MASK+:DATA-DRIVEN REGIONS SELECTION FOR ACOUSTIC FINGERPRINTING

Lucas Ondel[1,2]*, Xavier Anguera*[1] *and Jordi Luque*[1]

[1]Telefonica Research, Barcelona, Spain
[2]Faculty of Information Technology, Brno University of Technology, Czech Republic
iondel@fit.vutbr.cz, {xanguera, jls}@tid.es

ABSTRACT

Acoustic fingerprinting is the process to deterministically obtain a compact representation of an audio segment, used to compare multiple audio files or to efficiently search for a file within a big database. Recently, we proposed a novel fingerprint named MASK (**M**asked **A**udio **S**pectral **K**eypoints) that encodes the relationship between pairs of spectral regions around a single spectral energy peak into a binary representation. In the original proposal the configuration of location and size of the regions pairs was determined manually to optimally encode how energy flows around the spectral peak. Such manual selection has always been considered as a weakness in the process as it might not be adapted to the actual data being represented. In this paper we address this problem by proposing a unsupervised, data-driven method based on mutual information theory to automatically define an optimal MASK fingerprint structure. Audio retrieval experiments optimizing for data distorted with additive Gaussian white noise show that the proposed method is much more robust than the original MASK and a well known acoustic fingerprint.

Index Terms— Audio fingerprinting, content recognition

1. INTRODUCTION

With the increasing availability of vast quantities of audio-visual content online, it is very important to develop techniques allowing for an efficient representation and effective comparison of such content in search of duplicates or derivate works. In the audio domain this can be achieved through acoustic fingerprinting, which is the process by which audio content can be deterministically encoded into a compact representation that can be then used for search and retrieval applications. As described in [1], a good acoustic fingerprint can be measured in (at least) four main dimensions: discriminatory power, robustness, compactness and efficiency. Discriminatory power measures how different the fingerprints extracted from versions of the same audio are versus other audio. Robustness measures how much the fingerprint gets affected when distorting the original audio. Compactness refers to the size needed to encode the audio, and efficiency regards to how fast the fingerprint can be obtained from the audio and, equivalently, how fast it can be applied in the task it is used for.

Although the literature already offers many possible fingerprinting algorithms (see [2] for an early review of the area) most are not suitable for real-life situations. Probably the most well known fingerprints are Wang (Also known as Shazam) [3], Haitsma-Kalker (also Known as Philips) [4], Burges et al. (named RARE) [5] and Baluja-Covell (named Waveprint) [6]. The fingerprint used in this work builds mostly on inspiration from Wang and Haitsma fingerprints, which we briefly review next. On the one hand, in [3] they select relevant salient points in the short-term spectrum of the signal and encode their relationship in pairs. The strength of this method lies on the intelligent selection of robust salient points to be robust to typical audio distortions and additive noise. On the contrary, by encoding pairs of points they double the probability to miss at matching time when either of the original salient points disappears in the modified audio. On the other hand, in [4] they propose a binary representation of the acoustic signal obtained at fixed time intervals by encoding energy differences between adjacent frequency bands. This method allows for a thorough (but still compact) representation of the signal that is resilient to light distortions and noises. When using the fingerprint in very noisy conditions it looses most of its performance. To quantify this, in [7] the authors estimate an upper bound to the Haitsma fingerprint performance under noisy conditions and validate such model with real data. In [8] they perform a complete analysis of the Haitsma fingerprint under strong noisy conditions and propose the use of a power mask during matching process to improve its performance.

Recently we proposed MASK (**M**asked **A**udio **S**pectral **K**eypoints) [1] as an efficient and effective representation of the audio signal, improving upon some of the shortcomings observed in previously proposed fingerprints. For instance, in MASK the acoustic signal is represented by a series of compact low-dimensional binary fingerprints, obtained by encoding the energy flow around most relevant spectral energy peaks (whose desired density can be set as a parameter). Both fingerprint extraction and matching among fingerprints is very efficient and it was shown in [1] to be quite robust to most typical acoustic transformations, as well as useful at discriminating between acoustic signals. A known shortcoming of the initial MASK proposal is that it requires to manually define a spectro-temporal "mask" indicating what spectral region pairs to be compared. While a reasonable mask was proposed in [1] it was not proven whether it was optimal for the given test data. In addition, later tests distorting the input signal with additive Gaussian white noise showed that MASK's matching accuracy (as well as other fingerprints) still suffers in very noisy conditions.

