Blind Source Separation for a Robust Audio Recognition Scheme in Multiple Sound-Sources Environment

Wei Han[1,2,3], Songbin Zhou[1,2,3], Chang Li [1,2,3], Yisen Liu[1,2,3], Zhe Liu[1,2,3]
[1]Guangdong Institute of Automation
[2]Key Laboratory of Modern Control Technology of Guangdong Province
[3]Public Laboratory of Modern Control and Manufacture Technology of Guangdong Province
[1,2,3]Guangzhou, China
e-mail: w.han@gia.ac.cn

Abstract—In multiple sound-sources environment, robustness is a major challenge for audio recognition system based on audio fingerprinting, because mixed audio signals may make recognition rate has a significant decline. This paper proposes a novel audio fingerprinting method, which uses blind source separation to divide mixed audio signals into independent components and each is close to its original sound-source, then the classical Philips scheme can perform accurately identifying. Experimental results show that novel scheme is quite robust in noisy conditions where uncertain audio signals mixed by various numbers of sound-source, even though the feature of each original sound-source and their mixed model are unknown.

Keywords—audio recognition; audio fingerprinting; multiple sound-sources environment; blind source separation

I. INTRODUCTION

Now, audio fingerprinting is a very common way to recognize an unknown audio clip. It has been reported that there already are available services not only for providing music search such as Shazam [1], but also for monitoring broadcast for advertisement tracking [2] and integrity checking for audio content [3]. As the applications of audio fingerprinting on mobile devices are becoming more and more widely, it urgently needs to possess more robust against multiple sound-sources environment, especially for working in various public places.

Some excellent audio fingerprinting schemes have been proposed to satisfy audio recognition. Philips scheme proposed by Haitsma and Kalker [4] is proven to be the most accurate audio fingerprinting scheme in a relatively noise-free environment. Wooram [5] uses predominant pitch extraction to devise an approach of sub-fingerprint masking, which improves the robustness of Philips scheme. The system developed by Wang [6] has become a successful commercial application. Based on the idea of Wang's method, Jun-Yong Lee [7] proposes an adaptive audio fingerprinting extraction method based on the constant Q transform (CQT) to enhance the robustness of audio fingerprinting in a real noisy environment for real-time TV advertising identification.

In practice, however, it still needs further improvement to be used in multiple sound-sources environment. In this paper, blind source separation (BSS) is used to segregate unknown mixed audio signals to get independent components which are close to their original sound-sources, and then the classical Philips scheme can perform exactly identifying.

II. BLIND SOURCE SEPARATION BASED ON FASTICA ALGORITHM

In practical applications, the recorded signals are often polluted by other sound-sources. And worse still, all the original sound-sources and their mixed way are "blind", only indistinct mixed audio signals can be observed. But BSS, which can divide mixed signals into independent components, is an efficient way to restore original signals from their mixed signals. FastICA algorithm [9] is the most mutual implementation method for BSS.

A. Background of the FastICA Algorithm

Assume that the mixed audio signals matrix is $\boldsymbol{X}$ defined as

$$\boldsymbol{X} = \boldsymbol{AS} \tag{1}$$

where $\boldsymbol{X} = (x_1, x_2, \mathrm{L}\ , x_n)^T$ have n observed acoustical signals which mixed by n unknown independent original sound-sources $\boldsymbol{S} = (s_1, s_2, \mathrm{L}\ , s_n)^T$, and $\boldsymbol{A}$ is a full-rank n by n mixing matrix.

The goal of FastICA algorithm is to recover independent original audio signals from their mixed signals by finding a linear transformation matrix $\boldsymbol{W}$ that maximizes the mutual independence of sound-mixture. The decomposition model is shown in equation (2).

$$\boldsymbol{Y} = \boldsymbol{WX} = \boldsymbol{WAS} = \boldsymbol{GS} \tag{2}$$

Thus separation can be achieved when $\boldsymbol{G}=\boldsymbol{E}$ ($\boldsymbol{E}$ is a nth-order identity matrix) results from repeatedly learning.

FastICA measures non-Gaussianity using kurtosis to find independent components from their mixtures. FastICA algorithm based on the fixed-point iteration scheme is to find the maximum of the non-Gaussianity of $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ as measured by negentropy. The unit vector $\boldsymbol{W}$ is substituted into the projection $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ such that the negentropy is maximized. The fixed-point iteration operations, of the FastICA algorithm using an

approximate negentropy and Newton iteration are addressed as [10].

