International Journal of Innovative Computing, Information and Control
Volume **11**, Number **6**, December **2015**

 ISSN 1349-4198
pp. 2159–2175

ROBUST SPEECH PERCEPTUAL HASHING ALGORITHM BASED ON LINEAR PREDICATION RESIDUAL OF G.729 SPEECH CODEC

Qiuyu Zhang[1], Zhongping Yang[1], Yibo Huang[2]
Shuang Yu[1] and Zhanwei Ren[1]

[1]School of Computer and Communication
Lanzhou University of Technology
No. 287, Langongping Road, Lanzhou 730050, P. R. China
zhangqylz@163.com; zpyang90@126.com; {782809502; 1092813613}@qq.com

[2]College of Physics and Electronic Engineering
Northwest Normal University
No. 967, Anning East Road, Lanzhou 730070, P. R. China
huangyibo1982@163.com

Received May 2015; revised September 2015

Abstract. *In this paper, in order to meet the requirements of speech authentication for low amount of data and low complexity in the bandwidth resource-limited speech communication environment, we present a new robust perceptual audio hashing algorithm combined with the G.729 coding standards. This algorithm extracts the linear prediction residual (LPR) as feature of speech perceptual in the processing of the G.729 speech codec, and extracts the perceptual hash values. The experimental results illustrate the effectiveness of the proposed algorithm in terms of better robustness and discrimination to common speech transmission operation, low time complexity and high efficiency of the proposed algorithm as well, and it can effectively distinguish between the permissible operating and malicious tampering. In the meantime, it has a higher precision for tamper localization.*

Keywords: Speech authentication, Perceptual hashing, G.729 codec, Linear prediction residual, Tamper localization

1. **Introduction.** Speech communication is one of the most basic, direct and important ways in communication. Compared with the analog signal, the digital transmission and storage of speech signals not only have good reliability, anti-jamming capability and the characteristic of rapid transformation, but also are convenient, flexible and easy to keep secret. In order to guarantee the safety and reliability of speech communications, speech authentication must be proposed [1,2]. The traditional digest authentication algorithm has poor robustness and its requirement for resource is very high, its computation is huge, and it cannot be effectively applied to speech mobile communication terminal also. Through verification of the integrity and authenticity of multimedia information content, perceptual audio hashing authentication [3] technology protects multimedia information, makes multimedia audio information services more secure and reliable, and can be used to realize retrieval and authentication of content integrity of audio and broadband audio signal. Therefore, it has been widely researched.

Speech perceptual feature extraction is the key of speech perceptual authentication, and the existing algorithms about speech perceptual feature extraction and processing are based on the human ear psychoacoustics model. The speech perceptual feature extraction is mainly for the logarithmic spectrum coefficient [4], linear prediction coefficient

(LPC) [5], linear frequency spectrum [6], Mel-frequency cepstral coefficients (MFCC) [7], and so on. Chen et al. [8] first obtain the speech Mel cepstral coefficients, and then use the nonnegative matrix decomposition to analyze spectrum coefficient. M. Pavithra et al. [9] using sparse kernel principal component analysis to maximize the reduction of the model data, the original data can reduce the amount of data, thereby improving operation efficiency. In [10] Hilbert transform spectrum estimation method is used to implement robust speech feature extraction, and construct hash function perception. In order to improve the algorithm reliability in the transmission processing, it is combined with the speech coding standards. Jiao et al. [11] proposed a speech perceptual hashing algorithm in compressed domain based on MELP coding standards, Wu and Kuo [12] proposed a content authentication algorithm based on CELP coding standards, Wu and Kuo [13] proposed a content authentication algorithm based on ITU G.723.1 speech coding standards, and Jiao et al. [14] proposed a speech content authentication algorithm based on G.729 speech coding standards. All of the above-mentioned perceptual hashing algorithms based on speech coding standards which extract the related parameters as perceptual values in the process of encoding are of robustness to common speech transmission operations. However, the robustness of them to the white Gaussian noise and the low-pass filter operations is poor and the data volume is huge, and time complexity is high. In addition, the detection and localization of the malicious attacks or tamper of them are not accurate.

From what has been discussed above, in order to reduce the amount of data and the time computational complexity of the perceptual hashing algorithm, we improve the robustness of the common operations of the algorithm, realize detection and localization of the speech during transmission under malicious attacks or tamper. In this study, a robust perceptual audio hashing algorithm based on linear predication residual of G.729 speech codec is proposed. G.729 coding scheme is the telephone bandwidth of speech signal coding standards. G.729 protocol uses CS-ACELP algorithm which is based on the CELP coding model. This algorithm extracts the linear prediction residual as perceptual feature value in the processing of the encoding. On the one hand, small data volume of perceptual feature makes it easy to process the digest of the perceptual hashing function, and ensures that the perceptual hashing can satisfy the requirements of the nature of the hash function. On the other hand, this algorithm has low time complexity, and can realize accurate detection and localization in the processing of speech transmission by malicious tamper. It can satisfy the requirements real-time, robustness and security of speech information in the mobile computing environment.

The rest of this paper is organized as follows. Section 2 describes the basic theory of G.729 speech coding standards and linear prediction residual (LPR). A detailed Speech Perceptual Hashing Authentication scheme is described in Section 3. Subsequently, Section 4 gives the experimental results as compared with other related methods. Finally, we conclude our paper in Section 5.

2. **The Preliminaries.**

2.1. **G.729 speech coding standards.**

The G.729 speech coding standard was published in 1996 by the ITU-T 8 kbps/s speech coding protocol [15], using conjugate structure algebraic code excited linear prediction (CS-ACELP) method to 8 kbps bit rate speech coding, the coding method based on the code excitation linear prediction. The G.729 speech coding standard mainly consists of seven parts: (1) Pre-processing; (2) Linear prediction analysis and quantization of the LPC coefficients; (3) Open loop pitch estimation; (4) Close loop pitch estimation; (5) The search of adaptive codebook; (6)

The search of fixed codebook; (7) Codebook gain quantization. The G.729 belongs to low speech coding, and it can solve the insufficient bandwidth problem in the processing of speech transmission and reduce the coding rate in the case of ensuring high speech quality.

2.2. **Linear prediction residual.** The basic idea behind linear prediction is forming a value of a speech clip (frame) with a weighted linear combination of several past linear predictors. Linear prediction analysis, the most effective method of speech signal analysis, can be applied to estimation of many basic speech parameters, such as the pitch period, and spectrum signature. To speech signal $s(n)$, p-order linear prediction can be defined as follows:

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) + r(n), \quad i = 1, 2, \cdots, p \tag{1}$$

where $\hat{s}$ is predicted value of $s(n)$, $r(n)$ is the linear prediction residual, a_i is the linear prediction coefficient, and p is the order of prediction.