In this paper we propose an extension of the MASK fingerprint composed of a data-driven optimization method to automatically derive the spectro-temporal mask to be optimal under certain acoustic conditions. In this paper we focus on optimizing for data affected by strong additive Gaussian noise and show how, while the original MASK and Haitsma [4] fingerprints get quickly degraded, the proposed data-driven MASK performs much better at low Signal-to-Noise Ratio (SNR) levels.

The rest of the paper is organized as follows: first, in Section 2 we review how the MASK fingerprint is generated. Then, in Section 3 we introduce the use of Mutual Information (MI) as a measure

L. Ondel was visiting Telefonica Research during the time of this work

of robustness. Next, in Section 4 we explain how we use MI as an optimization method to automatically select the spectral regions in MASK. The fingerprint is then tested and compared to others in Section 5. Finally, we conclude and propose some next steps.

2. MASK ACOUSTIC FINGERPRINT

In this section we review how the MASK acoustic fingerprints [1] are extracted. The key idea of MASK is to encode the energy flow around salient points of the time-frequency spectrum using a binary descriptor. In order to increase robustness against noisy spectra, regions of the spectrum (instead of single time-frequency bins) are used to set each bit in the fingerprint. The MASK extraction process can be summarized in four steps as shown in figure 1, and described below.

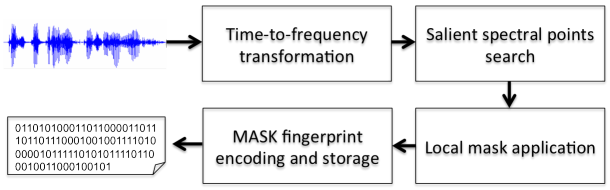


Fig. 1: MASK fingerprint extraction block diagram.

2.1. Time-to-Frequency Transformation

First, the signal is band-pass filtered between 300Hz and 3KHz, sampled in segments of 100 ms at a fixed sampling rate of 10 ms. Then, we apply a Hamming window to each segment, compute the FFT and apply a MEL filter bank of 21 triangular filters to the spectrum with equispaced filters. Note that on the contrary of the traditional MFCC or PLP features, neither equal loudness preemphasis nor compression is applied in this case. We repeat this process for all the signal to obtain a spectrogram representation used in the next step.

2.2. Selection of Salient Key Points

Next, we search for the spectral energy peaks in the short term spectrum. Like in previous works [3], we have observed that spectral energy peaks are robust to typical audio distortions and transformations, which makes them good "anchor" points to base the fingerprint on. In this work we consider a salient key point as a point in the time-frequency plane whose energy is greater than the energy of its direct neighbors (both in time and frequency). In order to limit the number of selected salient key points we apply a post-detection filtering to select only the peaks whose energy stays above a given temporal masking threshold defined according to its distance to the previously selected peak in the same frequency band. The threshold is defined as follow:

$$Thr[n] = \alpha^{\Delta t} E[n-1] e^{-\frac{(\Delta t)^2}{2\sigma^2}}$$

where Δt is the distance in frames between the previous spectral peak and the considered one, $E[n-1]$ is the energy of the previously selected peak and α, σ are two free parameters used to set the threshold falling rate and its width, respectively. Throughout all our experiments we set $\alpha = 0.98$ and $\sigma = 40$. In addition, because of the fixed size of the mask used in the next step, spectral peaks that are detected in the top or lower-most spectral bands are discarded to avoid edge problems in the mask.

