A Reliable Voice Perceptual Hash Authentication Algorithm

Li Li, Yang Li(✉), Zizhen Wang, Xuemei Li, and Guozhen Shi

Beijing Electronic Science and Technology Institute, Beijing 100070, China

Abstract. In order to protect the authenticity and integrity of voice, this paper proposes a reliable voice perceptual hash authentication algorithm. This algorithm considers the dynamic characteristics of voice signals and proposes a scheme to construct a perceptual feature matrix containing voice static and dynamic characteristics based on the Mel frequency inverted spectrum coefficient and its first- and second-order difference parameters. In the process of perceptual hash chain construction, the feature matrix is degraded by two-dimensional non-negative matrix factorization method, and the result of decomposition is quantified to construct the perceptual hash chain. In the process of voice authentication, hamming distance method is used to measure the distance between perceptual hash chains. Experiments show that the algorithm has good distinguishability and robustness, and the voice perceptual hash authentication can be carried out accurately and reliably.

Keywords: Perceptual hash · Non-negative matrix factorization · MFCC

1 Introduction

Because of its convenience, voice is more suitable for communication in the case of inconvenient text input, as an indispensable communication method in modern society, its security problems can not be ignored. Due to the diversity of audio editing tools and simplification of operation, voice authenticity authentication and integrity protection are particularly important. Traditional crypto-hash authentication technology because of its high sensitivity to content changes, has not been applicable to voice content authentication, while voice perceptual hash authentication technology has a good distinguishability and robustness, can be well certified voice content, so as to complete the protection of voice authenticity and integrity.

Perceptual hash technology uniquely maps multimedia digital represents with the same perceptual content to a digital digest, was first applied in the field of image authentication [1–4], which completes image authentication and recognition by short summary of image perceptual information and summary-based comparison matching. It is now widely used in image authentication, image retrieval, voice content authentication and tamper detection and other fields [5–7]. The common method of voice perceptual hash feature extraction is to project a set of signals into another domain, with the main methods being wavelet transform [8, 9], discrete cosine transformation [10], short-term Fourier

J. Xiong et al. (Eds.): MobiMedia 2021, LNICST 394, pp. 253–263, 2021.
https://doi.org/10.1007/978-3-030-89814-4_19

transformation [11], etc. Then select feature values in the transformation domain, such as Mel-Frequency Cepstral Coefficient (MFCC), short-term zero-pass rate, short-term energy, etc. The extracted eigenvalues are analyzed, the perceptual hash chain is constructed, and the voice perceptual hash authentication is completed. By analyzing the extracted eigenvalues, perceptual hash chains are constructed, and use them to complete voice perceptual hash authentication. The paper [12] proposes a scheme for constructing a perceptual hash chain using short-term energy and short-term zero-pass rate to perceive audio content in different formats, with good robustness and security, and good computational efficiency. In paper [13], MFCC and LPCC feature parameters are fused to construct feature matrix, and the complexity of the feature matrix is reduced by using two-dimensional non-negative matrix factorization, effectively improve the robustness of hash authentication, but the matrix block decomposition method is lacking. The paper [14] uses MFCC parameters as perceptual features, which has good robustness, but the security of the algorithm depends on pseudo-random sequences, and the robustness to resampling and other operations is poor. The paper [15] presents 2 voice perceptual hash authentication algorithms based on voice spectral graph and pseudo-harmonic model, the former which can meet the higher demand for real-time performance, and the latter, which has relatively poor operational efficiency and differentiation but has better robustness.

In this paper, combining the auditory characteristics of human ear and fully considering the dynamic characteristics of voice signal, a reliable voice perceptual hash authentication algorithm is proposed: the algorithm extracts the MFCC feature parameters of voice and calculates its first- and second-order differences parameters, constructs the perceived feature matrix using these three sets of feature parameters, the feature matrix is degraded by two-dimensional non-negative matrix factorization, so as to construct the perceptual hash chain, and the normalized Hamming distance is used as a certification standard for perceptive hash authentication. Finally, the distinguishability and robustness of the algorithm are verified by experiments.

