http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (16): 3591-3595, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.3591.3595
© 2013 Asian Network for Scientific Information

Algorithm for Evaluating Speech Perceptual Hash Similarity after Slight Tampering Occurs

[1]Huang Yi-bo, [2]Zhang Qiu-yu and [1]Yuan Zhan-ting
[1]College of Electrical and Information Engineering,
[2]School of Computer and Communication, Lanzhou University of Technology,
Lanzhou, Gansu, 730050, China

Abstract: While evaluating the performance of the speech perceptual hash algorithms, we need to test their robustness, safety and real-time properties, as well as judging the perceived similarity of the detected speeches. But the existing algorithms are so insensitive to slight speech tampering that the tampered speeches are mistakenly considered to be semantically unchanged. Therefore, we present an algorithm for measuring the perceived similarity. By displaying desirable sensitivity to slight speech tampering, the proposed algorithm can detect slight speech tampering and judge whether the meanings have changed. The proposed algorithm first divides the speech signals to many segments and then performs correlation coefficient test on each segment in order to compute the similarity. The experiment results show that the proposed algorithm can effectively detect quality changes of the speech signals and the similarity of the slightly tampered speeches. Its performance in evaluating perceived similarity is superior to the popular similarity evaluation algorithms.

Key words: Speech hashing, slight tampering occurs, similarity, evaluating

INTRODUCTION

The perceptual speech hashing conducts unidirectional mapping based on speech features to obtain short digit sequences and it can be used in speech labeling, speech content authentication, speech query, speaker verification, etc. As it processes extracted speech features, the perceptual speech hashing requires only a small amount of data for authentication, incurs small-scale redundancy and achieves great verification results (Niu and Jiao, 2008; Gu *et al*., 2012). Compared with traditional message encryption and message authentication code, the perceptual speech hashing is fit for modern applications that require huge amounts of speech data and impose high authentication demands (Grutzek *et al*., 2012). In real-life applications, normal operations (Gaussian white noises increased or decreased volume of the sound, delay, etc.) are usually performed to keep the content of speech signals unchanged. The perceptual speech hashing is expected to be highly robust, because no substantial modifications are made to the speech contents. The typical perceptual speech hash algorithms should satisfy the following requirements: (1) robustness: For two segments of perceptually similar speeches, the perceived hash value should be within the matching thresholds;(2) Sensitivity: When the contents of the speech signals have changed, the perceived hash value should change correspondingly; (3) safety: The possibility that the derived perceived hash value is faked approaches zero (Gupta *et al*., 2012).

Currently speech perception hash algorithm has rapid development (Zmudzinski *et al*., 2012), speech perception feature extraction(Cano *et al*., 2005) (Jin and Yoo, 2007) and constitutes (Jiao *et al*., 2010; Chen and Wan, 2009; Chen and Wan, 2010) algorithm has a variety of research findings, For speech perception hash compression format has been research (Jiao *et al*., 2008) .But for two differences in speech hash perception evaluation method not many.

The schemes for evaluating the difference in speech contents are either subjective or objective. The subjective approaches involve the participation of the humans and the evaluation results vary from person to person. In the objective methods it is the algorithms rather than the humans that evaluate the speech differences. However, as robustness is a concern for existing algorithms, most of them are insensitive to slight tampering. Therefore, we need to devise a scheme to precisely determine whether there is any slight tampering of speech signals.

The algorithm for measuring similarity based on the perceptual speech hashing after slight tempering occurs can be used in: (1) Speech communication authentication.

Corresponding Author: Huang Yi-bo, College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, Gansu, 730050, China

The speech signal is likely to be maliciously tampered during the transmission for special purposes. Messages that the tampered speeches convey to the receivers might be totally different from the original meanings. (2) Protection of the speech records. Keep the perceived hash values of the original speeches to ensure that the speech records have not been tampered.

Existing methods for measuring speech similarity: (1) Bit Error Rate (BER), (2) Tampering Evaluation Function (TAF), (3) Correlation Coefficient (NC).

The above-mentioned evaluation metrics consider the difference between speeches as a whole and cannot measure the deviation in similarity after the speech is slightly tampered. Therefore, in addition to the scheme to robustly evaluate the operations that keep contents unchanged, we need a special algorithm that is sensitive to slight tampering.

The proposed algorithm provides a method for evaluating perceived similarity based on block calculation. By estimating regional difference after slight tampering of speech signals it can desirably determine whether the speech has been tampered. This scheme can also represent the influence of the operations that keep contents unchanged (e.g. Gaussian while noise, increased or decreased volume of the sound, echo, etc) on speech signals.

