World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:5, No:7, 2011

A Robust Audio Fingerprinting Algorithm in MP3 Compressed Domain

Ruili Zhou, Yuesheng Zhu

***Abstract*—**In this paper, a new robust audio fingerprinting algorithm in MP3 compressed domain is proposed with high robustness to time scale modification (TSM). Instead of simply employing short-term information of the MP3 stream, the new algorithm extracts the long-term features in MP3 compressed domain by using the modulation frequency analysis. Our experiment has demonstrated that the proposed method can achieve a hit rate of above 95% in audio retrieval and resist the attack of 20% TSM. It has lower bit error rate (BER) performance compared to the other algorithms. The proposed algorithm can also be used in other compressed domains, such as AAC.

***Keywords*—**Audio Fingerprinting, MP3, Modulation Frequency, TSM

I. INTRODUCTION

WITH the development of multimedia technology and the advent of massive music information, some applications like music identification, music copyright verification, music searching and audio monitoring are developing as well. Audio fingerprinting is a compact signature based on the content of audio signal, which represents an important acoustic feature and the essence of the music signal. Audio fingerprinting on raw audio format has been studied [1]. A block diagram [2] of audio fingerprinting in uncompressed domain is shown in Fig.1, which extracts audio fingerprint from the energy difference between two adjacent bands and can resist most of the distortions except time scale modification (TSM) above 2%. A audio fingerprinting in entropy domain is proposed in [3], its robustness to low pass filtering and @32kbps compression is better than that of [2], but other distortions are not considered in this method. A noise robust fingerprinting method [4] uses long-window analysis strategy to resist some unknown distortions. The time-frequency variations based on a transformation with efficient time-scale localization is analyzed in [5] and two fingerprints are created for authentication and recognition purposes, respectively. In [6], time-frequency theory integrated with psychoacoustic results on modulation frequency perception is used for audio fingerprinting. It not only contains short-term information about the signals, but also provides long-term information representing patterns of time variation. A cross entropy approach is considered to classify music signals and a better performance under TSM distortion is achieved.

As compressed audio signals are increasing and have become the dominant fashion of audio file storage in personal electronic equipments and transmission on the Internet, it is important to extract the audio feature in compressed domain directly. It is noted that only a few work of music retrieval in compressed domain are reported and most of the algorithms are directly transplanted from those in raw data domain. Modified Discrete Cosine Transform (MDCT) coefficients and their energy derivation are frequently used to extract the fingerprint. Fig.2 shows the general procedures of audio fingerprinting in compressed domain. A robust compressed domain audio fingerprinting algorithm is developed in [7], which takes the ratio between the sub-band energy and the full-band energy of a segment as intra-segment feature and difference between continuous intra-segment features as inter-segment feature, and robustness to some distortions except TSM is shown. Compressed-domain spectral entropy is utilized as the audio feature in an audio fingerprinting algorithm [8], but it can resist TSM up to 2% only. In [9]- [11] the beat, rhythm and tempo information of the music signal is extracted respectively. A video retrieval system(VRS) is developed in [12] for Interactive-Television in AC3 domain, in which long-term logarithmic modified DCT modulation coefficients(LMDCT-MC) are used for audio indexing and retrieval and a two-stage search (TSS) algorithm for fast searching is also proposed.

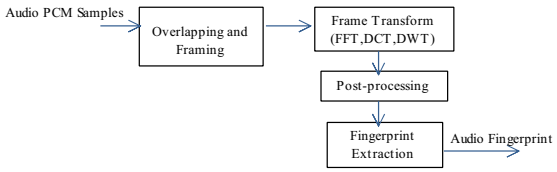

Fig.1.Block diagram of audio fingerprinting in uncompressed domain

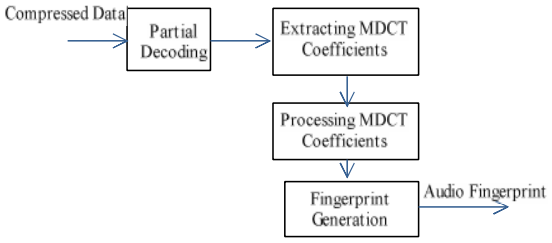

Fig.2.Block diagram of audio fingerprinting in compressed domain

In summary, most of the current algorithms in compressed domain are fragile to TSM distortions. It is our motivation to design an algorithm in compressed domain robust to TSM. Therefore, a new robust audio fingerprinting algorithm in MP3 compressed domain is proposed with high robustness to time scale modification (TSM) based on the method in [6]. The remaining of this paper is organized as follows: in Section II a

Ruili Zhou and Yuesheng Zhu are with Communication & Information Security Lab, Shenzhen Graduate School, Peking University, Shenzhen, CHINA. Corresponding author e-mail add: zhuys@szpku.edu.cn.

