The Speech multi features fusion perceptual hash algorithm based on tensor decomposition

Y B Huang¹, M H Fan¹ and Q Y Zhang²

¹ College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, Gansu, China
² School of Computer and Communication, Lanzhou University of Technology, Lanzhou, Gansu, China
E-mail: huang_yibo@foxmail.com

Abstract. With constant progress in modern speech communication technologies, the speech data is prone to be attacked by noise or maliciously tampered. In order to make the speech perception hash algorithm has strong robustness and high efficiency, this paper put forward a speech perception hash algorithm based on the tensor decomposition and multi features is proposed. This algorithm analyses the speech perception feature acquires each speech component wavelet packet decomposition. LPCC, LSP and ISP feature of each speech component are extracted to constitute the speech feature tensor. Speech authentication is done by generating the hash values through feature matrix quantification which use mid-value. Experimental results showing that the proposed algorithm is robust for content to maintain operations compared with similar algorithms. It is able to resist the attack of the common background noise. Also, the algorithm is highly efficiency in terms of arithmetic, and is able to meet the real-time requirements of speech communication and complete the speech authentication quickly.

1. Introduction

Speech messages as human communication means is the most natural, effective and convenient. Hash function is based on the perception of human psychology multimedia information processing theory, multimedia data is unidirectional mapped to a data set of multimedia perception summary sets. Only the perceptions of simple and fast hash function to meet multimedia massive data analysis applications. Therefore, hash function is very suitable for application in the field of information security¹.

At present, the main methods of speech perception hash feature extraction and processing are as follows: Mel frequency cepstrum coefficient(MFCC), Hilbert transformation, linear prediction coefficient, the buck with energy², the energy ratio of wavelet features³ and spectrum energy feature⁴ can also be used as the characteristics of the speech perception hash values. In literature [5], a perceptual hashing algorithm based on Mel cepstral coefficient combined with linear prediction cepstral coefficient is proposed, which has good robustness and tampering localization, but it is not effective in differentiating and content preserving operations. The literature [6] proposes a speech-aware hashing authentication algorithm based on the MFCC correlation coefficient, which considers security, it has good robustness and discrimination, but is not efficient. The literature [7] Hilbert transform spectrum estimation method is used to implement robust speech feature extraction, construct hash function perception. In literature [8], a perceptual hash algorithm based on RT and DCT is proposed, which has good robustness and high computational efficiency, but is poor in distinction. The literature [9] proposes a multi-format audio perceptual hashing algorithm based on dual-tree complex wavelet transform, which
has higher efficiency and achieves audio authentication in five different audio formats, including original domain and compressed domain, and achieves small-scale Tamper detection and positioning, but without considering the security issues, weak robustness.

The perceptual hashing must satisfy discrimination and robustness. Discrimination is an indicator which should reflect the data content in a unique way. The speech have different content should yield different hash sequences. Robustness is used to evaluate the identity of the speech after the content maintain operation, such as low-pass filtering, volume improvement, volume reduction, resample. Robustness is used to evaluate whether the speech after the operation is the same as the original speech.

2. Methodology
2.1. Tensor decomposition
Tensor in the field of information processing has been widely used. Tensor can be considered as a product of vector space, and it is a higher order generalization of vector and matrix. The order of the tensor can be expressed as $X \in \mathbb{R}^{R \times N_1 \times N_2 \times \cdots \times N_d}$. The tensor decomposition includes two kinds of decomposition methods: CANDECOMP/PARAFAC and Tucker\cite{10}. CANDECOMP/PARAFAC decomposition and Tucker decomposition are the high order generalization of the singular value decomposition in the matrix decomposition\cite{11}.

2.2. Wavelet packet transform
Wavelet Packet Transform (WPT) as the further expansion of wavelet analysis theory. Wavelet packet decomposition can reflect the feature and nature of the signal, very suitable for the analysis and processing of speech signal types of non-stationary signal\cite{12}.

2.3. LPCC Feature Coefficient Extraction
The recurrence relation between LPC coefficient and LPCC coefficient is below:

$$
\begin{align*}
  c_0 &= a_i \\
  c_n &= a_n + \sum_{k=1}^{n-1} c_n a_{n-k}, \quad 1 \leq n \leq N \\
  c_n &= \sum_{k=1}^{n-1} c_n a_{n-k}, \quad n > N
\end{align*}
$$

$c_0$ is the DC component, $a_n$ is the LPC coefficient, $c_n$ is the LPCC coefficient.