In the linear prediction coding, in order to improve the robustness of the linear prediction coefficients, there are many representing methods equivalent to linear prediction coefficients. There are many derived parameters generated from linear prediction parameters, such as linear prediction cepstrum coefficients (LPCC) [16], linear spectrum pair (LSP) [14], linear spectral frequency (LSF) [11], and linear prediction residual (LPR) [17,18].

In this study, the speech signal $s(n)$ which is regarded as the convolution of the glottis excitation signal $e(n)$ and the channel impulse response signal $v(n)$ is as follows:

$$s(n) = e(n) \otimes v(n) \tag{2}$$

The main idea behind linear prediction analysis of speech signal is relationship between the speech sampling signal $s(n)$ and the glottis excitation signal $e(n)$, which can be represented with difference equation, as follows:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Ge(n) \tag{3}$$

where p is the order of linear predictor, a_k is the prediction coefficient, and G is the amplitude factor. The first item on the right is the output of p-order linear predictor assuming that G is 1 in the ideal state, bringing it down as follows:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{4}$$

According to (4), the LPR signal $R(n)$ can be expressed as (5).

$$R(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) = e(n) \tag{5}$$

In order to achieve the smallest prediction error, the minimum mean square error criterion is usually used in the solving process of linear prediction coefficient a_k. The Levinson-Durbin algorithm is used in this study. The mean square error is expressed as follows.

$$E_p = \sum_{-\infty}^{+\infty} [s(n) - \hat{s}(n)]^2 \tag{6}$$

Namely

$$E_p = E\left[e(n)\left\{s(n) - \sum_{k=1}^{p} a_k s(n-k)\right\}\right] = r(0) - \sum_{k=1}^{p} a_k r(k) \tag{7}$$

Due to

$$E\left[e^2(n)\right] = E\left[\left[s(n) - \sum_{k=1}^{p} a_k s(n-k)\right]^2\right] \tag{8}$$

assuming

$$\frac{\delta E\left[e^2(n)\right]}{\delta a_k} = 0, \quad k = 0, 1, \cdots, p \tag{9}$$

substituting (5) into (9) gets (10)

$$E\left[s(n)s(n-k) - \sum_{k=1}^{p} a_k s(n)s(n-k)\right] = r(k) - \sum_{k=1}^{p} a_k r(k-j) \tag{10}$$

In (10)

$$r(k) = E\left[s(n)s(n-j)\right] \tag{11}$$

The $r(k)$ in (11) means the autocorrelation function of speech signal sequence. Solve (7) and (10) to get linear prediction coefficients a_k, whose matrix equations form can be expressed as follows.

$$\begin{bmatrix} r(0) & r(1) & \dots & r(p) \\ r(1) & r(0) & \dots & r(p-1) \\ r(2) & r(1) & \dots & r(p-2) \\ \vdots & \vdots & \vdots & \vdots \\ r(p) & r(p-1) & \dots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{bmatrix} = \begin{bmatrix} E_p \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{12}$$

Substituting a_k, the linear prediction coefficients, which can be achieved by using the Levinson-Durbin algorithm, into (5), then get the linear prediction residual signal.

3. The Proposed Scheme.

3.1. Speech perceptual feature value extraction.

Compared with general audio signals, the data volume of speech signal is small, first of all. The G.729 firstly pre-processes the input signal, and then conducts 10-order linear prediction analysis for every 10ms as a frame. Considering that the order of linear predictor filter in G.729 is too low, there are still a lot of channel information in residual signal.

In order to reduce the complexity and authentication data volume of algorithm, the algorithm proposed in this study extracts the linear prediction residual as speech perceptual feature value in the processing of the G.729 coding. The speech perceptual feature value extraction process is shown in Figure 1.

The detailed steps of the speech perceptual feature value extraction process are showed as follows.

Step 1: Pre-processing. First of all, the G.729 pre-processes the input speech signal, including signal calibration and high-pass filtering, and, as the main part in the G.729

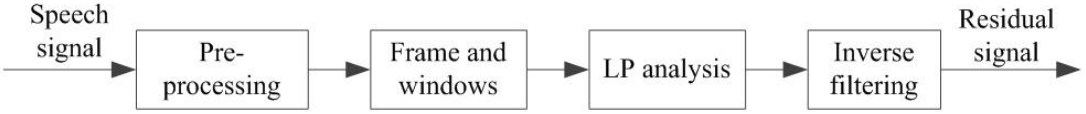


FIGURE 1. Speech feature extraction block diagram

coding pre-processing, the signal calibration makes the signal amplitude reduced by half, thereby reducing overflow probability in the fixed point operation. The high-pass filtering is aimed to filter out the low frequency part of signal, to improve the useful high frequency spectrum, to reduce the edge effects and to eliminate noise. The high-pass filter is a 2-order filter of pole/zero type with a cutoff frequency of 140 Hz. Associate signal calibration with high-pass filter, and the transfer function is as follows.

$$H(z) = \frac{0.46363718 - 0.92724705z^{-1} + 0.46363718z^{-2}}{1 - 1.9059465z^{-1} + 0.9114024z^{-2}} \tag{13}$$

where after $H(z)$ filtering of signal is called the pre-processing signal, denoted by $S(n)$.

Step 2: Windowing and framing. G.729 uses 10-order linear prediction filter to obtain the pre-processing signal $S(n)$ by pre-processing, and adds window function to speech signal $S(n)$ in order to ensure the short-time of the linear prediction analysis, and for after the pre-processing of speech signal $S(n)$ adding window function. The window function consists of two parts: the first part is a half of the hamming window, and the second part is a quarter of the cosine function window, as follows.

$$W(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{399}\right) & n = 0, \cdots, 199 \\ \cos\left[\frac{2\pi(n-200)}{159}\right] & n = 200, \cdots, 239 \end{cases} \tag{14}$$

The speech after windowing is:

$$S_w(n) = W(n)S(n), \quad n = 0, \cdots, 239 \tag{15}$$

Frame the speech signal $S_w(n)$ after windowing, and speech signal for per frame T_i can be obtained, as follows.

$$T_i = \{T_i(k) | i = 1, 2, \cdots, n, k = 1, 2, \cdots, l\} \tag{16}$$

where l is the length of every frame, and n is the number of frames.

Step 3: Linear prediction residual feature extraction. Conduct linear prediction analysis to each frame T_i speech signal. Use Levinson-Durbin algorithm to get results during linear prediction analysis and regard linear prediction residual as the speech perceptual feature with the extraction process shown in Table 1.

