A Review of Audio Features and Statistical Models Exploited for Voice Pattern Design

Ngoc Q. K. Duong, Hien-Thanh Duong

A Review of Audio Features and Statistical Models Exploited for Voice Pattern Design

Ngoc Q. K. Duong

Technicolor
975 avenue des Champs Blancs
35576 Cesson Sévigné, France
Email: quang-khanh-ngoc.duong@technicolor.com

Hien-Thanh Duong

Faculty of Information Technology
Hanoi University of Mining and Geology
Hanoi city, Vietnam
Email: duongthihienthanh@humg.edu.vn

***Abstract*—Audio fingerprinting, also named as audio hashing, has been well-known as a powerful technique to perform audio identification and synchronization. It basically involves two major steps: fingerprint (voice pattern) design and matching search. While the first step concerns the derivation of a robust and compact audio signature, the second step usually requires knowledge about database and quick-search algorithms. Though this technique offers a wide range of real-world applications, to the best of the authors' knowledge, a comprehensive survey of existing algorithms appeared more than eight years ago. Thus, in this paper, we present a more up-to-date review and, for emphasizing on the audio signal processing aspect, we focus our state-of-the-art survey on the fingerprint design step for which various audio features and their tractable statistical models are discussed.**

Keywords–Voice pattern; audio identification and synchronization; spectral features; statistical models.

I. INTRODUCTION

Real-time user interactive applications have emerged nowadays thanks to the increased power of mobile devices and their Internet access speed. Let us consider applications like music recognition [1][2], e.g., people hear a song in a public place and they want to know more about it, or personalized TV entertainment [3][4], e.g., people want to see more service and related content on the Web in addition to the main view from TV; both require a fast and reliable audio identification system in order to match the observed audio signal with its origin stored in a large database. For these purposes, several research directions have been studied, such as audio fingerprinting [5], audio watermarking [6], and timeline insertion [4]. While watermarking and timeline approaches both require to embed signature into the original media content, which is sometimes inconvenient for the considered applications, fingerprinting technique allows directly monitoring the data for identification. Hence, audio fingerprinting has been widely investigated in the literature and already been deployed in many commercialized products [1][7][8][9][10][11]. This technique has recently been exploited for other applications such as media content synchronization [12][13], multiple video clustering [14], repeating object detection [15], and live version identification [15].

A general architecture for an audio fingerprinting system, which can be used for either audio identification or audio synchronization purpose, is depicted in Fig. 1. The fingerprint extraction derives a set of relevant audio features followed by an optional post-processing and feature modeling. Fingerprints of the original audio collection and its corresponding metadata (e.g., audio ID, name, time frame index, etc.) are systematically stored in a database. Then given a short recording from the user side, its feature vectors (i.e fingerprints) are computed in the same way as they were for the original data. Finally, a searching algorithm will find the best match between these fingerprints with those stored in the database so that the recorded audio signal is labeled by the matched metadata.

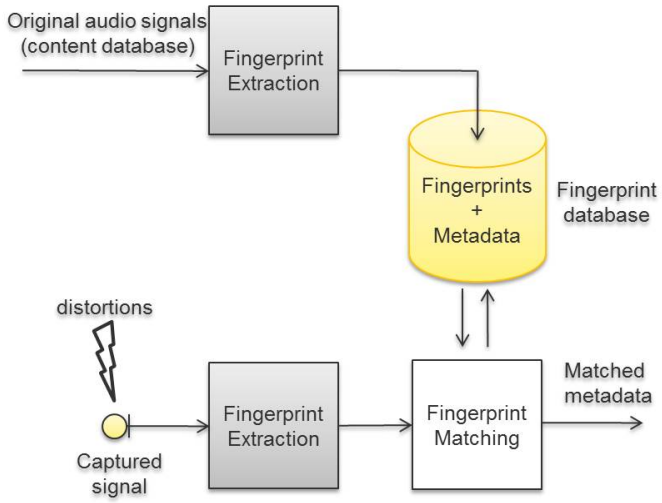

Figure 1: General architecture of an audio fingerprinting system.

