4}$$

Generated FFMAP is as shown in Figure 2. The horizontal axis of the FFMAP is the length of music. We call this width. And the vertical axis of it is the pitch value of the music. Only the pitch value except the octave value is indicated, but it has a higher octave value as it is placed higher location.

FFMAP extracted by proposed method has features that its width can be different from audio to audio. The height of FFMAP is a range of fundamental frequency fixed between 100Hz and 2000Hz, whereas the width of it depends on the length of music. It means that even the same audio can have different FFMAP's width through the tempo or speed modifications. And it leads to inaccuracy matching. Therefore, we fix the width of extracted FFMAP by applying affine transform. The expression of affine transform is as follow in equation (5) and the cubic spline interpolation is used as an interpolation method. Based on preliminary experiments, we set the width to 2000.

$$I(\mathrm{x},\mathrm{y})\begin{pmatrix} w & 0 \\ 0 & H \end{pmatrix} = I^*(x,y) \tag{5}$$

Where I is an original image, I* is a scaled image, and w and H are the width and the height of the music.

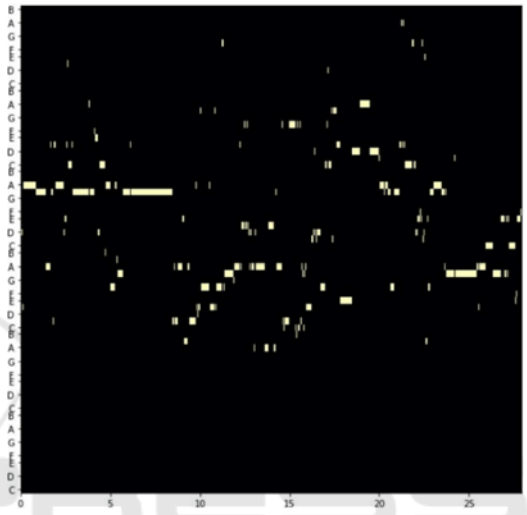


Figure 2: Extracted FFMAP.

2.2 Recognition Algorithm

To perform a search, the above fingerprinting step is performed on an audio file to generate a set of pitch array. This is done over for the entire music. Each array from the audio is used to the database for matching with unlabeled audio file to match. To increase the accuracy of the search, Pearson's correlation score is used in our calculations. The expression of Pearson's correlation score is as follows.

$$\rho_{\mathrm{X,Y}} = \frac{\mathrm{cov}(\mathrm{X},\mathrm{Y})}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{6}$$

$$\mathrm{E}[(\mathrm{X}-\mu_{\mathrm{X}})(\mathrm{Y}-\mu_{\mathrm{Y}}) = \frac{\sum_{i=i}^{m}(X_i-\mu_X)(Y_i-\mu_Y)}{m} \tag{7}$$

Where each μ_X and μ_Y are mean of population X and Y, σ_X and σ_Y are the standard deviation of population X and Y, and m is a number of population objects. Pearson's correlation score is correlation score based on changed in y according to x, unlike the Euclidian score, which is a correlation score based on the distance between x and y. Therefore, robust

fingerprints can be extracted for pitch changes, speed changes and tempo changes.

3 EXPERIMENTAL RESULTS

In this section we evaluate the proposed FFMAP fingerprinting method. We compare it to the algorithm proposed by Sonnleitner and Widmer (2014), henceforth referred to as "Quad-based"

To carry out the comparison, we select a total of 100 audios from different genres. Each audio file is set to mono channel and the sampling frequency is set to 8000Hz. For each song we generated multiple attacked versions by modifying it in terms of tempo, pitch, and speed and noise level. To clarify the terms given as experimental conditions, if only the frequency scale is modified, this is called "pitch" modification. If both time scale and pitch of music are changed proportionally, it is called "speed" modification. And changing only the time scale is referred as a "tempo" modification.

We evaluate the uniqueness of the extracted fingerprinting. To do so, a database of all the fingerprints extracted from all the 100 song is created. The fingerprint database is searched to find potential matches to fingerprints extracted from each query. The threshold value of similarity, which is determined that two audio files generated from the same source data file are the same data, is set to 0.4. The matching criteria used by Quad-based are different (they set the threshold to 0.5). And, we define one performance measures: Precision is the proportion of cases, out of all cases where the system claimed to have identified the reference, where its prediction is correct. Thus, high precision means low number of false positives.

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{8}$$

We tested proposed algorithm against different distortions including modification of tempo, pitch and speed rate which are range from 70% to 130% in steps of 10%. The other distortion environment is the noise adding. In this environment, we generate noisy audio by adding white noise to original audio in SNR level ranges from 0dB to 50dB in steps of 5dB. SNR is a measure that compared the level of a desired signal to the level of background noise. It is defined as the ratio of signal power to the noise power, expressed in decibel(dB). Therefore, the higher the SNR value, the higher the signal than the noise. In case of the speed modification and noise adding, we compare the performance of proposed algorithm to Quad-based under the same experimental conditions.