TABLE 1. The process of the LPR features extraction

Input: the each frame speech signal of frame and window
Output: the each frame LPR signal

for $i = 1$ to p
  $R^{[0]} = r(0)$
  $k_i = \left[r(i) - \sum_{j=1}^{i-1} a_j^{i-1} r(i-j)\right] \Big/ R^{i-j}, \ 1 \le i \le p$
  $a_i^{[i]} = k_i$
  for $j = 1$ to $i - 1$
    $a_j^{[i]} = a_j^{[i-1]} + k_i a_{i-j}^{[i-1]}$
  end
  $R^{[i]} = (1 - k_i^2) R^{(i-1)}$
end

The $R^{[i]}$ is the linear prediction residual in Table 1, the superscript is the number of predictor order, and k_i is reflection coefficient or partial correlation coefficient.

3.2. **Hashing generation.** In the analysis of speech signal, the simple structure and small data volume of binary data make it easy to be analyzed and operated. Therefore, we quantify the extracted perceptual features using median quantization, as follows.

$$H(n) = \begin{cases} 1 & R(n) \geq \hat{R} \quad 1 \leq n \leq L \\ 0 & R(n) < \hat{R} \quad 1 \leq n \leq L \end{cases} \tag{17}$$

where $\hat{R}$ is the median of $R(n)$, and L is the number of frames.

Conduct median quantization for residual signal $R(n)$ to obtain perceptual hash sequence $H(n)$. The hash generation process in the algorithm proposed in this study is shown in Table 2.

TABLE 2. Hashing generation of proposed algorithm

Input: the speech feature value **Output**: the perceptual hashing sequence
$\hat{R} = median(R)$; $\hat{R}$ is median of sequences R
for $i = 1$ to L
if $R(i) \geq \hat{R}$
$H(i) = 1$
else
$H(i) = 0$
end
end

3.3. **Transmission and matching.** In the speech authentication, the first problem needed to be considered is transmission. Adoption of different transmission models has a great influence on precision and accuracy of authentication. The algorithm proposed in this study is a perceptual audio hashing authentication algorithm based on linear prediction residual parameters of G.729 speech coding standards. Extracting the linear prediction residual as the authentication data at the encoding end can send authentication data and bit stream at the same time. Authentication data can be obtained from bit stream according to the relevant decoding operation at the decoding end. On the sending side, the algorithm extracts the linear prediction residual parameters in the G.729 encoding process to obtain perceptual hash H_1 through calculating, and then sends it with speech coding bit stream. At the receiver, the algorithm extracts the linear prediction residual parameters again when receiving bit stream to obtain perceptual hash H_2 by conducting hashing generation. Calculate the difference between the perceptual hashes on the two sides to get the absolute value, hash mathematical distance D_h, as the match value. This is defined by:

$$D_h(H_1, H_2) = \sum_m |h_1(m) - h_2(m)|, \quad m = 1, 2, \cdots, n \tag{18}$$

Comparing the match value above and the match threshold in (18), if the match value is less than the match threshold, it will pass; if the match value is greater than the match

threshold, it will not pass, as follows.

$$\begin{cases} NoPass & D_h(i) \geq \tau \quad 1 \leq i \leq n \\ Pass & D_h(i) < \tau \quad 1 \leq i \leq n \end{cases} \tag{19}$$

where n is the number of speech clips, and τ is the match threshold in (19).

The algorithm proposed in this study, combined with G.729 coding standards, extracts the linear prediction residual as the perceptual feature in the processing of coding in encoder, sends it with coding bit stream at the same time, extracts linear prediction residual again through different kinds of usual operations in transmission, calculates the mathematics distance between the two linear prediction residual perceptual features to get the match value, and finally compares the match value and the match threshold to get the authentication result.

The block diagram of the algorithm proposed in this study is shown in Figure 2.

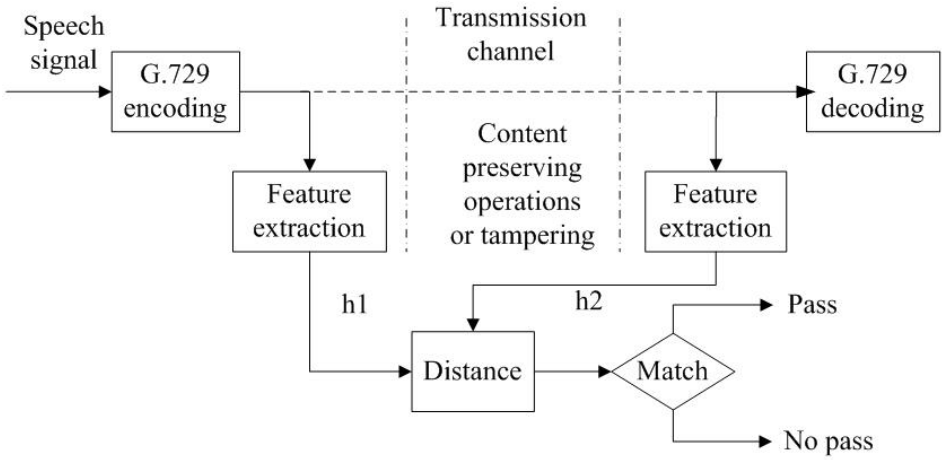


FIGURE 2. Proposed algorithm block diagram

4. **Experimental Results and Analysis.**

4.1. **Experimental environment.** Experiments in this study use the TIMIT speech library, including 400 clips at the length 4 s, and conduct encoding and decoding through the G.729 coding standards. Followings are experimental speech parameters: the coding standard is G.729, the sampling rate is 8000 Hz, the channel is mono, the sampling precision is 16 bits and the format is WAV.

The experimental hardware platform is Inter Core i3, 2450M, 2 G, 2.27 GHz, and software environment is the MATLAB R2012b under Windows 7 OS.

4.2. **Bit error rate, false accept rate and false reject rate.** This paper uses binary perceptual hash sequences to calculate, the evaluation parameter is bit error rate (BER), percentage of error bits in total bits, and the calculation formula is as follows.

$$BER = \frac{\sum_{i=1}^{N} (|h_1 - h_2|)}{N} = \frac{\sum_{i=1}^{N} (h_1 \oplus h_2)}{N} \tag{20}$$

where N is the number of speeches, and h_1 and h_2 are respectively perceptual hashing in (20).

The error recognition rate, which is also known as false accept rate (FAR), is the proportion of speech of different perceptual contents accepted by the system mistakenly.

The calculation formula is as follows.

$$FAR(\tau) = \int_{-\infty}^{\tau} f(\alpha\,|\mu,\sigma)d\alpha = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\alpha-\mu)^2}{2\sigma^2}} d\alpha \tag{21}$$

where μ is the mean, and σ is the standard deviation in (21).