SEGMENT WISE EVALUATION OF SPEECH SIMILARITY

For a given speech signal, different tampering will lead to different speech segments. Let A denote the speech signal, h_a denote the perceived speech hashing algorithm, H_a denote the set of perceived speech hash values. $H_z = A \rightarrow H_a$. Let dis (a, b) and τ represent the distance between the segments of A and B and decision threshold, resp. If the perceived hashing distance between the two segments is greater than the threshold, we conclude that the two segments are perceptually different. The probability that an event occurs is denoted by.$P^{(\bullet)}$ Suppose a, b, c, ϵ, M, h_a, h_b, $h_c \epsilon H_p$, $h_a = h(a)$, $h_b = h(b)$, $h_b = h(b)$.

The ideal algorithm for evaluating the perceived speech similarity should have the following properties:

- **Perceptual robustness:** After the operations that leave contents unchanged, speeches that have the same perceived contents should have the same perceived hash values. Suppose the operations are denoted by $O_{cp}(\bullet)$, $a' = O_{cp}(a)$, $h_a = h_p(x')$, then we have dis $(h_k, h_x) < \tau$
- **Tampering sensitivity:** The speeches that have the same perceived contents originally should have different perceived hash values after the tampering. Suppose tampering is denoted by $R_{cp}(\bullet)$, $a' = R_{cp}(a)$, $h_a = h_p(x')$, then we have dis $(h_k, h_x) < \tau$
- **Symmetry:** If the input positions of A and B are switched, the output similarity remains unchanged, i.e., dis $(h_x, h_x) =$ dis (h_x, h_x)

Because the tampered contents are typically concentrated within certain ranges, the metrics designed for evaluating perceived speech similarity after slight tampering must be able to distinguish differences of regional contents. To achieve this, we have to amplify the regional differences of the speeches.

Divide the speech I_1 into n segments and the ith segment is denoted by, $B_i^{(1)}$, $I_1 = B_i^{(1)} \cup B_2^{(1)} \cup ... \cup B_n^{(1)}$ The speech I_2 is segmented similarly. Assume I_2 is obtained by slightly tampering I_1 and both speeches are divided, then if a segment is seriously tampered, the difference between the corresponding segments should be enormous.

As a two-variable relevance analysis scheme, the Pearson correlation coefficient detection method is adopted to measure the similarity between blocks of $B_i^{(1)}$ and $B_i^{(2)}$:

$$r = \frac{\sum (X - \overline{X})(X - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} \quad (1)$$

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N}\left(B_i^{(1)}(n) - \mu B_i^{(1)}\right)\left(B_i^{(2)}(n) - \mu B_i^{(2)}\right)}{\sigma B_i^{(1)} \sigma B_i^{(2)}} \quad (2)$$

where $\mu_{BI}^{(1)}$ and $\mu_{BI}^{(2)}$ denote the means of $B_i^{(1)}$ and $B_i^{(2)}$, resp. The numerator and denominator is the covariance and standard deviation of each segment, resp. Set r = 1 in the case of blank speech segments where both of the standard deviations are 0; set r = 0 in the case of evaluating two speech segments where one standard deviation is 0 and the other is not. Normalize r into:

$$\rho = \frac{r_i - \min(r_i)}{\max(r_i) - \min(r_i)} \quad (3)$$

where max (r_i) denotes the maximum of ρ_i while min (r_i) is the minimum.

While representing the correlation between two variables, the correlation coefficient r is bounded by [-1, 1], where r>0 means positive correlation, r<0 means negative correlation and r = 0 means no correlation. The value of |r| closer to 1 indicates the two speeches are more correlated.

The perceived speech similarity can be defined as:

$$\rho_{small} < \hat{\rho} < \rho_{big} \tag{4}$$

$$S(l_1, l_2) \frac{\prod \rho_{small}}{\prod \rho_{big}} \tag{5}$$

With decreased difference between speech segments, both the similarity and correlation coefficient for the two segments increase. The equation () shows the relation between the most dissimilar m segments and most similar m segments of the to-be-evaluated image. Because the Pearson correlation coefficient is independent of the speech inputs, the symmetry conditions are satisfied. A high similarity between to-be-evaluated speeches leads to small difference between the correlation coefficients of ρ. If the speeches I_1 and I_2 are absolutely identical, we have S = 1.

EXPERIMENTS

The speeches from the TIMIT database and recording studio are used in the experiment to verify the proposed algorithm. We have 1,189 segments in total, each of which lasts 4 seconds, including the Chinese and English speeches with different contents as well as the speeches that share the same contents but are read by different speakers. The used speech parameters: 16000 Hz of the sampling rate, 256 kbps of the bit rate, mono, 16 bit of the sampling accuracy and the wav of the format. Set the frame size as 20 ms and frame shift as 10 m sec for frame segmentation. Forms of attacks include adding Gaussian white noise, increasing or decreasing the volume of the sound by 50%, re-sampling, delaying the echo by 300ms and slight tampering of 5%. The speeches are tampered by 5 and 10% to verify the algorithms' sensitiveness.

