electronics

MDPI

Article

Audio Fingerprint Extraction Based on Locally Linear Embedding for Audio Retrieval System

Maoshen Jia [1,*], Tianhao Li [1] and Jing Wang [2,*]

[1] Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; litianhao29@126.com

[2] School of Information and Electronic, Beijing Institute of Technology, Beijing 100081, China

* Correspondence: jiamaoshen@bjut.edu.cn (M.J.); wangjing@bit.edu.cn (J.W.); Tel.: +86-150-1112-0926 (M.J.)

Received: 3 August 2020; Accepted: 8 September 2020; Published: 10 September 2020

Abstract: With the appearance of a large amount of audio data, people have a higher demand for audio retrieval, which can quickly and accurately find the required information. Audio fingerprint retrieval is a popular choice because of its excellent performance. However, there is a problem about the large amount of audio fingerprint data in the existing audio fingerprint retrieval method which takes up more storage space and affects the retrieval speed. Aiming at the problem, this paper presents a novel audio fingerprinting method based on locally linear embedding (LLE) that has smaller fingerprints and the retrieval is more efficient. The proposed audio fingerprint extraction divides the bands around each peak in the frequency domain into four groups of sub-regions and the energy of every sub-region is computed. Then the LLE is performed for each group, respectively, and the audio fingerprint is encoded by comparing adjacent energies. To solve the distortion of linear speed changes, a matching strategy based on dynamic time warping (DTW) is adopted in the retrieval part which can compare two audio segments with different lengths. To evaluate the retrieval performance of the proposed method, the experiments are carried out under different conditions of single and multiple groups' dimensionality reduction. Both of them can achieve a high recall and precision rate and has a better retrieval efficiency with less data compared with some state-of-the-art methods.

Keywords: audio fingerprint; sub-regions; dimensionality reduction; audio retrieval

1. Introduction

With the development of internet technology, people are in an age of information explosion. The amount of audio information increased sharply in 30 years, which means people are exposed to various information every day [1]. For individuals, only a small fraction of it needs to be used, thus, how to obtain the content we want from the numerous audio information sources we are exposed to, namely audio retrieval, is an important problem worth studying [2,3].

In the field of audio retrieval, high-dimensional audio features reduce the searching efficiency because of their redundancy of information and large storage. These features are computed in the time or frequency domain [4]. For instance, short time energy, linear prediction coefficients (LPC), perceptual linear predictive (PLP), and Mel frequency cepstral coefficient (MFCC) are commonly adopted as audio features [5,6]. Additionally, the human auditory system (HAS) reflects similarities in auditory perception, so pitch is also used in related research. With the development of audio compression and storage technology, quantities of digital audio information have appeared on the internet which puts forward a higher requirement for the efficiency of audio features and gives rise to audio fingerprinting [7].

Audio fingerprinting can be seen as a short summary of audio signals that is a compact representation of acoustic characteristics [8]. Therefore an audio fingerprint with a limited number of

bits can represent an audio signal consisting of a large amount of data. It is extracted from the most important part of HAS, which has similarities of perception. Meanwhile audio fingerprinting shows higher search efficiency in audio retrieval works. It is widely used in music identification, copyright protection, integrity verification, and other aspects [9].

In a variety of algorithms, the most classic one is Philips' audio fingerprinting algorithm proposed by Haitsma and Kalker [10]. It computes fingerprints by comparing the energies in adjacent bands. The algorithm has high robustness to different kinds of signal degradations and the retrieval performance can be maintained when the audio signal is interfered. Moreover, it has a short granularity of about 3 s, which is the minimum length of an audio segment required for retrieval. The algorithm has low resistance to linear speed changes of the audio signal [11,12]. Another representative algorithm, named Shazam fingerprinting, proposed by Wang encodes the relationship between every two spectral maxima by using their time and frequency information [13]. It gains better robustness especially under noisy condition, but it uses a large number of bits in its encoding, which reduces the retrieval speed. In addition to the above two common methods, other researchers also introduced wavelet transform into fingerprint extraction [14,15]. This kind of method can provide time-frequency analysis of the signal which has a good resolution both in time and frequency [16]. It has a high recall rate in audio retrieval, but its accuracy drops when the audio signal is disturbed.

In recent years, Anguera combined the two methods mentioned above and proposed a novel fingerprint called Masked Audio Spectral Keypoints (MASK) [17]. The algorithm builds regions around selected salient spectral points and generates binary fingerprints by comparing the band energy. Although it has excellent performance of audio retrieval, its fingerprint dimension has room for further reduction. Thus, this paper optimizes the method by extracting a smaller fingerprint.

In this paper, an audio fingerprint extraction based on locally linear embedding (LLE) is proposed for audio retrieval system. The method uses LLE to reduce the number of sub-region energies and the reduction reflects in the dimension of fingerprints on account of comparing the energies [18]. In addition, a matching strategy based on dynamic time warping (DTW) is introduced into the system to gain a high resistance against linear speed changes of audio signals [19]. The two process above jointly realize an audio retrieval system and the framework includes two main parts: audio fingerprint extraction and audio fingerprint retrieval. The former consists of pre-process and fingerprint computation. The latter contains the establishment of database and fingerprint matching. The novel fingerprint has less data and improves the retrieval efficiency.

Figure 1 shows a common audio fingerprinting system including extraction, database establishment, and matching of fingerprints.

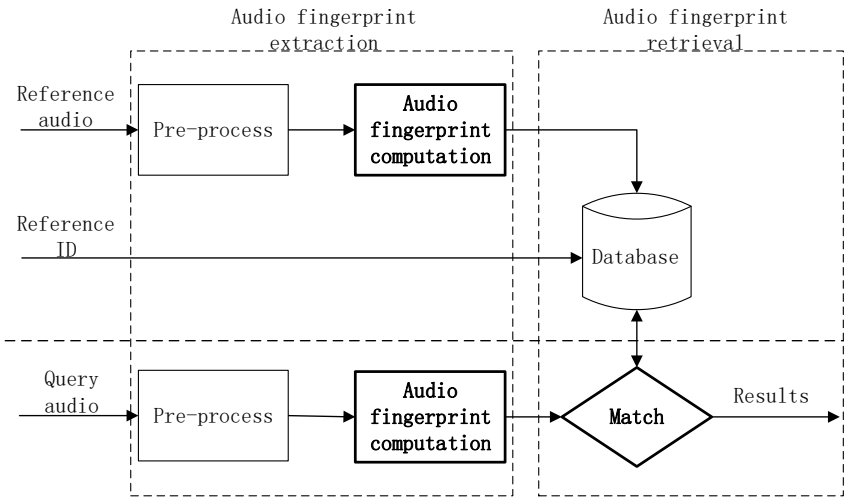


Figure 1. Audio fingerprinting system framework.

