Generalized Time-Series Active Search With Kullback–Leibler Distance for Audio Fingerprinting

Hui Lin, Zhijian Ou, *Member, IEEE*, and Xi Xiao

***Abstract*—In this letter, a new audio fingerprinting approach is presented. We investigate to improve robustness by more precise statistical fingerprint modeling with common component Gaussian mixture models (CCGMMs) and Kullback–Leibler (KL) distance, which is more suitable to measure the dissimilarity between two probabilistic models. To address the resulting complexity, generalized time-series active search is proposed, which supports a wide variety of distance measures between two CCGMMs, including L_1, L_2, KL, etc. Experiments show that the new approach with KL distance increases robustness to distortions (including low-quality MP3 compression, small room echo, and play-and-record) while achieving efficient search.**

***Index Terms*—Audio fingerprinting, audio search.**

I. Introduction

AUDIO fingerprinting includes a wide variety of applications (such as identification, verification, and retrieval of audio materials) and has received a lot of attention recently. Various audio fingerprinting algorithms have been proposed [1]–[3]. They differ mainly in the type of acoustic features, the fingerprint modeling, the distance measure for comparing fingerprints, and the search method for efficient matching. An ideal fingerprinting system should be able to identify different versions of the same audio content consistently, regardless of the distortions due to compression, transmission, and so on. It should also be computationally efficient.

In this letter, the task of finding given audio clips in an audio stream where the stream may be corrupted by distortions is used as a testbed (termed audio search) for studying these issues. Typical applications include analysis of broadcast music/commercials, copyright management over the Internet, or finding metadata for unlabeled audio.

While previous studies explored various features that were robust to distortions [2], [3], we focus on improving robustness by more precise statistical fingerprint modeling and better distance measure. In our approach, rather than roughly summarized as a vector or a histogram via hard quantization [1], the acoustic feature vectors from an audio segment are modeled by a Gaussian mixture model (GMM). We further assume that the GMMs for all the audio segments share a common set of Gaussian components, thus becoming common component GMMs (CCGMMs), which was introduced in [4] (also known as tied-mixture [5]). Since only the weights of the components are to be estimated when using CCGMM fingerprint modeling, poor estimation due to insufficient data is avoided. This is important for precise modeling of short audio segments (e.g., 2–3 s). Moreover, the use of CCGMMs simplifies the Kullback–Leibler (KL) distance calculation between two GMMs [4] and, more importantly, makes it possible for us to derive generalized time-series active search.

The active search proposed in [6] worked with L_1 distance between histograms. In our approach, generalized time-series active search is proposed, which supports a wide variety of distance measures between CCGMMs, including L_1, L_2, KL, etc. The idea is to skip unnecessary hypothesis matches while mathematically guaranteeing no false dismissals.

The main contribution of this letter is to show how we could increase robustness to distortions while achieving efficient search, which benefits from the joint use of refined statistical modeling (CCGMMs), KL distance measure, and generalized active search. The theory and experimental evaluations of the proposed approach are presented in Sections II and III, respectively, and the conclusions are given in Section IV.

II. Theory

Fig. 1 outlines the audio search system. First, the feature vector sequences are extracted from both the audio clip and the audio stream. Then, windows of the same length as the clip are applied to both feature sequences. The feature vectors within the window are used to estimate a CCGMM model. Also, the distance is calculated between the two CCGMMs estimated, respectively, from the clip and the windowed stream. If the distance is below a threshold, the audio clip is considered to be detected in the audio stream. Finally, the window on the audio stream is shifted forward in time, and the search proceeds.

A. Fingerprint Modeling: Common Component GMMs

In our approach, any audio segment is modeled by a GMM, as its compact fingerprint. Specifically, the feature vectors x extracted from an audio segment k are represented by a GMM, which consists of M Gaussian components with

Manuscript received August 16, 2005; revised February 21, 2006. This work was supported by the China Ministry of Information Industry. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm D. Macleod.