The error rejection rate, which is also known as false reject rate (FRR), is the proportion of speech of the same perceptual content rejected by the system mistakenly. The calculation formula is as follows.

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} f(\alpha\,|\mu,\sigma)d\alpha = 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\alpha-\mu)^2}{2\sigma^2}} d\alpha \tag{22}$$

4.3. **Analysis of robustness.** Conduct operations such as volume adjustment, resampling, echo added, noise added, cutting, low-pass filtering and MP3 compress. Various content keeping operations are shown in Table 3.

The BER of content keeping operation above is shown in Table 4.

As shown in Table 4, the biggest BER of different kinds of content keeping operations above is below 0.28.

The average BER under various content keeping operations of the algorithm is shown in Table 5.

As shown in Table 5, the average BER values of the algorithm proposed under various operations described in Table 3 are less than the ones of the LSP algorithm [11], implying the better robustness of the algorithm proposed to the content keeping operations.

TABLE 3. Operating means and corresponding level

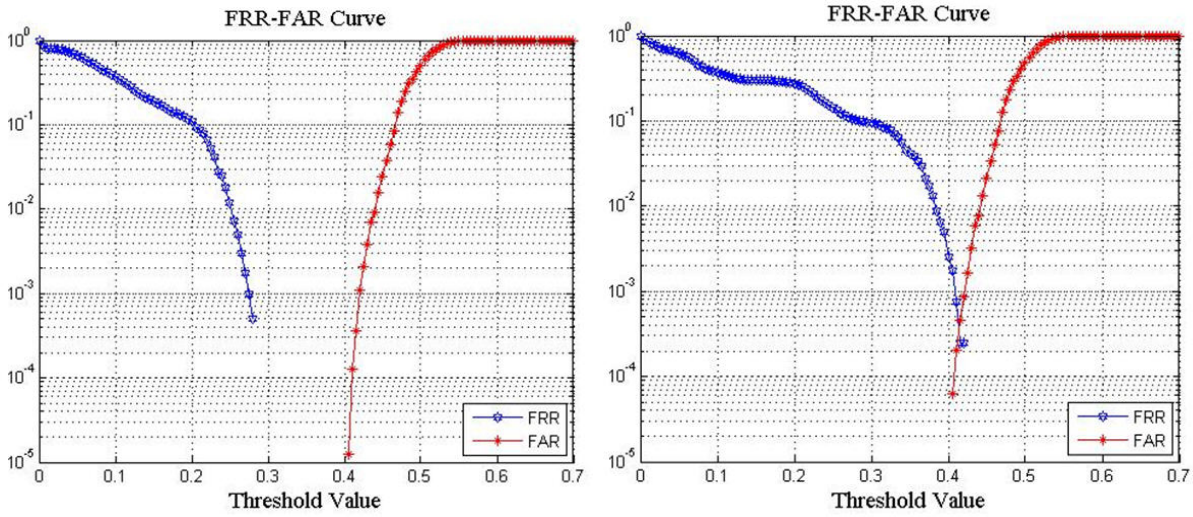
Operating means	Level
Adjust volume	150%
Adjust volume	50%
Resampling	8-4-8 bits/sample
Resampling	8-16-8 bits/sample
Addition echo	300 ms, 10%
Addition noise	50 dB
Cut	300 ms
Low-pass filter	4 kHz 6-order
MP3 compress	32 kbps
MP3 compress	128 kbps

TABLE 4. The BER of the proposed algorithm

Parameters	Average BER	Standard deviation	Width BER
Increase volume	0.1168	0.0290	0.2300
Decrease volume	0.1678	0.0411	0.2800
Resampling (4-8)	0.0041	0.0033	0.0175
Resampling (16-8)	0.0007	0.0014	0.0100
Addition echo	0.2208	0.0218	0.2800
Addition noise	0.0421	0.0204	0.1375
Cut	0.0653	0.0091	0.0950
Low-pass filter	0.1155	0.0212	0.2050
MP3 compress (32 kbps)	0.0887	0.0144	0.1315
MP3 compress (128 kbps)	0.0479	0.0103	0.0819

TABLE 5. Average BER of the proposed algorithm

Operating means	Proposed algorithm	LSP algorithm
Parameters	Average BER	
Increase volume	0.1168	0.0664
Decrease volume	0.1678	0.0229
Resampling (4-8)	0.0007	0.0018
Resampling (16-8)	0.0041	0.0112
Addition echo	0.2208	0.2211
Addition noise	0.0421	0.0486
Cut	0.0653	0.0655
Low-pass filter	0.1155	0.1048
MP3 compress (32 kbps)	0.0887	0.2352
MP3 compress (128 kbps)	0.0479	0.3447