2.3. Spectrogram Masking Around Salient Key Points

Similarly to [4] we encode the information in the spectrogram by using a set of binary comparisons between spectral values. To implement this, in MASK we define a spectro-temporal "mask" centered around each selected key point. Such mask identifies pairs of spectral regions (i.e. groups of spectro-temporal energy points in the spectrogram) to obtain a good representation of how energy flows around each key point. In practice, the energy values inside each region are averaged and compared to define each bit in the final fingerprint. This process can also be seen as similar to the widely used Δ features in speech but in this particular case the Δ are computed along the time and/or the frequency axes and at different scales (i.e. the size of the regions).

Choosing the regions wisely is fundamental to obtain a robust and highly discriminant final MASK fingerprint. Previously [1] we had defined the regions by hand trying to optimally cover the space around the key point. The resulting mask was thus derived from experience and not as a result of an optimization process to the target acoustic condition. As explained in Section 4, in this paper we propose a data-driven method to obtain an optimal mask according to a mutual information criterion. Figure 2 shows the mask we used in the experiments in this paper, optimized for low SNR signals. Position (0,0) in the mask is centered on the spectral salient point. The gray levels in the mask indicate how many times each particular time-frequency location was involved in defining any of the comparison regions, i.e. the importance of each point in the fingerprint. It is interesting to see that in this case both frequencies below and times before the salient point become most important.

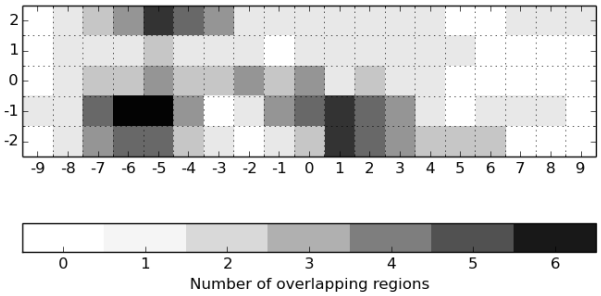


Fig. 2: Mask obtained by the proposed method. The darkness level indicates the number of overlapping regions.

2.4. Fingerprint Construction, Indexing and matching

Finally, the MASK fingerprint is obtained by joining together the information about the band location of the key points (in binary form) and the set of bits obtained from applying the mask. On the one hand, the band location is a fixed length binary string that indicates in which filter bank the spectral peak appears. In our configuration we used 5 bits to encode 21 band locations. On the other hand, each difference between region pairs in the mask is stored as a single bit in the final fingerprint. In this paper we used 20 region pairs, thus accounting to a 25bits fingerprint.

Each fingerprint is then indexed into an inverted file index where the fingerprint acts as the key and each content field contains a list of all content IDs and time locations (in frames) where this fingerprint was found.

For this paper we implemented a simple matching algorithm between query and reference. For each query file (i.e. a distorted

version of a file whose fingerprints have been pre-indexed) we extract its fingerprints and retrieve for each one the matching content ID and its time within the file. Then a histogram of counts is built per reference content ID to obtain a matching score. The histogram counts how many query-to-reference time-differences are equal (i.e. Hist($\Delta_t = T_r - T_q$), where T_r and T_q are the times where reference and query contain the same fingerprint). The maximum of this histogram is used as the matching score.

3. MUTUAL INFORMATION AS A MEASURE OF ROBUSTNESS

Next we review the mutual information (MI) measure and explain how we use it as a measure of robustness in order to formulate a data-driven region selection method for MASK. Let X be a random variable whose realisations are the vectors $\mathbf{x} = (x_1, ..., x_n)^T$ where each x_k represents the difference of energies computed, as explained above, from audio references. Let Y be another random variable whose realisations are defined as $\mathbf{y} = (f(x_1), ..., f(x_n))^T$ where f is a quantization function shown in Eq. 1.

$$f(x_k) = \begin{cases} 0 & \text{if } x_k < 0 \\ 1 & \text{if } x_k \geq 0 \end{cases} \quad (1)$$

Thus $\mathbf{y}$ is a multidimensional binary variable. In a similar way, let the random variables X' and Y' be the counterpart of X and Y for the audio queries. We make the assumption that audio queries are transmitted via a Gaussian channel at a specific SNR. Given that the random variable X is defined as a linear transformation of the time varying spectrum (averaging and taking the difference) the random variable X' is also defined as a Gaussian channel with input X, as $X' = X + N$, where N is the noise sample from a normal distribution with mean 0 and is independent of X. It is important to note that the particular realizations of Y (and consequently of Y') greatly depend on the choice on the regions to be compared in building the fingerprint. Because we would like our fingerprint to be robust to general signal transformations and to noise we wish to select the configuration θ that yields the lowest error for a given SNR.