The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: linhui99@mails.tsinghua.edu.cn; ozj@tsinghua.edu.cn).

Digital Object Identifier 10.1109/LSP.2006.874394

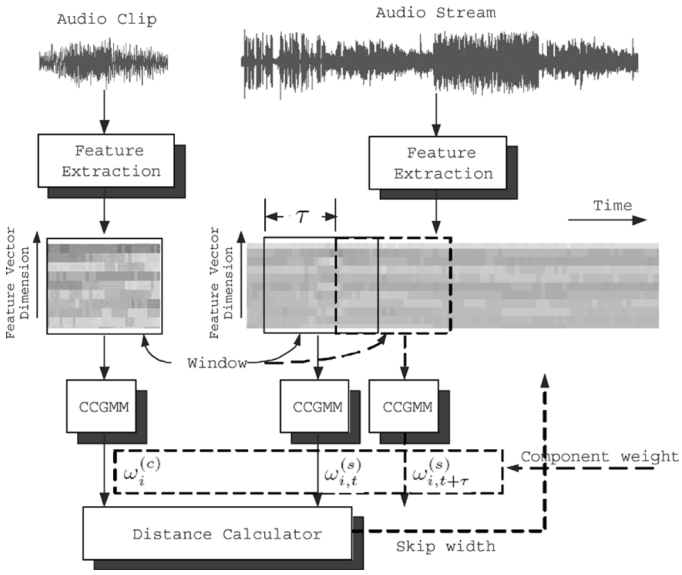

Fig. 1. Overview of the audio search system.

component weights $\omega_i^{(k)}$, means $\mu_i^{(k)}$, and covariances $\Sigma_i^{(k)}$, $i = 1, 2, \ldots, M$

$$p^{(k)}(x) = \sum_{i=1}^{M} \omega_i^{(k)} \mathcal{N}\left(x|\mu_i^{(k)}, \Sigma_i^{(k)}\right). \tag{1}$$

To avoid poor estimation and simplify KL distance calculation, we further assume that the GMMs for all the audio segments share a common set of Gaussian components. It is pre-trained and stored as a universal GMM: $p^{(u)}(x) = \sum_{i=1}^{M} \omega_i^{(u)} \mathcal{N}(x|\mu_i^{(u)}, \Sigma_i^{(u)})$, Thus, (1) becomes (with $\mu_i^{(k)}$, $\Sigma_i^{(k)}$ being constrained to $\mu_i^{(u)}$, $\Sigma_i^{(u)}$, respectively)

$$p^{(k)}(x) = \sum_{i=1}^{M} \omega_i^{(k)} \mathcal{N}\left(x|\mu_i^{(u)}, \Sigma_i^{(u)}\right). \tag{2}$$

Now we only need to estimate the component weights to specify a CCGMM for an audio segment. Given T feature vectors $x_t (t = 1, 2, \ldots, T)$ for an audio segment k, the weights are estimated by the following formula:

$$\omega_i^{(k)} = \frac{1}{T} \sum_{t=1}^{T} \frac{\omega_i^{(u)} \mathcal{N}\left(x_t|\mu_i^{(u)}, \Sigma_i^{(u)}\right)}{\sum_{j=1}^{M} \omega_j^{(u)} \mathcal{N}\left(x_t|\mu_j^{(u)}, \Sigma_j^{(u)}\right)}. \tag{3}$$

B. Generalized Time-Series Active Search

As the window applied over the audio stream shifts forward in time, the distances show a certain continuity from one time-step to the next. Time-series active search [6] takes advantage of this continuity by computing a lower bound of the distance measure as a function of the time-step and skipping all intermediate distance calculations until this lower bound is below the detection threshold. Here we generalize and prove the active search rigorously to various distance measures for CCGMMs.

