Audio-Fingerprinting via Dictionary Learning

Christina Saravanos
Computer Enginnering and Infromatics Department, University of Patras, Greece
saravanou@ceid.upatras.gr

Dimitris Ampeliotis
Computer Enginnering and Infromatics Department, University of Patras, Greece
ampeliot@ceid.upatras.gr

Kostas Berberidis
Computer Enginnering and Infromatics Department, University of Patras, Greece
berberid@ceid.upatras.gr

***Abstract*— In recent years, several successful schemes have been proposed to solve the song identification problem. These techniques aim to construct a signal's audio-fingerprint by either employing conventional signal processing techniques or by computing its sparse representation in the time-frequency domain. This paper proposes a new audio-fingerprinting scheme which is able to construct a unique and concise representation of an audio signal by applying a dictionary, which is learnt here via the well-known K-SVD algorithm applied on a song database. The promising results which emerged while conducting the experiments suggested that, not only the proposed approach preformed rather well in its attempt to identify the signal content of several audio clips –even in cases this content had been distorted by noise - but also surpassed the recognition rate of a Shazam-based paradigm.**

Keywords—song identification, audio-fingerprinting, sparse coding, dictionary learning, K-SVD, OMP

I. INTRODUCTION

Identifying the signal content of an audio signal has always been a challenging problem. Over the last three decades, several solutions have been proposed with the most promising one being the so-called audio-fingerprinting scheme and its various alternates. In the early 2000's, Wang et al. proposed a novel audio-fingerprinting paradigm – which has become one of the most widely used song identification applications known as Shazam, that aims to identify the signal content of an audio clip by manipulating its content in the frequency domain [1], [2]. Notably, these works suggest that, a promising approach should construct the audio-fingerprint by extracting the most robust and descriptive points within the signal's spectrogram, which are often referred to as the keypoints or peaks.

As compressive sensing (CS) and the associated sparse coding [3]-[13] and dictionary learning (DL) methods [14]-[18] became widely popular, many sought to propose novel fingerprinting techniques or to solve similar problems via the signals' sparse representations. In particular, [25] aspires to identify environmental sounds, while [29] and [30] aim to recognize similar audio-visual events in videos via the signal's sparse representation in the time-frequency domain. Furthermore, [31] proposes an innovative audio-fingerprinting scheme which aims to determine the signal content of a 5-seconds long audio segment against a database which comprises several thousands of 5-seconds long audio clips via their time-frequency sparse representations. Note that, the signals' sparse representations were computed by combining the Matching Pursuit (MP) algorithm [12], [14], [15] along with multi-scale, shift-invariant Gabor or modified discrete cosine transform (MDCT) dictionaries. These dictionaries are often used to solve audio-related problems thanks to their ability to capture the acoustic/perceptual attributes of an audio signal.[24]-[28].

On the other hand, in the context of image-based information retrieval, [32] suggests an image-fingerprinting scheme which attempts to identify an image segment against a database of complete images. This paradigm differs from the ones, which were previously mentioned, in that the images' sparse representation is computed by combing the Orthogonal Matching Pursuit (OMP) algorithm [12], [15]-[19] over a dictionary that was learnt by applying the K-SVD algorithm [23] onto the images which compose the database.

Nonetheless, [41] –[44] aspire to match cover songs to their original recordings by employing convolutional neural networks, while [40] aims to identify a cover's original song by combining several similarity techniques. Furthermore, [45]-[47] suggest three novel audio-to-score alignment scores which apply dynamic time warping (DTW) whereas the paradigm advocated in [48] relies on recurrent neural networks (RNNS).

In this paper, we propose a new audio-fingerprinting scheme which combines sparse coding and DL techniques. The advocated approach consists of two subsequent, yet equally important steps: In the first step, DL is employed to construct a global dictionary by employing the K-SVD algorithm onto the songs which maintain the database. On the other hand, the second step aims to construct the signals' fingerprints via their sparse representations which are computed via the OMP algorithm and the dictionary which was, previously, learnt. The rather promising results, which emerged while conducting the experiments suggest that the proposed method is superior in performance and robust to noise than most Shazam-based approaches.

The remaining of this paper is organized as follows: Section 2 presents the proposed audio-fingerprinting scheme. Section 3 elaborates on the results which emerged while conducting the experiments and compares them to a Shazam-based audio-fingerprinting technique. Finally, Section 4 draws the final conclusions.

II. THE PROPOSED AUDIO-FINGERPRINTING SCHEME

This section elaborates on the fundamental steps of the proposed audio-fingerprinting scheme which are employed to construct an audio signal's unique and concise representation. Figure 1 illustrates an overview of the suggested paradigm, which comprises two important and equally important steps: Shortly speaking, the first step aims to learn a global dictionary by concatenating and feeding the content of the concatenated songs, which compose the database to the K-SVD algorithm. On the other hand, the second phase of the recommended scheme aspires to compute an audio signal's sparse representation by combining its original content, the

This work was supported in part by the University of Patras and the RPF, Cyprus, under the project INFRASTRUCTURES/1216/0017 (IRIDA).