2 Voice Perceptual Hash Construction Algorithm

The voice perceptual hash construct in this paper is based on MFCC feature extraction, and the construction process is shown in Fig. 1.

Fig. 1. Voice perceptual hash construction

2.1 MFCC Extraction

Mel-Frequency Cepstral Coefficient, MFCC is one of the important characteristic parameters used in voice recognition technology, and its physical meaning is the distribution of energy in the signal spectrum in different frequency bands. The MFCC is designed

based on the auditory properties of the human ear, as a result of the frequency of sound heard by the human ear is not linear, the growth of the Mel-frequency is consistent with the auditory characteristics of the human ear, the actual frequency is linearly distributed below 1000 Hz, and increases logarithmically above 1000 Hz, the specific relationship between them is shown in formula (1):

$$f_{mel}(f) = 2595 \lg(1 + \frac{f}{700}) \tag{1}$$

Among them, the f is the actual frequency, f_{mel} is the Mel-frequency.

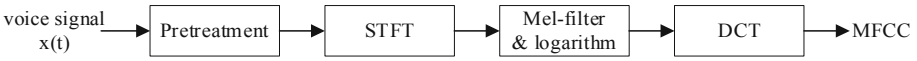

Fig. 2. MFCC extraction process

The specific process of MFCC extraction, as shown in Fig. 2, is divided into the following steps:

1> Pretreatment. Windowing and framing of the input voice signal, add Hanning window to the voice signal, and set frame length to 2048, frame shift to 512 for framing.

2> Short-term Fourier transform, STFT. Use fast Fourier transform (FFT) to calculate the results obtained in the previous step, to get spectrum of each frame to get the spectrum distribution information. This process of frame-by-frame fast Fourier transform is called short time Fourier transform. The FFT formula is as follows:

$$X_a(k) = \sum_{n=0}^{N-1} x(t) e^{-j2\pi k/N}, 0 \le k \le N \tag{2}$$

Among them, the $x(t)$ is the input voice signal, N is the number of Fourier transform points, and a is the frame index number.

3> Mel-filter and logarithm. The amplitude spectrum $|X_a(k)|$ is obtained by taking the mode of the spectrum $X_a(k)$ obtained in the previous step, and then the power spectrum $|X_a(k)|^2$ is obtained by squaring it. The power spectrum is passed through a group of M Mel-filter banks $H_m(k)$, and the dot product operation is performed and the logarithm is taken. In this paper, the number of Mel-filters M is set to 128, and the logarithm of the output of each filter is calculated as shown in formula (3):

$$S(m) = \ln(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)), 0 \le m \le M \tag{3}$$

Among them, the $S(m)$ is the logarithmic energy.

4> Discrete Cosine Transform, DCT. The MFCC coefficient is obtained by DCT:

$$MFCC(n) = \sum_{m=0}^{N-1} S(m) \cos(\frac{\pi n(m - 0.5)}{M}), n = 1, 2, ...L \tag{4}$$

Among them, the L is the order of MFCC. Taking the logarithmic energy $S(m)$ into DCT, the parameters of L-order MFCC are obtained. The purpose of DCT is to change the data distribution and separate the redundant data. Because most of the signal data will be concentrated in the low frequency region after DCT, only the low order coefficients of MFCC after DCT are needed. In this algorithm, $L = 13$ is set, the standard MFCC coefficients of order 13 are taken.

2.2 Feature Matrix Construction

The standard MFCC parameters can only reflect the static characteristics of voice, and the dynamic characteristics of voice can be described by the difference function of these static features, combining dynamic and static characteristics can effectively improve the recognition performance of the system. Therefore, in addition to the above 13-order standard MFCC feature parameters, the perceptual feature matrix constructed in this paper also includes the first- and second-order difference parameters of the 13th-order MFCC parameters used to describe dynamic characteristics in order to improve recognition performance. After framing a voice signal, the total number of frames obtained is m, then the perceptual feature parameters include 13 × m-dimensional standard MFCC parameters, 13 × m-dimensional first-order difference parameters and 13 × m-dimensional second-order difference parameters, the perceptual feature matrix construction is shown in Fig. 3.