To check the robustness of the algorithms in evaluating perceived speech similarity under slight tampering, we plot the normal distribution probabilities of BER, NC, TAR and our algorithm under 5 and 10% tampering. Fig. 1 shows distribution maps of normal probabilities. From these maps it can be seen that The perceived speech hash similarity computed under the slight tampering overlaps with the similarity obtained after the operations that keep contents unchanged and misidentification results. The proposed algorithm outshines others in detecting slight tampering. The perceived speech hash similarity is distributed in [0.15, 0.30] for 5% tampering and in [0.10, 0.22] for 10% tampering. It is distinguishable from the similarities

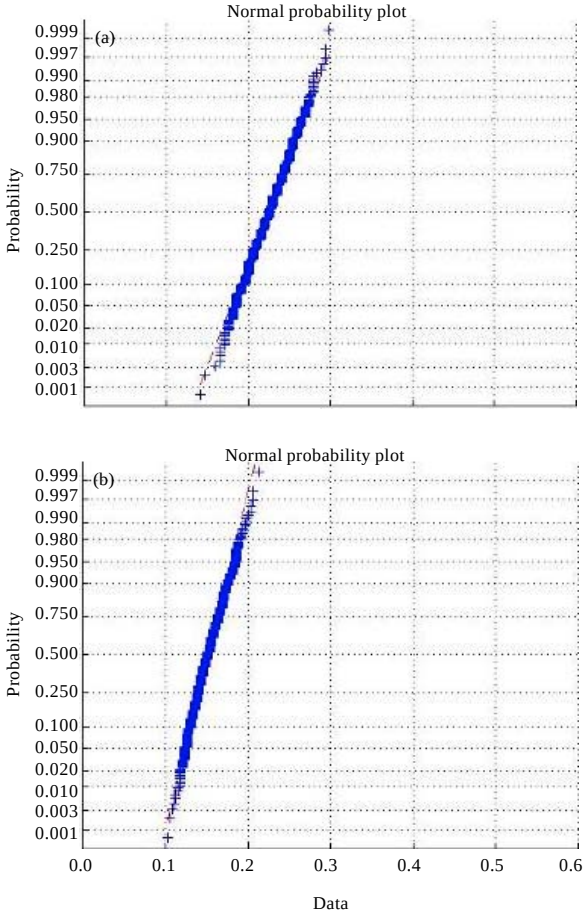

Fig. 1: Algorithm Evaluation similarity of replaced 5% and 10% normal map

obtained by the operations that keep contents unchanged and thus achieves enormously reduced misidentification probabilities.

As shown in Fig. 2, the FAR-FRR curve of our algorithm has no overlapping and is highly distinct. Our algorithm also improves the evaluation range of existing algorithms. The evaluation threshold is [0.29, 0.38]. Compared with the algorithm in Fig. 2 its threshold range is quite obvious and can better detect speech contents. Randomly select 100 segments to check the similarity distribution of the speeches that are processed by operations that keep contents unchanged, speeches with different contents and slightly tampered speeches, as shown in Fig. 3. It can be seen from the Figure that the speech that are processed by operations that keep contents unchanged is above the decision threshold and both regions have evident decision thresholds. The curves of the speeches with different contents and slightly tampered speeches almost overlap with each