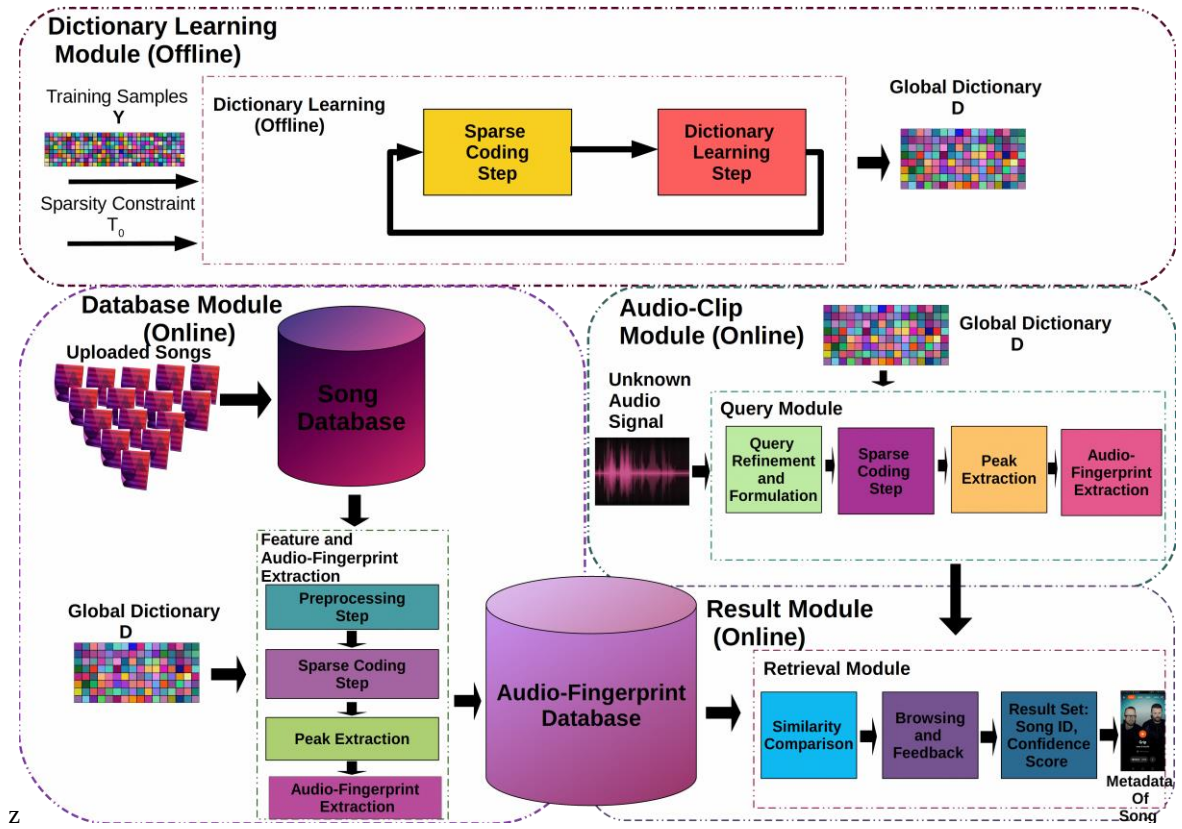

Figure 1: An Overview of the Proposed Audio-Fingerprinting Scheme

dictionary which was previously created and the OMP algorithm. The audio signal either corresponds to a track within the database or a several seconds long audio clip. The signal's fingerprint is constructed by extracting the most robust and descriptive atoms from its respective sparse representation. Finally, the advocated approach aims to identify the audio clip by matching its fingerprint to a fingerprint of a song, within the database. In the following paragraphs, each partial step of the suggested scheme is described with more details.

A. *Dictionary Learning*

Computing a dictionary which can correctly capture the attributes of several signals is a key-problem in most sparse coding and DL applications. The advocated audio-fingerprinting technique - unlike the ones which were mentioned in [25], [29]-[31] - relies on correctly learning an over-complete, global dictionary $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, .. \boldsymbol{d}_k] \in R^{2^k \times K}$ via the songs within the database and which can capture the acoustic and perceptual attributes of the audio signals.

The signal X_i corresponding to song i – within the database – is initially down-sampled at $8kHz$ and segmented into audio frames - which comprise 2^k samples where $k > 6$, with 50% overlap. This procedure results in creating the audio signal matrix $\boldsymbol{Y}_i \in \mathbb{R}^{2^k \times N_i}$, where N_i denotes the number of audio frames of song i. The audio signal matrices $\boldsymbol{Z}_i \in \mathbb{R}^{2^k \times N_i}$ are then concatenated, thus creating the overall audio signal matrix $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_M] \in \mathbb{R}^{2^k \times M}$, where $M = \sum_i N_i$ is the total number of frames in the database.