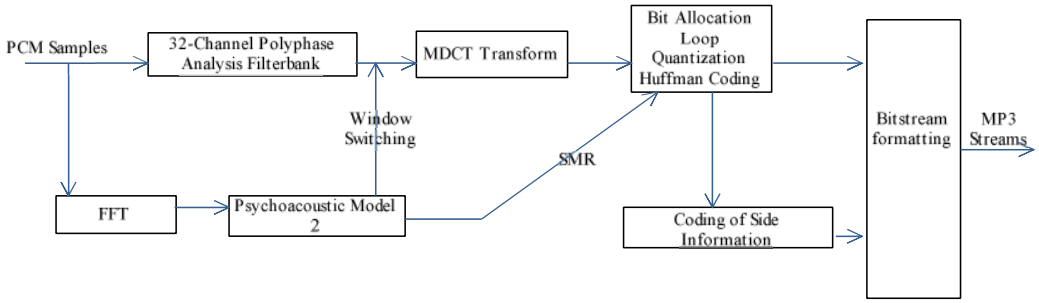

Fig.3.The flowchart of MP3 encoding

brief introduction to the principles of MP3 encoding is given; in Section III the proposed method is described, the experimental results are discussed in Section IV; and the conclusions is gives in Section V.

II. Principles Of MP3 Encoding

Fig.3 shows the encoding process of MP3. Audio data is encoded frame-by-frame. An MP3 frame consists of 2 granules, where each granule contains 576 samples per channel. Every sample is 16 bit. Firstly, The input PCM signal is mapped into 32 equal-bandwidth sub-bands by a poly-phase filter bank, which simulates the critical bands in the human auditory system (HAS). Then each sub-band is equally divided into 18 sub-bands by using MDCT. At last according to the SMR provided by Psychoacoustic Model 2, the MP3 bit stream is generated through bit allocation, quantization and Huffman coding for each sub-band signal.

MP3 decoding is the inverse process of encoding. In this paper, only partially decoding is needed to retrieve the MDCT coefficients for audio fingerprinting so that both the computation complexity and the storage requirement can be reduced, which is beneficial for practical applications.

III. Modulation Frequency Analysis

A. The principle of modulation frequency analysis

A detailed theory of modulation frequency analysis and its applications is given in [13]. Modulation transform are used as the second dimension transform in [14], enabling the audio coding algorithm to outperform traditional MP3 coding. The audio fingerprint on PCM format with modulation frequency analysis is extracted in [6], which shows its good robust performance under TSM and time misalignment distortions.

Short data window analysis technique is used in most of the algorithms for audio fingerprinting. But it ignores the long-term signal variation. Compared to the traditional methods, a new component, Modulation frequency analysis, is added as shown in Fig.4. In this paper, wavelet Transform is used for modulation frequency analysis. The mathematical process is detailed as follows:

Assuming $P_{sp}(t,\omega)$ as the result of Fourier transform in Fig.4, it's calculated in (1) :

$$P_{sp}(t,\omega)=\frac{1}{2\pi}\left|\int x(u)w^{*}(u-t)e^{-j\omega u}du\right|^{2} \quad (1)$$

Taking Continuous Wavelet Transform (CWT) as the modulation transform, $P_{sp}(s,\eta,\omega)$ is gotten in (2) with s, $\psi(\,)$, η representing wavelet filter, discrete scales, wavelet translation respectively:

$$P_{sp}(s,\eta,\omega)=\int P_{sp}(t,\omega)\psi^{*}(\frac{t-\xi}{s})dt \quad (2)$$

A sum across the wavelet translation axis η is performed to produce a joint frequency representation with non-uniform frequency resolution on the modulation frequency axis as shown in (3).

$$P_{sp}(s,\omega)=\int\left|P_{sp}(s,\eta,\omega)\right|^{2}d\eta \quad (3)$$

$P_{sp}(s,\omega)$ is the representation of modulation frequency of audio signal that we need in audio fingerprinting. To simplify the mathematical process above, a simple function y1(t) is applied to (1):

$$y1(t)=\cos(10\pi t)+\cos(50\pi t) \quad (4)$$

The spectrogram of y1(t) is shown in Fig.5. Fig.6 shows the result of modulation frequency analysis. Compared to Fig.5, high-energy values are concentrated in low modulation frequencies, which represent the energy more compactly and are better for feature extraction.