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{9}$$

3.1 Tempo Modification

In Figure 3, vertical axis presents the precision of the proposed algorithm, and horizontal axis signifies the relative length of the modified audio compared to that of the original audio. The proposed method has been 100% detected for all tempo-change attacks. Therefore, the proposed fingerprint is very invariant about the tempo change.

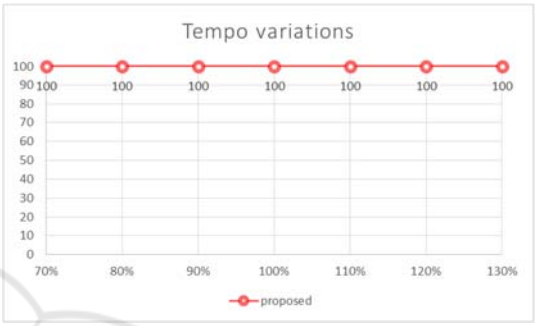


Figure 3: Precision according to tempo variation.

3.2 Pitch Modification

Here, as an additional factor, we put 'pitch change attack' in an experiment taken in Quad-based and run the experiment. The precision for the decrease of semitone was relatively higher than the detection rate for the increase of the original data. And as precision shows in Figure4, high detection rates are shown in all conditions except for the 3 semitones increase attack. Also, the horizontal axis in Figure 4 signifies the relative pitch degree of the modified audio compared to the pitch degree of the original one.

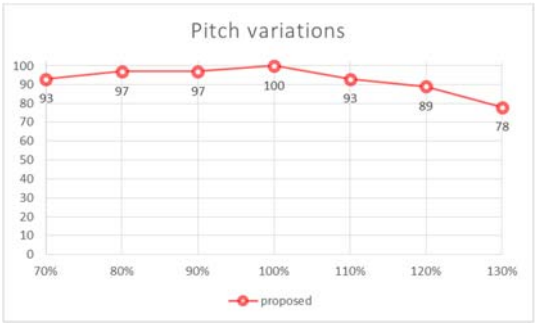


Figure 4: Precision according to pitch variation.

3.3 Speed Modification

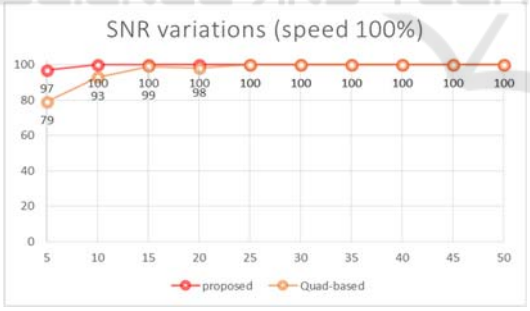
In speed modification, tempo distortion and pitch

distortion occur simultaneously, so the robustness for both should be guaranteed. In Figure 5, horizontal axis indicates the relative degree both length and pitch of the modified audio compared to those of the original audio. And Figure 5 shows that the proposed algorithm can be seen to be much more robust to the speed change than Quad-base.

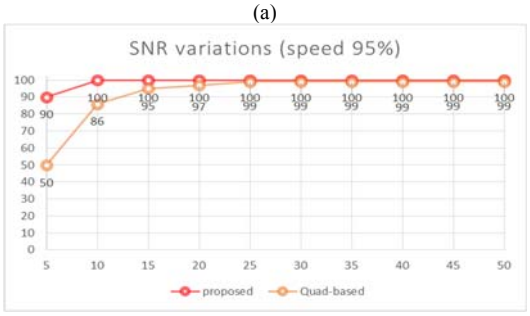
Figure 5: Precision according to speed variation.

3.4 Noise Adding

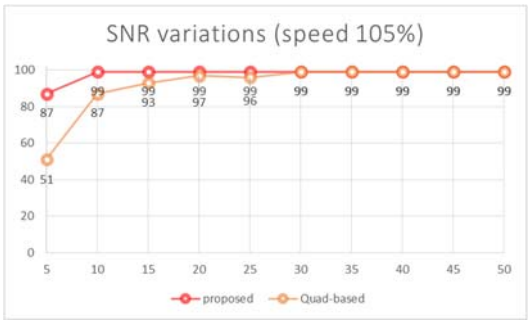
Last experiment is to confirm the robustness against complex attacks, which is noise adding and speed changing. In the experimental environment, we use the original file and speed changed files (speed decrease/increase in steps of 5%) applied same noise adding conditions. As shown in Figure 6 (a), (b) and (c), we still get better results compared to Quad-based.