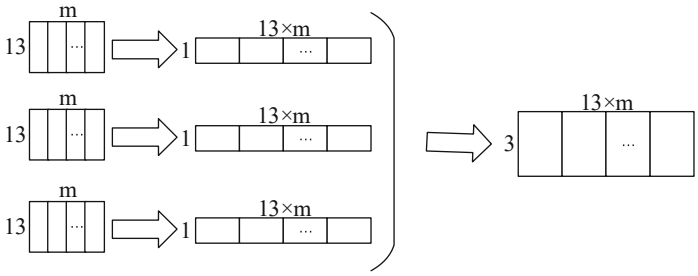

Fig. 3. Feature matrix construction process

The MFCC parameter and its first-order second-order differential parameters are 13 × m dimensional matrixes, each of which tiles it by column, obtains 3 row matrixes which is 1 × 13m-dimensional, and then stacks the 3 row matrixes into a matrix, the 1st is standard MFCC characteristic parameter, the 2nd is its first-order differences parameters, and the 3rd is its second-order differences parameters. Then we have a 3 × 13m dimensional feature matrix which can be used for the construction of voice perceptual hash chain.

2.3 Feature Matrix Factorization

In this paper, use two-dimensional non-negative matrix factorization (2DNMF) to degraded reduction the feature matrix by using the non-negative matrix factorization

(NMF) method twice. Non-negative matrix decomposition exerts non-negative constraints on the data matrix, so that only addition combinations are allowed in the factorization process [16], resulting in the whole data being superimposed by part without positive or negative offsetting, and achieving the effect of the part expressing the whole, which coincides with the perceptual basis of human brain, and has the characteristics of fast convergence and small storage space. 2DNMF was first proposed by the paper [17] and applied to the field of image processing. The specific process of this method is shown in Fig. 4.

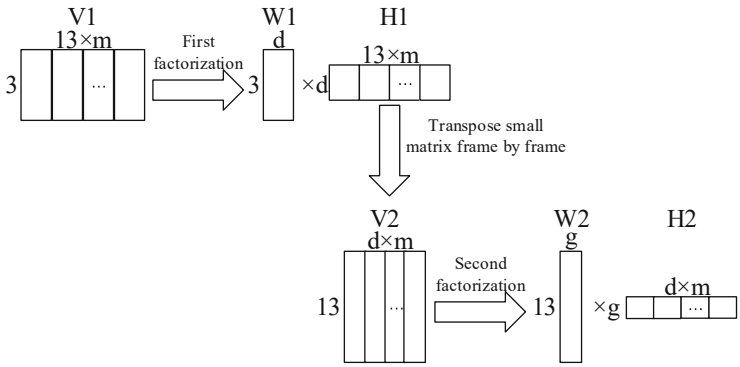

Fig. 4. Feature matrix factorization

The feature matrix constructed in Sect. 2.2 is denoted as $V1_{3\times 13m}$, which is decomposed into $W1_{3\times d}$ and $W1_{3\times d}$ by NMF:

$$V1_{3\times 13m} \approx W1_{3\times d} H1_{d\times 13m} \tag{5}$$

Among them, d is the decomposition rank, also called decomposition order. Take $H1_{d\times 13m}$, transpose each $d \times 13$ matrix frame by frame according to the number of total frames m, get $V2_{13\times dm}$, and then continue the matrix decomposition, that is

$$V2_{13\times dm} \approx W2_{13\times g} H2_{g\times dm} \tag{6}$$

Among them, g is the decomposition rank.

In the above decomposition process, the values of d and g need to satisfy $d < 3$, $g < 13$. In order to facilitate the subsequent calculation, the algorithm in this paper sets $d = 1$, $g = 5$, that is, after the decomposition of formula (5) and (6), the coefficient matrix is $H2_{5\times m}$, which is recorded as the result matrix C.