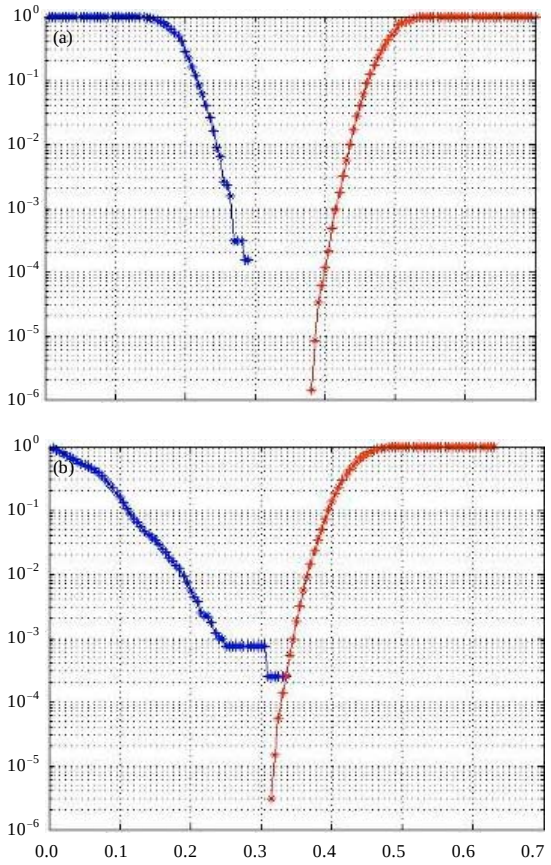

Fig. 2: Algorithm and CHEN's algorithm FAR-FRR curve

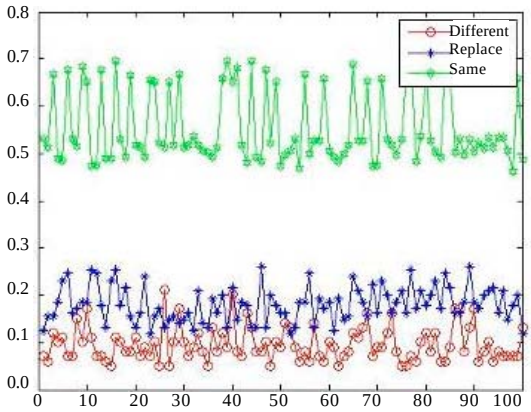

Fig. 3: Distribution interval graphs of same speech, different speech and tampering speech

other, meaning that our similarity evaluation algorithm is so powerful in distinguishing speeches that it is able to map the slightly tampered speeches into regions totally irrelevant to the original speeches.

CONCLUSIONS

We propose a scheme to objectively evaluate the perceived speech similarity under slight tampering. By computing the segment wise similarity of the perceived speech hash sequence, our algorithm can sensitively detect the changes of the sequence after it is segmented and ensure the sensitivity to slight tampering. The experiment results show that our algorithm can detect significant changes that can possibly occur to speeches, even if it is slightly tampered. As it has yet to consider the perceived similarity of the speeches that have passed the 1KHZ low-pass filter or attacked by long delay echoes, our algorithm is not highly robust under these two forms of attacks. However, these two kinds of attacks are popular in speech signal transmission. So in future research, we will investigate the evaluation of the perceived speech similarity in these two cases.

ACKNOWLEDGMENT

The authors world 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. 1212RJZA006)

REFERENCES

Cano, P., E. Batlle, E. Gomez, L.C.T. Gomes and M. Bonnet, 2005. Audio Fingerprinting: Concepts and Applications. In: Computational Intelligence for Modelling and Prediction, Halgamuge, S.K. and L. Wang (Eds.). Springer-Verlag, Berlin, Heidelberg, Germany, pp: 233-245.

Chen, N. and W.G. Wan, 2009. Speech hashing algorithm based on short-time stability. Proceedings of the 19th International Conference on Artificial Neural Networks, September 14-17, 2009, Limassol, Cyprus, pp: 426-434.

Chen, N., W.G. Wan, 2010. Robust speech hash function. ETRI J., 32: 345-347.

Grutzek, G., J. Strobl, B. Mainka, F. Kurth, C. Poerschmann and H. Knospe, 2012. Perceptual hashing for the identification of telephone speech. Proceedings of the Speech Communication, September 26-28, 2012, Germany, pp: 1-4.

Gu, J., L. Guo, H. Liang and L. Cheng, 2012. Effective robust speech authentication algorithm based on perceptual characteristics. J. Chinese Comput. Syst., 4: 1461-1466.

Gupta, S., S. Cho and C.C.J. Kuo, 2012. Current developments and future trends in audio authentication. IEEE Multimedia, 19: 50-59.

Jiao, Y.H., L.P. Ji and X.M. Niu, 2010. Perceptual speech hashing and performance evaluation. Int. J. Innovative Comput. Inform. Control, 6: 1447-1458.

Jiao, Y.H., Q. Li and X.M. Niu, 2008. Compressed domain perceptual hashing for MELP coded speech. Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing, August 15-17, 2008, Harbin, Germany, pp: 410-413.

Jin, M. and C.D. Yoo, 2007. Temporal dynamics for spectral sub-band centroid audio fingerprints. Proceedings of the IEEE International Conference on Multimedia and Expo, July 2-5, 2007, Beijing, China, pp: 180-183.

Niu, X.M. and Y.H. Jiao, 2008. An overview of perceptual hashing. Acta Electron. Sin., 36: 1405-1411.

Zmudzinski, S., B. Munir and M. Steinebach, 2012. Digital audio authentication by robust feature embedding. Proceedings of the SPIE, Media Watermarking, Security and Forensics, February 9, 2012, USA., pp: I1-I7.