In real-world recording, the audio signal often undergoes many kinds of distortion: acoustical reverberation, background noise addition, quantization error, etc. Thus, the derived fingerprints must be robust with respect to these various signal degradations. Beside, the fingerprint size should be as small as possible to save memory resources and to allow real-time matching. The details of general properties of the audio fingerprint was well-discussed in [1][16][17]. In order to fulfil those requirements, audio sample signal is often transformed into Time-Frequency (T-F) domain via the Short Time Fourier transform (STFT) [16] where numerous distinguishable characteristics such as high-level musical attributes, e.g., predominant pitch, harmony structure, or low level spectral features, e.g., mel-frequency cepstrum, spectral centroids, spectral note onsets, etc., are exploited. To further compact the fingerprints, some approaches continue to fit the spectral feature vectors to a statistical model, e.g., Gaussian Mixture Model (GMM) [18], Hidden Markov Model (HMM) [19], so that in the end only the set of model parameters are used as fingerprints.

Though diverse fingerprinting algorithms have been proposed in the literature, the number of review papers remains limited where, to the best of the authors' knowledge, a comprehensive review of fingerprinting algorithms was presented more than eight years ago [5][16], and a more recent survey [20] only focusing on computer vision based approaches (e.g., methods proposed in [21][22]). In this paper, we present a more up-to-date review of the domain, with particular focus concerning the fingerprint extraction block in Fig. 1, where various audio spectral features and their statistical models are summarized systematically. The presentation would particularly benefit new researchers in the domain and engineers in the sense that they would easily follow the described steps to implement different audio fingerprints.

The structure of the rest of the paper is as follows. We first present a general architecture for fingerprint design in Section II, we then review various audio features, which have been extensively exploited in the literature, in Section III. The detail of some statistical feature models is introduced in Section IV. Finally, we conclude in Section V.

II. General architecture of fingerprint design

Fig. 2 depicts a general workflow of the fingerprint design. The purpose of each block is summarized as follows:

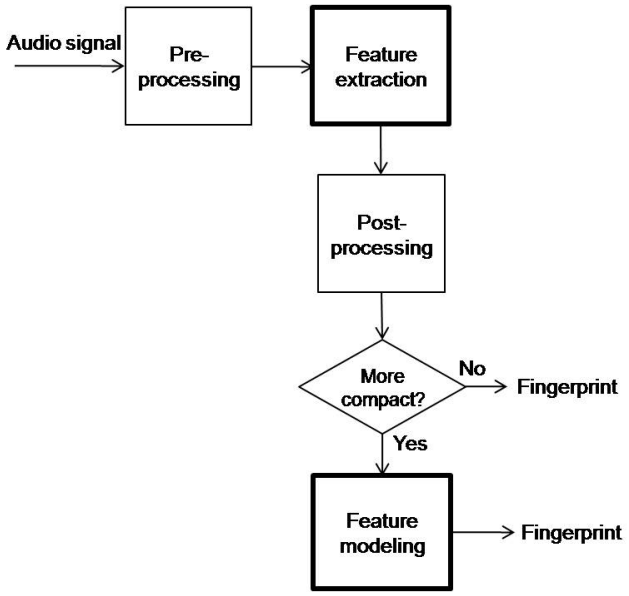

Figure 2: General workflow of the fingerprint design.