(a)



(b)



(c)

Figure 6: (a) Precision according to SNR variation with 100% audio speed (b) Precision according to SNR variation with 95% audio speed (c) Precision according to SNR variation with 105% audio speed.

4 CONCLUSIONS

In this paper, we proposed a new audio identification method that is highly robust to tempo, pitch, speed and noise modification. The proposed method is based on a modified spectrogram representation of audio signal as frame-fundamental frequency representation called FFMAP. The audio fingerprint which is generated from FFMAP achieved high performance with all experimental conditions compared to Quad-based. Future research will focus on other requirements the ideal fingerprinting system.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00189, Voice emotion recognition and indexing for affective media service)

REFERENCES

Anguera, X., Garzon, A. & Adamek, T. 2012, "Mask: Robust local features for audio fingerprinting", *2012 IEEE International Conference on Multimedia and Expo*IEEE, pp. 455.

Arnold, M. 2000, "Audio watermarking: Features, applications and algorithms", *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)* IEEE, pp. 1013.

Bellettini, C. & Mazzini, G. 2010, "A framework for robust audio fingerprinting.", *JCM,* vol. 5, no. 5, pp. 409-424.

Cano, P., Batle, E., Kalker, T. & Haitsma, J. 2002, "A review of algorithms for audio fingerprinting", *2002 IEEE Workshop on Multimedia Signal Processing.*IEEE, , pp. 169.

Chen, M., Xiao, Q., Matsumoto, K., Yoshida, M., Luo, X. & Kita, K. 2013, "A fast retrieval algorithm based on fibonacci hashing for audio fingerprinting systems", *2013 International Conference on Advanced Information Engineering and Education Science (ICAIEES 2013)* Atlantis Press, .

Ewert, S., Muller, M. & Grosche, P. 2009, "High resolution audio synchronization using chroma onset features", *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* IEEE, pp. 1869.

Haitsma, J. & Kalker, T. 2002, "A highly robust audio fingerprinting system.", *Ismir*, pp. 107.

Haitsma, J., Kalker, T. & Oostveen, J. 2001, "Robust audio hashing for content identification", *International Workshop on Content-Based Multimedia Indexing* Citeseer, pp. 117.

Hu, H. & Hsu, L. 2015, "Robust, transparent and high-capacity audio watermarking in DCT domain", *Signal Processing,* vol. 109, pp. 226-235.

Jiang, N., Grosche, P., Konz, V. & Müller, M. 2011, "Analyzing chroma feature types for automated chord recognition", *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*Audio Engineering Society.

Jiao, Y., Yang, B., Li, M. & Niu, X. 2007, "MDCT-based perceptual hashing for compressed audio content identification", *2007 IEEE 9th Workshop on Multimedia Signal Processing* IEEE, pp. 381.

Lang, D., Hogg, D.W., Mierle, K., Blanton, M. & Roweis, S. 2010, "Astrometry. net: Blind astrometric calibration of arbitrary astronomical images", *The astronomical journal,* vol. 139, no. 5, pp. 1782.

Malekesmaeili, M. & Ward, R.K. 2014, "A local fingerprinting approach for audio copy detection", *Signal Processing,* vol. 98, pp. 308-321.

Malekesmaeili, M. & Ward, R.K. 2012, "A novel local audio fingerprinting algorithm", *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*IEEE, pp. 136.

Milaš, I., Radović, B. & Janković, D. 2016, "A new audio watermarking method with optimal detection", *2016 5th Mediterranean Conference on Embedded Computing (MECO)* IEEE, pp. 116.

Müller, M., Kurth, F. & Clausen, M. 2005, "Audio Matching via Chroma-Based Statistical Features.", *ISMIR*, pp. 6th.

Seo, J.S. 2014, "An asymmetric matching method for a robust binary audio fingerprinting", *IEEE Signal Processing Letters,* vol. 21, no. 7, pp. 844-847.

Sonnleitner, R. & Widmer, G. 2016, "Robust quad-based audio fingerprinting", *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 3, pp. 409-421.

Sonnleitner, R. & Widmer, G. 2014, "Quad-Based Audio Fingerprinting Robust to Time and Frequency Scaling.", *DAFx* Citeseer, pp. 173.

Wang, A. 2003, "An Industrial Strength Audio Search Algorithm.", *Ismir*Washington, DC, pp. 7.

Wu, C., Su, P. & Kuo, C.J. 2000, "Robust and efficient digital audio watermarking using audio content analysis", *Security and Watermarking of Multimedia Contents II International Society for Optics and Photonics*, pp. 382.

Yang, G., Chen, X. & Yang, D. 2014, "Efficient music identification by utilizing space-saving audio fingerprinting system", *2014 IEEE International Conference on Multimedia and Expo (ICME)* IEEE, pp. 1.