Thereupon, the audio signal matrix $\boldsymbol{Z}$, as shown in Figure 1, is fed into the K-SVD algorithm. The K-SVD algorithm [23], similar to most DL schemes, comprises two subsequent and iterative steps: The first, which is referred to as the Sparse Representation step, aims to compute the current sparse representation matrix for $\boldsymbol{Y}, \boldsymbol{Z} \in R^{K \times N}$, while keeping the dictionary fixed. On the other hand, the second step, which is known as the dictionary update step, aims to update the dictionary $\boldsymbol{D}$ while keeping $\boldsymbol{Y}$ fixed. Once the dictionary has been learnt, the sparse representation of $\boldsymbol{Y}, \boldsymbol{Z}$ will be discarded.

Fig. 1 hints that the dictionary $\boldsymbol{D}$ is learnt offline and as a result, the computational cost of the K-SVD algorithm won't be taken into account in calculating the overall computational complexity of the proposed scheme.

B. *Constructing the Audio-Fingerprint*

This step of the proposed audio-fingerprinting scheme aspires to construct the signal's unique and concise representation by combining the OMP algorithm and the global dictionary $\boldsymbol{D} \in R^{2^k \times K}$, which was previously learnt. The fingerprints of the audio tracks, which cultivate the database, and the clip are similarly created. Nonetheless, Figure 1 suggest that these fingerprints are constructed in two different modules – one which corresponds to the songs within the database, while the other coincides with the excerpt, respectively.

A discrete and finite-length audio signal $\boldsymbol{x}$ – in a similar fashion as in the DL process – is down-sampled at $8kHz$ and segmented into several audio frames, which consist of 2^k samples with 50% overlap, therefore creating the audio signal matrix $\boldsymbol{y} \in R^{2^k \times N}$, where N denotes the number of audio

frames for signal $\boldsymbol{x}$. Subsequently, the sparse representation matrix of $\boldsymbol{y}$, denoted as $\boldsymbol{z} \in R^{K \times N}$ is computed by employing the OMP algorithm [12], [15]-[19] along with the dictionary $\boldsymbol{D}$, which was, previously, learnt.

Afterwards, the most active atoms employed in the sparse representation matrix $\boldsymbol{z}$ are extracted to construct the audio-fingerprint. These atoms are analogous to the peaks or keypoints which are extracted from $\boldsymbol{y's}$ respective spectrogram [1], [2] thanks to their descriptiveness and robustness. In particular, the selected atoms best represent the signal's dominant original content in the dictionary's domain and are robust to distortion (e.g. noise) which may appear in $\boldsymbol{z}$. The atom's weight value (sparse representation coefficient), its frame and atom indices are, then, stored onto an $R \times 3$ matrix, where R denotes the number of the extracted most active atoms, as a representation of the audio extract x.

Thereupon, the landmark pairs, i.e. four-point peak – or in this case, most active atom - pairs, are created. The landmark pairs are extracted in a similar manner to the approaches described in [1], [2] and [30].

C. Matching the Audio Clip Against the Database

The final step of the proposed scheme aspires to match the audio clip to a song within the database by comparing their respective fingerprints. The landmark pairs of the audio signal, which were previously hashed and stored onto a hash map. The hash functions which are applied to compute the hash key and value of a landmark pair are similar to the ones described in [1], [2], [30], [36],[37]

Afterwards, the proposed scheme aims to match the audio clip against the songs which uphold the database by computing the *Jaccard similarity coefficient* [38] or, simply, *Jaccard coefficient* between their respective hash keys. The Jaccard coefficient measures the similarity between two finite sample sets and is defined as the number of samples which compose their intersection, divided by the number of samples, which uphold their union. Mathematically, the Jaccard coefficient is defined by

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{1}$$

where A, B denote the two finite sample sets. The proposed technique sorts the emerged Jaccard coefficients in descending order and, finally, returns the metadata of the audio track with the highest coefficient value.

D. Computational complexity

The computational complexity of calculating the sparse representation of matrix $\boldsymbol{z}$ via the OMP algorithm is proportional to

$$C_{OMP} = O(N \cdot 2^k \cdot q \cdot T_0), \tag{2}$$

where $N, 2^k, q$ and T_0 denote the number of audio frames of signal x, the length of each audio frame, the length of each sparse audio frame $\boldsymbol{y_i}$ and the number of iterations required to compute $\boldsymbol{z}'$s sparse representation, respectively. Let $p_r = (f_r, t_r)$ be an active atom which is not only a peak, but also serves as an anchor point. The advocated approach will search for its m closest atoms within the target zone which results in forming a cluster - whose centroid is p_r – which comprises $m + 1$ points. Furthermore, let N_{anchor} be the total number of anchor points which are present in each second of the audio clip. The computational cost of storing the landmark pairs is proportional to

$$C_{STORAGE} = O(N_{anchor} \cdot m), \tag{3}$$

which is related to the computational complexity of selecting the N_{anchor} anchor points-which is proportional to

$$C_{SELECTION} = O(N_{anchor}), \tag{4}$$

As also mentioned in the previous, the global dictionary $\boldsymbol{D}$, as illustrated in Fig. 1, is learnt offline and therefore, the computational complexity of the K-SVD algorithm is discarded. Resultantly, the complexity of the proposed scheme is proportional to

$$C_{PROPOSED} = O(N \cdot 2^k \cdot q \cdot T_0), \tag{5}$$

which is more computationally expensive than most Shazam-based approaches [33]. The computational complexity of the advocated paradigm could be reduced by replacing the batch OMP algorithm with its parallelized version [34], [35]. Other approaches to reduce the complexity associated with the sparse coding step are the subject of ongoing work. For instance, substantial reduction in complexity may be achieved by properly exploiting the fact that neighboring audio frames have similar support sets.