Denote $\omega_{i,t}^{(s)}$, $\omega_{i,t+\tau}^{(s)}$, $i = 1, 2, \ldots, M$, as the weights estimated for the windowed stream at time-step t and $t+\tau$, respectively. Here we use the superscript s to explicitly express that the weights are for the windowed stream. Denote T as the window length. Then we have

$$\sum_{i=1}^{M} \left|\omega_{i,t+\tau}^{(s)} - \omega_{i,t}^{(s)}\right| \le \frac{2\tau}{T}. \tag{4}$$

Given any convex function $d_i(\cdot)$, $i = 1, 2, \ldots, M$, we have

$$d_i\left(\omega_{i,t+\tau}^{(s)}\right) \ge d_i\left(\omega_{i,t}^{(s)}\right) - \left|\omega_{i,t+\tau}^{(s)} - \omega_{i,t}^{(s)}\right| \left|d_i'\left(\omega_{i,t}^{(s)}\right)\right| \tag{5}$$

where $d_i'(\cdot)$ denotes the first derivative of $d_i(\cdot)$.

Denote $\omega_i^{(c)}$, $i = 1, 2, \ldots, M$ as the weights estimated for the audio clip, using superscript c. Fix $\omega_i^{(c)}$ and view the distance measures between $\omega_i^{(c)}$ and $\omega_{i,t}^{(s)}$ as a function of $\omega_{i,t}^{(s)}$. Then for any distance evaluated at time-step t, which has the form

$$d(t) = \sum_{i=1}^{M} d_i\left(\omega_{i,t}^{(s)}\right) \tag{6}$$

combining (4)–(6), we have

$$d(t+\tau) \ge d(t) - \frac{2\tau}{T} \max_i \left|d_i'\left(\omega_{i,t}^{(s)}\right)\right|. \tag{7}$$

If the distance at time-step $t+\tau$ is to drop below a given threshold θ, that is $d(t+\tau) \le \theta$, then we must have

$$\tau \ge \frac{T\left(d(t) - \theta\right)}{2 \cdot \max_i \left|d_i'\left(\omega_{i,t}^{(s)}\right)\right|} \stackrel{\text{def}}{=} \tau_{min}. \tag{8}$$

So the skip width for the window can be derived as

$$\tau_{skip} = \begin{cases} \lfloor \tau_{min} \rfloor + 1, & \tau_{min} > 0 \\ 1, & \text{otherwise} \end{cases} \tag{9}$$

where $\lfloor x \rfloor$ denotes the greatest integer less than x. This tells us that it is guaranteed that we will not miss any section that will give distance less than the threshold θ, even if we skip the width τ_{skip} given by (9). Thus, we accelerate the search by avoiding unnecessary computations. The larger τ_{min} is, the more will the search be speeded up.

This generalized active search is suitable for any distance with the form of (6). For KL distance between CCGMMs [4]

$$d_{KL}(t) = \sum_{i=1}^{M} \left(\omega_{i,t}^{(s)} - \omega_i^{(c)}\right) log \frac{\omega_{i,t}^{(s)}}{\omega_i^{(c)}} \tag{10}$$

we have

$$\tau_{min-KL} = \frac{T\left(d_{KL}(t) - \theta\right)}{2 \cdot \max_i \left|1 + log \frac{\omega_{i,t}^{(s)}}{\omega_i^{(c)}} - \frac{\omega_i^{(c)}}{\omega_{i,t}^{(s)}}\right|}. \tag{11}$$