- *Pre-processing*: in this step, input audio signal is often first digitalized (if necessary), re-sampled to a target sampling rate, and bandpass filtered. Other types of processing includes decorrelation and amplitude normalization [16]. Then the processed signal is segmented into overlapping time frames where a linear transformation, e.g., Fast Fourier Transform (FFT), Discrete Cosine transform (DCT), or wavelet transform [16], is applied to each frame. At this stage, the input time-domain signal is represented in a feature domain, and the most popular feature domain is time-frequency representation given by the STFT.
- *Feature extraction*: this is a major process since the choice of "which feature is used" will directly affect the performance of the entire fingerprinting system. A great diversity of features have been investigated targeting the reduction of dimensionality as well as the invariance to various distortions. For summary, most approaches first map the linear time-frequency representation given by the STFT to an auditory-motivated frequency scale, i.e., Mel, Bark, Log, or Cent scale, via filterbanks [2][23]. This mapping step greatly reduces the spectrogram size since the number of filterbanks is usually much smaller than the FFT length. Then a feature vector such as Mel-Frequency Cepstral Coefficients (MFCC), spectral centroids of all subbands, etc., are computed for each time frame. In some systems, the first and second derivatives of the feature vectors are also integrated to better track the temporal variation of audio signals [16][18]. Other types of feature that worth mentioning are e.g., time localized frequency peak [24], time-frequency energy peak location [1], or those developed in image processing based approaches such as top-wavelet coefficients computed on the spectral image [22] and multiscale Gabor atoms extracted by Matching Pursuit algorithm [14]. Recently, a general framework for dictionary based feature learning has also been introduced [25].
- *Post-processing*: the feature vectors computed in the previous step are often real-valued and the absolute range depends on the signal power. Therefore when Euclidean distance is used in the matching step, mean substraction and component wise variance normalization are recommended [26][27]. Another popular post-processing is quantization where each entry of the feature vectors is quantized to a binary number in order to gain robustness against distortions and, more importantly, to obtain memory efficiency [2][28][22] [15]. In many existing system, fingerprint is achieved after this step.
- *Feature modeling*: this block is sometimes deployed in order to further compact the fingerprint. In this case, a large number of feature vectors along time frames is fitted to a statistical model so that an input audio signal is well-characterized by the model parameters, which are then stored as a fingerprint [29][18][30]. Popular model includes Gaussian Mixture Model (GMM), Hidden Markov Model (HMM). Other approaches used decomposition techniques, e.g., Non-negative MAtrix Factorization (NMF), to help decreasing data dimension and therefore to reduce the local statistical redundancy of the feature vectors [31][32].

Since the pre-processing and post-processing steps are quite straightforward, in the following of the paper we will present more detail only on the feature extraction and the feature modeling blocks.

III. Feature extraction

Summarizing numerous types of audio features used for the fingerprint design so far will certainly go beyond the scope of

this paper. Thus in this section, we select to present the most popular low level features in the spectral domain only.

A. MFCC

MFCC is one of the most popular feature considered in speech recognition where the amplitude spectrum of input audio signal is first weighted by triangular filters spaced according to the Mel scale, and DCT is then applied to decorrelate the Mel-spectral vectors. MFCC was shown to be applicable for music signal also in [33]. Examples of fingerprinting algorithms used MFCC feature are found in [33][18]. In [34], MFCC was used also for clustering and synchronizing large scale audio-video sequences recorded by multiple users during an event. Matlab implementations for the computation of MFCC are available [35][36].

B. Spectral Energy Peak (SEP)

SEP for music identification systems was described in [37][1] where a time-frequency point is considered as a peak if it has higher amplitude than its neighboring points. SEP is argued to be intrinsically robust to even high level background noise and can provide discrimination in sound mixtures [38]. In well-known Shazam's system [1] time-frequency coordinates of the energy peaks was described as sparse landmark points. Then by using pairs of landmark points rather than single points, the fingerprints exploited the spectral structure of sound sources. This landmark feature can also be found in [14] and [39] for multiple video clustering. Ramona *et al.* used start times of the spectral energy peaks, referred to as onsets, for the automatic alignment of audio occurrences in their fingerprinting system [23][40].

C. Spectral Band Energy (SBE)

Together with spectral peak, SBE has been widely exploited in fingerprinting algorithms. Let us denote by $s(n, f)$ a STFT coefficient of an audio signal at time frame index n and frequency bin index f, $1 \leq f \leq M$. Let us also denote by b an auditory-motivated subband index, i.e., in either Mel, Bark, Log, or Cent scale, and l_b and h_b the lower and upper edges of b-th subband. SBE is then computed, with normalization, in each time frame and each frequency subband range by

$$F_{n,b}^{\text{SBE}} = \frac{\sum_{f=l_b}^{h_b} |s(n,f)|^2}{\sum_{f=1}^{M} |s(n,f)|^2}. \quad (1)$$

Haitsma et al. proposed a famous fingerprint in [2] where SBEs were first computed in a block containing 257 time frames and 33 Bark-scale frequency subbands, then each $F_{n,b}^{\text{SBE}}$ was quantized to a binary value (either 0 or 1) based on its differences compared to neighboring points. Other fingerprinting algorithms exploiting SBE feature were found for instance in [41][18]. Variances of this subband energy difference feature can be found in more recent approaches [28][21].