2.4 Perceptual Hash Chain Construction

The dimension reduced result matrix is used as the basis to construct the perceptual hash chain. Sum matrix C column by column according to the following formula (7) to get

the element sum of each frame of the corresponding voice signal, which is recorded as $S(a)$:

$$S(a)=\sum_{i=1}^{g=5} C_{ia}, 1 \le a \le m \tag{7}$$

Among them, i is the row index of the matrix and a is the frame index. The average value of $S(a)$ is calculated as $\overline{S}$, and the voice perceptual hash chain H is constructed according to the following formula:

$$h_a = \begin{cases} 1, S(a) > \overline{S} \\ 0, other \end{cases}, 1 \le a \le m \tag{8}$$

Finally, the voice perceptual hash chain is $H = [h_1 h_2 ... h_m]$.

3 Perceptual Hash Authentication

In this algorithm, the authentication of voice is essentially the authentication of the perceptual hash chains of voice signals. In the certification process, the similarity of the two perceptual hash chains is measured by normalizing Hamming distance, and the normalized Hamming distance is defined as follows:

$$MFCC(n) = \sum_{m=0}^{N-1} S(m) \cos(\frac{\pi n(m-0.5)}{M}), n = 1, 2, ...L \tag{9}$$

Among them, H^x and H^y are perceptual hash chains of two segments of voice respectively, L is the length of perceptual hash chain.

Normalized Hamming distance can also be defined as Bit Error Ratio (BER), and the hypothetical test using BER describes the perceptual hash chain authentication as follows: for two voice segments $x(t)$ and $y(t)$, construct the perceptual hash chain H^x and H^y separately, if the voice segments' perceptual content is same, then $D(H^x, H^y) \le \tau$; if the voice segments' perceptual content is not same, then $D(H^x, H^y) > \tau$. Among them, τ is authentication threshold. Therefore, the perceptual hash chain authentication criteria are as follows: set the authentication threshold τ, compare the perceptual hash chains' mathematical distance D, if the mathematical distance less than the threshold, then the corresponding two voice segments are considered to be the same as the perceptual content, and the authentication is passed; otherwise, the authentication is not passed.

The specific authentication process is shown in Fig. 5, $x(t)$ is the original voice signal and $y(t)$ is the voice signal to be authenticated.

To evaluate the above authentication algorithm, define the false accept rate (FAR). FAR refers to the percentage that is incorrectly accepted, i.e., the percentage misjudged as passed by the perceptual hash chain that should not be passed:

$$\gamma_\tau = P(D \le \tau) \tag{10}$$

Among them, γ_τ is FAR, $P(\bullet)$ is probability, τ is authentication threshold. The smaller τ, the lower the FAR, the better the distinguishability of the perceptual hash, and the higher the authentication accuracy.

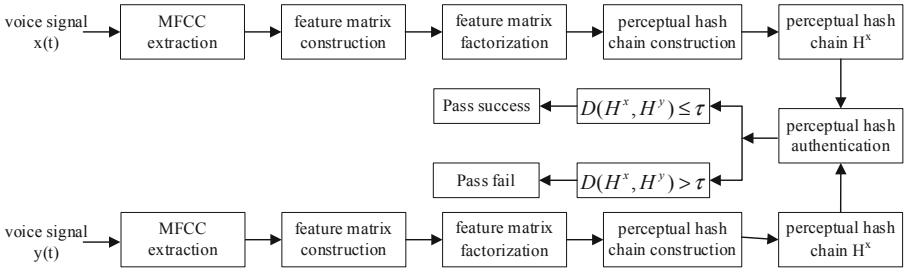

Fig. 5. Perceptual hash authentication process

4 Experimental Results and Analysis

In order to verify the distinguishability and robustness of the algorithm, the results of simulation experiments are carried out and analyzed. The hardware platform used in the experiment is Intel(R) Core(TM) i5-8300H CPU @ 2.30 GHz, the software environment is Python-3.6.9, and the audio data used in the experiment is 13388 audio files in wav format in THCHS-30 (Tsinghua University 30 h Chinese voice library). The main parameters of the experiment are set as follows: the frame length is 2048 and the frame shift is 512.