(a) Proposed algorithm (b) LSP algorithm

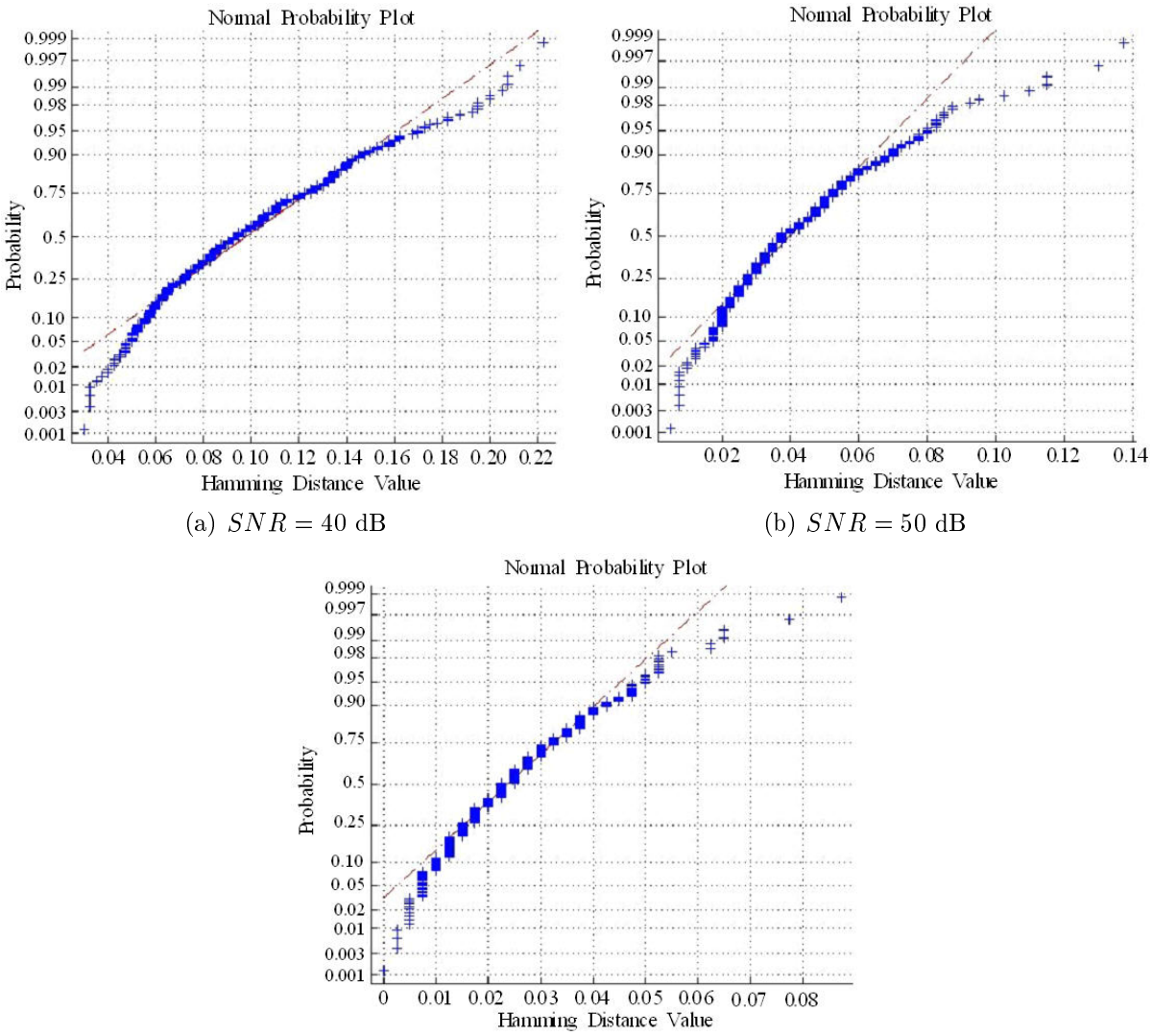
FIGURE 3. The algorithm FRR-FAR curves

The FRR-FAR curve of the LSP algorithm is shown in Figure 3(b), and the FRR-FAR curve of the proposed algorithm is shown in Figure 3(a).

In Figure 3(a), the perceptual hash is extracted from speeches with the same content, whose BER values are below threshold $\tau = 0.28$.

Experimental results show that the curves of FRR and FAR do not intersect, and the FRR curve has obvious convergence, and has a relatively broad decision interval. When the decision threshold τ is between 0.28 to 0.4, the algorithm can conduct authentication both same speech clips and different clips of speeches accurately, at the same time, it can conduct authentication among speech clips through content keeping operations and content malicious attacks, demonstrating that the algorithm proposed has good discrimination and robustness at the same time. As what can be seen from Figure 3, the robustness and discrimination of the algorithm in this study are better than the LSP algorithm.

4.4. **The robustness of white Gaussian noise.** The speech signal is easily affected by noise in the processing of transmission, and the normal distribution curves and FRR-FAR curves of algorithm proposed in this study under white Gaussian noise operations with different SNR are shown in Figure 4 and Figure 5.



(a) $SNR = 40$ dB

(b) $SNR = 50$ dB

(c) $SNR = 55$ dB

FIGURE 4. Normal distribution curves under addition white Gaussian noise

The SNR in Figure 4(a) is 40 dB, the SNR in Figure 4(b) is 50 dB, and the SNR in Figure 4(c) is 55 dB. As what can be seen from Figure 4 clearly, the maximum value of the normal distribution curve under the white Gaussian noise adding different SNR values is less than 0.28. As what can be seen from Figure 5 obviously, the algorithm proposed in this study still keeps good robustness and discrimination under the white Gaussian noise operations adding different SNR values.

4.5. **The robustness under low-pass filtering.** The operational robustness to low-pass filtering [8-11] in the existing perceptual audio hashing algorithms is pretty poor. The normal distribution curves and the FRR-FAR curves of the algorithm proposed in this study under different cut-off frequency FIR low-pass filtering operations are shown in Figure 6 and Figure 7.

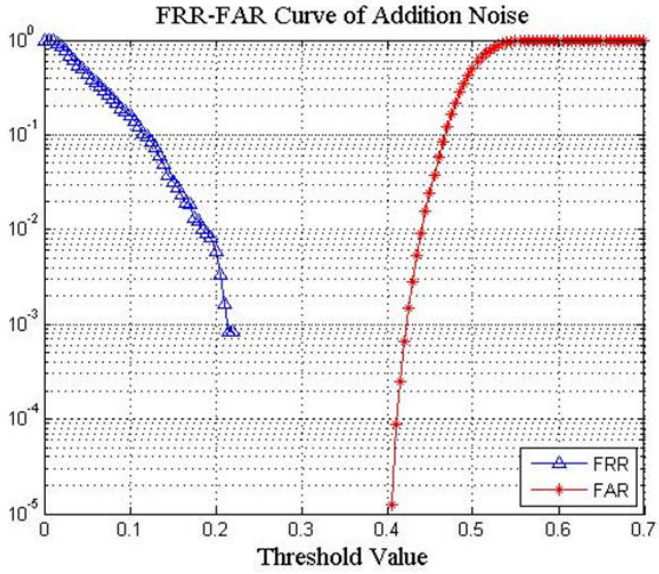
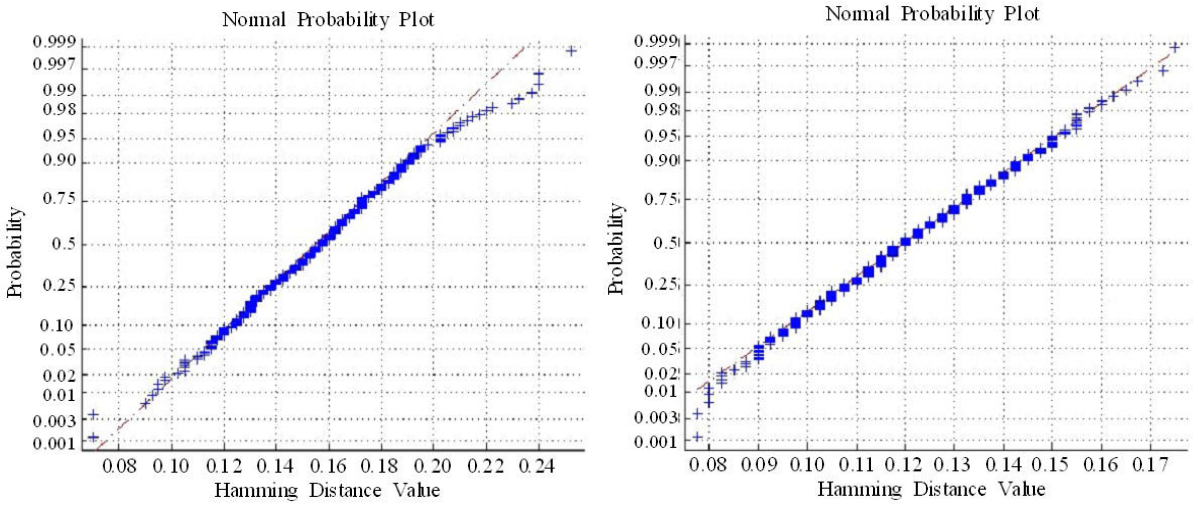
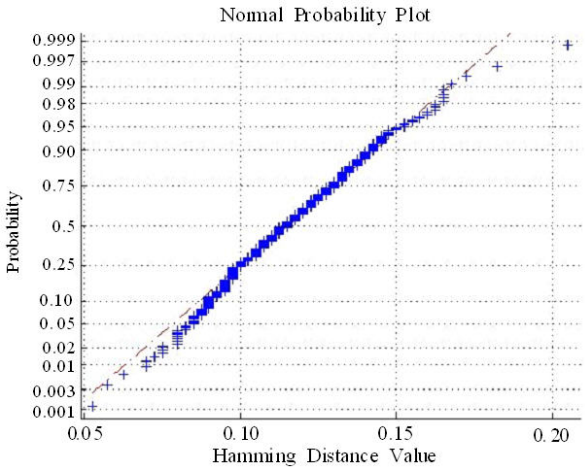