D. Spectral Flatness Measure (SFM)

SFM, also known as Wiener entropy, relates to the tonality aspect of audio signals and it is therefore often used to distinguish different recordings. SFM is computed in each time-frequency subband point (n, b) as

$$F_{n,b}^{\text{SFM}} = \frac{\left(\prod_{f=l_b}^{h_b} |s(n,f)|^2\right)^{\frac{1}{h_b - l_b + 1}}}{\frac{1}{h_b - l_b + 1} \sum_{f=l_b}^{h_b} |s(n,f)|^2}. \quad (2)$$

A high SFM indicates the similarity of signal power over all frequencies while a low SFM means that signal power is concentrated in a relatively small number of frequencies over the full subband.

A similarly feature to SFM, which is also a measure of the tonal-like or noise-like characteristic of audio signal and was exploited as fingerprint, is spectral crest factor (SCF). SCF is computed by

$$F_{n,b}^{\text{SCF}} = \frac{\max_{f \in [l_b, h_b]} \left(|s(n,f)|^2\right)}{\frac{1}{h_b - l_b + 1} \sum_{f=l_b}^{h_b} |s(n,f)|^2}. \quad (3)$$

SFM and SCF were found to be the most promising features for audio matching with common distortions in [42] and were both considered in other fingerprinting algorithms [41][18].

E. Spectral Centroid (SC)

SC is also a popular measure used in audio signal processing to indicate where the "center of mass" of a subband spectrum is. It is formulated as

$$F_{n,b}^{\text{SC}} = \frac{\sum_{f=l_b}^{h_b} f.|s(n,f)|^2}{\sum_{f=l_b}^{h_b} |s(n,f)|^2}. \quad (4)$$

SC was argued to be robust over equalization, compression, and noise addition. It was reported in [26] and [18] that SC-based fingerprints offered better audio recognition than MFCC-based fingerprints with 3 to 4 second length audio clips. In our preliminary experiment with speech utterances distorted by reverberation and real-world background noise, we also observed that SC-based fingerprints resulted in higher recognition accuracy than MFCC-, SBR-, and SFM-based fingerprints without post-processing.

Given one of the feature parameters $F_{n,b}$ computed in each time-frequency subband point (n, b) as described above, a d-dimensional feature vector $\mathbf{F}_n = [F_{n,1}, ..., F_{n,d}]^T$ is formed to describe the corresponding characteristic of the signal at time frame n, where T denotes vector transpose and d is the total number of subbands. When the first and second derivatives of the feature vectors are additionally considered, for better characterizing the temporal variation of audio signal, $\mathbf{F}_n$ will then be a $3d$-dimensional vector [18] before passing to the post-processing block shown in Figure 2.

IV. Feature Modeling

In some systems, in order to further compact the fingerprint the feature vectors $\mathbf{F}_n$ can be adapted to a statistical model. This step allows to reduce the global redundancy of spectral features. As a result, a long sequence of feature vectors $\mathbf{F}_n, n = 1, ..., N$ is characterized by a significantly smaller number of the model parameters while ensuring the discriminative power. In this section we review the use of three popular models, namely gaussian mixture model (GMM), hidden Markov model (HMM), and nonnegative matrix factorization (NMF), for the fingerprint design.