III. Experimental setup and results

This section elaborates on the experimental setup and the results which emerged while evaluating the recommended audio-fingerprinting paradigm. Note that, the experiments were conducted in a preliminary stage and we plan to conduct experiments using larger databases in the near future.

A. Audio databases

While conducting the experiments two databases, whose size differ and which comprise different audio tracks were employed. The first database consists of 12 audio tracks performed by similar artists, while the second is upheld by 70 songs produced by a wider range of performers.

B. Dictionary learning

While evaluating the proposed audio-fingerprinting scheme, four dictionaries -whose dimensions vary- were learnt via the spectral and temporal content of the songs which maintain the database. The latter were learnt by concatenating the songs' respective spectrogram. Note that, the dictionaries, which were learnt via the songs' content in the time domain had dimensions 128×256 and 256×512 while the other two were of dimensions 129×258 and 129×516 respectively. Note that, to further assess the dictionaries' ability to capture the signals' acoustic and perceptual attributes, half of the songs were, randomly, selected from each database to partake in the DL process.

C. Keypoints selection

While conducting the experiments, two distinct pruning techniques were applied to extract the most active atoms from

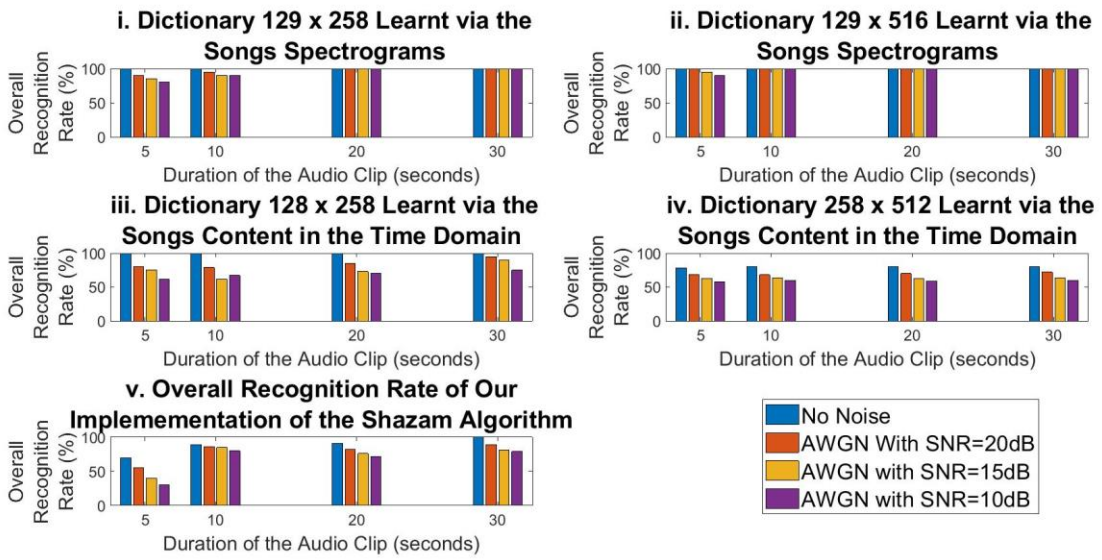

Figure 2: The overall recognition rate of the proposed scheme after identifying the signal content of several hundreds of 30-, 20-, 10- and 5-seconds long audio clips which were extracted from the database which comprises 12 songs and whose content was, additionally, distorted by 3 AWGNs. The signal's fingerprints were constructed by employing the dictionaries: i. 129×258 *ii.* 129×516, *iii.* 128×256, *iv.* 256×512 *and extracting the dominant atom from each audio frame and v. the Shazam-based approach, which is mentioned in [36], [37]*

a signal's sparse representation. The former aspires to elicit the dominant atom, i.e. the atom with the highest weight value, from each audio frame, whereas the latter excerpts the q most active atoms from the signal's overall sparse representation. The value of q ranged from 150 to 10,000 depending, primarily, on the duration of the audio clip. Each audio which was extracted via the former pruning technique corresponds to an audio frame. On the other hand, the atoms which were elicited by employing the latter approach can either be accumulated to a single audio frame or scattered within the signal's sparse representation.