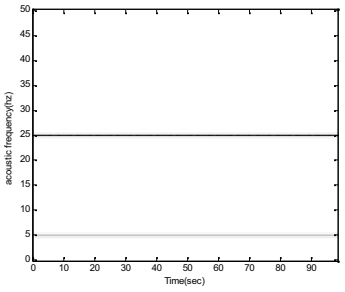

Fig.5.The spectrogram of signal (4)

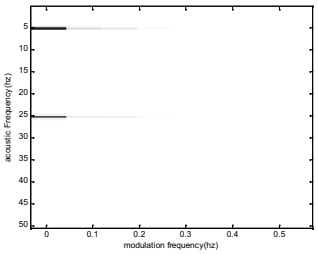

Fig.6.Modulation frequency representation of signal (4)

Fig.4. Modulation frequency analysis

As an example, y2(t) is a simple AM signal with modulation frequencies of 4Hz, 20Hz and carrier frequency of 2000Hz respectively given in (5).

$$y2(t)=(1+\cos(8\pi t)+\cos(40\pi t))\cos(4000\pi t) \quad (5)$$

Fig.7 (a), 7(b) shows the result of FFT, CWT respectively. The energy of the signal are concentrated around 2000Hz of acoustic frequency in both figures, but the two modulation frequencies are separated more clearly in Fig.7 (b) than in Fig. 7(a), which means FFT cannot resolve these modulation frequencies. Therefore CWT is chosen in our algorithm.

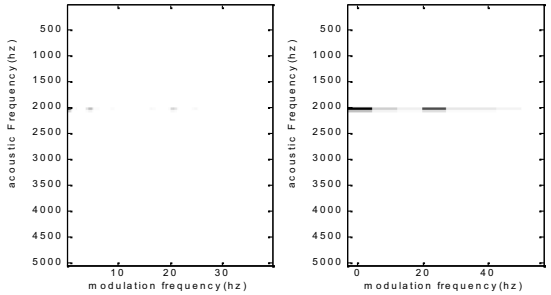

Fig.7. (a) Modulation frequency representation based on FFT.
(b) Modulation frequency representation based on CWT

B. The proposed feature extraction method

In our algorithm, the modulation frequency analysis is used for audio fingerprinting in MP3 domain, as it has been successful applied to the raw audio format in [6]. Each song is with a 10-s digital audio passage. MDCT coefficients derived from partially decoded MP3 songs are windowed into multiple blocks with 4-s block length and 1-s block rate. Each block is further grouped into overlapping sub-blocks with 20 granules and an overlap factor of 17/20. Then the MDCT coefficients in each sub-block are divided into 18 new sub-bands [15] as shown in TABLE I. The sum energy of the MDCT coefficients is calculated for each new sub-band in the sub-block as in (6), where SBE(i,j,k) denotes the k^{th} new sub-band energy in the j^{th} sub-block of i^{th} block, $mdctk_{down}$ and $mdctk_{up}$ represent the top and bottom index of MDCT coefficients which belong to the k^{th} subband respectively.

$$SBE(i,j,k)=\sum_{m=1}^{20}\sum_{n=mdctk_{down}}^{mdctk_{up}}\left|mdct(i,j,m,n)\right|^2 \quad (6)$$

With (2) applied along j axis of SBE(i,j,k) and a sum across the wavelet translation shown in (3),we get the final modulation frequency expression $P_i(\eta,\omega)$ for the i^{th} block.