A. GMM-based fingerprint

GMM has been used to model the spectral shape of audio signals in many different applications ranging from speaker identification [43] to speech enhancement [30], etc. It was also investigated for audio fingerprinting by Ramalingam and Krishnan [18], where spectral feature vectors $\mathbf{F}_n$ are modeled as a multidimensional K-state Gaussian mixture with probability density function (pdf) given by

$$p(\mathbf{F}_n) = \sum_{k=1}^{K} \alpha_k \mathcal{N}_c(\mathbf{F}_n|\mu_k, \mathbf{\Sigma}_k) \tag{5}$$

where α_k, which satisfies $\sum_{k=1}^{K} \alpha_k = 1$, μ_k and $\mathbf{\Sigma}_k$ are the weight, the mean vector and the covariance matrix of the k-th state, respectively, and

$$\mathcal{N}_c(\mathbf{F}_n|\mu_k, \mathbf{\Sigma}_k) = \frac{1}{|\pi\mathbf{\Sigma}_k|} e^{-(\mathbf{F}_n-\mu_k)^H \mathbf{\Sigma}_k^{-1}(\mathbf{F}_n-\mu_k)} \tag{6}$$

where ^{H} and $|.|$ denote conjugate transpose and determinant of a matrix, respectively. The model parameters $\theta = \{\alpha_k, \mu_k, \mathbf{\Sigma}_k\}_k$ are then estimated in the maximum likelihood (ML) sense via the expectation-maximization (EM) algorithm, which is well-known as an appropriate choice in this case, with the global log-likelihood defined as

$$\mathcal{L}_{ML} = \sum_{n=1}^{N} \log p(\mathbf{F}_n|\theta). \tag{7}$$

As a result, the parameters are iteratively updated via two EM steps as follow:

- E-step: compute the posterior probability that feature vector $\mathbf{F}_n$ is generated from the k-th GMM state

$$\gamma_{nk} = \frac{\alpha_k p(\mathbf{F}_n|\mu_k, \mathbf{\Sigma}_k)}{\sum_{l=1}^{K} \alpha_l p(\mathbf{F}_n|\mu_l, \mathbf{\Sigma}_l)}. \tag{8}$$

- M-step: update the parameters

$$\alpha_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk} \tag{9}$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{F}_n}{\sum_{n=1}^{N} \gamma_{nk}} \tag{10}$$

$$\mathbf{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} (\mathbf{F}_n - \mu_k)(\mathbf{F}_n - \mu_k)^H}{\sum_{n=1}^{N} \gamma_{nk}}. \tag{11}$$

With GMM, N d-dimensional feature vectors $\mathbf{F}_n$ are characterized by K set of GMM parameters $\{\alpha_k, \mu_k, \mathbf{\Sigma}_k\}_{k=1,...,K}$ where K is often very small compared to N. However, since GMM does not explicitly model the amplitude variation of sound sources, signals with different amplitude level but similar spectral shape may result in different estimated mean and covariance templates. *To overcome this issue, another version of GMM called spectral Gaussian scaled mixture model (GSMM) could be considered instead. Though GSMM has been used in speech enhancement [30] and audio source separation [44], it has yet been applied in the context of fingerprinting.*

B. HMM-based fingerprint

HMM is a well-known model in many audio processing applications [45]. When applied for audio fingerprinting, pdf of the observed feature vector $\mathbf{F}_n$ can be written as

$$p(\mathbf{F}_n) = \sum_{q_1,q_2,...,q_d} \pi_{q_1} b_{q_1}(F_{n,1}) a_{q_1 q_2} b_{q_2}(F_{n,2}) \\ ...a_{q_{d-1}q_d} b_{q_d}(F_{n,d}) \tag{12}$$

where π_{q_i} denotes the probability that q_i is the initial state, $a_{q_i q_j}$ is state transition probability, and $b_{q_i}(F_{n,i})$ is pdf for a given state.

Given a sequence of observations $\mathbf{F}_n, n = 1, ..., N$ extracted from a labeled audio signal, the model parameters $\theta = \{\pi_{q_i}, a_{q_i q_j}, b_{q_i}\}_{i,j}$ are learned via e.g., EM algorithm (detail formulation can be found in [45]) and stored as a fingerprint. Cano et al. modeled MFCC feature vectors by HMM in their AudioDNA fingerprint system [29]. In [19] HMM-based fingerprint was shown to achieve a high compaction by exploiting structural redundancies on music and to be robust to distortions.

Note that when applying GMM or HMM for the fingerprint design, a captured signal at the user side is considered to be matched with an original signal fingerprinted by the model parameter θ in the database if its corresponding feature vectors $\widehat{\mathbf{F}}_n$ are most likely generated by θ.

