A POWER MASK BASED AUDIO FINGERPRINT

Bob Coover, Jinyu Han

Gracenote, Emeryville, CA {bcoover, jhan}@gracenote.com

ABSTRACT

The Philips audio fingerprint[1] has been used for years, but its robustness against external noise has not been studied accurately. This paper shows the Philips fingerprint is noise resistant, and is capable of recognizing music that is corrupted by noise at a -4 to -7 dB signal to noise ratio. In addition, the drawbacks of the Philips fingerprint are addressed by utilizing a "Power Mask" in conjunction with the Philips fingerprint during the matching process. This Power Mask is a weight matrix given to the fingerprint bits, which allows mismatched bits to be penalized according to their relevance in the fingerprint. The effectiveness of the proposed fingerprint was evaluated by experiments using a database of 1030 songs and 1184 query files that were heavily corrupted by two types of noise at varying levels. Our experiments show the proposed method has significantly improved the noise resistance of the standard Philips fingerprint.

Index Terms— Audio Fingerprint, Music Recognition

1. INTRODUCTION

An audio fingerprint is a compact representation of an audio signal that can be easily stored, indexed, and used for comparisons between audio documents in a database of very large scale. In recent years many audio fingerprinting systems have been proposed. The well-known fingerprinting algorithms include the Philips fingerprint which encodes the spectral differences[1], the Shazam system which encodes the spectral peaks[2], the RARE fingerprint based on Distortion Discriminant Analysis[3], and the Waveprint which uses the wavelet transform[4].

The Philips Fingerprint [1] has been used commercially by Gracenote for many years. One challenge facing the authors is to improve the noise robustness of the Philips Fingerprint without changing the underlying fingerprint. The proposed method in this paper addresses this challenge.

The Philips fingerprint consists of two main parts: a fingerprint representation and a fingerprint index. Most extended works such as [5, 6] studied the indexing part of the algorithm. This paper instead focuses on the fingerprint representation itself. While the work in [7, 8, 9] can be viewed as an improvement to [1] by using localized binarization and more robust filters, these methods change the underlying fingerprint feature. This would require re-computing all the existing fingerprints from millions of audio documents for any services that are currently using the Philips fingerprint. The proposed extension in this work not only improves its noise robustness, but also can be seamlessly incorporated into the Philips fingerprint scheme so that any existing fingerprints can still be used. We achieve this by adding a simple extra matching step under the original Philips framework.

The Philips fingerprint is well suited to find a match in the presence of codec disortion, GSM encoding, compression, filtering, echo, and the addition of low level noise[1]. However, its robustness against noise that is equally loud or louder than the underlying music is largely unknown. The authors found no existing work in this area. Unfortunately, music identification services largely deal with queries that have dominant noise from many sources (wind noise, car noise, talking and crowd noise, etc.). In this work, we conduct extensive experiments to show that the Philips fingerprint is capable of recognizing music that is corrupted by noise and improve further on this ability with the Power Mask method.

Our method improves the noise resistance of the Philips fingerprint by addressing its disadvantage when compared to the spectral peak-based algorithms [2, 10]. Spectral peak based algorithms try to match audio documents by considering peaks in the audio spectrum. On the other hand, the Philips algorithm is looking at all parts of the spectrum at all times. As a result, if an audio signal is corrupted by noise in parts of the audio spectrum where there is not much energy in the music, the Philips fingerprint tends to represent the noise rather than the music. This disadvantage is addressed in this paper.

This paper describes an addition to the Philips fingerprint that allows us to determine the part of the spectrum that is more relevant to the fingerprint. This new addition is a binary mask that encodes the region of the Philips fingerprint that contains the more noise resistant part of the audio. This mask is determined by a particular type of power measurement across the audio spectrum, so we call it a "Power Mask". By applying the Power Mask to the original Philips fingerprints during the matching process, the parts of the fingerprint that are more noise resistant are penalized to a larger degree if a mismatch occurs, which improves the fingerprint's robustness to dominant noise significantly. The proposed Power Mask is easy to calculate, small in size, efficient to apply during the matching process, and can be used with the Philips' indexing method [5, 6].

The remainder of this paper is organized as follows. In section 2 we describe the Philips fingerprint and motivate the Power Mask based scheme by describing the weakness of the Philips fingerprint. In section 3 we introduce the Power Mask based fingerprint as our main contribution. In section 4, the effectiveness of the proposed method is evaluated by experiments using heavily corrupted query files. Finally we conclude the paper in section 5.