(7) is the modulation frequency expression for the whole music segment with i ,M representing the block index and the total block num respectively.

$$P=\{P_1,P_2,...P_i...P_M\} \quad (7)$$

Finally we derive the feature hash vector S in (9) by applying (8) to each P in (7).

$$S_i(s_d,k)=P_i(s_d,k+1)-P_i(s_d,k) \quad (8)$$

$$S=\{S_1,S_2,...S_i...S_M\} \quad (9)$$

TABLE I
NEW SUBBANDS DIVISION

Subband Index	MDCT Coefficients (Long Window)	Frequency Range(Hz)
1	1-4	-175
2	5-8	175-328
3	9-12	328-481
4	13-16	481-634
5	17-20	634-830
6	21-24	830-983
7	25-30	983-1213
8	31-36	1213-1443
9	37-44	1443-1749
10	45-52	1749-2098
11	53-62	2098-2481
12	63-74	2481-2983
13	75-90	2983-3639
14	91-110	3639-4491
15	111-134	4491-5495
16	135-162	5495-6610
17	163-196	6610-8084
18	197-238	8084-9950

C. Matching

A feature set is defined to represent a query song in (10)

$$Q_i=\{Q_{i1},Q_{i2},...Q_{im}...Q_{iN}\},\quad N\geq M \quad (10)$$

where *i* is the i^{th} reference song in the database. Select an M-dimension vector Q_i^j from Q_i iteratively, where $1\leq j\leq N-M+1$, calculate the hamming distance between S and Q_i^j, then the minimum one is selected as the distance between the two. Finally bit error rate (BER) is calculated as the ratio of distance to the total number of bits. If BER is below the predefined threshold, they are identified as the similar songs, otherwise different.

IV. Experiment Results And Discussion

In our database, there are 1019 mono songs with 44.1 kHz sample rate and 128kbps bit rate. Each song is 10s long.

A. BER threshold determination

500 songs are selected as the query examples to determine the threshold. Twelve signal degradations are applied to each

song using Adobe Audition and Gold Wave, resulting in 12 distorted copies of them. The setting is as follows:

·MP3 encoding: @64kbps

·Equalization:

Freq.(Hz)	31	62	125	250	500	1k	2k	4k	8k	16k
Gain(dB)	-3	3	-3	3	-3	3	-3	3	-3	3

·Echo Addition: 100ms, 50% echo addition.

·Noise Addition: noise interference with 15dB SNR.

·Change Volume:-6.0206dB, 3.5218dB respectively

·Resampling: downsampling to 22.05 kHz and then upsampling back.

Band-pass Filtering: cut-off frequencies are 100Hz and 6000Hz respectively.

·Time Scale Modification: ±2% and ±4% .Only the tempo change, the pitch remains unaffected.

In our experiment, the BER of all possible pairs between the fingerprint sequences of 12 distorted copies and those stored in the database are calculated, total 611400 BER values are further classified as BER of the same song in Fig.8 and BER of different songs in Fig.9.

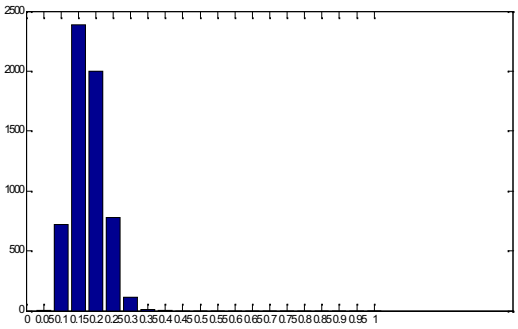

Fig.8. BER distribution of the similar songs

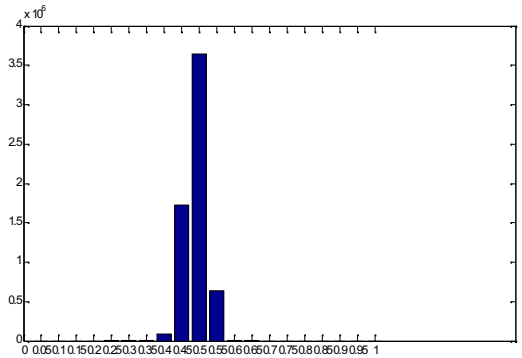

Fig.9. BER distribution of the distinct songs

According to Fig.8 and Fig.9, the BER threshold is set in a range of 0.3 to 0.35. TABLE II shows the false positive rate (FPR) varying with the different thresholds. According to the FPR proposed in [6]-[8], 0.32 is chosen as the BER threshold, its corresponding FPR equals 4.4e-6 that is adequate for practical audio identification applications.

