Robust Speech Hash Function

Ning Chen and Wanggen Wan

In this letter, we present a new speech hash function based on the non-negative matrix factorization (NMF) of linear prediction coefficients (LPCs). First, linear prediction analysis is applied to the speech to obtain its LPCs, which represent the frequency shaping attributes of the vocal tract. Then, the NMF is performed on the LPCs to capture the speech's local feature, which is then used for hash vector generation. Experimental results demonstrate the effectiveness of the proposed hash function in terms of discrimination and robustness against various types of content preserving signal processing manipulations.

Keywords: Speech hash function, linear prediction coefficients (LPCs), non-negative matrix factorization (NMF).

I. Introduction

The speech hash function (or fingerprinting), which maps speech to a short binary string based on the speech's perceptual properties, is proposed as a new solution for automated speech indexing and speech content authentication. Unlike the traditional cryptographic hash function, which is extremely sensitive to the input data, the speech hash function allows for some modification of the speech while distinguishing one speech clip from another. In general, the speech hash function needs to have two properties: discrimination, which means that perceptually distinct speech clips must have different hash vectors, and perceptual robustness, which means that perceptually identical speech clips must have the same hash vector.

Due to wide application in automated audio indexing and audio content authentication, the perceptual audio hash function has been widely studied recently [1]-[4]. However, there are few speech hash functions available. In [5], a compressed domain speech hash scheme integrated with a mixed excitation linear prediction codec is proposed. It utilizes partial bits of the speech bit stream, the linear spectral frequencies, for hash vector generation. In this letter, a new speech hash function for uncompressed speech signal based on non-negative matrix factorization (NMF) [6] of linear prediction coefficients (LPCs) is proposed.

II. Proposed Speech Hash Function

1. Hash Vector Generation

In the proposed speech hash function, the hash vector can be generated by following five steps:

Step 1. The original speech, denoted by $\boldsymbol{a}$, is segmented into M equal and non-overlapping frames, denoted by $\boldsymbol{f}_i$, $i=1,\cdots,M$.

Step 2. Linear prediction analysis is performed on each frame $\boldsymbol{f}_i$ to get its N-th order LPCs, denoted by $\boldsymbol{c}_i=\left\{c_i(n)\middle|n=1,\cdots,N\right\}$.

Step 3. Make $\hat{\boldsymbol{c}}_i=\left\{\left|c_i(n)\right|\middle|n=1,\cdots,N\right\}$ and generate M×N matrix $\boldsymbol{C}$:

$$\boldsymbol{C}=\begin{pmatrix}\hat{\boldsymbol{c}}_1\\ \hat{\boldsymbol{c}}_2\\ \vdots\\ \hat{\boldsymbol{c}}_M\end{pmatrix}. \tag{1}$$

Step 4. Apply r rank NMF to $\boldsymbol{C}$:

$$\boldsymbol{C}\approx\boldsymbol{W}\times\boldsymbol{H}, \tag{2}$$

where $\boldsymbol{W}$ is $M\times r$ and $\boldsymbol{H}$ is $r\times N$. Generate the vector $\hat{\boldsymbol{h}}=\left\{\hat{h}(n)\middle|n=1,\cdots,r\cdot N\right\}$ by concatenating the rows of $\boldsymbol{H}$.

Step 5. Generate the hash vector, denoted as

Manuscript received July 27, 2009; accepted Dec. 18, 2009.

This work was supported by the National Natural Science Foundation of China (60872115, 60903186), the Shanghai's Key Discipline Development Program (J50104), the China Post-doctoral Research Program (20090450570), and the Shanghai Post-doctoral Research Program (09R21412800).

Ning Chen (phone: +86 136 11758166, email: chenning@shu.edu.cn) and Wanggen Wan (email: wanwg@staff.shu.edu.cn) are with the School of Communication and Information Engineering, Shanghai University, Shanghai, China.

doi:10.4218/etrij.10.0209.0309

$\boldsymbol{h} = \left\{ h(n) \middle| n = 1, \cdots, r \cdot N \right\}$, as

$$h(n) = \begin{cases} 0, & \hat{h}(n) \geq \tilde{h} \\ 1, & \hat{h}(n) < \tilde{h} \end{cases}, \tag{3}$$

where $\tilde{h}$ is the median value of $\hat{h}(n),\ n = 1, \cdots, r \cdot N$.