C. NMF-based fingerprint

NMF is well-known as an efficient decomposition technique which helps reducing data dimension [46]. It has been widely considered in audio and music processing, especially for audio source separation [47][48]. When applying in the context of audio fingerprinting, a $d \times N$ matrix of the feature vectors $\mathbf{V} = [\mathbf{F}_1, ..., \mathbf{F}_N]$ is approximated by

$$\mathbf{V} = \mathbf{WH} \tag{13}$$

where $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices of size $d \times Q$ and $Q \times N$, respectively, modeling the spectral characteristics of the signal and its temporal activation, and Q is much smaller than N. The model parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ can be estimated by minimizing the following cost function:

$$C(\boldsymbol{\theta}) = \sum_{bn} d_{IS}\left([\mathbf{V}]_{b,n}|[\mathbf{WH}]_{b,n}\right), \tag{14}$$

where $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$ is Itakura-Saito (IS) divergence, and $[\mathbf{A}]_{b,n}$ denotes an entry of matrix $\mathbf{A}$ at b-th row and n-th column. The resulting multiplicative update (MU) rules for parameter estimation write [49]:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}\right)}{\mathbf{W}^T (\mathbf{WH})^{\cdot -1}} \tag{15}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}\right) \mathbf{H}^T}{(\mathbf{WH})^{\cdot -1} \mathbf{H}^T} \tag{16}$$

where $\odot$ denotes the Hadamard entrywise product, $\mathbf{A}^{\cdot p}$ being the matrix with entries $[\mathbf{A}]_{ij}^p$, and the division is entrywise. Fingerprints are then generated compactly from the basis matrix $\mathbf{W}$, which has much smaller size compared to the feature matrix $\mathbf{V}$.

NMF was applied to the spectral subband energy matrix in [32] and to the MFCC matrix in [50]. The resulting fingerprint was shown to better identify audio clips than another decomposition technique namely singular value decomposition (SVD).

V. Conclusion

In this paper, we presented a review of the existing audio fingerprinting systems which have been developed by numerous researchers during the last decade for a range of practical applications. We described a variety of audio features and reviewed state-of-the-art approaches exploiting them for the fingerprint design. Furthermore, the use of statistical models and decomposition techniques to reduce the global statistical redundancy of feature vectors, and therefore to decrease fingerprint size, was also summarized. As a result, the combination of different presenting features and/or the deployment of a statistical feature model afterward are both applicable to obtain a robust and compact audio signature.

References

[1] A. L.-C. Wang, "An industrial-strength audio search algorithm," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, 2003, pp. 1–4.

[2] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, 2002.

[3] M. Fink, M. Covell, and S. Baluja, "Social- and interactive-television. applications based on real-time ambient-audio identification," in *Proc. European Interactive TV Conference (Euro-ITV)*, 2006.

[4] C. Howson, E. Gautier, P. Gilberton, A. Laurent, and Y. Legallais, "Second screen tv synchronization," in *Proc. IEEE Int. Conf. on Consumer Electronics (ICCE)*, 2011.

[5] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," in *IEEE Workshop on Multimedia Signal Processing*, 2002, pp. 169–173.

[6] H.J.Kim, Y.H.Choi, J.W.Seok, and J.W.Hong, "Audio watermarking techniques," in *Intelligent Watermarking Techniques*, 2004, ch. 8, pp. 185–218.

[7] R. Macrae, X. Anguera, and N. Oliver, "Muvisync: Realtime music video alignment," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2010.

[8] T. website, "http://www.tvplus.com/."

[9] I.-N. application from Yahoo, "http://www.intonow.com/ci/soundprint."

[10] M.-S. website, "http://media-sync.tv/."

[11] C. website, "http://www.civolution.com."

[12] N. Q. K. Duong, C. Howson, and Y. Legallais, "Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation," in *Proc. IEEE International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, 2012, pp. 241–244.

[13] N. Q. K. Duong and F. Thudor, "Movie synchronization by audio landmark matching," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 3632–3636.

[14] C. V. Cotton and D. P. W. Ellis, "Audio fingerprinting to identify multiple videos of an event," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 2386–2389.