TABLE II
FPR VS. BER THRESHOLD

BER threshold	FPR
0.3	4.9e-7
0.31	1.47e-6
0.32	4.4e-6
0.33	1.36e-5
0.34	4.06e-5
0.35	1.28e-4
0.36	3.85e-4

B. Retrieval Results

During the retrieval test, other 509 songs are used to generate the distorted songs by using the 12 operations, some distortions are further included to test the robustness to unseen operations as shown bellow.

Amplitude Compression: with the following compression ratios: 8.94:1 for|A|≥-28.6 dB; 1.73:1 for -46.4 dB<|A|<-28.6 dB; 1:1.61 for |A|≤-46.4 dB.Time distortion: 0.5 second and 1 second time misalignmentTime Scale Modification: ±8%, ±12%, ±16%, ±20%, and ±30% respectively.The detection rate is the ratio of the num of songs bellow the BER threshold to the total num of test songs for each distortion and the accuracy rate is the ratio of correct ones to detection results for each distortion, and the results under known distortions are above 95% as shown in TABLE III, which is similar to other MP3 domain algorithms [7] [8], but more robust to TSM distortions.

TABLE III
RESULTS UNDER KNOWN DISTIORTIONS

Processing	Detection Rate	Accuracy Rate
equalization	100%	100%
MP3@64kbps	100%	100%
Volume change(-6.0206dB-)	100%	99.8%
Volume change(+3.5218dB)	100%	99.8%
Echo addition	100%	99.8%
Noise addition	100%	100%
Bandpass filtering	99.8%	100%
Downsampling	100%	100%
TSM+2%	99.8%	99.2%
TSM-2%	100%	99.8%
TSM+4%	99.8%	99.2%
TSM-4%	99.8%	99.4%

TABLE IV shows the detection rate and the accuracy rate under unseen distortions. The algorithm in this paper can resist Amplitude Compression which is not considered in most of the MP3 domain algorithm like [7] [8]. The results under 1s time misalignment is not as good as that in [6], but it suggests good robustness under 0.5s time misalignment. Also a strong robustness against 20% TSM is obtained, which outperforms other MP3 domain algorithms [7] [8].

TABLE IV
RESULTS UNDER UNKNOWN DISTORTIONS

Processing	Detection rate	Accuracy
Amplitude compression	94.5%	99.8%
1s time misalignment(start)	75.8%	99.7%
1s time misalignment(end)	76%	99.2%
0.5s time misalignment(start)	99%	100%
0.5stime misalignment(end)	99.4%	99.8%
TSM+8%	99.8%	99.4%
TSM-8%	99.4%	99.2%
TSM+12%	99.2%	99.2%
TSM-12%	99.6%	99.2%
TSM+16%	98.4%	97.8%
TSM-16%	99%	99%
TSM+20%	97.6%	98.2%
TSM-20%	97.6%	98.2%
TSM+30%	92%	90%
TSM-30%	62.3%	96.1%

C. Comparison Results

A comparison is made between our algorithm and the algorithms in [7] [8] with four songs of different musical genres: “Nothing In The World” by Atomic Kitten, “Canon in D” by Johann Pachelbel, ”Yellow” by Cold play, and ”Ring a ling” by The Black Eyed Peas. The results are shown in TABLE V, and II, III represents the algorithms in [7] and [8] respectively. We can see that the BER values in [7] [8] are higher than those in our algorithm, which verifies the effectiveness of the proposed algorithm.

V. Conclusion

In this paper, an audio fingerprinting algorithm is proposed based on the modulation frequency analysis of MP3 streams. Both short-term information about the signals and long-term information representing patterns of time variation are contained in the fingerprint. It shows strong robustness to TSM distortion up to 20%, which is better than most other algorithms in the MP3 domain at the cost of a higher computation complexity.