2. Hash Matching

The problem of hash matching can be formulated as the hypothesis testing using the hash function $H(\cdot)$ and the distance measure $D(\cdot,\cdot)$.

L0: Two speech clips $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are from the same speech if

$$D\left(H\left(\boldsymbol{a}_1\right), H\left(\boldsymbol{a}_2\right)\right) < \tau. \tag{4}$$

L1: Two speech clips $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are from different speech if

$$D\left(H\left(\boldsymbol{a}_1\right), H\left(\boldsymbol{a}_2\right)\right) \geq \tau, \tag{5}$$

where τ is a predetermined threshold, which can be obtained for a given false accept rate (FAR). FAR, denoted by R_{FA}, is the probability that L0 is accepted when L1 is true.

In the proposed scheme, the square of the Euclidean distance (see (6)) is utilized to measure the distance between any two hash vectors $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$.

$$x = D\left(\boldsymbol{h}_1, \boldsymbol{h}_2\right) = \frac{1}{L_h} \sum_{n=1}^{L_h} \left[h_1(n) - h_2(n)\right]^2, \tag{6}$$

where L_h is the length of the hash vector. By the central limit theorem, the above distance measure has a normal distribution if L_h is sufficiently large and the contributions in the sums are sufficiently independent. Assuming that the distance measure can be approximated as the normal distribution $N(\mu, \sigma)$, the FAR is given as

$$R_{\text{FA}} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\tau} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] dx. \tag{7}$$

Then, for a given R_{FA}, the threshold τ can be determined by (7), theoretically.

III. Experimental Results

To verify the discriminative and robust nature of the proposed hash function, it was applied to 1,000 speech clips (16 bits signed, 8 kHz, 6 seconds long) with various contents. The setting of the feature parameters were M=360, N=12, and r=1.

1. Discrimination

For each speech clip, the distance of its hash vector and the hash vector of each of the remaining 999 speech clips was calculated by (6), and 499,500 distance values were obtained. A comparison between the probability density distribution of these distance values and the normal distribution shown in Fig. 1 indicates that the distance value has an approximately normal distribution. The expected value and standard deviation are $\mu = 0.4997$ and $\sigma = 0.0412$, respectively.

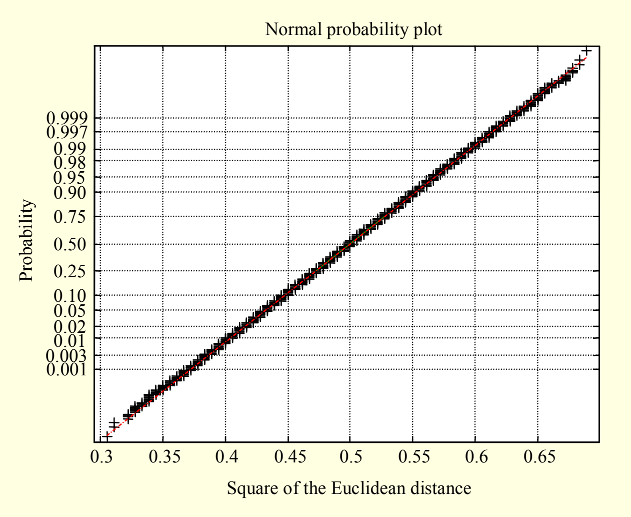

Fig. 1. Comparison of probability density distribution of the distance values plotted as '+' and normal distribution.

Table 1. FAR varying with threshold.

τ	R_{FA}
0.10	1.4795×10^{-22}
0.15	1.0479×10^{-17}
0.20	1.7347×10^{-13}
0.25	6.7540×10^{-10}
0.30	6.2548×10^{-7}

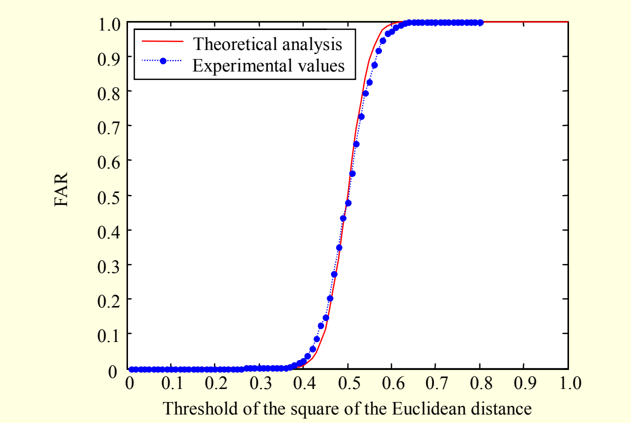

Fig. 2. Comparison of FAR curves obtained by theoretical analysis and experimental values.