The rest of paper is structured as follows: Section 2 explains the proposed audio fingerprint extraction based on LLE. Section 3 describes the process of audio fingerprint matching by bit error rate (BER). Then, in Section 4 we show the experiments that we test the fingerprint and comparisons with other algorithm. Finally, we draw a conclusion in Section 5.

2. Audio Fingerprint Extraction Based on LLE

A typical audio fingerprint is encoded by comparing the energy of bands or sub-regions in the frequency domain. This paper introduces LLE into audio fingerprint extraction which maps the energy vector to a lower dimension so that smaller fingerprint can be obtained. The newly-acquired fingerprint can represent the audio signal uniquely with fewer bits and it is more efficient in audio retrieval. The extraction process is as follows: first the data is pre-processed, then band division is implemented in the frequency domain and a set of energies is computed. Finally, after being processed by LLE, the audio fingerprint is encoded by comparing the adjacent energy.

As seen in Figure 2 the dimensionality reduction consists of three steps. First, K frames with the nearest energy vector are selected for each frame according to the shortest Euclidean distance. Then, all the neighbors' reconstruction weights are computed. Finally, we obtain an energy vector mapping in a lower dimensional space. More details are described in the following sections.

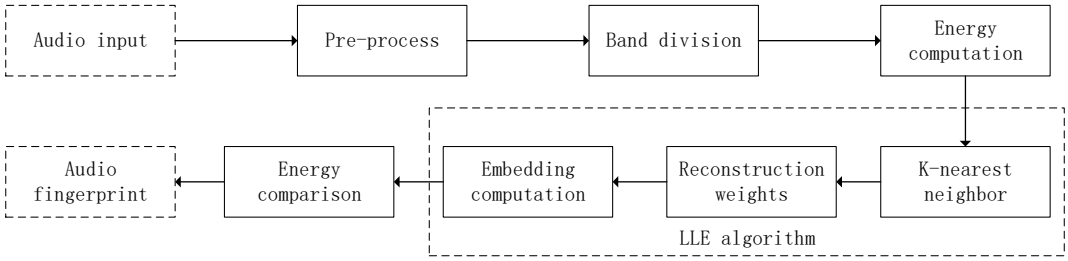


Figure 2. Block diagram of audio fingerprint extraction based on LLE.

2.1. Pre-Processing

As the audio signal has short-time stability, pre-processing steps, such as pre-emphasis, are required before the fingerprint extraction. First, we down-sample the input audio signal to 8 KHz because the sensitive frequency area of human is concentrated on 4 KHz and lower which can also reduce the amount of data involved in calculation. Audio fingerprinting needs high frequency resolution and wide time coverage so we use frame of length 100 ms overlapped with an interval of 10 ms.This framing strategy can reduce the error caused by the inconsistency of the boundary between test and sample and have a good retrieval efficiency even in the worst case (5 ms off). In order to reduce the poor impact of muted segments in the matching process we sift the original signal roughly using voice activity detection (VAD) with frame energy.

2.2. Audio Fingerprint Extraction

(1) Maximum Spectral Points Extraction

The discrete Fourier transform (DFT) is applied to the pre-processed audio signal and the spectrum is divided into 18 bands. In this work, we select a peak which has the maximum amplitude for each frame recording its current frame number and band number. To maintain a wide time coverage the number of selected peaks should be in the range of 70 to 100 samples.

(2) MASK Region Construction

After the peak selection step, two index sets of data is obtained, namely the frame index and frequency band index where the selected peak of each frame is located, which is equivalent to the coordinate information of each salient point. Then these data are used to construct MASK regions centered by each selected peak, so there should be one MASK region for each frame. The region covers

five frequency bands (one located band, two bands above, and two bands below) and the time span is 19 frames (one located frame, nine frames before, and nine frames after). It is worth mentioning that overlaps are allowed between regions with different peaks during the construction process. To prevent peaks from appearing at the band boundary, the first and last two bands are not used as a selection area so there is no case where a region is built out of scope.

Supposing that the peak is in n-th band of m-th frame and its MASK region is shown in Figure 3. Note that the frequency bands and time frames are distributed symmetrically. The MASK region is represented as a grid range of 5×19, and the ordinate has five frequency bands and the abscissa consists of 19 time frames.

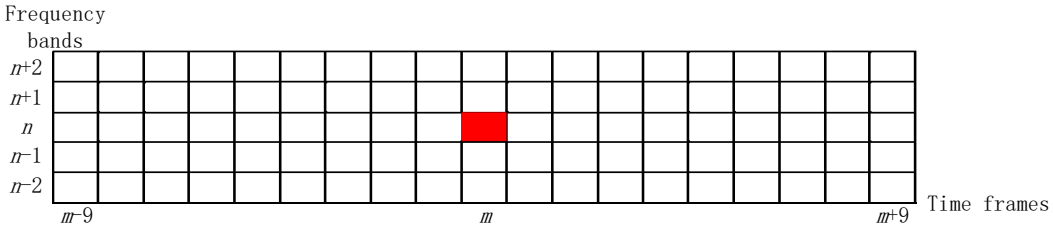
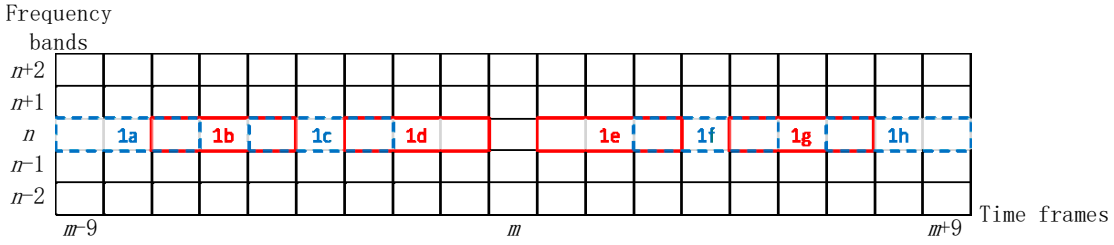


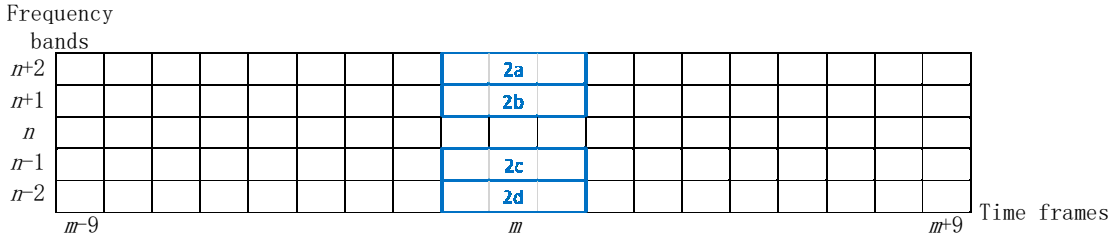
Figure 3. A MASK region centered by a selected peak (red block).