D. Numerical results

The recommended audio-fingerprinting scheme was, initially, evaluated against its ability to, correctly, identify the signal content of several hundreds of 30, 20, 10 and 5 seconds long audio clips, which were, randomly, extracted from each database, respectively. The signals' sparse representations were computed by combining the OMP algorithm along with one of the four dictionaries, which were previously mentioned. Furthermore, the "keypoints" were elicited by applying either one of the two pruning techniques, which were described in the previous section. Figs. 2-5 illustrate the overall recognition rate of the proposed approach while employing these eight dictionary/pruning technique combinations to construct the signals' respective fingerprints. Figs.2-5 point out that the overall recognition rate of the proposed scheme was equal to 100% when employing either one of the two dictionaries which were learnt via the songs' spectral content and the dictionary with dimensions 128×256. On the other hand, the overall recognition rate of the advocated paradigm was slightly lower when applying the dictionary with dimensions 256×512 suggesting that some perceptual information regarding the audio signals may be lost due to the rather large length of the audio frames. The rather satisfying results -which emerged while conducting the experiments on two non-extensive database whose size differ significantly, suggest that the proposed scheme might have a similar performance rate when employing a larger database.

The proposed audio-fingerprinting scheme was, additionally, evaluated against its robustness to noise. To this end, three distinct $AWGNs$ corresponding to $SNR = 20, 15, 10dB$ were added onto the signal content of the audio clips. The results point out that, the proposed audio-fingerprinting scheme had the best performance when combining either one of the two dictionaries, which were learnt via the songs' spectral content along with extracting the dominant atom from each audio frame of a signal's sparse representation. Moreover, the proposed audio-fingerprinting scheme preformed quite well when employing the dictionary 128×256 along with extracting the dominant atom from each sparse audio frame. In this case, the emerged results suggest that even though the audio clip's content has been denoised via the OMP a few pieces of information could possibly have become lost due to its distortion from the $AWGN$. On the other hand, the advocated scheme had a slightly poorer overall recognition rate -similar to the one which was previously described, when employing the dictionary 256×512 regardless of the pruning technique used to extract the signals' most active atoms.

Furthermore, the emerged results of the advocated approach were, additionally, compared to the respective results of a Shazam-based approach [36], [37]. Figs. 2-5 hint that the advocated approach surpassed the Shazam-based audio-fingerprinting scheme when employing either one of the two dictionaries which were learnt via the songs' spectral content and the dictionary 128×256 which was learnt via the songs' content in the spectral content. These results suggest that the most active atoms which have been extracted from the signal's sparse representation are more descriptive

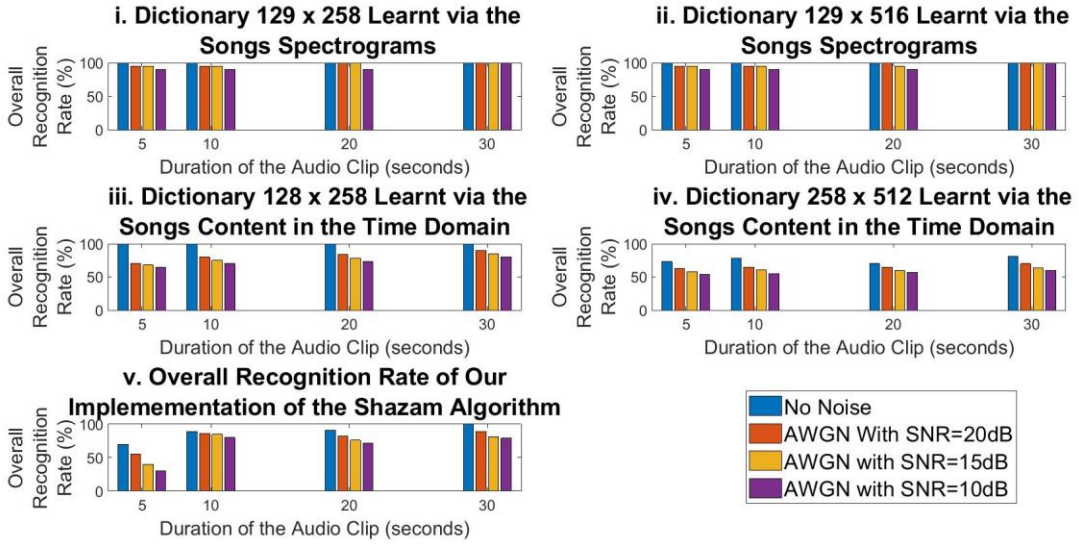

Figure 3: The overall recognition rate of the proposed scheme after identifying the signal content of several hundreds of 30-, 20-, 10- and 5-seconds long audio clips which were extracted from the database which comprises 12 songs and whose content was, additionally, distorted by 3 AWGNs. The signal's fingerprints were constructed by employing the dictionaries: i. 129×258 *ii.* 129×516, *iii.* 128×256, *iv.* 256×512 *and extracting the q most active atoms from the signals' overall sparse representation and v. the Shazam-based approach, which is mentioned in [36], [37]. The value of q ranges from 150 to 10,000 depending on the signal's duration.*