B. *The Effectiveness of BSS*

Almost all schemes [4-8] extract audio fingerprinting from spectrum feature of audio signals. Therefore, the difficulty for identifying mixed audio signals by audio fingerprinting can be deduced from analyzing the discrepancy between mixed signals' spectrums and original signals' spectrums. Randomly selecting and mingling arbitrary three audio clips A1, A2 and A3, their mixed signals are A4, A5, and A6, as shown in Fig .1; their own spectrums and their mixed signals' spectrums are shown in Fig .2; the separated independent components from mixed signals are AA1, AA2 and AA3, their spectrums are shown in Fig .3.

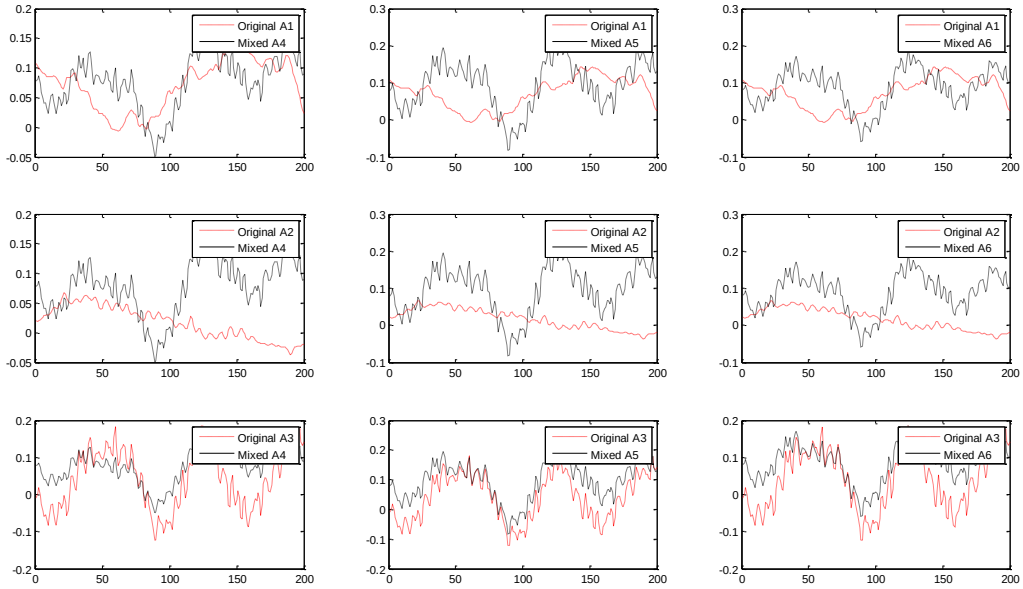

Figure 1. Original audio signals and their mixed signals

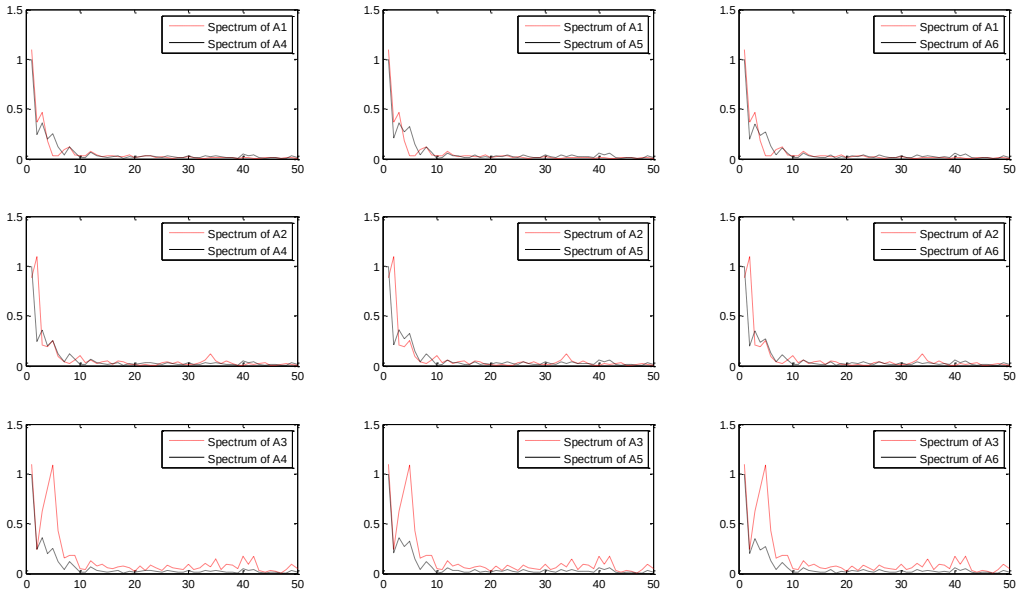

Figure 2. Spectrums of original audio signals and their mixed signals

In practice, however, we can not confirm the corresponding relation between separated independent components and original sound-sources in general, i.e., the AA1, AA2 and AA3 are not doubtless respectively corresponding to A1, A2 and A3. Therefore, in Fig .3, the spectrums' relationship between A1 and AA1, AA2, AA3 is listed separately, so is A2 and A3.