The FAR (calculated by (7)) varying with the threshold is shown in Table 1.

Figure 2 shows a comparison of the FAR curve obtained by theoretical analysis (see (7)) with that obtained by the experimental values. The probability density distribution of the distance values follows the normal approximation fairly well; thus, it is verified that the threshold obtained by (7) can be used in practice with reasonable accuracy.

2. Perceptual Robustness

On each of the 1,000 speech clips, 8 kinds of content preserving manipulations (amplitude boost/cut: 3 dB; re-quantize: 32 bits/sample; re-sample: 16 kHz; low-pass filtering: 4 kHz; normalize: 90%; echo addition: 100 ms, 10%; and invert) were performed one by one to generate 8,000 processed clips. Given a reference speech clip, the other 999 speech clips and the 8,000 processed clips were classified as either intra or inter processed speech clips depending on whether they had been derived from a reference speech clip. Then, the distance between the hash vector of the reference speech clip and that of each of the remaining 8,999 speech clips was computed. The intra (inter) matching denotes the distance computation with an intra (inter) speech clip. The worst intra distance and the best inter distance for each speech clip are shown in Fig. 3. If the threshold τ is set in the range [0.10, 0.25], the proposed hash function can reliably decide whether two speech clips have similar content.

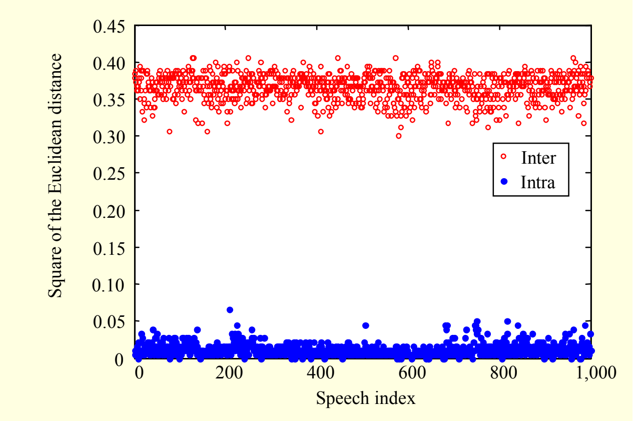

Fig. 3. Worst intra matching and best inter matching for each speech clip.

IV. Conclusion

For a reliable hash function, the feature extracted should be both discriminative and robust. In this letter, linear prediction analysis and non-negative matrix factorization were investigated for speech hash function. Linear prediction analysis was performed to extract the frequency shaping attributes of the vocal tract to realize the perceptual robustness of the proposed scheme. The non-negative constraints of NMF were utilized to capture the local feature of the obtained LPCs to classify speech clips with distinct content. Experimental results demonstrated the effectiveness of the proposed hash function in terms of discrimination and perceptual robustness.

References

[1] P. Cano et al., “A Review of Audio Fingerprinting,” *J. VLSI Signal Process.*, vol. 41, no. 3, 2005, pp. 271-284.

[2] A. Ramalingam and S. Krishnan, “Gaussian Mixture Modeling of Shorttime Fourier Transform Features for Audio Fingerprinting,” *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 4, 2006, pp. 457-463.

[3] M. Park, H. Kim, and S.H. Yang, “Frequency-Temporal Filtering for a Robust Audio Fingerprinting Scheme in Real-Noise Environments,” *ETRI J.*, vol. 28, no. 4, 2006, pp. 509-512.

[4] Y. Jiao et al., “Key-Dependent Compressed Domain Audio Hashing,” *Proc. ISDA*, 2008.

[5] Y. Jiao, Q. Li, and X. Niu, “Compressed Domain Perceptual Hashing for MELP Coded Speech,” *Proc. IIHMSP*, 2008, pp. 410-413.

[6] D.D. Lee and H.S. Seung, “Learning the Parts of Objects by Non-negative Matrix Factorization,” *Nature*, vol. 401, no. 6755, 1999, pp. 788-791.

ETRI Journal, Volume 32, Number 2, April 2010 Ning Chen et al. 347