4.1 Distinguishability

Distinguishability is used to evaluate the reliability of algorithms to distinguish between different voice content by different or identical people. The BER for perceptual hash values for different content audio basically obeys the normal distribution. The audio data used in the experiment included 60 different speaker, 2 different voice segments for each speaker, and ensured that all voice contents are different, taking a total of 120 different audio files as test data for this section, as follows.

Construct the perceptual hash chain of these 120 audio segments separately, and the normalized Hamming distance D between each two hash chains is calculated by the formula (9), which is 7140 sets of data in total, the probability distribution histogram is shown in Fig. 6.

Assuming that the result conforms to a normal distribution, its mathematical expectations μ_D is 0.4497, its standard deviation σ_D is 0.0523, and theoretically its FAR is the probability integral of the normal distribution:

$$\gamma_\tau = P(D \le \tau) = \int_0^\tau \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{11}$$

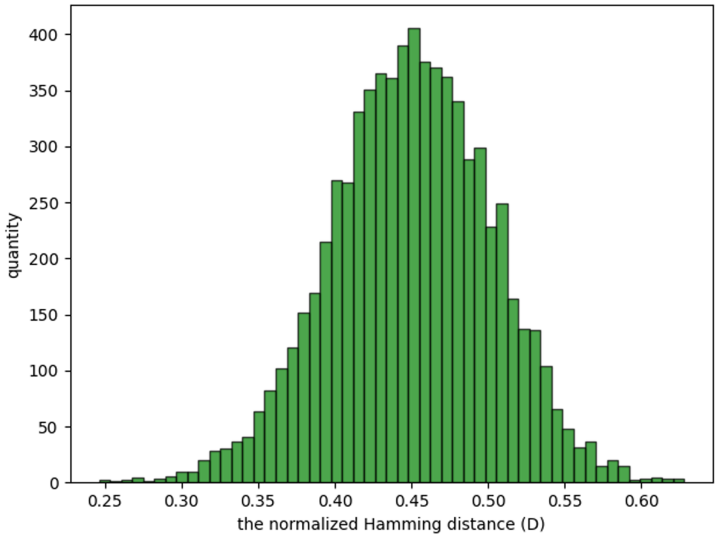

Fig. 6. The probability distribution histogram of the normalized Hamming distance

FAR reflects the distinguishability of the algorithm, the higher the threshold, the larger the FAR, the lower the distinguishability and the weaker the anti-collision ability of the algorithm. Figure 7 shows the comparison between the theoretical FAR and the experimental FAR. It can be seen from Fig. 7(a) that the two curves basically coincide, which proves that the experimental results are in line with the expectation and belong to normal distribution. Figure 7(b) enlarges the local part of the two curves. When the value τ is less than 0.3, the FAR is at a low level. At this time, the algorithm can be almost completely distinguished, so the threshold τ can be set to a value less than 0.3.