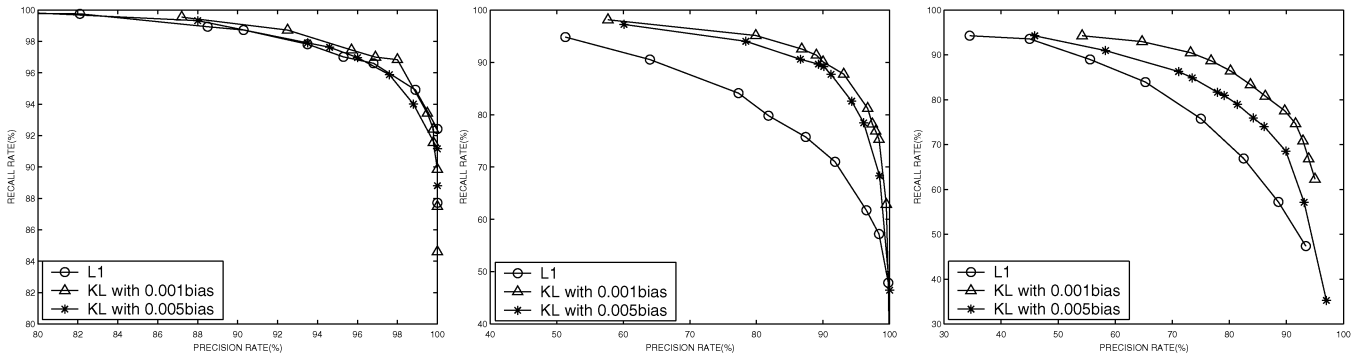

Fig. 2. Recall and precision rate curves for L_1 distance and KL distance with 0.001/0.005 bias under three distortions. (a) Low-quality MP3 compression. (b) Small room echo. (c) Play-and-record.

For L_1 distance, which simply treats the component weights as vectors in M dimension space

$$d_{L_1}(t) = \sum_{i=1}^{M} \left| \omega_{i,t}^{(s)} - \omega_i^{(c)} \right| \tag{12}$$

we have

$$\tau_{min-L_1} = \frac{T\left(d_{L_1}(t) - \theta\right)}{2}. \tag{13}$$

This reduces to the algorithm in [6].

Note that there is some problem with directly applying (11), since if any mixture weight $\omega_{i,t}^{(s)}$ is near 0, the denominator of (11) will become so large that the resulting skip width is often not greater than 1. To solve this problem, the component weights estimated for the clip and the windowed stream are both added with a positive bias Δ. Under such adjustment, (4) remains valid, and so does the generalized active search. Also, KL distance is modified as

$$d_{KL}(t) = \sum_{i=1}^{M} \left(\omega_{i,t}^{(s)} - \omega_i^{(c)} \right) log \frac{\omega_{i,t}^{(s)} + \Delta}{\omega_i^{(c)} + \Delta}. \tag{14}$$

Experiments below show that this modification slightly reduces the accuracy but greatly increases the efficiency.

III. Experiments

A. Experimental Setup

In the experiments, audio signals were first resampled at 8 KHz and then framed with 25 ms length and 10 ms shift. Twelve mel-frequency cepstral coefficients (MFCCs) were extracted from each frame as feature vectors. The 1997 Mandarin Broadcast News corpus (Hub4-NE), which consists of 30 h of recorded broadcasts from different sources such as CCTV and VOA, was used to train the universal GMM with 128 diagonal Gaussian components via EM algorithm by HTK [5]. We recorded another 10 h of broadcast radio from VOA as the test audio stream and took 1000 3-s audio segments from this audio recording as the test audio clips. Three different distortions were applied to the test stream: a) low-quality VBR MP3 compression (40–50 Kbps); b) small room echo added using CoolEdit software; and c) played back through low-quality speaker and recorded with low-quality microphone. The task was to find the selected audio clips in these distorted 10-h audio streams. These experiments were challenging since the selected clips were short, with similar styles, and even from the same speakers. The task is different from those in [3] and [7], where the fingerprints are extracted and compared from a known location. Moreover, here we make continuous decisions to identify every short segment, rather than to identify a song [2]. The false positives were more strictly counted in our experiments.

B. Results

Fig. 2 shows the recall and precision rate curves for L_1 distance and KL distance with biases. Although KL distance only slightly outperforms L_1 distance under low-quality MP3 compression, it achieves significant advantages under the other two distortions. Table I shows the accuracy and efficiency for different schemes. Here, the accuracy value is the precision rate (or the recall rate) when the precision rate equals the recall rate; the efficiency value is the percentage of the number of matching calculations that the active search skips compared with exhaustive search. The CPU time indicates the average time of searching a 3-s clip in the 10-h stream and was measured on a Pentium4 1.4-GHz PC. On average over the three distortions, using KL distance with 0.005 bias could skip 93.47% of the exhaustive matches, take 3.04 s to search a 3-s clip in the 10-h stream, and reduce the error rate by 31.7% compared with using L_1 distance. By adjusting the bias for KL distance, we can achieve a trade-off between accuracy and efficiency.