Note that the naive solution to the problem of selecting the regions configuration giving the lowest error between Y and Y' could result in an inefficient robust solution. Indeed if Y is composed of robust variables having a low entropy, the fingerprint configuration will be robust but will carry little information, and hence resulting into a weak fingerprint. Therefore our problem needs to be thought of as a constrained minimization problem where we try to minimize the error while keeping the entropy of Y as high as possible. Another remark is that the configuration in the fingerprint also determines the dimension of the variable Y, hence our goal here is to find a subset of bits (each bit will be one dimension of Y) that satisfies the aforementioned constraints.

The Mutual Information (MI) is a measure of dependence between two random variables and it is defined as

$$I(X;Y) = H[X] + H[Y] - H[X,Y] \quad (2)$$

Where $H[X]$ is the entropy of X, $H[Y]$ the entropy of Y and $H[X,Y]$ is the joint entropy of X and Y. This measure is widely used in features selection for classification where it offers a way to quantify how informative is a feature when inferring a class. An important property of this measure is that $I(X;Y) \geq 0$ always, equality being reached if and only if the two random variables are independent. For instance, for a low SNR Y' will tend to be independent of Y thus leading to MI close to zero whereas at high SNR Y' will be correlated to Y, resulting in a high mutual information. The regions selection problem can be therefore turned into a problem of finding the configuration that maximizes the mutual information between Y and Y'. This interpretation is possible thanks to the Gaussian channel assumption and because Y and Y' are (multivariate) binary variables. A situation where this method would yield a poor solution would be if Y and Y' are inverted, i.e. if $Y = 101...$ then $Y' = 010...$ and vice-versa. Such situation would yield high mutual information while making the fingerprint completely inefficient. This is however an unrealistic case since adding Gaussian noise can only result in a decorrelation of the two variables. Note that using MI as a regions selection criterion respects the constraint that the final fingerprint configuration should have an entropy as high as possible to keep the search of fingerprint in the database efficient.

4. DATA-DRIVEN REGION SELECTION IN MASK

In this section we describe the method followed to select the bits (i.e. the region-pairs comparisons) of the MASK fingerprint so that, for a given noise level, those bits are robust and are not altered between clean and noisy signals.

First of all, we generate all possible fingerprints in a training set. The fingerprints are extracted as defined in Section 2 but obtaining all possible comparisons of block pairs of size 3x1, 3x2 and 4x2 (we limited our search to these block topologies to constraint computational cost). Block pairs overlapping each other for more than 50 % of their size were discarded. This produced a set of fingerprints of about 6500 bits each. In addition, we generated another set of fingerprints with the same configuration and the same audio data but with white noise added at an SNR of 0db. Following the model described above, Y will be the multidimensional random binary variable which takes as values the fingerprints from the clean data (the references) and Y' the multidimensional random binary variable which takes as values the fingerprints from the noisy data (the queries). We modeled the probability of these two variables with a mixture of Bernoulli distributions [9] as shown in Eq. 3

$$p(\mathbf{y}, \mathbf{y'}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\mu'}) = \sum_k \pi_k p(\mathbf{y}, \mathbf{y'}|\boldsymbol{\mu}, \boldsymbol{\mu'}) \quad (3)$$

with

$$p(\mathbf{y}, \mathbf{y'}|\boldsymbol{\mu}, \boldsymbol{\mu'}) = \prod_{i=1}^{D} \mu_i^{y_i}(1-\mu_i)^{(1-y_i)} \prod_{i=1}^{D} {\mu'}_i^{y'_i}(1-\mu'_i)^{(1-y'_i)}$$