References

[1] P. Cano, E. Batlle, T. Kalker and J. Haitsma, “A review of algorithms for audio fingerprinting”, Journal of VLSI Signal Processing, pp.271–284, 2005.
[2] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” Proc. ISMIR, International Conference on Music Information Retrieval, pp.107–115, Oct.2002.
[3] A. C. Ibarrola and E. Chavez, "A robust entropy-based audio fingerprint," proceeding of the IEEE international conference on multimedia and expo, pp.1729-1732, 2006,.
[4] C. J. C.Burges, J.C.Platt and S.Jana, “Distortion discriminant analysis for audio fingerprinting,” IEEE Trans. Speech and Audio Processing, vol.11, pp. 165–174, Mar. 2003.
[5] LuCS, ”Audio fingerprinting based on analyzing time—frequency Localization of signals,” IEEE International Workshop on Multimedia Signal Processing, USA, pp.174-177,2002 .
[6] S. Sukittanon, L.E.Atlas and J.W.Pitton, “Modulation-scale analysis for content identification,” IEEE Trans. Signal Process, vol.52, no.10, pp.3023–3035, Oct.2004.
[7] Y.H.Jiao, B.Yang, M.Y.Li and X.M.Niu, "MDCT-based perceptual hashing for compressed audio content identification," proc. of the IEEE workshop on multimedia signal processing, pp.381-384, 2007.
[8] Wei Li, Yaduo Liu and Xiangyang Xue, "Robust audio identification for MP3 popular music,” ACM SIGIR 2010, pp.27-634. 2010
[9] Y.Wang and M.Vilermo,"A compressed domain beat detector using MP3 audio bit streams," proceeding of the ACM international conference on multimedia (ACM Multimedia 2001), pp.-202. 2001
[10] Jarina R, OConnor N, Marlow S, Murphy N, "Rhythm Detection for Speech-Music Discrimination in MPEG Compressed Domain," the 14th Intl Conf on Digital Signal Processing,Greece,pp.1-3 July 2002.
[11] A.D'Aguanno and G.Vercellesim, "Tempo induction algorithm in MP3 compressed domain," proceeding of the ACM international conference on multimedia information retrieval, pp.153-158. 2007.
[12] Kim, H.G., Kim,J.Y.&Park,T.”Video bookmark based on soundtrack identification and two-stage search for interactive-television.” IEEE Transactions on Consumer Electronics, 53(4), pp.1712–1717. 2007
[13] S.Schimmel, “Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices,” Ph.D. dissertation, Univ. of Washington, Seattle, 2007.
[14] M. Vinton and L. Atlas, “A Scalable And Progressive Audio Codec,” in Proc. of ICASSP 2001, pp. 3277-3280, 2001.
[15] T Y Chang. ”Research and implementation of MP3 encoding algorithm,” Ph.D. dissertation, Taiwan: National Chiao Tung University, 2002.

TABLE V
.COMPARISON OF THE PROPOSED ALGORITHM AND THE ONE IN

Processing	Nothing in the world			Canon in D			Yellow			Ring a ling		
	our	II	III	our	II	III	our	II	III	our	II	III
MP3@64kbps	0.1111	0.1442	0.1809	0.0847	0.1698	0.2149	0.097	0.1698	0.2149	0.1614	0.079	0.1317
downsampling	0.0697	0.1364	0.1372	0.0626	0.134	0.1956	0.0697	0.134	0.1956	0.1526	0.0674	0.1204
Noise addition	0.0838	0.1364	0.138	0.1032	0.1499	0.2091	0.075	0.1499	0.2091	0.2028	0.0855	0.1349
Equalization	0.172	0.1455	0.1628	0.1005	0.1857	0.2102	0.1129	0.1857	0.2102	0.172	0.1062	0.1368
Amplitude compression	0.112	0.126	0.1577	0.0785	0.1578	0.1871	0.0908	0.1578	0.1871	0.1966	0.0816	0.1196
Echo addition	0.149	0.1909	0.1973	0.1667	0.2308	0.2336	0.1446	0.2308	0.2336	0.1993	0.158	0.1949
Volume change(-6.0206)	0.1093	0.1299	0.1426	0.075	0.1512	0.1824	0.0829	0.1512	0.1824	0.1658	0.0764	0.1161
Volume change(+3.5218)	0.0908	0.1299	0.1423	0.0776	0.1446	0.181	0.1005	0.1446	0.181	0.1711	0.0894	0.1151
Bandpass filtering	0.1146	0.3065	0.2763	0.0908	0.2321	0.2163	0.1323	0.2321	0.2163	0.2099	0.2396	0.2642
1s time misalignment	0.2522	0.5078	0.5615	0.2055	0.504	0.4901	0.2002	0.504	0.4901	0.2945	0.4521	0.4841