FIGURE 5. *FRR-FAR* curves under addition white Gaussian noise



(a) 6-order 2 kHz

(b) 6-order 3.4 kHz



(c) 6-order 4 kHz

FIGURE 6. Normal distribution curves under FIR low-pass filter

As what can be seen from Figure 6 clearly, the 6-order 2 kHz FIR filter in Figure 6(a), the 6-order 3.4 kHz FIR filter in Figure 6(b) and the 6-order 4 kHz FIR filter in Figure 6(c), the maximum value of the normal distribution curve under different low-pass filtering operations is less than 0.28.

As what can be seen from Figure 7 obviously, the algorithm proposed in this study has good robustness and discrimination under the FIR low-pass filtering operations of different cut-off frequency and improves robustness of the algorithm for low-pass filtering operations.

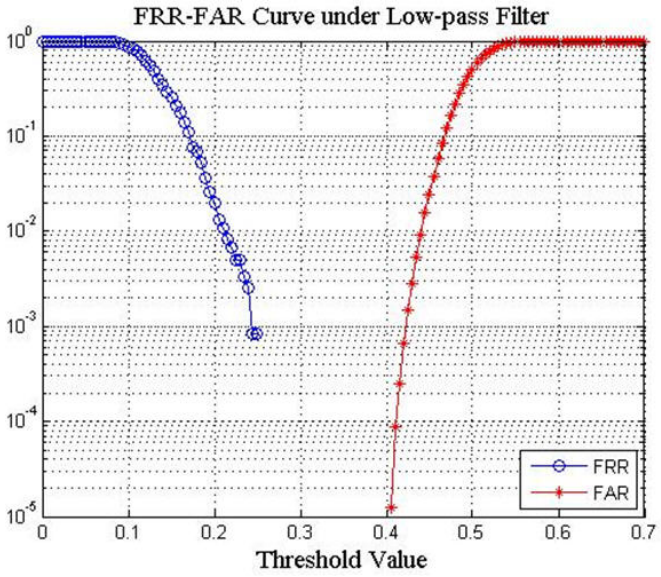


FIGURE 7. FRR-FAR curves under FIR low-pass filter

4.6. **Analysis of discrimination.** This paper totally gets 79800 BER data by conducting pairwise comparison between perceptual hashing values from 400 different speech clips, with normal distribution curves of BER shown in Figure 8.

The perceptual hash mathematical distance in the algorithm proposed in this study can be approximately seen as hamming distance. According to central-limit distance, the BER approximately obeys the normal distribution with mean $\mu = 0.5$ and standard deviation $\sigma = \sqrt{1/4N}$. The length of perceptual hash sequence in the algorithm proposed in this study is 400, namely $N = 400$. According to the theoretical calculation, theory parameters are ($\mu = 0.5$, $\sigma = 0.0250$). The parameters of the experimental measuring are ($\mu = 0.4984$, $\sigma = 0.0252$), which is very close to the theoretical values. The FAR curve drawn is shown in Figure 9.

According to the theory parameter values μ and σ, the FAR curve is shown in Figure 9 with dotted line, the smaller the FAR value is, the better the discrimination is. As what can be seen from Figure 8 and Figure 9, the algorithm proposed in this study has good discrimination.

According to (21) FAR values of algorithm can be obtained under the different match threshold τ, as shown in Table 6, and the FAR values are very small. Compared with the LSP algorithm, the algorithm proposed is worse than the LSP algorithm, but when $\tau = 0.3$, the $FAR = 1.9199\text{e-}15$, which means that if $\tau = 0.3$, there will be approximately two speech clips which is wrong in 10^{15} speech clips. So it can meet requirement of the people who ask for speech perceptual authentication; therefore, the algorithm has good discrimination.