2.4. LSP feature parameter extraction
Linear spectral pair (LSP) is mostly used as the feature vector. LPS can be obtained by solving the conjugate complex root of $P + 1$ order symmetric and antisymmetric polynomials.

$$
\begin{align*}
P(z) &= A(z) + z^{-(P+1)}A(z^{-1}) \\
Q(z) &= A(z) - z^{-(P+1)}A(z^{-1})
\end{align*}
$$

The roots of the above two polynomials are on the unit circle and appear alternately. Where $P(z)$ has a root of $z = -1$, $Q(z)$ has a root of $z = 1$. $P(z)$ and $Q(z)$ have $M/2$ conjugate complex roots located on the unit circle. So get a new polynomial.

$$
P(z) = (1 + z^{-1}) \prod_{i=1}^{p/2} (1 - 2 \cos \omega_i z^{-1} + z^{-2})
$$

$$
Q(z) = (1 - z^{-1}) \prod_{i=1}^{p/2} (1 - 2 \cos \omega_i z^{-1} + z^{-2})
$$

Where $\cos \omega_i, i = 1, 2, \cdots, p/2$ is the representation of the LSP feature in the cosine domain.

2.5. ISP feature parameter extraction
Immittance Spectral Pairs (ISP) is an important equivalent representation of linear prediction feature,
which can effectively reduce the computational complexity and ensure the calculation precision at the same time. According to the transfer function of linear prediction filter $A(z) = \sum_{i=0}^{p} a_i z^{-i}$, $a_0 = 1$.

$$F(z) = A_p(z) + A_p^\ast(z) = A_p(z) + z^{-p} A_p(z^{-1})$$

$$G(z) = A_p(z) - A_p^\ast(z) = A_p(z) - z^{-p} A_p(z^{-1})$$

3. Speech perception hash authentication scheme

The flow chart of speech perception hash authentication is shown in figure 1.

Step 1: preprocessing, framing and windowing: conduct pre-emphasis on the speech in the speech library to be tested, enhance the useful frequency spectrum of high frequency, reduce the edge effect and eliminate noise. in order to eliminate the inter frame loss during framing, conduct framing and add the Hamming window for speech $x(t)$; during framing, the frame length is $L$; when the frame moves at $L/2$, $s(n)$ can be obtained; later, add Hamming window for $s(n)$ to obtain $s_w(n)$. $n$ is the frame number.

Step 2: construction of speech feature tensor and decomposition of speech tensor: conduct wavelet package decomposition for the speech frame, calculate the MFCC and $\Delta$MFCC features of each frame, conduct the tensor construction of feature feature to obtain the speech feature tensor $\chi$. Conduct Tucker decomposition for the feature tensor $\chi$ to obtain the low-dimensional core tensor $G$.

Step 3: quantization: Reconstruct the core tensor $G$ to form the two-dimensional feature matrix, calculate the sum of each column of matrixes.

$$R_h(j) = \sum_{i=1}^{r} H^{(n)}_{ij} \quad 1 \leq j \leq k$$

In the formula above, $H^{(n)}_{ij}$ signifies the feature feature in row $j$ and line $i$; $k$ is the number of rows of feature matrix. Quantize the feature formed and row matrix to form the Hash value $h(j)$ of speech segment, $\hat{R}_h$ is the mid-value.

$$h(j) = \begin{cases} 1 & R_h(j) > \hat{R}_h, 1 \leq j \leq k \\ 0 & \text{other} \end{cases}$$

Step 4: calculation and matching of perception Hash distance: suppose that there are two speech segments $\alpha$ and $\beta$; define the Hash mathematic distance is $D_h(\alpha, \beta)$, which is shown below:

$$D_h(H_{\alpha}, H_{\beta}) = \sum_{j=1}^{n} |h_\alpha(j) - h_\beta(j)|, j = 1, 2, \ldots, n;$$

if the perception contents of the two speech segments $\alpha$ and $\beta$ are the same,

$$D_h(H_{\alpha}, H_{\beta}) \leq \tau$$

if the perception contents of the two speech segments $\alpha$ and $\beta$ are different,

$$D_h(H_{\alpha}, H_{\beta}) > \tau$$

In the formulas above, $\tau$ is the matching threshold.
4. Experimental analysis

4.1. The experimental environment

As for the experiment in the paper, the MATLAB 2010b simulation realization is adopted. The speech of the length 4 seconds 1189 segments, in which different content of speech contains Chinese and English and the same content of different people read speech. Speech parameters used for the sampling rate 16000Hz, bit rate is 256kbps, the number of channels is mono, sampling precision is 16bit, and format is wav. Frame length is 20ms, frame shift is 10ms.

4.2. Discrimination

The perception hash gain 816003 bit error rate data. The obtained bit error rate of the normal distribution as shown in figure 2.

The bit error rate of perception hash value of different speech content mainly obey normal distribution, where probability distribution parameter is $\mu (\mu=0.4971)$, and Standard deviation $\sigma=0.0274$.

As can be seen from table 1, the algorithm deterrent rate in this paper has no more obvious improvement than the deterrent rate based on LPC speech perception hash algorithm.

4.3. Robustness

The average bit error rate can be obtained due to the attack to speech library.

As shown in figure 2, FRR and FAR curve has obtain BER, extracted from the content of the same speech perception of the hash value and BER are below the threshold value of 0.35, the experimental results show that this algorithm has higher robustness. And FRR - FAR curve in the figure is not crossed, this algorithm also has good distinguishability and robustness. It can accurately identify their content and content of malicious operation. According to table 2 shows that when the threshold $\tau= 0.35$, FAR=$3.8185e^{-008}$.