(3) Sub-region Division

The MASK region is divided into 4 groups of sub-regions as shown in Figure 4. The four groups of sub-regions are defined as follows: the first group is horizontal sub-regions which corresponds to 1a–1h; the second group is vertical sub-regions including 2a–2d; the third group represents central sub-regions with 3a–3d; the fourth group represent marginal sub-regions in the form of 4a–4h. It can be seen that the horizontal and marginal groups have eight sub-regions, while the vertical and central groups have four sub-regions. All of them correspond to the same number of energies which represented as L (L = 4 or 8). Then the energies of these sub-regions are calculated and the energy vector of a single group is denoted as $\boldsymbol{e}_m^{\mathbf{H}}$, $\boldsymbol{e}_m^{\mathbf{V}}$, $\boldsymbol{e}_m^{\mathbf{C}}$ and $\boldsymbol{e}_m^{\mathbf{M}}$ which means the energy vector of horizontal, vertical, central and marginal group in that order. We take $\boldsymbol{e_m} = [e_{m1}, e_{m2}, \ldots, e_{mL}]$ to uniformly present $\boldsymbol{e}_m^{\mathbf{H}}$, $\boldsymbol{e}_m^{\mathbf{V}}$, $\boldsymbol{e}_m^{\mathbf{C}}$, and $\boldsymbol{e}_m^{\mathbf{M}}$ for simplicity, where e_{ml} is the energy of l-th sub-region of m-th frame and the value of L depends on the group type.

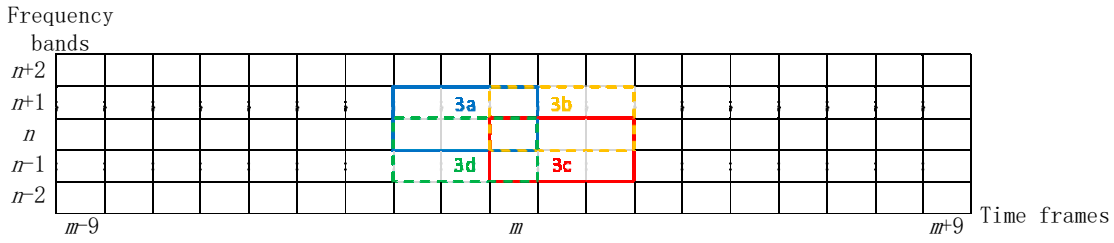


(**a**) Horizontal sub-regions.

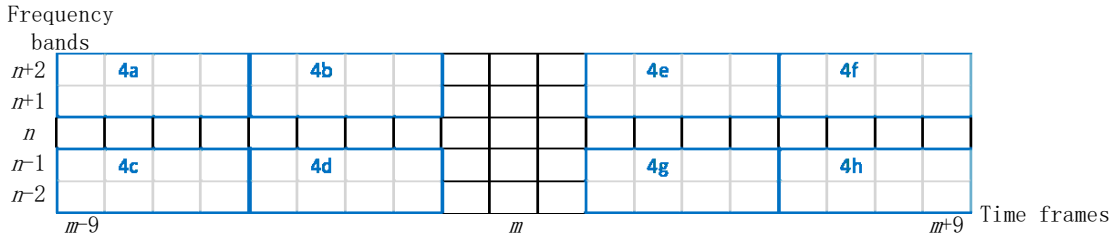


(**b**) Vertical sub-regions

Figure 4. *Cont.*



(**c**) Central sub-regions



(**d**) Marginal sub-regions

Figure 4. 4 layers of sub-region division.

(4) Audio Fingerprint Extraction based on LLE

The dimension of the energy vectors obtained above is still high and the redundancy can be reduced. Thus, in consideration of the relationship of the energies, this paper use LLE to reduce the number of energies of each group respectively and finally extracts smaller fingerprints. LLE algorithm is a typical nonlinear dimensionality reduction method, which considers that the observed data is actually a result of mapping from a low-dimensional space to a high-dimensional space. Due to the internal characteristics of the data, there is redundancy in some high-dimensional data which can be uniquely represented with less data. The algorithm needs to first obtain the reconstruction weights of the sample frame in the neighborhood, and then solve the mapping of the sample in the low-dimensional space according to the principle that the reconstruction weights remains unchanged equally to minimize the loss. The method in this paper fully considers the linear relationship between sample frames which corresponds to the time sequence structure of audio signals, and describes the essential characteristics of the data in the low dimension with the minimal feature loss. We reduce the dimension of the energy vector of audio signals belonging to the same group of sub-regions which also reduce the amount of energy involved in fingerprint calculation. The new energy vector is used to extract low dimensional fingerprints by comparing the adjacent energies.

(i) Selection of K nearest frames

The calculation is closely related to the group of its sub-regions. Therefore, the processed data is the energy vector composed of all the sub-regions' energies of an audio frame in a same group which is denoted as $\boldsymbol{e_m}$ mentioned above.

Assuming that there are M frames in the given audio data, the MASK region of each frame has four groups of sub-regions and L energies are calculated in a single group (L = 4 or 8) which forms an energy matrix with M L-dimensional data $\mathbf{E} = [\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_M}]$. The algorithm selects K nearest frames by finding the results with minimum Euclidean distance between the m-th energy vector $\boldsymbol{e_m}$ and other data in matrix $\mathbf{E}$. The results are arranged in ascending order of Euclidean distance and represented as $\boldsymbol{\Theta} = \{\boldsymbol{\eta_1}, \boldsymbol{\eta_2}, \ldots, \boldsymbol{\eta_K}\}$ where $\boldsymbol{\eta_K}$ is the k-th nearest energy vector of $\boldsymbol{e_m}$ and the value range of K is from 3 to 8 which can ensure the efficiency and prevent the complexity from becoming too high. However, if we only use Euclidean distance to select the proximities, it inevitably occurs that two frames are too far apart in time. In practical conditions, it is obviously unreasonable for two frames that are not related in information to be determined as proximity. In order to ensure that the times span of the nearest selection results is not too long and in view of the upper end of the K's value range is 8, this paper limits the range of selection within 10 frames (five before and five after) which

considers the correlation between audio frames and reduces the possibility of proximities appearing out of the matching segment, making a contribution to reducing the error of retrieval.