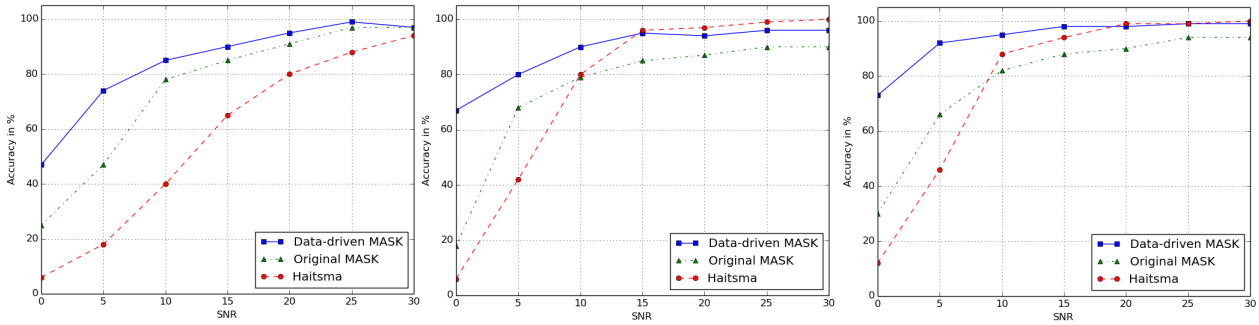
We used 100 components in the mixture and we estimated the parameters with the Expectation-Maximization (EM) algorithm. The number of components directly reflects the capacity of the model to capture the covariance across bits. Because of the high number of bits of the fingerprint, we found that 100 components was a good trade-off between computational requirements and accuracy of the model. The bit selection process (i.e. the selection of the most relevant spectral pairs) is then performed as described in Alg. 1.

Note that evaluating the MI for a high dimensional vector can quickly become computationally infeasible. To avoid this computational barrier we iteratively ran several times the algorithm in blocks of 5 bits and then concatenated those chosen bits all together yielding an approximate solution of our problem.

5. EXPERIMENTAL EVALUATION

5.1. Experimental Setup

In this section we evaluate the proposed fingerprint and compare it to the original MASK [1] and the Haitsma [4] fingerprints.



(a) T4 = mix with speech. (b) T5 = T4 + multiband compression. (c) T6 = T5 + bandpass filter.

Fig. 3: Accuracies for 10 seconds queries for different SNR levels and acoustic transformations(T)

Algorithm 1 data-driven bit selection process based on MI

Input: $\mathbf{y}, \mathbf{y'}$
Output: set of selected bits
$i \leftarrow (\text{argmax}_i \, \text{MI}(y_i, y'_i))$
$\mathbf{y} \leftarrow [y_i]$
$\mathbf{y'} \leftarrow [y'_i]$
for number of desired bits **do**
$i \leftarrow (\text{argmax}_i \, \text{MI}([\mathbf{y}, y_i], [\mathbf{y'}, y'_i]))$
$\mathbf{y} \leftarrow [\mathbf{y}, y_i]$
$\mathbf{y'} \leftarrow [\mathbf{y'}, y'_i]$
end for
return $\mathbf{y}$

Tests were performed on a database, composed of 150 audio excerpts of 30 seconds each, obtained from various sources (approx. 45% from TV and movies, 40% from music files and 15% from clean speech). Each file in the database was modified using the acoustic transformations defined in the NIST-TRECVID benchmark evaluations [10], obtaining 7 versions of each file: the original file (transform 0) + 6 transformations (transforms 1-6).

After fingerprints are extracted for all files in all conditions, the task consists on retrieving the correct "clean" file given one of the "transformed" files. Accuracy was measured by checking whether the best match corresponds to the same file. After preliminary experimentation, we observed that both MASK and Haitsma fingerprints performed very well on all transformations (obtaining matching accuracies between 98% and 100%). To make the task harder we further deteriorated all signals by adding Gaussian noise at SNR levels between 0 and 30dB (depending on the test). In all cases we used a single data-driven MASK, trained on the original data (transform 0), not specifically adapted to any of the audio transformations.