[15] Z. Rafii, B. Coover, and J. Han, "An audio fingerprinting system for live version identification using image processing techniques," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[16] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *Journal of VLSI Signal Processing*, no. 3, pp. 271–284, 2005.

[17] G. Richard, S. Sundaram, and S. Narayanan, "An overview on perceptually motivated audio indexing and classification," *Proceedings of the IEEE*, no. 9, pp. 1939–1954, 2013.

[18] A. Ramalingam and S. Krishnan, "Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting," *IEEE Trans. on Information Forensics and Security*, no. 4, pp. 457–463, 2006.

[19] E. Batlle, J. Masip, E. Guaus, and P. Cano, "Scalability issues in hmm-based audio fingerprinting," in *Proc. IEEE Int. Conference on Multimedia and Expo*, 2004, pp. 735–738.

[20] V. Chandrasekhar, M. Sharifi, and D. A. Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications," in *12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 801–806.

[21] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 597–604.

[22] S. Baluja and M. Covell, "Audio fingerprinting: Combining computer vision and data stream processing," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[23] M. Ramona and G. Peeters, "Automatic alignment of audio occurrences: application to the verification and synchronization of audio fingerprinting annotation," in *Proc. DAFX*, 2011, pp. 429–436.

[24] E. Dupraz and G. Richard, "Robust frequency-based audio fingerprinting," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 2091–2094.

[25] M. Moussallam and L. Daudet, "A general framework for dictionary based audio fingerprinting," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 3077–3081.

[26] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo, "Audio fingerprinting based on normalized spectral subband centroids," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 213–216.

[27] K. Seyerlehner, M. Schedl, P. Knees, and R. Sonnleitner, "A refined block-level feature set for classification, similarity and tag prediction," in *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.

[28] X. Anguera, A. Garzon, and T. Adamek, "Mask: Robust local features for audio fingerprinting," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 455–460.

[29] P. Cano, E. Batlle, H. Mayer, and H. Neuschmied, "Robust sound modeling for song detection in broadcast audio," in *Proc. 112th Audio Engineering Society Convention (AES)*, 2002.

[30] J. Hao, T.-W. Lee, and T. J. Sejnowski, "Speech enhancement using gaussian scale mixture models," *IEEE Trans. on Audio Speech and Language Processing*, no. 18, pp. 1127–1136, 2010.

[31] C. J. Burges, J. C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. on Audio Speech and Language Processing*, no. 3, pp. 165–174, 2003.

[32] J. Deng, W. Wan, X. Yu, and W. Yang, "Audio fingerprinting based on spectral energy structure and nmf," in *Proc. Int. Conf. on Communication Technology (ICCT)*, 2011, pp. 1103–1106.

[33] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, 2002.

[34] A. Bagri, F. Thudor, A. Ozerov, and P. Hellier, "A scalable framework for joint clustering and synchronizing multi-camera videos," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.

[35] W. 1, "http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html."

[36] W. 2, "https://engineering.purdue.edu/malcolm/interval/1998-010/."

[37] C. Yang, "Macs: music audio characteristic sequence indexing for similarity retrieval," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 123–126.

[38] J. Ogle and D. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 233–236.

[39] N. J. Bryan, P. Smaragdis, and G. J. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 2389–2392.

[40] M. Ramona and G. Peeters, "Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 477–480.

[41] E. Allamanche, J. Herre, and O. Hellmuth, "Content-based identification of audio material using mpeg-7 low level description," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, 2002.

[42] J. Herre, E. Allamanche, and O. Hellmuth, "Robust matching of audio signals using spectral flatness features," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 127–130.

[43] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, no. 1, pp. 72–83, 1995.

[44] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[45] L. R. Rabiner, "A tutorial on hmm and selected applications in speech recognition," *Proceeding of the IEEE*, vol. 17, no. 2, pp. 257–286, 1989.

[46] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[47] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[48] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on nonnegative matrix factorization," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1586–1590.

[49] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[50] N. Chen, H.-D. Xiao, and W. Wan, "Audio hash function based on nonnegative matrix factorisation of mel-frequency cepstral coefficients," *IET Information Security*, no. 1, pp. 19–25, 2011.