Fig .1 and Fig .2 demonstrate that the obvious differences in original signals' spectrums and mixed signals' spectrums, which will result in imparity between mixed signals' audio fingerprinting and original signals' audio fingerprinting, even though the mixed signals composited by original signals.

Fig .3 shows the spectrums of separated independent signals. It can be easily to perceive that the spectrums of independent signals are very approximate to their original signals' spectrums, from which the enormous degree of closeness of their audio fingerprinting can be

concluded. And actually, it can be clearly seen at least that the spectrums of A2, A3 is similar to AA1's spectrum, AA2's spectrum separately.

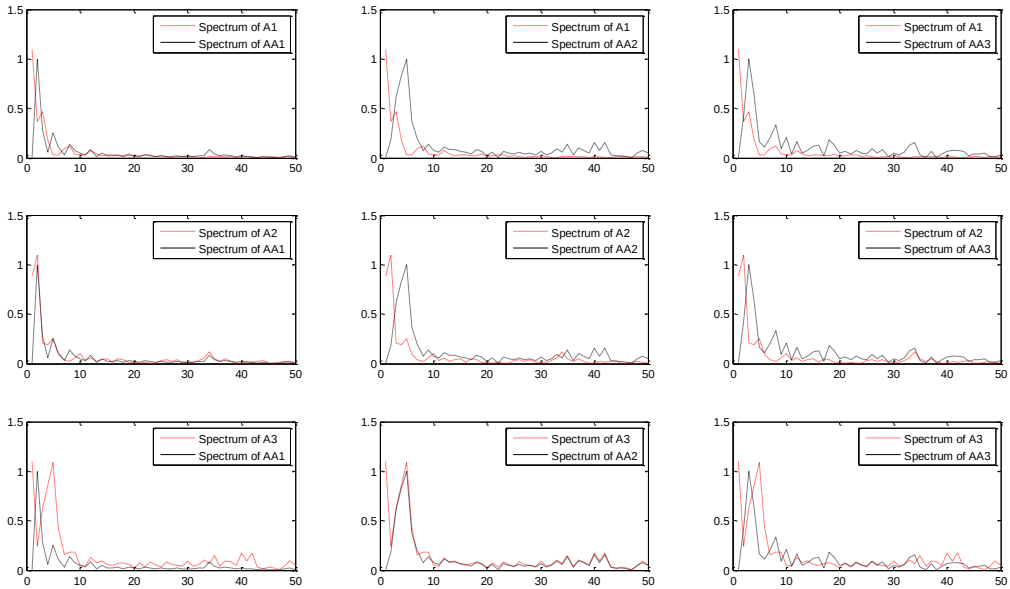

Figure 3. Spectrums of original audio signals and separated independent signals

III. AUDIO FINGERPRINTING SCHEME

The proposed audio fingerprinting system (BFP scheme) is based on the Philips' hashing algorithm. This section is divided into two modules to describe in detail.

A. *Philips Scheme*

The particulars of Philips' hashing algorithm are given in [4]. The audio signal is sampled at the rate of 44100 Hz and segmented into overlapping frames, each of which contains 512 non-overlapped samples and 15872 overlapped samples. Each frame of 16384 samples is then Fast Fourier Transformed. By logarithmically dividing the obtained audio spectrum, 33 non-overlapping frequency bands from 300 Hz to 2000Hz are acquired. Then total of 32 hash bits are assigned for each frame to become a single sub-fingerprint. A single sub-fingerprint for frame *n*th frame is defined as a bit sequence of F(n,m) for 0⩽m ⩽31 where F(n,m) is defined as equation (3).

$$F(n,m)=\begin{cases}1 & if\,(E(n,m)-E(n,m+1)) \\ & -(E(n-1,m)-E(n-1,m+1))>0 \\ 0 & if\,(E(n,m)-E(n,m+1)) \\ & -(E(n-1,m)-E(n-1,m+1))\le 0\end{cases} \quad (3)$$

B. *BFP Scheme*

As shown in Fig .4 is the overview of BFP scheme, for the robust fingerprinting extraction in multiple sound-sources environment, we propose to use N microphones (N should more than the number of original sound-source in general [11]) to collect mixed audio signals, then divide mixed signals into independent components by BSS. Each independent component is very approximate to its original. Due to it is hard to exactly confirm the sequence of independent components and their corresponding relation with the original signals, i.e., it is uncertain that which is the needed independent component, thus every independent component has to be put into fingerprinting database to query.