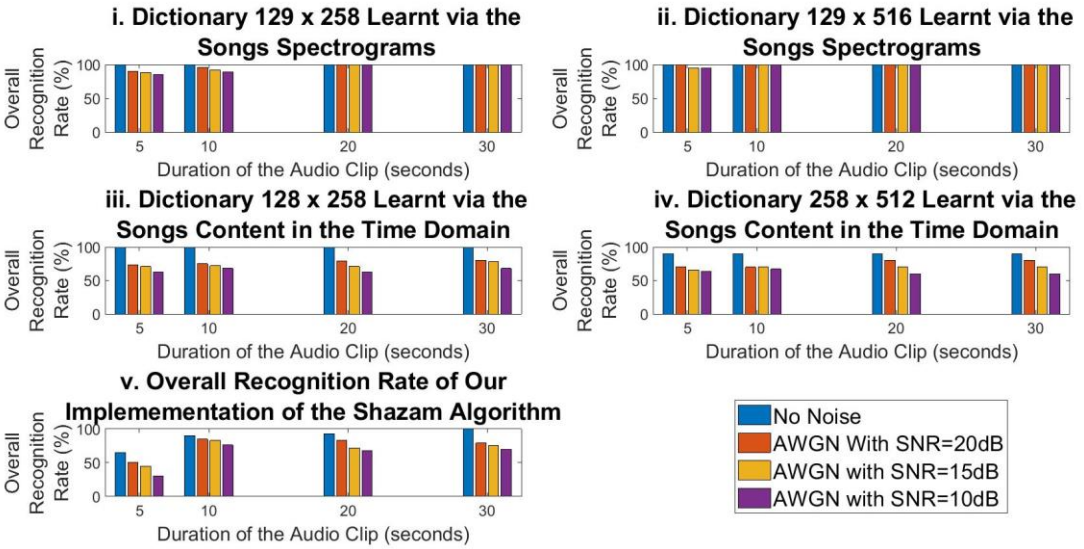

Figure 4: The overall recognition rate of the proposed scheme after identifying the signal content of several hundreds of 30-, 20-, 10- and 5-seconds long audio clips which were extracted from the database which comprises 70 songs and whose content was, additionally, distorted by 3 AWGNs. The signal's fingerprints were constructed by employing the dictionaries: i. 129×258 *ii.* 129×516, *iii.* 128×256, *iv.* 256×512 *and extracting the dominant atom from each audio frame and v. the Shazam-based approach, which is mentioned in [36], [37]*

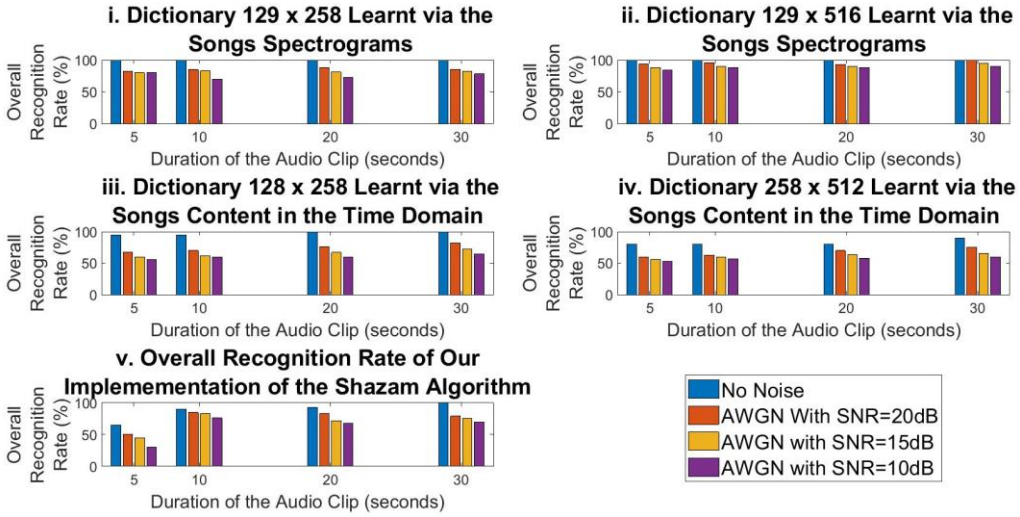

Figure 5: The overall recognition rate of the proposed scheme after identifying the signal content of several hundreds of 30-, 20-, 10- and 5-seconds long audio clips which were extracted from the database which comprises 70 songs and whose content was, additionally, distorted by 3 AWGNs. The signal's fingerprints were constructed by employing the dictionaries: i. 129×258 *ii.* 129×516*, iii.* 128×256*, iv.* 256×512 *and extracting the q most active atoms from the signals' overall sparse representation and v. the Shazam-based approach, which is mentioned in [36], [37]. The value of q ranges from 150 to 10,000 depending on the signal's duration.*

and robust than the peaks which are extracted from its respective spectrogram.

Lastly, the rather promising results of the proposed audio-fingerprinting scheme point out the dictionaries' ability to capture the acoustic and perceptual attributes of most audio signals. The dictionaries, as we mentioned earlier, were learnt by using only half of the songs extracted from each database, respectively, whereas the audio clips were, randomly, extracted from the tracks within the database, regardless of whether they participated in learning the dictionaries. This means that, the employed dictionaries could be considered equivalent to the multi-scale, shift-invariant dictionaries which are, often, used to solve audio-related problems.