5.2. Experimental Results

Table 1 shows results of searching for the correct clean audio file by using each of the transformed files (around 30s/query) at SNR = 0 dB. We see that the data-driven MASK is almost not affected by the noise whereas we observe severe performance degradation on the Haitsma fingerprint and on the original MASK.

To further investigate the effects of noise on the acoustic fingerprints, we kept the same search index but we shortened the length of the query to 10 seconds and computed the matching accuracies at different SNR values, ranging from 0 to 30db. Figures 3a, 3b and 3c show the results for the acoustic transformations 4, 5 and 6, corresponding to the most challenging cases.

Table 1: Accuracy in percent of the MASK system for the different audio transformations and SNR = 0 dB

Transform	Haitsma	Original MASK	data-driven MASK
1	19.04	81.75	**100**
2	58.50	92.56	**100**
3	78.91	96.62	**100**
4	6.12	45.27	**100**
5	11.56	52.02	**94.55**
6	22.44	66.89	**97.95**

We see that Haitsma fingerprint performs quite well for high SNR values but it quickly degrades in noisy signals. On the contrary, MASK does not degrade so abruptly because energy comparisons are performed between bigger spectro-temporal regions (not only on very few spectral points). When automatically training the mask using the proposed data-driven algorithm we obtain an even more robust fingerprint, whose performance does not collapse at low SNR rates. Note that although the mask has been trained and tested on the same data, the high volume of fingerprints extracted (in the order of millions) avoids the system to overfit to the data, but instead learns how low SNR levels affected the data.

6. CONCLUSIONS AND FUTURE WORK

In this paper we propose an improvement to MASK, a recently proposed acoustic fingerprint that has been shown to be effective at compactly representing an acoustic signal using binary descriptors. In particular, we propose a data-driven method to define the structure of the MASK fingerprint to be optimal for a desired target acoustic condition. To do so, we use mutual information as an optimization criterion to select which spectral region pairs around each salient spectral peak are most robust when the signal is deteriorated. We test the algorithm optimizing for noise at 0dB SNR and applying it in an audio search task using signals with various distortion and SNR levels. Results show that the proposed data-driven MASK clearly outperforms the original MASK implementation as well as an implementation of the well known Haitsma fingerprint, specially at low SNR levels. Our next steps include the understanding of how different sorts of audio can affect the automatic selection of spectral regions. We also plan to train and test the fingerprint with bigger datasets and for particular use-case scenarios (e.g. for music-only, TV or speech-only).

7. REFERENCES

[1] Xavier Anguera, Antonio Garzon, and Tomasz Adamek, "Mask: Robust local features for audio fingerprinting," in *Proc. International Conference on Multimedia and Expo (ICME)*, 2012, pp. 455–460.

[2] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma, "A review of audio fingerprinting," *J. VLSI Signal Process. Syst.*, vol. 41, no. 3, pp. 271–284, Nov. 2005.

[3] Avery Li chun Wang and Th Floor Block F, "An industrial-strength audio search algorithm," in *Proceedings of the 4 th International Conference on Music Information Retrieval*, 2003.

[4] Jaap Haitsma and Ton Kalker, "A highly robust audio fingerprinting system.," in *ISMIR*, 2002.

[5] Christopher J. C. Burges, John C. Platt, and Soumya Jana, "Distortion discriminant analysis for audio fingerprinting.," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 165–174, 2003.

[6] Shumeet Baluja and Michele Covell, "Audio fingerprinting: Combining computer vision and data stream processing," in *In International Conference on Acoustics, Speech, and Signal Processing (ICASSP*, 2007, pp. 213–216.

[7] Flix Balado, Neil J. Hurley, Elizabeth P. McCarthy, and Guenole C. M. Silvestre, "Performance analysis of robust audio hashing.," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 2, pp. 254–266, 2007.

[8] B. Coover and Jinyu Han, "A power mask based audio fingerprint," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1394–1398.

[9] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[10] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.