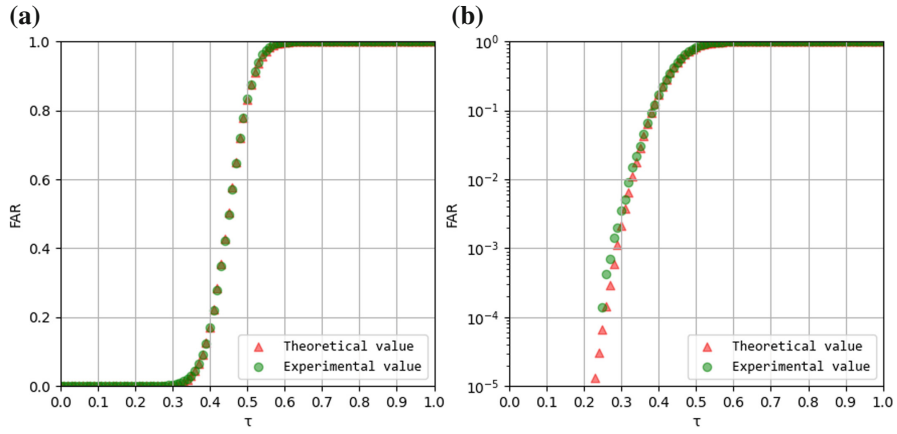

Fig. 7. The FAR

As shown in Fig. 7, the theoretical values of the FAR of this algorithm are 0.0021 when τ is 0.3, 6.717×10^{-5} when τ is 0.25 and 1.143×10^{-11} when τ is 0.2. That is, when there are enough voice sample data, the threshold value is set to 0.2, and about one of every 10^{11} voice segments will be wrongly recognized, which indicates that the algorithm in this paper has strong anti-collision ability and good distinguishability. However, compared with the literature [18, 19], there is still a large room for improvement.

4.2 Robustness

In order to verify the robustness of the algorithm proposed in this paper, that is, the same voice can still be effectively certified reliability after the content is maintained, this section uses some actions that do not change the voice content to interfere with the test audio, as shown in Table 1.

Table 1. Maintain content operations

The type of action	How it works	Symbol abbreviation
Turn down the volume	Reduce the volume by 50%	V.↓
Turn up the volume	Increase the volume by 50%	V.↑
Resampling	The sample rate is reduced to 0.8×10^4 and then to 1.6×10^4	R.8 → R.16
Echo	Overlays produce echoes	E.
Noise	50 dB Gauss noise	N.
Butterworth filtering	6th order Butterworth low pass filter, cutoff frequency 3400	B.W
FIR filtering	6th order FIR low pass filter, cutoff frequency 3400	F.I.R

In the experiment, all 13,388 audio test files are respectively operated as shown in Table 1, each test file generates 7 corresponding audio files, constructing the voice perceptual hash chain after the operation and the original voice perceptual hash chain respectively, and calculating the Hamming distance, which is worth 7 sets of BER averages, as shown in Table 2 and compared to the paper [12, 18, 19].

As can be seen from the data listed in Table 2, the algorithm proposed in this paper has a maximum value of 0.2450 for the audio BER mean with the same perceptual content, and the gap with paper [12] is small, indicating that the algorithm in this paper has good robustness. And by comparing with the data of paper [18, 19], for the operation of noise and low-pass filtering, the BER mean of this algorithm is obviously lower, which shows that the algorithm has better robustness in combating these operations.

Table 2. BER average

Action	This paper	Paper [12] I	Paper [12] II	Paper [12] III	Paper [12] IV	Paper [12] V	Paper [18]	Paper [19]
V.↓	0.0310	0.0202	0.0173	0.0346	0.1580	0.1097	0.0004	0.0002
V.↑	0.0821	0.0088	0.0095	0.0188	0.1013	0.1055	0.0116	0.0173
R.8 → R.16	0.0879	0.0841	0.0143	0.0230	0.0906	0.0508	0.0002	0.0026
E.	0.2450	0.1947	0.2002	0.1831	0.2327	0.2324	–	0.1137
N.	0.0260	0.0488	0.1564	0.1285	0.1834	0.0731	0.0581	–
B.W	0.0813	0.1450	0.1641	0.1449	0.1996	0.2109	0.1057	0.1412
F.I.R	0.0488	0.1551	0.1755	0.1611	0.2187	0.2056	0.1214	0.1520

The above experimental results show that the algorithm proposed in this paper has a high accuracy of voice perceptual hash authentication, and has good distinguishability and robustness. It can accurately and reliably authenticate the voice that has been operated by content hold. The disadvantage is that the FAR of this algorithm has not been reduced to the most ideal level.

5 Summary and Prospect

In this paper, a reliable voice perceptual hash authentication algorithm is proposed, which uses the 2DNMF method to degrade the MFCC feature matrix, and finally constructs the perceptual hash chain according to the decomposition results, and the perceptual hash authentication is carried out by normalizing hamming distance measurement. The experimental results show that the algorithm has good distinguishability and robustness, and has better robustness for low-pass filtering, noise and other disturbances. The next research direction is to refine the authentication results, analyze the relationship between the change of perceptual hash chain and content tampering, and detect and locate the tampering of the test voice after authentication.