(ii) Reconstruction weights computation

The purpose of this step is to represent each energy vector by its K nearest frames with linear weights, called reconstruction weights. For the regression problem of calculating the reconstruction weights, this paper uses the mean square error as the loss function of the algorithm:

$$J(\boldsymbol{w_m}) = \sum_{m=1}^{M} \|\boldsymbol{e_m} - \sum_{k=1}^{K} w_{mk}\boldsymbol{\eta_k}\|_2^2 \tag{1}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\eta_1}, \boldsymbol{\eta_2}, \ldots, \boldsymbol{\eta_K}\}$ is the set of energy vectors of K nearest frames, the loss function $J(\boldsymbol{w_m})$ represents the construction error which aimed to minimize. $\boldsymbol{w_m} = [w_{m1}, w_{m2}, \ldots, w_{mK}]$ where w_{mk} is the construction weight between the energy vector $\boldsymbol{e_m}$ of the m-th frame and its k-th proximity $\boldsymbol{\eta_k}$ with the following constraint:

$$\sum_{k=1}^{K} w_{mk} = 1 \tag{2}$$

The normalized representation of the reconstruction vector $\boldsymbol{w_m}$ can be obtained by solving the Equation (1) with the Lagrange multiplier method:

$$\boldsymbol{w_m} = \frac{\mathbf{Z_m^{-1}}\boldsymbol{c_K}}{\boldsymbol{c_K^T}\mathbf{Z_m^{-1}}\boldsymbol{c_K}} \tag{3}$$

where $\boldsymbol{c_K}$ is K-dimensional 1 vector, $\mathbf{Z_m}$ is the local covariance matrix and its formula is as follows:

$$\mathbf{Z_m} = (\boldsymbol{e_m} - \boldsymbol{\eta_k})^T(\boldsymbol{e_m} - \boldsymbol{\eta_k}) \tag{4}$$

Finally, the reconstruction vectors of all frames are combined into a reconstruction matrix, denoted as $\mathbf{W} = [\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_M}]$.

(iii) Low dimensional mapping

Suppose the mapping result of the low dimension is $\mathbf{E'} = [\boldsymbol{e_{1'}}, \boldsymbol{e_{2'}}, \ldots, \boldsymbol{e_M}']$, and the linear relationship is expected to be consistent with that of the high dimension. The loss function $J(\boldsymbol{w_m})$ also needs to be minimized which is solved by the Lagrange multiplier method. In this step the dimensionality reduction problem can be simplified as the eigenvalue decomposition of the target matrix $\mathbf{D}$:

$$\mathbf{D} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W}) \tag{5}$$

where the constraint is:

$$\mathbf{I} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{e_m}'(\boldsymbol{e'_m})^T = \frac{1}{M}\mathbf{E'}\left(\mathbf{E'}\right)^T \tag{6}$$

where $\boldsymbol{e_m}' = [e'_{m1}, e'_{m2}, \ldots, e'_{md}]$ is the energy vector in the low dimension.

To figure out $\boldsymbol{e_m}'$, the loss function should be also minimized in the low dimension and it can be rewritten as follow according to matrix theory:

$$J\left(\mathbf{E'}\right) = trace\left[\mathbf{E'}\mathbf{D}\left(\mathbf{E'}\right)^{\mathrm{T}}\right] \tag{7}$$

Then the Lagrange multiplier method is adopted, then the equation can be represented as:

$$\mathrm{Lag}\left(\mathbf{E'}\right) = \mathbf{E'}\mathbf{D}\left(\mathbf{E'}\right)^{\mathrm{T}} + \boldsymbol{\lambda}\left[\mathbf{E'}\left(\mathbf{E'}\right)^{\mathrm{T}} - M\mathbf{I}\right] \tag{8}$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier. Finally the derivative of Equation (8) is taken and set it equal to 0:

$$\frac{\partial \mathrm{Lag}\left(\mathbf{E'}\right)}{\partial \mathbf{E'}} = 2\mathbf{D}\left(\mathbf{E'}\right)^T + 2\,\boldsymbol{\lambda}\left(\mathbf{E'}\right)^T = 0 \tag{9}$$

Equation (9) formally conforms to the definition of matrix eigenvalues and eigenvectors. Thus, the optimization result of this paper is a matrix composed of the eigenvectors corresponding to the smallest d non-0 eigenvalues of matrix $\mathbf{D}$ ($d < L$). Until now, the algorithm has obtained a new energy vector $\boldsymbol{e'_m} = [e'_{m1}, e'_{m2}, \ldots, e'_{md}]$ and the number of vector is still M corresponding the M frames audio signal. But the dimension of the elements reduces from L to d reflecting the reduction of sub-regions' energy. Finally, the audio fingerprints are extracted by comparing the energies in adjacent sub-regions as follow.

$$F(m,i) = \begin{cases} 1, \ if \ e'_{m,i+1} - e'_{m,i} > 0 \\ 0, \ if \ e'_{m,i+1} - e'_{m,i} < 0 \end{cases}, \ i = 1, 2, \ldots, d-1 \tag{10}$$

where $e'_{m,i}$ is the energy of i-th sub-region in the m-th frame, $F(m,i)$ is the i-th bit of the fingerprint in one of the sub-region group and the fingerprints in other groups can also be computed in the same way. By changing the d and observing the retrieval results, the limitation of dimensionality reduction in a single group can be determined.

3. Audio Fingerprint Matching

3.1. Similarity Measurement

In this paper, the similarity measurement used in audio fingerprint matching is the bit error rate:

$$BER = \frac{\sum_{r=1}^{R} \sum_{s=1}^{S} F(r,s) \bigoplus F'(r,s)}{R \times S} \tag{11}$$

where $\bigoplus$ is XOR operation, *F(r,s)* is the reference fingerprint in database and *F'(r,s)* is the actual input which called the query fingerprint, s is bit index and r is frame index, R is the total number of frames in the matching segment, S is the number of the fingerprint dimensions.

The threshold is set based on the experimental results. If the *BER* of the query is less than the reference which means they are very similar, the current reference should be as one of the candidate result, whereas there is a big difference between the query and the reference, the latter would not be selected as a retrieval result.

3.2. Dynamic Time Warping

For improving the resistance against linear distortion of the audio signal in the matching process, the DTW is adopted. The DTW is mainly used to solve the case that the two sequences are not equal in length of the comparison, which exactly corresponds to the change of audio signal in linear speed. The retrieved objects include speech and music. If the query is a recording or someone's speech received by microphone, the playback speeds depends on the speaking speeds which varies from person to person. The same goes for retrieving music by humming. Especially in the copyright protection, differences in audio versions can result in different playback speeds. When the reference is compared with the query, the length of two sequences is no longer equal, denoted R and R'. The algorithm calculates the similarity between every two fingerprints, then plans a matching path from the first to the last comparison according to the minimum accumulative error. The path planning strategy ensures the alignment in time and reduces the errors caused by time dislocation. The specific method to achieve the following:

(1) Similarity Matrix Computation

Before aligning the query and the reference, first a similarity matrix of size $R \times R'$ needs to be constructed, denoted as $\mathbf{X}$ where R is the length of reference and R' is the length of query. The elements in the matrix is the similarity of every two fingerprints in the query and the reference. The row and column represent the positions of two fingerprints in the respective sequences of the comparison. To be expressed more clarity, an example of matrix $\mathbf{X}$ is given in Figure 5.