2. THE PHILIPS FINGERPRINT

The standard Philips fingerprint [1] is created by taking the spectrum of an audio signal once every 11.6 milliseconds. The overlapping spectrum frames have a length of 0.37 seconds. This spectrum is then grouped into 33 bands that are logarithmically spaced from 300 Hz to 2000 Hz. A fingerprint for each spectrum frame, referred to as a sub-fingerprint, is a 32-bit number that represents the sign of energy differences along the frequency and time axes:

$$F(n,m) = \begin{cases} 1 & \textbf{if } P(n,m) - P(n,m+1) - \\ & (P(n-1,m) - P(n-1,m+1)) > 0 \\ 0 & \textbf{otherwise} \end{cases} \quad (1)$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

where $F(n, m)$ is the m-th bit of the sub-fingerprint at time frame n and $P(n, m)$ denotes the power of the band m at time frame n of the audio spectrum. Sub-fingerprints are grouped together sequentially to form 3 second blocks, which consist of 256 sub-fingeprints.

During the matching process, two audio signals of 3 seconds in length are declared a match if the Bit Error Rate, (BER), between the two derived fingerprint blocks is below a certain threshold θ. In [1], $\theta = 0.35$ is shown to produce a very low false positive rate. The BER between two fingerprint blocks Q and R is calculated as follows:

$$BER = \frac{\sum_{n=1}^{256} \sum_{m=1}^{32} (R(n, m) \oplus Q(n, m))}{TB} \quad (2)$$

where $\oplus$ is the XOR logic operation and TB is the total number of bits in a fingerprint block, ($256 \times 32 = 8192$).

This paper focuses on improving the representation and matching process of the Philips fingerprint. The indexing part of the Philips fingerprint is out of the scope of this work. See [1] for a complete explanation.

The Philips fingerprint encodes the sign of the power difference for each band, which has proven to be very robust to many kinds of light distortion [1]. However, the Philips fingerprint only keeps the sign of a power difference, completely discarding the information containing the amount of the power difference. This drawback is addressed in Sec. 3 by introducing a Power Mask to the matching process, which weights different bits based on their relevance to the fingerprint.

3. THE POWER MASK BASED FINGERPRINT

The region of the audio spectrum where the power differences are close to zero is the part of the fingerprint that is most vulnerable to external noise. We call this region of the spectrum the "weak-bit region" because the fingerprint bits calculated from this region, (referred to as the "weak bits"), are easily corrupted by noise, (i.e. fingerprint bits can be changed by adding a small amount of noise energy). Similarly, fingerprint bits extracted from the "strong-bit region", where the absolute power differences of the audio spectrum are large, are referred to as "strong bits".

The concept of weak bits has been utilized in [1] as a candidate gathering method in the index of the Philips fingerprint to generate a list of probable candidates. By flipping the most unreliable bits in the query, candidates with either bit pattern can be pulled in from the database. However this information about the reliability of fingerprint bits has never been utilized during the matching process. We incorporate a variant of this information in the matching process for the proposed fingerprint scheme.

When noise becomes dominant, weak bits in the Philips fingerprint tend to be representative of the external noise rather than the music signal itself. In this case, a significantly higher BER between a noise-corrupted query and its target will result in a mismatch. By giving the same weights to the "weak bits" and "strong bits" during the matching process, the reliability of the bits is not taken into consideration. We now discuss how to correct this.

3.1. The Power Mask

There are many ways to determine if a bit $F(n; m)$ plays a relevant part in the fingerprint. Peaks are a common measurement in many fingerprints, but in this work we use strong bits. This is a bit that has a large power difference across both time and frequency. For each sub-fingerprint of 32 bits, a Power Mask is a second 32-bit number, which encodes a strong bit by 1 and a weak bit by 0:

$$PM(n, m) = \begin{cases} 1 & \textbf{if } F(n, m) \text{ is a strong bit} \\ 0 & \textbf{if } F(n, m) \text{ is a weak bit} \end{cases} \quad (3)$$

where $PM(n, m)$ is the Power Mask for the m-th bit of the sub-fingerprint at time frame n. Now let $DIF(n, m)$ denote the absolute power difference along the time and frequency axes at band m and frame n:

$$\begin{aligned} DIF(n, m) = &|P(n, m) - P(n, m + 1) \\ &- (P(n - 1, m) - P(n - 1, m + 1))| \end{aligned} \quad (4)$$

We set the strong bits to be the bits that correspond to the largest T absolute power differences, $DIF(n, m)$, of the power spectrum at time n. In this work, we use $T = 24$. This number is further explained in Sec. 4.1. The reason for using a fixed number of strong bits per sub-fingerprint is explained in Sec. 3.2.