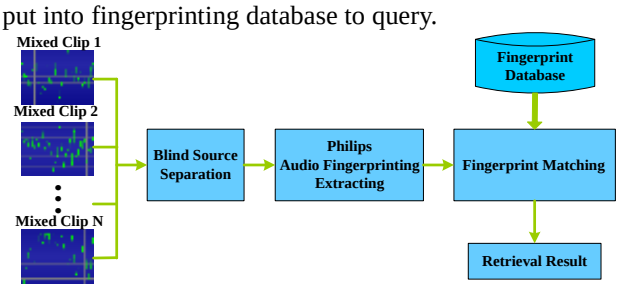

Figure 4. The overview of BFP scheme

IV. EXPERIMENTS

To evaluate the performance of BFP scheme, we implement the following three schemes including the proposed algorithm to compare: 1) Our fingerprinting scheme (BFP); 2) Wooram's fingerprinting scheme (MBM) [5]; 3) Philips scheme [4].

A. *Experimental Data*

Experiments were performed using a music database containing 1000 songs randomly selected from worldwide popular songs of various genres such as DJ, electronic, classic, blues, jazz, folk, light music, hip-hop, country, rock and so on. All the audio data are stored in PCM format with mono, 16 bit depth and 44.1 kHz sampling rate. Fingerprinting database is composed of these 1000 songs' audio fingerprinting. From the selected songs, 1000 randomly created audio query clips of three, six and nine seconds. And in the following experiments, the mixture of M (M=2, 3, 4) sound-sources refers to an unknown audio clip mixed by arbitrary M audio clips in

these 1000 fragments.

B. *Experimental Results*

Tab.I and Fig .5 show the results of the audio retrieval experiments performed on the database based on three different schemes, which are BFP, MBM and Philips scheme. In the experiment, the length of audio query clips is 6s, and for MBM scheme, the bit-mask used in our experiment has seven bits set to 1. These results clearly show that BFP scheme outperforms other two schemes in retrieval accuracy in the conditions of sound-commixture mixed by various numbers of sound-source, including the most common white noises.

TABLE I. THE ACCURACY OF THREE SCHEMES

Process approach	BFP Scheme	MBM Scheme 7-Bit	Philips Scheme
Mixture of two sound-sources	95.1%	65.7%	63.2%
Mixture of three sound-sources	84.3%	30.8%	29.5%
Mixture of four sound-sources	68.4%	7.1%	6.5%
Mixture of three sound-sources and white noise	67.6%	5.4%	5.3%

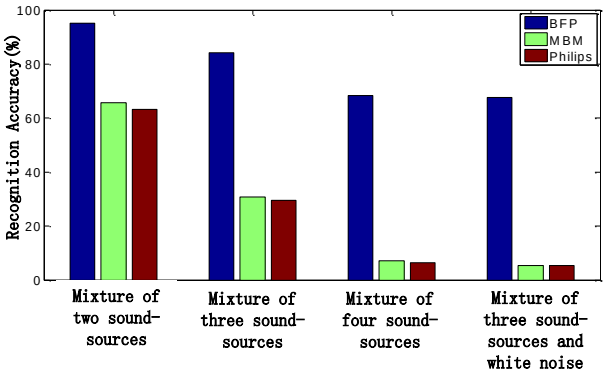

Figure 5. Recognition performance evaluation of BFP, MBM and Philips

Tab.II and Fig .6 show the recognition performance of BFP scheme when query length is changed. This result indicates that the accuracy increases as the length of the query prolongs. Also, the proposed scheme shows satisfactory performance with just three seconds long query.