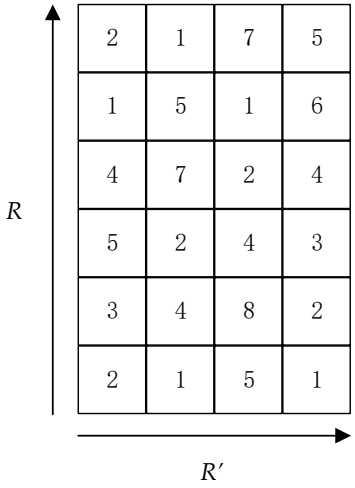


Figure 5. Similarity matrix.

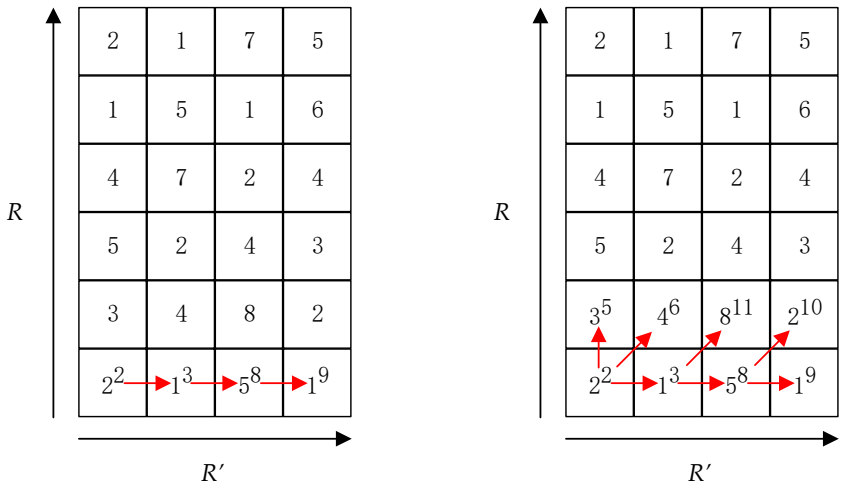
The matrix **X** is shown schematically in Figure 5 where the numbers in the grid is the similarities between the reference and query fingerprints computed by the Equation (11). The length of the reference and the query could be different or be set artificially to gain a better resistance against linear speed changes.

(2) Similarity Accumulation Matrix Computation

Accumulating the above similarity matrix **X** from the starting point (1, 1) to (R, R'). There are three kinds of paths for each accumulation: transverse, longitudinal, or oblique. The direction with the minimum similarity is chosen for each accumulation. The computation is carried out first in row and then in column, and the result is recorded as matrix **Y**. Its formula can be express as:

$$\mathbf{Y}(i,j) = \mathbf{X}(i,j) + \min\{\mathbf{Y}(i-1,j-1), \mathbf{Y}(i-1,j), \mathbf{Y}(i,j-1)\} \tag{12}$$

The similarity accumulation of the first row is computed from left to right as shown in Figure 6a and the second row is computed in the same order represented in Figure 6b. The computation rule of the similarity accumulation is the number in grid plus the former accumulation according to the minimum of the sum as mentioned in Equation (12). All the accumulations are labeled in the upper right corner of each number in grid and these superscripts form the matrix **Y**.



(**a**) Similarity accumulation of the first row (**b**) Similarity accumulation of the second row

Figure 6. Similarity accumulation.

(3) The Shortest Path Planning

Once the accumulation matrix **Y** is obtained, the DTW works backwards from the end point to the starting point to plan the path with minimum similarity accumulation and count the number of matching points it passes served as compensation. The path planning process is shown in Figure 7:

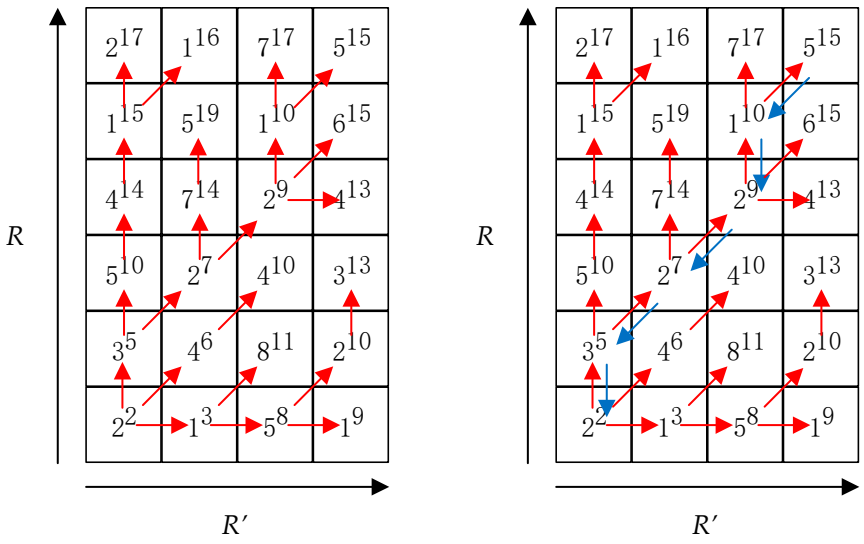
The superscripts in Figure 7a shows the accumulation results and the red arrow is the accumulation direction. The blue arrow in Figure 7b is the matching path obtained by looking back from the end to the beginning. It is important to notice that three conditions must be met for path planning:

a. The path planning begins from (1, 1) and ends at (R, R');
b. The path has continuity, that is, the fingerprints cannot be matched across each other, but can only be aligned with their adjacent fingerprints, so as to ensure that each fingerprint appears in the path;
c. Monotonicity. The path selection must be carried out monotonically in time order to prevent any crossover occurred in the alignment process.

Finally, the minimum similarity accumulation and the number of matching points though the path are obtained and now the similarity measure can be represented as follow:

$$\gamma = \frac{Y(R, R')}{p} \tag{13}$$

where $Y(R, R')$ is the minimum similarity accumulation, p is the total number of the matching points on the path and γ is the final similarity measurement. If γ is less than the threshold value, the matching is considered successful and the corresponding reference is taken as a candidate. Otherwise, the matching fails and the query continues to match with other references. When the matching process is over, all the candidates are arranged in ascending order of γ.