In our implementation, a Power Mask is an array of 32-bit unsigned integers each of which has a corresponding sub-fingerprint. Each 32-bit unsigned integer has a bit set to 1 for each of the T frequency bands that corresponds to a strong bit in the sub-fingerprint.

The Power Mask is not created for a query signal, but is created on the reference side only. The reason for this is that the reference signal, which does not contain any noise or distortions, has the true representation of the absolute spectral differences that are characteristic of the music signal we are trying to match to. On the other hand, a query could have additional noise or distortions at any part of the spectrum at any time during the fingerprint. So making a Power Mask from the query would not accurately show which bits are relevant to the fingerprint. If the power differences from the spectrum of the external noise are large, the Power Mask will set strong bits that represent the noise instead of the music signal.

3.2. Matching with the Power Mask

Given two fingerprint blocks Q and R that are derived from a query and a reference signal respectively, and the Power Mask PM that is calculated from the reference signal, the BER between Q and R is calculated using Eq. 5:

$$\begin{aligned} BER = &\frac{\sum_{n=1}^{256} \sum_{m=1}^{32} \alpha \times (R(n, m) \oplus Q(n, m)) \& (\neg PM(n, m))}{\alpha \times WB + \beta \times SB} \\ &+ \frac{\sum_{n=1}^{256} \sum_{m=1}^{32} \beta \times (R(n, m) \oplus Q(n, m)) \& PM(n, m)}{\alpha \times WB + \beta \times SB} \end{aligned} \quad (5)$$

where $WB = 256 \times (32 - T)$ and $SB = 256 \times T$ are the number of weak bits and strong bits respectively in a fingerprint block, α and β are the penalty weights given to the weak bits and strong bits, and $\&$ and $\neg$ are the logic operations AND and NOT respectively.

A strong bit is more noise resistant than a weak bit due to its large absolute power difference. A mismatch at a strong bit of the reference fingerprint R lends more weight to the assumption that R is not a match to the query fingerprint Q. Based on this assumption, we set the ratio for $\frac{\beta}{\alpha}$ to be greater than 1, giving a larger penalty to the mismatches occurring at the strong-bit locations.

Similar to Philips, two audio signals of 3 seconds long are considered a match if the BER calculated using Eq. 5 between the two derived fingerprint blocks is below a certain threshold θ.

In this work, the number of strong bits per sub-fingerprint is a fixed number, T. However, this does not need to be the case. An adaptive threshold can be applied to each time slice of spectrum that

allows the mask to vary the number of strong bits per sub-fingerprint. By only selecting the number of bits that have an absolute difference value that is greater than an adaptive threshold, slightly better results have been obtained as compared to setting a fixed T. However, this variable number of strong bits per sub-fingerprint comes at the expense of efficiency, which is explained below.

The match logic for the Philips fingerprint looks for a BER under a certain threshold between the query and reference fingerprint blocks. According to Eq. 5, a BER is calculated by dividing the weighted number of mismatched fingerprint bits between the query and reference fingerprint block by the weighted total number of fingerprint bits in a fingerprint block. By fixing the number of strong bits per sub-fingerprint to a constant value T, the weighted total number of fingerprint bits in a fingerprint block becomes a constant. In this way, a BER can be calculated quickly by multiplying the weighted number of mismatched bits with a constant.

On the other hand, if we allow the number of strong bits to vary from sub-fingerprint to sub-fingerprint, we have to count the weighted total number of strong bits and weak bits in each fingerprint block, and divide by this number for every BER calculation. This is not a large concern when performing a single comparison, but when we perform this operation millions of times to locate the correct match from among thousands of possible candidates returned by the Philips indexing method, this becomes a serious performance hit. Dividing and bit counting using the variable number of strong bits, even when using hardware calls or the popcount algorithm[11], takes significantly more cycles than the single multiplication that is achieved with a fixed number of strong bits per sub-fingerprint.