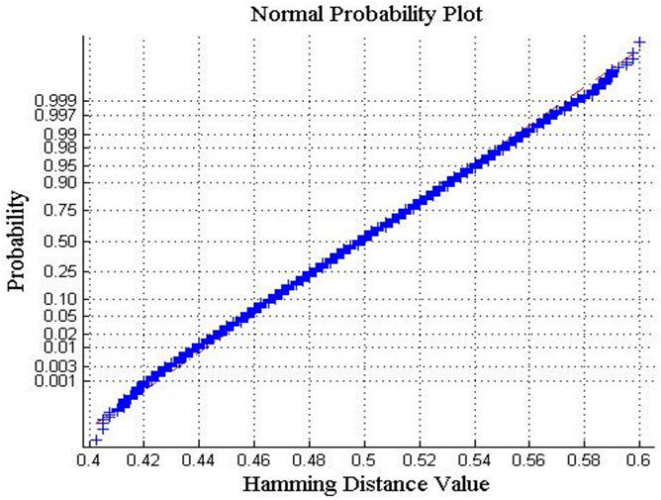


FIGURE 8. Normal distribution curves of the algorithm

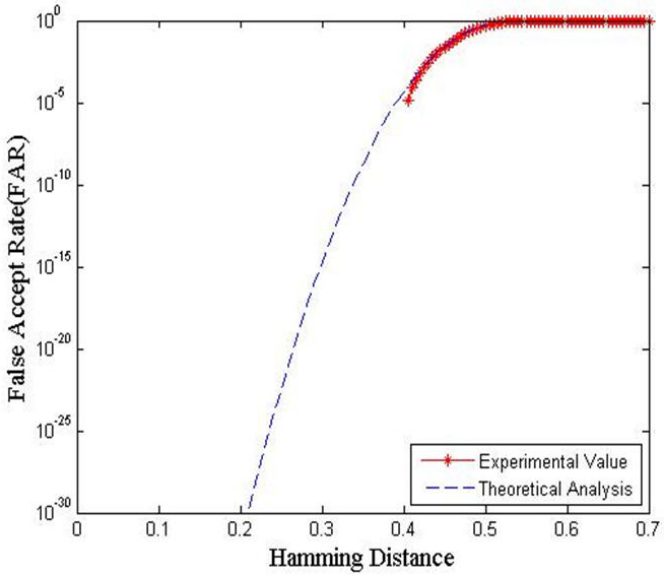


FIGURE 9. FAR curves of the algorithm

TABLE 6. FAR of the proposed algorithm and LSP algorithm

Threshold	Proposed algorithm	LSP algorithm
τ	FAR	
0.35	2.0583e-09	1.3858e-09
0.30	1.9199e-15	1.1060e-15
0.25	3.7610e-23	1.8256e-23
0.20	1.5155e-32	6.1069e-33
0.15	1.2417e-43	4.0930e-44

4.7. **Tamper detection and localization.** The speech signal is vulnerable to illegal tamper attacks in the processing of transmission. In order to guarantee the reliability of the speech content, the perceptual audio hashing algorithm should have sensitive and accurate tamper detection ability. To test sensitivity to content tamper of the algorithm

proposed, different speeches from the same speaker are used to replace parts in original speech clips, tampered at the length of about 500 ms. According to the standard sound speed 600 words per minute, the content of the replacement is about three words. The experimental cumulative probability curve of the bit-error-rate got from 400 comparisons is shown in Figure 10 with dotted line, and the experimental cumulative probability curve of the bit-error-rate got from content keeping operations is shown in Figure 10 with solid line.

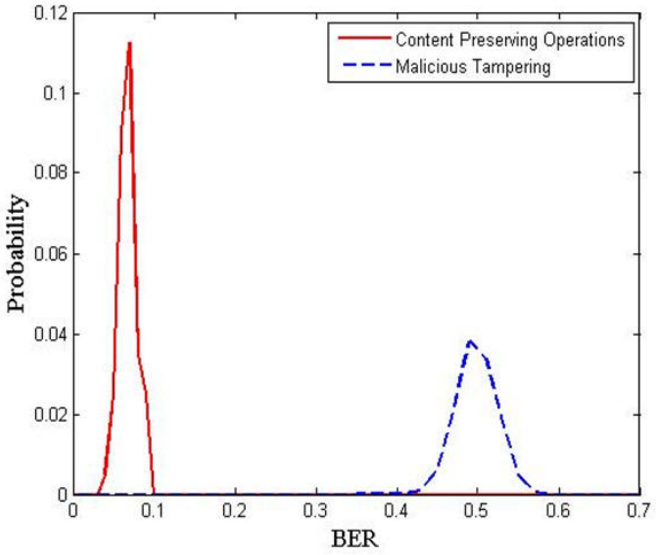


FIGURE 10. Probability plot of BER

As shown in Figure 10, the two curves separated from each other, meaning that the algorithm proposed has the ability of tamper detection.

Following is analysis of tamper localization. Due to the binary perceptual hash used in this algorithm, we can determine whether the information has been tampered or not by comparing the perceptual hash. Assuming that the perceptual hash of the original speech signal is h_{orig} and the perceptual hash of the tamper speech signal is h_{tam}, the determination of tamper localization according to the perceptual hash is as (23) and (24):

$$h_{orig}(i) = h_{tam}, \quad 1 \leq i \leq l \tag{23}$$

$$h_{orig}(i) \neq h_{tam}, \quad 1 \leq i \leq l \tag{24}$$

where l is the length of perceptual hash.

If the perceptual hash h_{orig} and h_{tam} meet (23), it is not tampered. If the perceptual hash h_{orig} and h_{tam} satisfy (24), the h_{tam} is tampered, and the tamper location is $h_{tam}(i)$, and the tamper location can be determined according to (24). The experiment randomly selects one 4 s speech clip, and randomly replaces three places in it greater than 10 frames. The schematic diagrams of tamper localization perceptual hash in time domain are shown in Figure 11, the tampered area are areas included in elliptic curves. It can be seen that the algorithm proposed has the ability of tamper detection, and it can be able to accurately realize tamper detection and localization at one or more point.

4.8. **Analysis of efficiency.** The characteristics of the algorithm proposed in this study are small authentication data, low time complexity and high efficiency. In terms of perceptual hashing algorithm, the algorithm proposed in this study extracted the linear prediction residual as perceptual feature in the processing of G.729 codec. The proposed

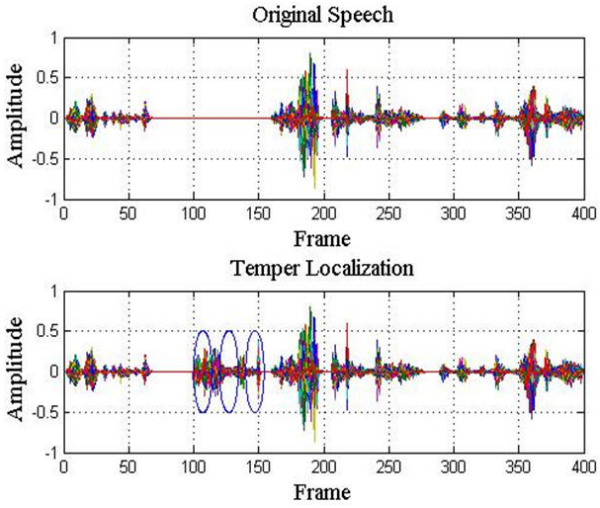


FIGURE 11. Tamper localization schematic diagrams

TABLE 7. Run time

Algorithm	Proposed algorithm	LSP algorithm
Operating means	Average run time (s)	
File lengths	4 s	4 s
Platform working frequency	2.27 GHz	2.27 GHz
Feature extraction	0.076473	0.700739
Hashing structure	0.008508	0.007888
Total	0.084981	0.708627

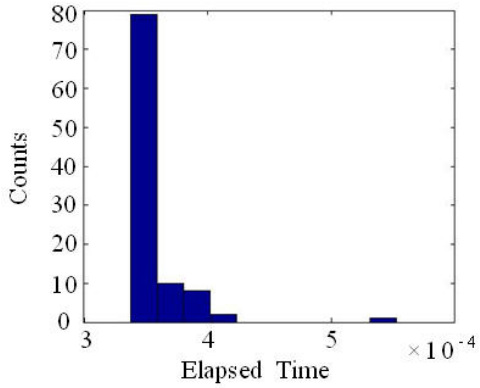


FIGURE 12. Histogram of authentication time

algorithm in this study greatly reduced the time cost of the extraction of perceptual feature and the time complexity of the algorithm, and improved the efficiency of the algorithm. The experiment randomly extracted 100 speech clips from the speech library and did statistics on algorithm running time for totally 100 times and the average run time is shown in Table 7.