References

1. Tang, Z., Zhang, X., Huang, L., et al.: Robust image hashing using ring-based entropies. Sig. Process. **93**(7), 2061–2069 (2013)
2. Niu, X., Jiao, Y.: An overview of perceptual hashing. Acta Electron. Sin. **07**, 1405–1411 (2008)
3. Li, Z., Zhu, M., Chen, Z.: Object tracking algorithm based on perception hash technology. J. Image Graph. **20**(006), 795–804 (2015)
4. Zhang, W., Kong, X., You, X.: Secure and robust image perceptual hashing. J. SE Univ. (Nat. Sci. Ed.) **000**(S1), 188–192 (2007)
5. Qin, C., Sun, M., Chang, C.C.: Perceptual hashing for color images based on hybrid extraction of structural features. Sig. Process. **142**, 194–205 (2018)

6. Sabahi, F., Ahmad, M.O., Swamy, M.N.S.: Content-based image retrieval using perceptual image hashing and hopfield neural network. In: 2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 352–355. IEEE (2018)
7. Zhang, Q., Qiao, S., Huang, Y., et al.: A high-performance voice perceptual hashing authentication algorithm based on discrete wavelet transform and measurement matrix. Multimedia Tools Appl. **77**(16), 21653–21669 (2018)
8. Saikia, N., Bora, P.K.: Perceptual hash function for scalable video. Int. J. Inf. Secur. **13**(1), 81–93 (2014)
9. Yang, M., Tang, G., Liu, X., et al.: Low-light image enhancement based on Retinex theory and dual-tree complex wavelet transform. Optoelectron. Lett. **14**(6), 470–475 (2018)
10. Li, J., Wang, H., Jing, Y.: Audio perceptual hashing based on NMF and MDCT coefficients. Chin. J. Electron. **24**(3), 579–588 (2015)
11. Ramalingam, A., Krishnan, S.: Gaussian mixture modeling of short-time Fourier transform features for audio fingerprinting. IEEE Trans. Inf. Forensics Secur. **1**(4), 457–463 (2006)
12. Zhang, Q., Qiao, S., Zhang, T., Huang, Y.: Perceptual hashing authentication algorithm for multi-format audio based on energy to zero ratio. J. Huazhong Univ. Sci. Technol. (Nat. Sci. Ed.) **45**(09), 33–38 (2017)
13. Huang, Y., Zhang, Q., Yuan, Z., Yang, Z.: The hash algorithm of voice perception based on the integration of adaptive MFCC and LPCC. J. Huazhong Univ. Sci. Technol. (Nat. Sci. Ed.) **43**(02), 124–128 (2015)
14. Li, J., Wu, T., Wang, H.: Perceptual hashing based on correlation coefficient of MFCC for voice authentication. J. Beijing Univ. Posts Telecommun. **38**, 89 (2015)
15. Zhang, T.: Research on voice perceptual hashing authentication method and its application in mobile terminal. MS thesis. Lanzhou University of Technology (2018)
16. Bao, C., Bai, Z.: Voice enhancement based on nonnegative matrix factorization: an overview. J. Sig. Process. **36**(6), 791–803 (2020)
17. Zhang, D., Chen, S., Zhou, Z.-H.: Two-dimensional non-negative matrix factorization for face representation and recognition. In: Zhao, W., Gong, S., Tang, X. (eds.) AMFG 2005. LNCS, vol. 3723, pp. 350–363. Springer, Heidelberg (2005). https://doi.org/10.1007/11564386_27
18. Zhang, Y., Mi, B., Zhou, L., Zhang, T.: Speech perception hash authentication algorithm based on short-term autocorrelation. Radio Eng. **49**(10), 899–904 (2019)
19. Zhang, D.: Research on ciphertext voice content authentication and tampering recovery method based on perceptual hash in cloud environment. Lanzhou Univ. Technol. (2020)