4. EXPERIMENT

We now compare the proposed Power Mask fingerprint to the standard Philips fingerprint. The effectiveness of the Power Mask based fingerprint is demonstrated by experiments using a database of 1030 songs and 1184 query files.

4.1. Experiment Setup

The reference data set is made of approximately 60 hours of 1030 music pieces that are chosen from popular music, heavy metal, rap, country, classical, jazz, and international artists. The query test set is made from 37 song fragments from these same genres.

Many works [1, 5, 6] have shown that the Philips fingerprint is well suited to match with light distortions and noise. We are interested in the robustness of the Philips and the Power Mask based fingerprint against dominant noise. So each query song fragment has been corrupted by varying levels of two types of noise. The first type of interfering noise consists of a group of people talking along with background music and ambient noise that was recorded in a party environment. We refer to this type of noise as "crowd noise". The second type of noise is "pink noise", simulating steady state wind or car engine noise. The crowd and pink noise were chosen in our evaluation since they are most representative of the interfering noise that is experienced with real-world mobile queries submitted to Gracenote. They are considered the main source of error in a mobile device fingerprint match.

Each 15-second query fragment, which corresponds to 5 fingerprint blocks, is mixed with two types of noise at a Signal to Noise Ratio (SNR) ranging from 0 to -15 dB at one decibel decrements. Note that at 0 dB SNR, the distortion introduced by the noise is already very noticeable, beyond any codec or filtering distortions studied in many existing works. In most cases, SNR of -10 dB or worse drowns out the music to such a degree that the song is barely audible to humans, simulating the worst case scenario in music identification. Each type of noise, crowd or pink noise, is added at the 16 different SNR levels to the original music fragment, making a total of 16 different probe queries for each noise type and for each of the 37 music fragments. Thus, the total number of query files is 1184.

Each corrupted query signal is fingerprinted using the Philips method. Each reference audio document is fingerprinted with the proposed Power Mask fingerprint. Since a Power Mask is just an additional 32-bit mask that comes in with each Philips sub-fingerprint, this fingerprint set can be used for the standard Philips match as well.

As mentioned in Sec. 3, the concept of a weak bit has been utilized in [1] as a candidate gathering method for the Philips fingerprint. Given a sub-fingerprint extracted from a query, a number of weak bits are permuted to generate a list of probable candidate song tracks from the reference database. Philips and Gracenote experiments confirm that for low-quality audio query files, using 14 weak bits per sub-fingerprint achieves good candidate searching performance. With the Power Mask based fingerprint, experimental evidence shows that the number of strong bits needs to be even higher, at 24, if we are going to use a single static number.

For a query probe of 15-seconds, 5 consecutive fingerprint blocks, $F_Q := (Q_1, Q_2, Q_3, Q_4, Q_5)$, are extracted. Each fingerprint block $\{Q_i : i = 1, ..., 5\}$ consists of 256 sub-fingerprints representing 3 seconds of audio each. The identification task is to find from a database of reference tracks an audio document with a fingerprint representation of $\{R_j : j = 1, ..., K\}$ and a time offset $t \in \{1, ..., K\}$, such that five consecutive fingerprint blocks $F_{R_t} := (R_t, R_{t+1}, ..., R_{t+4})$ are similar to F_Q.

Given a query fingerprint F_Q and a 5-block reference fingerprint F_{R_t} at the time offset t, 5 BER values $\{B_{t+j-1} : j = 1, ..., 5\}$ are calculated, using Eq. 2 for Philips or Eq. 5 for the Power Mask, between each query fingerprint block Q_j and reference fingerprint block R_{t+j-1} in a time-aligned manner. A score S_t is given to the 5-block reference fingerprint F_{R_t} as follows:

$$S_t = \begin{cases} \min_j B_{t+j-1} & \textbf{if } \min_j B_{t+j-1} \leq 0.35 \\ \bar{B} & \textbf{else if } 2 \text{ or more BERs} \in (0.35, 0.43] \\ +\infty & \textbf{otherwise} \end{cases} \tag{6}$$

where $\bar{B}$ is the average value of the two or more BERs that are in the interval of $(0.35, 0.45]$.