The histogram of authentication time is shown in Figure 12. The authentication time of the proposed algorithm concentrates from 3.4e-04 s to 3.5e-04 s, and the average authentication time $t = 3.5682$e-04 s. The authentication time is steady and short, which means that the proposed algorithm can satisfy the requirement of real-time application.

5. **Conclusions.** A robust perceptual audio hashing algorithm combined with G.729 speech coding standards based on linear prediction residual parameters is proposed in this study. The proposed algorithm extracts the linear prediction residual parameters as perceptual feature value in the processing of the speech coding, and sends it with coding bit stream at the same time. At the receiver, when after receiving the speech bit stream, it extracts the perceptual feature value again and calculates the mathematics distance of perceptual feature values from both sender and receiver to conduct matching and authentication. The experimental results illustrate that compared with the LSP algorithm, the proposed algorithm not only has a good robustness and discrimination to common speech channel transmission operation, but also has a better robustness and discrimination under the white Gaussian noise and low-pass filtering operations, and low time computational complexity, small data volume of the perceptual hashing and high efficiency. In addition, it can be able to accurately realize tamper detection and localization at one or more point and satisfy the real-time and robustness requirement of the existing mobile speech communication.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (No. 61363078), the Natural Science Foundation of Gansu Province of China (No. 1212RJZA006, No. 1310RJYA004). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] G. Grutzek, J. Strobl, B. Mainka, F. Kurth, C. Pörschmann and H. Knospe, Perceptual hashing for the identification of telephone speech, *Proc. of the Speech Communication; 10. ITG Symposium; Proc. of VDE*, Braunschweig, Germany, pp.1-4, 2012.

[2] Z. H. Liu and H. X. Wang, A novel speech content authentication algorithm based on Bessel-Fourier moments, *Digital Signal Processing*, vol.24, no.1, pp.197-208, 2014.

[3] X. M. Niu and Y. H. Jiao, An overview of perceptual hashing, *Acta Electronica Sinica*, vol.36, no.7, pp.1405-1411, 2008.

[4] H. Özer, B. Sankur, N. Memon and E. Anarim, Perceptual audio hashing functions, *EURASIP Journal on Applied Signal Processing*, vol.2005, no.12, pp.1780-1793, 2005.

[5] P. Lotia and D. M. R. Khan, Significance of complementary spectral features for speaker recognition, *International Journal of Research in Computer and Communication Technology*, vol.2, no.8, pp.579-588, 2013.

[6] M. Nouri, N. Farhangian, Z. Zeinolabedini and M. Safarinia, Conceptual authentication speech hashing base upon hypotrochoid graph, *Proc. of the 6th IEEE International Conf. Symposium Telecommunications*, Tehran, Iran, pp.1136-1141, 2012.

[7] V. Panagiotou and N. Mitianoudis, PCA summarization for audio song identification using Gaussian mixture models, *Proc. of the 18th IEEE International Conf. Digital Signal Processing*, Fira, Greece, pp.1-6, 2013.

[8] N. Chen, H. D. Xiao and W. G. Wan, Audio hash function based on non-negative matrix factorisation of mel-frequency cepstral coefficients, *Information Security, IET*, vol.5, no.1, pp.19-25, 2011.

[9] M. Pavithra, G. Chinnasamy, A. Periasamy and S. Muruganand, Feature matching by SKPCA with unsupervised algorithm and maximum probability in speech recognition, *Journal of Management and Science*, vol.1, no.1, pp.11-15, 2011.

[10] H. Zhao, H. Liu, K. Zhao and Y. Yang, Robust speech feature extraction using the Hilbert transform spectrum estimation method, *International Journal of Digital Content Technology and Its Applications*, vol.5, no.12, pp.85-95, 2011.

[11] Y. H. Jiao, Q. Li and X. M. Niu, Compressed domain perceptual hashing for MELP coded speech, *Proc. of the IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Harbin, China, pp.410-413, 2008.

[12] C. P. Wu and C. C. J. Kuo, Speech content authentication integrated with CELP speech coders, *Proc. of the IEEE Int. Conf. Multimedia and Expo.*, Tokyo, Japan, pp.1009-1012, 2001.

[13] C. P. Wu and C. C. J. Kuo, Speech content integrity verification integrated with ITU G.723.1 speech coding, *Proc. of the IEEE Int. Conf. Info. Tech.: Coding and Computing*, Las Vegas, NV, USA, pp.680-684, 2001.

[14] Y. H. Jiao, Y. Tian, Q. Li and X. M. Niu, Content integrity verification for G.729 coded speech, *Proc. of the IEEE Int. Conf. Intelligent Information Hiding and Multimedia Signal Processing*, Kaohsiung, Taiwan, vol.2, pp.295-300, 2007.

[15] D. Yessad and A. Amrouche, Robust regression fusion of GMM-UBM and GMM-SVM normalized scores using G729 bit-stream for speaker recognition over IP, *International Journal of Speech Technology*, vol.17, no.1, pp.43-51, 2014.

[16] Y. J. Yuan, P. H. Zhao and Q. Zhou, Research of speaker recognition based on combination of LPCC and MFCC, *Proc. of the IEEE Int. Conf. Intelligent Computing and Intelligent Systems*, Xiamen, China, vol.3, pp.765-767, 2010.

[17] K. S. R. Murty, V. Boominathan and K. Vijayan, Allpass modeling of LP residual for speaker recognition, *Proc. of the IEEE Int. Conf. Signal Processing and Commun.*, Bangalore, India, pp.1-5, 2012.

[18] A. P. Prathosh, T. Ananthapadmanabha and A. Ramakrishnan, Epoch extraction based on integrated linear prediction residual using plosion index, *IEEE Trans. Audio, Speech, and Language Processing*, vol.21, no.12, pp.2471-2480, 2013.