(**a**) Similarity accumulations (**b**) Path planning

Figure 7. The shortest path planning.

4. Experimental Results and Analysis

The experimental data is introduced as follows:

Database 1 consists of 5000 audio files, including the data collected in the laboratory and the data collected from the internet. The length of each audio file is between 3 s and 5 min. The storage size of the data is 12.3 GB and the total length is 230 h.

Database 2 is the dataset generated by adding 40 dB white noise to database 1.

Database 3 is the dataset generated by adding 30 dB white noise to database 1.

Database 4 is the dataset generated by adding 20 dB white noise to database 1.

Database 5 is the dataset generated by adding 10 dB white noise to database 1.

Database 6 is a random interception of 1000 audio segments with the length of 3 s from database 1 and it is used as the query audio set.

Database 7 is a random interception of 200 audio segments with the length of 3 s from database 1. The playback speeds of the segments are in a random range of 95% to 105%.

4.1. Performance Evaluation

In this paper, recall and precision rate which commonly used in the field of information retrieval, are selected as the metrics to evaluate the performance of the algorithm. They are defined as follows:

Recall rate = the number of correct targets detected from the retrieval source/the number of targets that should have been retrieved × 100%

Precision rate = the number of correct targets detected from the retrieval source/the number of targets actually retrieved × 100%

4.2. Results and Analysis

It is preliminarily verified by experiments that if all groups of LLE fingerprint extraction is carried out at the same time, too much features are lost and the retrieval performance is poor. So as to ensure the efficiency of audio retrieval is not seriously affected by the feature lost in the dimensionality reduction process, the algorithm does not reduce the dimension of energy vectors in all groups. The rest of the energy vectors that have not been reduced by LLE are calculated as shown in Table 1.

Table 1. Non-reduced fingerprint extraction.

Dimension	Group Type	Comparison
1–7	Horizontal	1a−1b, 1b−1c, 1c−1d, 1d−1e, 1e−1f, 1f−1g, 1g−1h
8–10	Vertical	2a−2b, 2b−2c, 2c−2d
11–14	Central	3a−3b, 3b−3c, 3c−3d, 3d−3a
15–18	Marginal	4a−4b, 4c−4d, 4e−4f, 4g−4h
19–22	Extended	4a + 4b−(4c + 4d), 4c + 4d−(4e + 4f), 4e + 4f−(4g + 4h), 4g + 4h−(4a + 4b)

As can be seen in Table 1, 1–18 dimensions are computed in a single group and represent the energy distribution of horizontal, vertical, central, and marginal sub-regions. The extended sub-region is composed of marginal sub-region in pairs and then generates the fingerprint by the comparison of energy pairs corresponding to the 19–22 dimensions.

Once the fingerprint has extracted by the reduced energy vector, the computation has uncertain effect on the efficiency of retrieval. The maximum reduced dimensions in a single group and the group numbers can be reduced at the same time are presented in the rest of paper by giving the experimental results. The performance is evaluated by the metrics mentioned in Section 4.1.

4.2.1. Influence of Nearest Frames Number *K* on Retrieval Performance

The algorithm is sensitive to the selection of *K* and different *K* has great influence on the retrieval result. The influence of different *K* is analyzed in each group reduction and the optimal *K* is selected for the experiment. Generally *K* is between 3 and 8. Table 2 shows the retrieval results corresponding to different *K* under the condition that the horizontal sub-region is reduced by 1 dimension.

As can be seen in Table 2, when *K* is set to 3, the best retrieval performance is obtained. Through the same experimental steps, the selection of *K* value under different conditions is determined and the optimal *K* value is applied in the subsequent experiment. It is worth noting that the influence of *K* value on the retrieval results is not always monotonous, but the higher its value means the increase of

algorithm complexity. Therefore, this paper comprehensively considers both the retrieval results and the algorithm complexity to select the most appropriate K value.

Table 2. The retrieval result of different K (reduced dimension of horizontal group = 1).

K Value	Recall (%)	Precision (%)
3	99.6	98.6
4	97.2	95.4
5	93.3	89.3
6	89.1	85.4
7	82.6	76.5
8	75.2	70.8

4.2.2. Dimensionality Reduction in a Single Group

In order to verify the maximum dimensionality reduction that LLE can achieve, this paper increases the number of reduced dimension step by step on the single sub-region group. The increasing is implemented until the group has no space for further reduction or its retrieval performance cannot be accepted by the actual situation. However, since the construction of the fingerprint in the extended group is a combination of multiple sub-region energies, this group is not included in the consideration of dimensionality reduction in the paper. First, Table 3 shows the retrieval results of different reduction of horizontal group.

In the case that the reduced dimension is set to 4 in the horizontal group, the method can still maintain high recall and precision rate (Table 3). The original dimension of fingerprint in the horizontal group is 7 and the reduced dimension is raised from 1 to 4. It can be seen that the reduction proportion is more than 50% in the horizontal group and 20% in the overall fingerprint (22 dimensions). Similar results are obtained in other groups.

Table 3. Horizontal dimensionality reduction results.

Reduced Dimension	Optimal K	Proportion of Group	Proportion of Fingerprint	Recall (%)	Precision (%)
1	3	1/7	1/22	99.6	98.6
2	3	2/7	2/22	99.4	98.4
3	5	3/7	3/22	99.0	98.2
4	5	4/7	4/22	98.8	97.9

The original fingerprint dimensions in each group are different, so the maximum reduction are also different. The reduction of the three groups shown in Table 4 reached a proportion 50% of group and about 10% of the overall fingerprint. The result is mainly limited by the original fingerprint dimension. The original fingerprint dimension in the horizontal group is quite large (seven dimensions), while the other groups have no reduction space as large as that in the horizontal group, making the overall dimensionality reduction lower than that in the horizontal area.