TABLE II. ACCURACY EVALUATION ACCORDING TO QUERY LENGTH

Process approach	Query Length		
	3s	6s	9s
Mixture of two sound-sources	91.3%	95.1%	95.8%
Mixture of three sound-sources	80.8%	84.3%	85.1%
Mixture of four sound-sources	61.9%	68.4%	68.7%
Mixture of three sound-sources and white noise	60.5%	67.6%	68.0%

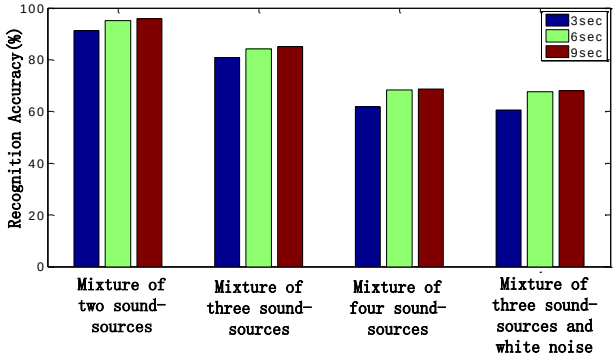

Figure 6. Accuracy evalution according to query length

V. CONCLUSIONS

This paper proposes a novel modified audio fingerprinting algorithm based on Philips scheme to recognize mixed audio signals in multiple sound-sources environment. The proposed algorithm enhances the Philips fingerprinting algorithm by dividing mixed audio signals into independent components which are close to their original sound-sources, which guarantees great similarity between separated independent component's audio fingerprinting and original signals' audio fingerprinting. It clearly outperforms original Philips algorithm in recognizing audio signals in multiple sound-sources environment. However, the corresponding relationship between separated independent signals from mixed audio signals and original signals is unknown — that is, it is uncertain that which is the needed independent signals. So we have to put every separated independent signals' audio fingerprinting into fingerprinting database to query, which will increases retrieval time undoubtedly. Although there already have some BSS algorithms with restrictive conditions to implement accurately separating, the effectiveness should be improved. Therefore, the improvement in exactly sound-sources separating to reduce retrieval time is considered for future work.

ACKNOWLEDGMENT

This work was supported by the Science and Technology Project of Guangdong Province (Grant no. 2013B091100013, Grant no. 2013B060100013), the Science and Technology Project of Guangzhou City (Grant no. 2013J2200062), the Science and Technology Project of Guangdong Institute of Automation (Grant no. A201406), and the Scientific Research Foundation of Guangdong Academy of Science for Young (Grant no. qnjj201306).

REFERENCES

[1] http://www.doreso.com/

[2] Cerquides, J.R. "A real Time Audio Fingerprinting System for Advertisement Tracking and Reporting in FM Radio," Radioelektronika, 2007. 17th International Conference , Apr. 224-25, 2007, Brno, The Czech Republic, pp. 203-206.

[3] E.G ́omez, P.Cano, C.T.Gomes, etc. "Mixed Watermarking-fingerprinting Approach for Integrity Verification of Audio Recordings," Proceedings of International Telecommunications Symposium—ITS2002, Sept. 2002, Natal, Brazil, pp. 271-284.

[4] Haitsma, J. and T. Kalker. “A highly robust audio fingerprinting system,” Proceedings of the 3rd International Conference on Music Information Retrieval, Oct. 13-17, 2002, Paris, France, pp. 107-115.

[5] Wooram Son, Hyun-Tae Cho, Kyoungro Yoon and Seok-pil Lee. “Sub-fingerprint Masking for a Robust Audio Fingerprinting System in a Real-noise Environment for Portable Consumer Devices,” IEEE Transactions on Consumer Electronics, vol. 56, pp. 156-160, Feb. 2010.

[6] Avery Wang. “The Shazam Music Recognition Service,” Communications of the ACM, vol. 49, pp.44-48, Aug. 2006.

[7] Jun-Yong Lee, Hyoung-Gook Kim. “Audio Fingerprinting to Identify TV Commercial Advertisement in Real-Noisy Environment,” 2014 International Symposium on communications and Information Technology (ISCIT), Sep. 24-26, 2014, Incheon, Korea, pp. 527-530.

[8] Chahid Ouali, Pierre Dumouchel and Vishwa Gupta. “A Robust Audio Fingerprinting Method for Content-Based Copy Detection,” 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), Jun. 18-20, 2014, Klagenfurt, Austria, pp. 1-6.

[9] Kuo-Kai Shyu, Ming-Huan Lee, Yu-Te Wu, and Po-Lei Lee. “Implementation of Pipelined FastICA on FPGA for Real-Time Blind Source Separation,” IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 19, pp. 958-970, June 2008.

[10] Lan-Da Van, Di-You Wu and Chien-Shiun Chen. “Energy-Efficient FastICA Implementation for Biomedical Signal Separation,” IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 22, pp. 1809-1822, Nov. 2011.

[11] Da-Peng Guo, Qiu-Hua Lin. “Fast decryption utilizing correlation calculation for BSS-based speech encryption system,” 2010 Sixth International Conference on Natural Computation (ICNC), Aug. 10-12, 2010, Yantai, Shandong, pp. 1428-1432.