According to Eq. 6, a 5-block reference fingerprint will have a non-infinity score if and only if one of its five fingerprint blocks has a BER below the threshold $\theta = 0.35$ as suggested by [1], or at least two fingerprint blocks have a BER below $\theta = 0.43$ based on experimantal evidence in running a production service.

An overall score for a reference song is the lowest score obtained from one of its 5-block fingerprints. The song with the lowest noninfinit score is returned as a match to a query probe. It is considered as a correct match if the music in the noisy query is from the returned song track. Otherwise, the returned song track is considered a false positive. If no track with a non-infinite score is returned, the fingerprint has failed to find a match to the query.

Since this work focuses on the underlying fingerprint and matching process, instead of the indexing method, the experiment is conducted by running a brute force match of a query against all the 1030 reference tracks. However, the candidate gathering method which uses the standard Philips fingerprint index method, can still be used with a very large database. The power mask only affects the final brute force match.

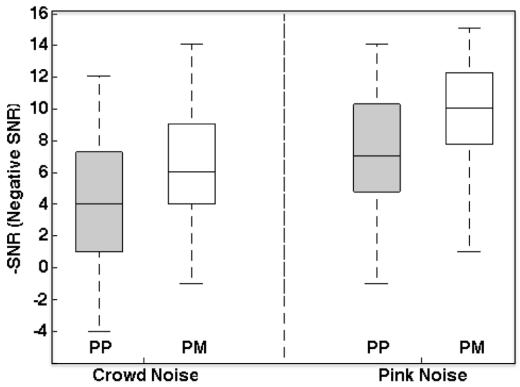

Fig. 1. The averaged Signal to Noise Ratio (SNR) below which the fingerprint fails to make a correct match. For illustration purpose, **negative SNR** ($-$SNR), which is equivalent to Noise to Signal Ratio, is used to label the Y-Axis. The higher the "$-$SNR" is, the more robust a fingerprint is against noise. PP stands for Philips. PM stands for Power Mask.

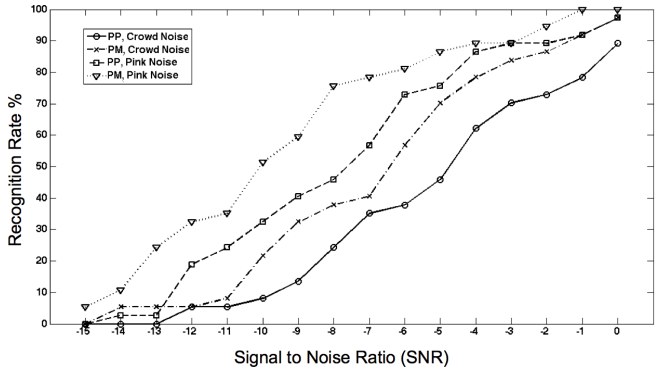

Fig. 2. Recognition Rate of the Philips fingerprint and the proposed Power Mask at different Signal to Noise Ratio (SNR) against a reference database of 1030 song tracks. X-Axis is the SNR. Y-Axis is the Recognition rate. PP stands for "Philips". PM stands for "Power Mask".

4.2. Experiment Results

To investigate the Power Mask's robustness against noise, we ran a fingerprint match, as described in Sec. 4.1, on each of the 1184 query song fragments with varying noise levels, and compared the results to standard Philips.

We first show the noise level above which the Philips fingerprint and the proposed method fail to make a correct match. This is averaged across the 37 query fragments. Fig. 1 shows the box plots of these results by Philips (PP) and the proposed Power Mask (PM) for crowd noise and pink noise. Each box plot is generated by the 1184 data points that represent the 37 query fragments. The lower and upper lines of each box show 25th and 75th percentiles of the sample. The line in the middle of each box is the sample median. The lines extending above and below each box show the extent of the rest of the samples, excluding outliers. The Y-Axis is the **negative Signal to Noise Ratio** ($-$SNR) which indicates the noise level over the music in the queries. The higher the "negative SNR" is, the louder the noise is compared to the music, and therefore the more noise resistant a fingerprint is. When the interfering noise is crowd noise, there are 3 cases with Philips and one case with the Power Mask where the noise must be below the music in order to be correctly matched. In Fig. 1 this is indicated by the negative portion of the Y axis.