C. Discussions

An important application of audio fingerprinting is to identify unknown audio, based on a reference database that contains a set of precomputed fingerprints (e.g., extracted from original songs) [1]. The fingerprints of the input audio are generated at

TABLE I
ACCURACY AND EFFICIENCY

Distortions	Distance	Bias	Acc.(%)	Eff.(%)	CPU time(s)
MP3 compression	L_1	–	95.7	98.93	0.82
	KL	0.001	97.0	88.45	5.32
		0.005	96.6	93.95	2.96
small room echo	L_1	–	81.2	99.32	0.63
	KL	0.001	90.6	86.06	6.26
		0.005	90.0	94.39	2.12
play-and-record	L_1	–	74.8	98.53	1.16
	KL	0.001	83.5	85.23	7.01
		0.005	80.5	92.07	4.03
AVERAGE	L_1	–	83.9	98.93	0.87
	KL	0.001	90.4	86.58	6.20
		0.005	89.0	93.47	3.04

repeated intervals (e.g., every 186 ms [2]). There are two different approaches to compare the serial input fingerprints with the reference fingerprints in the database. One approach is that the input fingerprints are separately compared with the database to find a match. To overcome exhaustive scan, the use of a simple criterion to quickly prune the search space is explored [7]. There are also methods that precompute some statistics (e.g., overlap test results [8]) and build an index. So the comparisons, while still exhaustive, can be more efficiently executed (e.g., bitwise [8]). Another approach, which is presented here, employs the time-series active search. The reference fingerprints in the database are separately compared against the input audio stream. The number of comparisons can be significantly reduced by skipping, and it is guaranteed that nothing is missed. While the indexing method makes use of the redundancy of the distances between an input fingerprint and the stored reference fingerprints to achieve efficient search [8], the active search makes use of the continuity of the distances between a reference fingerprint and the serial input fingerprints. The efficiency values for both methods are comparable (approximately 98%, or speedup over exhaustive scan by a factor of 50 [6], [8]). The active search method could be further enhanced by parallel searching given multiple reference fingerprints at the same time [6] or combining the indexing method to precompute distances.

IV. CONCLUSIONS

In this letter, a new audio search approach is proposed. The main feature is its joint use of refined statistical modeling (CCGMMs), KL distance measure, and generalized active search. Experiments show its advantages with increased robustness to distortions while achieving efficient search.

REFERENCES

[1] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," in *Proc. Int. Workshop Multimedia Signal Processing*, 2002.

[2] C. J. C. Burges, J. C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 165–174, May 2003.

[3] S. Sukittanon, L. E. Atlas, and J. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3023–3035, Oct. 2004.

[4] Y. Wang and C. Huang, "Speaker-and-environment change detection in broadcast news using the common component GMM-based divergence measure," in *Proc. INTERSPEECH*, 2004, pp. 1069–1072.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The htk Book Version 3.0*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[6] K. Kashino, T. Kurozumi, and H. Murase, "A quick search method for audio and video signals based on histogram pruning," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 348–357, Sep. 2003.

[7] V. Venkatachalam, L. Cazzanti, N. Dhillon, and M. Wells, "Automatic identification of sound recordings," *IEEE Signal Process. Mag.*, vol. 21, no. 2, pp. 92–99, Mar. 2004.

[8] J. Goldstein, J. C. Platt, and C. J. C. Burges, "Redundant bit vectors for quickly searching high-dimensional regions," in Deterministic and Statistical Methods in Machine Learning, J. Winkler, M. Niranjan, and N. Lawrence, Eds., Springer Lecture Notes on Computer Science, vol. 3635, pp. 137–158, 2005.