As can be seen from the figure 3, discriminant threshold of algorithm in this paper has arrange between 0.33 and 0.41. As can be seen from table 2, In addition to 20dB Gaussian white noise, the average bit error rate of the above several content keeping operation are below decision threshold 0.35, increasing and decrease the volume does not change the channel model. Therefore, the coefficients greater change will not happen, so adjust the volume does not produce the error rate. Due to the algorithm using parameters normalization which can improve the resistance to Gaussian white noise, the algorithm robustness in the paper against noise is great.
Figure 2. The algorithm of normal distribution graphics

Table.1 The threshold false accepted rate of the algorithm proposed algorithm

| τ | False Accepted Rate |
|---|---------------------|
| 0.25 | 8.5838e-020 |
| 0.30 | 2.9505e-013 |
| 0.35 | 3.8185e-008 |

Figure 3. The algorithm FRR - FAR curve

Table.2 The matching rate of speech authentication after the being kept operating content %

| Resampling | The proposed algorithm | Average bit rate |
|------------|------------------------|-----------------|
| 50% volume | 100                    | 0.0497          |
| +50% volume | 100                    | 0.1211          |
| Resampling 8kHz-16kHz | 100 | 0.1295 |
| Resampling 32kHz-16kHz | 100 | 0.0518  |
| 20dB Gaussian white noise | 72.3 | 0.3215 |
| 30dB Gaussian white noise | 87.1 | 0.2942 |
| 4kHz low-pass filtering | 99.7 | 0.1978 |

5. Conclusion

In order to solve the balance of authentication efficiency, the content remain operational robustness and the discriminative an efficient and robust speech authentication algorithm based on perceptive hash of tensor decomposition model is proposed. Through theoretical analysis and experimental analysis, the proposed algorithm has good anti-collision performance and robustness which can effectively resist the content preservation attacks and common background noise. The algorithm can authenticate speech information and speaker information more accurately.
We only consider the matching of speech in the speech library when the algorithm is built. Personal speech habits are different, and feature extraction is complex. Post-research needs to consider the extraction the unique characteristics of personal speech, the speech authentication which is not related to speaking time, the establishment of hash value database and so on. Therefore, the speech perception hash algorithm, which is unrelated to the speaker is the direction to be studied.

REFERENCES

[1] Z Q Liu, Q Li, J R Liu and X M Niu, 2012, ICDIP.Proc. of spie, A Novel Evaluation Method For Perceptual Hash Algorithm In Authentication Mode, Vol.8334 83444-1.
[2] M Ramona and G Peeters, 2011, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection, 477-480.
[3] J Z Wang, T L Xu, K G Xuan, 2011, Computer Applications and Software, Audio Digital Watermarking Algorithm Based on Energy Ratio of Wavelet Coefficient, 28(5): 288-291.
[4] J J Deng, W G Wan, S Ram, X Q Yu and X Q Pan, 2011, Smart and Sustainable City (ICSSC 2011), IET International Conference on, An audio fingerprinting system based on spectral energy structure, 6-8 July:1-4.
[5] Y B Huang, Q Y Zhang, Z T Yuan, Z P Yang, 2015, Huazhong University of Science and Technology (Natural Science Edition), The Robust Hash Algorithm of Speech Perception Based on the Integration of Adaptive MFCC and LPCC, 43(2), 124-128.
[6] J F Li, W Tao, H X Wang, 2015, Journal of Beijing University of Posts and Telecommunications, Perceptual Hashing Based on Correlation Coefficient of MFCC for Speech Authentication, (2) 89-93.
[7] H Zhao, H Liu, K Zhao and Y Yang, 2011, International Journal of Digital Content Technology and its Applications, Robust Speech Feature Extraction Using the Hilbert Transform Spectrum Estimation Method, 5(12):85-95.
[8] J F Li, T Wu, 2015, Proc of 2015 11th international Conference on computational intelligence and security, Shenzhen; IEEE, Perceptual audio hashing using RT and DCT in wavelet domain, 363-366.
[9] Q Y Zhang, P F Xing, Y B Huang, R H Dong, Z P Yang, 2016, Journal of Beijing University of Posts and Telecommunications, Perceptual Hashing Algorithm for Multi-Format Audio, 39 (4) 77-82.
[10] B W Bader and Kolda T G, 2006, Scient. Comput., Efficient MATLAB computations with sparse and factored tensors. SIAM J. 30: 205-231.
[11] J Li, L X Jin and G N Li, 2013. J. Electron. Inform. Technol., Hyper-spectral remote sensing image compression based on nonnegative tensor factorizations in discrete wavelet domain, 35: 489-493.
[12] Y Yang and S Nagarajaiah, 2014, Signal Process., Blind identification of damage in time-varying systems using independent component analysis with wavelet transform, 47: 3-20.

Acknowledgment

The authors would like to thank for the support by the National Nature Science Foundation of China (No. 61363078), the Natural Science Foundation of Gansu Province of China (No. 1606RJYA274).