It can be seen from Fig. 1 that the Philips fingerprint is able to recognize queries that are corrupted with crowd noise at an average SNR of -4.35 dB and with pink noise at an average SNR of -7.27 dB, respectively. This is a very impressive result considering the fact that the noise is approximately twice as loud as the underlying music at a SNR of -6 dB. The proposed Power Mask based method further improves the noise resistance to -6.22 dB for crowd noise and -9.14 dB for Pink Noise, respectively. On average, a 1.87 dB improvement was achieved for both the crowd noise and pink noise when using the Power Mask. Nonparametric paired sign tests show that the differences between these two methods are statistically significant at the 5% significance level for both types of noise.

In addition, we found that both methods show better performance against pink noise than crowd noise. Since the energy of crowd noise is less evenly distributed over the spectrum than pink noise, we believe that for the same signal to noise ratio, a fingerprint bit is more likely to be corrupted at the region of the spectrum where the energy of the crowd noise is concentrated.

Amongst the 592 query files that were corrupted with pink noise, 2 queries resulted in a false positive at a SNR of -9 dB and -14 dB, respectively, when using the proposed Power Mask method. Amongst the 592 query files that were corrupted with crowd noise, one false positive resulted at an SNR of -6 dB. No false positive was found using the standard Philips method. This implies that the default Philips BER threshold that we used for the Power Mask method is not the optimal one. The BER threshold might need to be moved lower, and this will be investigated in future work.

Fig. 2 shows the recognition rates using both fingerprint methods for the 37 query fragments mixed with both types of noise for SNRs ranging from 0 to -15 dB. As expected, the recognition rate goes up as the Signal to Noise Ratio increases. It also shows that both methods perform better with pink noise than with crowd noise, which is consistent with the result presented in Fig. 1. The Power Mask method achieves better results than the standard Philips fingerprint for both pink noise and crowd noise. With pink noise, the recognition rates of the Power Mask method and Philips are 100% and 97.30% at a SNR of 0 dB respectively, and they drop to 50% approximately at a SNR of -10 dB and -7.5 dB respectively. With crowd noise, Fig.2 shows the Power Mask with a 97.30% recognition rate at 0 dB, and a 50% rate at -6 dB. In comparison, the Philips method gets a recognition rate of 89.19% at 0 dB and a 50% recognition rate at approximately -5 dB.

5. CONCLUSIONS

In this paper we have studied the noise resistance of the widely used Philips fingerprint system and have shown that the Philips fingerprint performs well with significant levels of noise. This paper also introduces a method to weight the relevant bits in a Philips fingerprint, which results in the fingerprint being able to handle almost 2 decibels of additional noise. The proposed Power Mask method is easy to calculate, small in size , efficient to apply during the matching process, and can seamlessly work in conjunction with the original Philips scheme. Extensive experiments show that the proposed method improves the noise resistance of the Philips fingerprint significantly. Future work includes finding a better BER threshold for the Power Mask based fingerprint matching.

6. REFERENCES

[1] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. ISMIR*, 2002.

[2] A. Wang, "An industrial-strength audio search algorithm," in *Proc. ISMIR*, 2003.

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

[4] S. Baluja and M. Covell, "Audio fingerprinting: Combining computer vision & data stream processing," in *Proc. ICASSP*, 2007.

[5] H. Schreiber, P. Grosche, and Muller M., "A re-ordering strategy for accelerating index-based audio fingerprinting," in *Proc. ISMIR*, 2011.

[6] Qingmei Xiao, Motoyuki Suzuki, and Kenji Kita, "Fast hamming space search for audio fingerprinting systems," in *ISMIR*, 2011.

[7] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Proc. CVPR*, 2005.

[8] K. Kashino, A. Kimura, H. Nagano, and T. Kurozumi, "Robust search methods for music signals based on simple representation," in *Proc. ICASSP*, 2007.

[9] D. Jang, C. Yoo, S. Lee, S. Kim, and T. Kalker, "Pairwise boosted audio fingerprint," *IEEE Trans. Information Forensics and Security*, vol. 4, no. 4, pp. 995–1004, 2009.

[10] X. Anguera, A. Garzon, and T. Adamek, "Mask: robust local features for audio fingerprinting," in *Proc. ICME*, 2012.

[11] "Hamming weight," *http://en.wikipedia.org/wiki/Hamming_weight*.