IV. Conclusions

In the last three decades, several solutions have been developed to solve the song identification problem with the most promising being the audio-fingerprinting scheme. The state-of-the-art approaches aim to match rather short audio clips to a song within the database by constructing their respective fingerprints. The unique and concise representations of the audio signals are either created via conventional signal processing techniques [1], [2], [36], [37] or by combining sparse coding along with DL approaches [25], [29]-[31] employing structured dictionaries. This paper sought out to proposed an alternate to the conventional audio-fingerprinting scheme which, initially, constructs the sparse representation of a signal via the OMP algorithm and a global dictionary which has been learnt via the songs which compose the database. The rather promising results suggested that not only did the proposed scheme had very good performance - i.e. an overall recognition rate which ranged from 90 to 100%- in most cases, but was also able to surpass the respective results of a Shazam-based paradigm. The results, additionally, pointed out that a data-driven dictionary could be considered equivalent to dictionaries which are learnt via predefined functions. A drawback of the advocated scheme with respect to existing schemes is its higher computational complexity which, however, can be significantly reduced as pointed out is Section 2.D.

References

[1] Avery Li-Chun Wang, "An Industrial-Strength Audio Search Algorithm*", Ismir*, Vol. 2003, 2003.

[2] Avery Wang, "The Shazam Music Recognition Service", *Communications of the ACM* , Vol. 49, No. 8, August 2006.

[3] Michael Unser, "Sampling- 50 Years After Shannon", *Proceedings of the IEEE*, Vol. 88, No. 4, April 2000.

[4] David L. Donoho, "Compressed Sensing', *IEEE Transactions On Information Theory*, Vol. 52, No. 4, April 2006.

[5] Simon Foucart and Holger Rauhut, "A Mathematical Introduction to Compressive Sensing", *Bull. Am. Math*, Vol. 54, 2017.

[6] Emmanuel J. Candès and Michael B. Wakin, "An Introduction to Compressive Sampling", *IEEE Signal Processing Magazine*, Vol. 25, No. 2, 2008.

[7] Fatima Salahdine, Naima Kaabouch, Hassan El Ghazi, "A Survey On Compressive Sensing Techniques for Cognitive Radio Networks", *Physical Communication*, Vol. 20, 2016.

[8] Richard G. Baraniuk, "Compressive Sensing", *IEEE Signal Processing Magazine*, Vol. 24, No. 4, 2007.

[9] M.Amin Khajehnejad Alexandros G. Dimakis Weiyu Xu Babak Hassibi, 'Sparse Recovery of Positive Signals with Minimal Expansion", Preprint. *arXiv: 0902.4045*, June 2009.

[10] Irena Orovic, Vladan Papic , Cornel Ioana, Xiumei Li, and Srdjan Stankovic, "Compressive Sensing in Signal Processing: Algorithms and Transform Domain Formulations", *Mathematical Problems in Engineering*, Vol. 2016.

[11] Massimo Fornasier and Holger Rauhut, "Compressive Sensing", *Handbook of mathematical methods in imaging*, 2015.

[12] Zheg Zhang, Yong Xu, Jian Yang, Xuelong Li and David Zhang, "A Survey of Sparse Representation: Algorithms and Applications", *IEEE Access*, Vol. 3, 2015.

[13] Meenu Rani, S.B. Dhok and R.B. Demshmukh, "Systematic Review of Compressive Sensing:Concepts, Implementations and Applications", *IEEE Access*, Vol. 6, 2018.

[14] Y.C.Pati, R. Rezaiifar and P.S. Krishnaprasad, "*Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition*", Proceedings of 27th Asilomar conference on signals, systems and computers, 1993

[15] Shane F. Cotter and Bhaskar D. Rao, "Sparse Channel Estimation via Matching Pursuit With Application to Equalization", *IEEE Transactions on Communications*, Vol. 50, No. 3, March 2002.

[16] Joel E. Tropp, Stephen J. Wright, "Computational Methods for Sparse Solution of Linear Inverse Problems", *Proceedings of the IEEE*, Vol. 98, No. 6, June 2010.

[17] Michael Elad, "Sparse And Redundant Representations", Springer, 2010.

[18] Joel A. Tropp and Anna C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit", *IEEE Transactions on information theory*, Vol. 53, No. 12, 2007.

[19] T. Tony Cai and Lei Wang, "Orthogonal Matching Pursuit for Sparse Signal Recovery", Technical Report, 2010.

[20] Ivana Tosic and Pascal Frossad, "Dictionary Learning", *IEEE Signal Processing Magazine*, March 2011.

[21] Ron Rubinstein, Alfred M. Bruckstein and Michael Elad, "Dictionaries for Sparse Representation Modeling", *Proceedings of the IEEE*, Vol 98, No. 6, June 2010.

[22] Ke Huang and Selin Aviyente, "Sparse Representation for Signal Classification", *Advances in neural information processing systems*, 2007.

[23] Michael Aharon, Michael Elad and Alfred Brucktein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", *IEEE Transactions on Signal Processing*, Vol 54, No. 11, November 2006.

[24] Stephane G. Mallat and Zhifeng Zhand, "Matching Pursuits With Time-Frequency Dictionaries", *IEEE Transactions On Signal Processing*, Vol. 41, No. 12 December 1993.