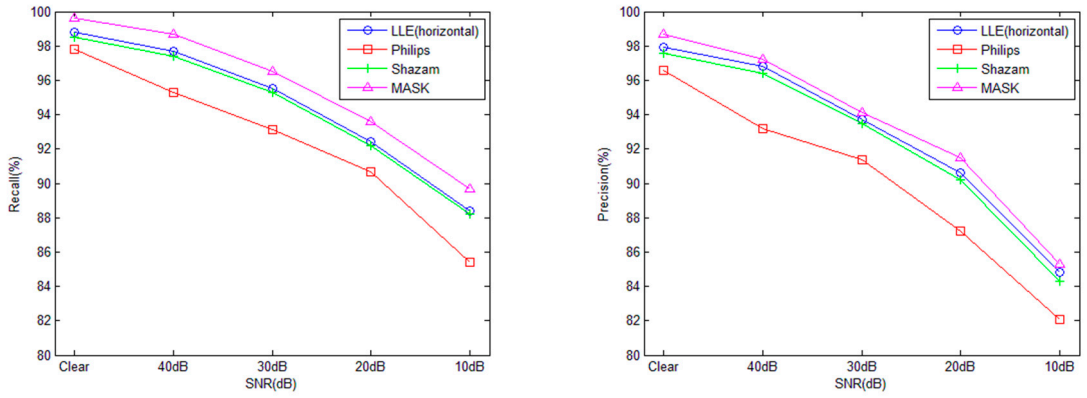
Table 4. None horizontal dimensionality reduction results.

Reduced Dimension	Group	Optimal K	Proportion of Group	Proportion of Fingerprint	Recall (%)	Precision (%)
1	Vertical	5	1/3	1/22	99.2	98.0
2	Vertical	5	2/3	2/22	98.1	96.5
1	Central	5	1/4	1/22	99.3	98.9
2	Central	5	2/4	2/22	98.5	97.6
1	Marginal	6	1/4	1/22	99.1	97.8
2	Marginal	4	2/4	2/22	97.4	94.2

4.2.3. Single Reduction Result Compared with Other Algorithm under Different SNR

The results of the horizontal group with the largest dimensionality (four dimensions) reduction are selected in the comparison. To verify its robustness under noisy conditions, Philips and Shazam algorithms are applied as reference methods. As classical algorithms, they are widely used in audio fingerprint retrieval and have good retrieval performance. In Philips' algorithm, the spectrum of audio signal is divided into 33 bands and generates a 32-bit binary fingerprint by comparing the energies in adjacent bands. On the other hand, Shazam uses the peak information extracted in the frequency domain to construct a hash value with the storage size of 32 bits. In addition, a comparison of the original MASK audio fingerprint is added in the experiment to prove the feasibility of the dimensionality reduction. The original MASK size is 22 bits and the reduced fingerprint size is 18 bits. Figure 8 shows the retrieval performance of each method under different SNR.

The comparison of the recall and precision rate of each method is shown in Figure 8a,b. Both of them represent the retrieval performance and the performance of the horizontal dimension reduced fingerprint in this paper is basically the same as the original MASK, and slightly higher than that of Philips and Shazam algorithms. With the decrease of SNR, the retrieval performance of the method presented in this paper decreases slowly which indicates that it has stronger resistance against noise. To sum up, the method in this paper reduces the amount of fingerprint data which saves storage space and improves retrieval speed on the premise that the robust performance is almost not affected.



(**a**) Comparison of the recall rate (**b**) Comparison of the precision rate

Figure 8. Comparison of horizontal reduction with other methods.

4.2.4. Dimensionality Reduction in Multiple Groups

Once the maximum reduction in single group is determined, this section presents the result of the reduction in multiple groups. The results are based on the central group and combine it with other groups to analyze the influence on retrieval performance. The reason for choosing the central group is that in the process of sub-region division, there are more or less overlaps between central group and other groups which means that they have partial similarity in features and the feature loss in the multiple reduction process should be small. In the multiple reduction experiment, the K value is the optimal value in each group and the reduction operation is relatively independent, so there is no need to adjust the K value again. The experiment is conducted to prove all cases of multiple reduction starting from 1 reduction to the maximum. The result of the central and horizontal group is firstly shown in Table 5.

The retrieval performance of multiple reduction is lower than the corresponding results of single group due to the increase of feature loss. When both groups reach the maximum reduction, the algorithm also has high retrieval efficiency which indicates that the reduction in central and horizontal group can be carried out simultaneously to achieve a total reduction of six dimensions with

a proportion of about 27%, more than 1/4 (Table 5). Moreover, the reduction in central and vertical group has similar results.

The reduction in central and vertical group can also reach a total dimensionality reduction of four dimensions with a reduction rate of about 18% (Table 6). However, there is a significant decrease in retrieval performance, and it cannot be applied to large databases.

In addition to the above two combinations, the fingerprint features in other cases are greatly different and the feature loss is quite large, which makes it difficult to guarantee the retrieval efficiency and fails to achieve ideal results. The horizontal group has more bits of fingerprint, that is to say, has more reduction space. Thus, the reduction in central and horizontal can get less data and better retrieval performance.

Table 5. Central and horizontal reduction results.

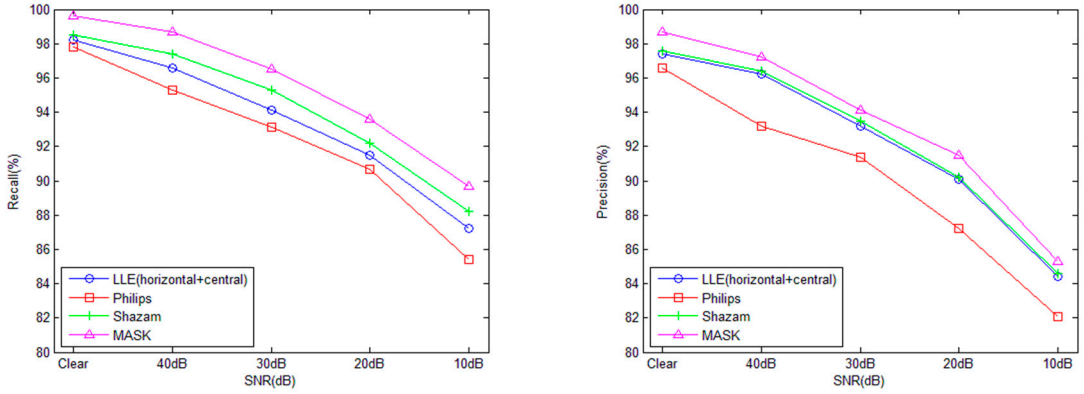
Central / Horizontal	Reduced Dimension = 1		Reduced Dimension = 2	
	Recall (%)	**Precision (%)**	**Recall (%)**	**Precision (%)**
Reduced dimension = 1	99.2	98.6	99.0	98.5
Reduced dimension = 2	98.9	98.3	98.7	98.1
Reduced dimension = 3	98.8	98.0	98.5	97.8
Reduced dimension = 4	98.5	97.7	98.2	97.4

Table 6. Central and vertical reduction results.