[25] Selina Chu, Shrikanth Narayanan and C.C. Jay Kuo, "Enviromental Sound Recognition With Time-Frequency Audio Features", *IEEE Transactions on Audio, Speech and Language Processing*, 2009.

[26] John P.Princen and Alan Bernad Bradley, "Analysis/ Synthesis Filter Bank Design Based on Time Domain Cancellation", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-34, No. 5, October 1986.

[27] Ye Wang and Mikka Vilermo, "Modified Discrete Cosine Transform-Its Implications for Audio Coding and Error Concealment," *Journal of the Audio Engineering Society*, Vol. 51, No. 1/2, 2003.

[28] Emmaneul Ravelli, Gal Richard and Laurent Daudet, "Union of MDCT Bases for Audio Coding", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16, No. 8, November 2008.

[29] Courtenay Cotton and Daniel P.W. Ellis, "*Finding Similar Acoustic Events Using Matching Pursuit and Locality-Sensitive Hashing*", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October, 2009.

[30] Courtenay Cotton and Daniel P.W. Ellis, "*Audio Fingerprinting to Identify Multiple Videos of an Event*", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010.

[31] Manuel Moussallam and Laurent Daudet, "*A General Framework for Dictionary Based Audio Fingerprinting*", IEEE Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[32] Yue nan Li, "Robust Content Fingerprinting Algorithm Based on Sparse Coding", *IEEE Signal Processing Letters*, Vol. 22, No. 9, September 2015.

[33] WS Moses, ED Demaine, "Computational Complexity of Arranging Music", *The Mathematics of Various Entertaining Subjects: Research in Games, Graphs, Counting, and Complexity*, Vol. 2, 2017.

[34] Sujuan Liu, Ning Lyu , and Haojiang Wang, "The Implementation of the Improved OMP for AIC Reconstruction Based on Parallel Index Selection", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 26, No.2, February 2018.

[35] Amey Kulkarni, Houman Homayoun, Tinoosh Mohsenin, "*A Parallel and Reconfigurable Architecture for Efficient OMP Compressive Sensing Reconstruction*", GLSVLSI'14, May 21–23, Houston, Texas, USA, 2014.

[36] Daniel P.W. Ellis, "Robust Landmark-Based Audio Fingerprinting web resource": http://labrosa.ee.columbia.edu/matlab/fingeprint/, 2009

[37] https://www.mathworks.com/matlabcentral/fileexchange/23332-robust-landmark-based-audio-fingerprinting

[38] S. Niwattanakul, J. Singthongchai, E. Naenudorn and S. Wanapu "*Using of Jaccard coefficient for keywords similarity*", in Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, No. 6, pp. 380-384) , March 2013.

[39] A. Gkillas, D. Ampeliotis and K. Berberidis, "*Fast Sparse Coding Algorithms for Piece-Wise Smooth Signals*," 28th European Signal Processing Conference (EUSIPCO 2020), 18-22 January 2021, 2020, Amsterdam, NL (to appear)

[40] Chen, N., Li, W. & Xiao, H. Fusing similarity functions for cover song identification. Multimed Tools Appl **77,** 2629–2652 (2018). https://doi.org/10.1007/s11042-017-4456-9

[41] Xu, Xiaoshuo, Xiaoou Chen and Deshun Yang. "Key-Invariant Convolutional Neural Network Toward Efficient Cover Song Identification" 2018 IEEE International Conference on Multimedia and Expo (ICME) (2018): 1-6.

[42] Jiang, Chaoya, Deshun Yang and Xiaoou Chen. "Similarity Learning For Cover Song Identification Using Cross-Similarity Matrices of Multi-Level Deep Sequences." ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020): 26-30

[43] M. Mehta, A. Sajnani and R. Chapaneri, "Cover Song Identification with Pairwise Cross-Similarity Matrix using Deep Learning," 2019 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2019, pp. 1-5, doi: 10.1109/IBSSC47189.2019.8973064.

[44] Qi X., Yang D., Chen X. (2018) Triplet Convolutional Network for Music Version Identification. In: Schoeffmann K. et al. (eds) MultiMedia Modeling. MMM 2018. Lecture Notes in Computer Science, vol 10704. Springer, Cham. https://doi.org/10.1007/978-3-319-73603-7_44

[45] Muñoz-Montoro, A.J., Cortina, R., García-Galán, S. et al. A score identification parallel system based on audio-to-score alignment. J Supercomput (2020). https://doi.org/10.1007/s11227-020-03185-2

[46] Andreas Arzt, Stefan Lattner. "Audio-to-Score Alignment using Transposition-invariant Features", 19th International Society for Music Information Retrieval Conference, Paris, France, 2018

[47] Alonso, P., Cortina, R., Rodríguez-Serrano, F.J. et al. Parallel online time warping for real-time audio-to-score alignment in multi-core systems. JSupercomput **73,** 126–138(2017). https://doi.org/10.1007/s11227-016-1647-5

[48] Kwon, Taegyun, Dasaem Jeong and Juhan Nam. "Audio-to-score alignment of piano music using RNN-based automatic music transcription." ArXiv abs/1711.04480 (2017): n. pag.