Central / Horizontal	Reduced Dimension = 1		Reduced Dimension = 2	
	Recall (%)	**Precision (%)**	**Recall (%)**	**Precision (%)**
Reduced dimension = 1	98.6	97.2	97.2	96.2
Reduced dimension = 2	96.8	94.6	95.4	93.8

4.2.5. Multiple Reduction Result Compared with Other Algorithm under Different SNR

The combination of central and horizontal group with the largest reduction is selected for the experiment which has a reduction of six dimensions and fingerprint size of 16 bits. The comparison methods are still Philips, Shazam, and MASK algorithms, and the results are shown in Figure 9.



(**a**) Comparison of the recall rate (**b**) Comparison of the precision rate

Figure 9. Comparison of multiple reduction with other methods.

The retrieval performance of the reduction combination in this paper is slightly lower than that of the Shazam and the original MASK, which indicates that the combination of groups aggravates the feature loss. But its performance is basically the same as that of Philips and Shazam algorithms, indicating that the method in this paper can meet the retrieval requirements to some extent. For example,

when dealing with small- or medium-sized audio retrieval, the multiple reduction proposed in this paper reduces the amount fingerprint data to the greatest extent and has basic retrieval efficiency. In large-scale audio retrieval, the performance of multiple reduction is difficult to maintain the accuracy, but single reduction can still be used for it.

To present the compression potential more clearly, a table of the fingerprint size of each method is shown below.

Table 7 includes the fingerprint size of three comparison methods and two proposed methods. All the methods have same length of the frame and extract a fingerprint from each frame, so the experiment only needs to compare the size of a single fingerprint. The multiple reduction has the smallest fingerprint size which is half that of the Philips and Shazam. The fingerprint size of proposed method is also a significant reduction compared with the MASK.

Table 7. Fingerprint size of different methods.

Method	Fingerprint Size (per Frame)
Philips	32 bits
Shazam	32 bits
MASK	22 bits
LLE (horizontal)	18 bits
LLE (horizontal + central)	16 bits

4.2.6. Resistance to Linear Speed Changes

In order to prove the validity of DTW, 200 audio segments are randomly selected and artificially modified playback speeds at a random rate (Database 7). These segments are used as a query set and the fingerprint with the smallest size is applied. An experiment is carried out not using DTW for the purpose of comparison.

The experiment obtains a higher recall and precision rate with using DTW (Table 8), so the DTW can indeed improve the resistance of the retrieval system to linear speed changes.

Table 8. Results of linear speed changes.

	Recall (%)	Precision (%)
Using DTW	100	99.6
Not using DTW	99.4	98.5

5. Conclusions

A fingerprint dimensionality reduction method based on LLE and a fingerprint matching method based on DTW are introduced in this paper. Firstly, the energy vectors of sub-regions in the frequency domain is reduced which involved in fingerprint calculation. Then the reduction of the fingerprint is obtained by comparing the adjacent energies. Additionally, the matching segment is adjusted by the DTW algorithm, which improved the resistance to linear speed changes of the audio signal. The experimental results show that this novel method has good retrieval performance both in single group and multiple group reduction, and the maximum fingerprint reduction ratio can reach 27%, which greatly reduces the fingerprint data.

Author Contributions: M.J. and T.L. contributed equally in conceiving the whole proposed codec architecture, designing and performing the experiments, collecting and analyzing the data, and writing the paper. J.W. collected the data and implemented the final revision. J.W. supervised all aspects of this research. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China under grant no. 61971015.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, X.; Zou, X.; Hu, Q.; Zhang, P. Audio retrieval method based on weighted DNA. *China Sci.* **2018**, *13*, 2295–2300.
2. Wang, Y.; Liu, Z.; Huang, J. Multimedia content analysis using both audio and visual clues. *IEEE Signal Process. Mag.* **2000**, *17*, 12–36. [CrossRef]
3. Li, G.; Wu, D.; Zhang, J. Concept framework for audio information retrieval. *J. Comput. Sci. Technol.* **2003**, *18*, 667–673. [CrossRef]
4. Klapuri, A.P. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 804–816. [CrossRef]
5. Jiang, X. An audio data retrieval method based on MFCC. *Comput. Digit. Eng.* **2008**, *9*, 24–26.
6. Jiang, X.; Li, Y. Audio data retrieval method based on LPCMCC. *Comput. Eng.* **2009**, *35*, 246–247.
7. Li, W.; Li, X.; Chen, F.; Wang, S. Review of digital audio fingerprinting. *J. Chin. Comput. Syst.* **2008**, *29*, 2124–2130.
8. Cano, P.; Batle, E.; Kalker, T.; Haitsma, J. A Review of algorithms for audio fingerprinting. In Proceedings of the Multimedia Signal Processing, St. Thomas, VI, USA, 9–11 December 2002; pp. 169–173.
9. Doets, P.J.O.; Lagendijk, R.L. Extracting quality parameters for compressed audio from fingerprints. In Proceedings of the International Conference on ISMIR, London, UK, 11–15 September 2005; pp. 498–503.
10. Haitsma, J.; Kalker, T. A highly robust audio fingerprinting system. In Proceedings of the International Conference on Music Information Retrieval, Paris, France, 13–17 October 2002; pp. 107–115.
11. Doets, P.J.O.; Lagendijk, R.L. Stochastic model of a robust audio fingerprinting system. In Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona, Spain, 10–14 October 2004; pp. 2–5.
12. Park, M.; Kim, H.; Yang, S. Frequency-temporal filtering for a robust audio fingerprinting scheme in real-noise environments. *IEEE Trans. Inf. Syst.* **2006**, *28*, 509–512. [CrossRef]
13. Wang, A.; Li, C. An industrial strength audio search algorithm. In Proceedings of the International Conference on Music Information Retrieval, Baltimore, MD, USA, 27–30 October 2003; pp. 7–13.
14. Baluja, S.; Covell, M. Waveprint: Efficient wavelet-based audio fingerprinting. *Pattern Recognit.* **2008**, *41*, 3467–3480. [CrossRef]
15. Pucciarelli, G. Wavelet analysis in volcanology: The case of phlegrean fields. *J. Environ. Sci. Eng. A* **2017**, *6*, 300–307.
16. Chen, D.; Zhang, W.; Zhang, Z. Audio retrieval based on wavelet transform. In Proceedings of the IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, 24–26 May 2017; pp. 531–534.
17. Anguera, X.; Garzon, A.; Adamek, T. MASK: Robust local features for audio fingerprinting. In Proceedings of the IEEE International Conference on Multimedia and Expo, Melbourne, Australia, 9–13 July 2012; pp. 455–460.
18. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2001**, *290*, 2323–2326. [CrossRef] [PubMed]
19. Vavrek, J.; Viszlay, P.; Lojka, M.; Juhar, J.; Pleva, M. Weighted fast sequential DTW for multilingual audio query-by-example retrieval. *J. Intell. Inf. Syst.* **2018**, *51*, 439–455. [CrossRef]