HUMMING-BASED HUMAN VERIFICATION AND IDENTIFICATION

Minho Jin, Jaewook Kim and Chang D. Yoo

Div. of EE, Dept. of EECS, Korea Advanced Institute of Science and Technology,
373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Korea

ABSTRACT

This paper considers humming-based systems for human verification and identification. Humming of a target person is modeled as a Gaussian mixture model, and the matching score between a target model and humming is computed as the likelihood of humming given a target model. Verification is performed by comparing the matching score to the likelihood given a universal background model, and identification is performed by selecting the best-matched model. The verification and identification performances are evaluated using various acoustical features. The experimental results show that linear prediction cepstral coefficients and perceptually linear prediction coefficients are conducive to verification and identification, respectively.

Index Terms— Biometrics, Humming, GMM-UBM

1. INTRODUCTION

This paper considers humming as a biometric characteristic. Biometric characteristics can be obtained from deoxyribonucleic acid (DNA), face shape, ear shape, fingerprint, gait pattern, hand-vein pattern, hand-and-finger geometry, iris scan, retinal scan, signature, speech, etc [1]. To the best of the authors' knowledge, this paper is the first work that evaluates the performance of humming as a biometric characteristic in verification and identification systems. Humming conveys little linguistic information, and thus for verification and identification, it is not expected for humming to outperform speech in verification and identification. However, a humming-based biometric system may be applicable to a person with speech disorder and an infant who is not able to speak [2,3]. In terms of universality, which is an essential criterion to be considered in a biometric recognition system [1], humming is more universally available on everyone than speech.

Our experimental results indicate that pitch information contained in humming is not very useful for human verification and identification. Pitch information has shown to be effective for humming-based music retrieval [4, 5] and speaker verification [6]. However, pitch contained in humming is highly dependent on the melody and not on the person who is humming. For this reason, the paper considers Mel-frequency cepstral coefficients (MFCCs), linear prediction coefficients (LPCs), linear prediction cepstra coefficients (LPCEPSTRA) and perceptual linear prediction coefficients (PLPs) which are well-known acoustical features for speech and speaker recognition.

This research was partly supported by the IT R&D program of MKE/IITA [2007-S017-01, Development of user-centric contents protection and distribution technology].

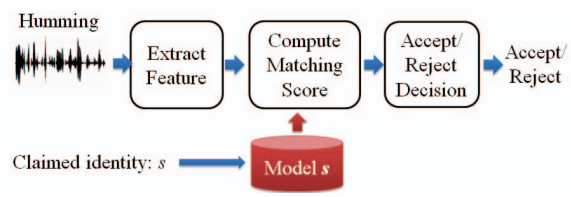

Fig. 1. Human verification with humming

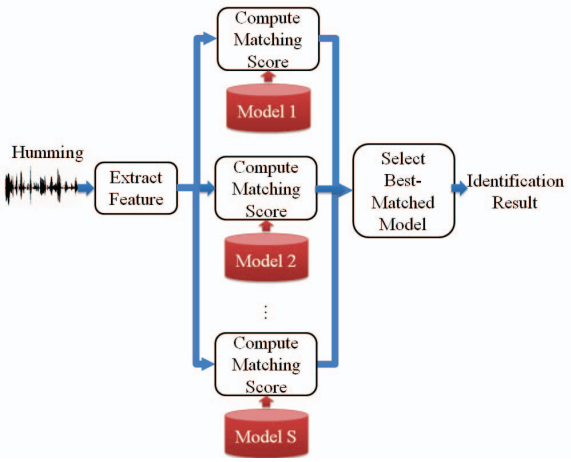

Fig. 2. Human identification with humming

The rest of this paper is organized as follows: Section 2 describes humming-based verification and identification systems. Section 3 demonstrates our experimental results, and Section 4 summarizes this paper.

2. HUMAN VERIFICATION AND IDENTIFICATION WITH HUMMING

In the system considered, voice activity detection is performed to segment query input into humming and non-humming intervals, and acoustical features mentioned above are extracted from the humming interval longer than 0.5s.

2.1. Human Verification with Humming

Fig. 1 illustrates a verification system considered in this paper. Let $\mathbf{x}_1^T = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$ be the T-length acoustical feature sequence. The verification is performed using following hypotheses:

- H_0: $\mathbf{x}_1^T$ is generated from a target person s.
- H_1: $\mathbf{x}_1^T$ is not generated from a target person s.

For the above hypothesis testing, the Gaussian mixture model and universal background model (GMM-UBM) [7] is used. The UBM λ_U is trained on humming from various people not included in the target and impostor list. The model of target person s, λ_s, is adapted from the UBM λ_U using enrollment data of s in the maximum a *posterior* [8] sense. In the system considered, an accept/reject decision is made as follows:

$$\frac{p(\mathbf{x}_1^T|\lambda_s)}{p(\mathbf{x}_1^T|\lambda_U)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau, \tag{1}$$

where τ is a preset threshold.

2.2. Human Identification with Humming

Fig. 2 illustrates an identification system considered in this paper. Given a set of S target person models $\{\lambda_s | s = 0, 1, \ldots, S-1\}$, $\mathbf{x}_1^T$ is identified as s^* as follows:

$$s^* = \underset{s}{\operatorname{argmax}}\, p(\mathbf{x}_1^T|\lambda_s). \tag{2}$$

3. EXPERIMENTS

3.1. Experimental Setup

The verification and identification performances were evaluated on a humming database of 321 min. from 22 male students: humming was recorded at 16kHz sampling rate using a microphone embedded in a laptop. The mean and the standard deviation of the number of hummed song per each person are 10.9 and 9.8, respectively. We selected 8 targets whose enrollment data is longer than 6 min. Unless specified, experimental results in this paper were obtained using the experimental setting in Table. 1. From humming, we extracted 39-dimensional MFCC, LPC, LPCEPSTRA, and PLP which consist of 13 coefficients (12 coefficients + energy), their delta and delta-delta time difference. In addition, we extracted 4-dimensional pitch information which consists of pitch, its delta and delta-delta time difference, and long span pitch [9]. The verification performance is measured in terms of equal-error rate (EER) and detection error trade-off (DET) curves, and the identification performance is measure in terms of identification error rate.

Table 1. Experimental Setup

Targets	8 male students 6 min. enrollment data/target person
Trials	1.11s long on average 2 512 and 29 851 true and impostor trials for verification experiments 1 648 trials from 8 targets for identification experiments
Acoustic model	GMMs of 64 kernels for verification and 1 024 kernels for identification UBM trained on 22 min. of humming data that are separate from true and impostor trials.

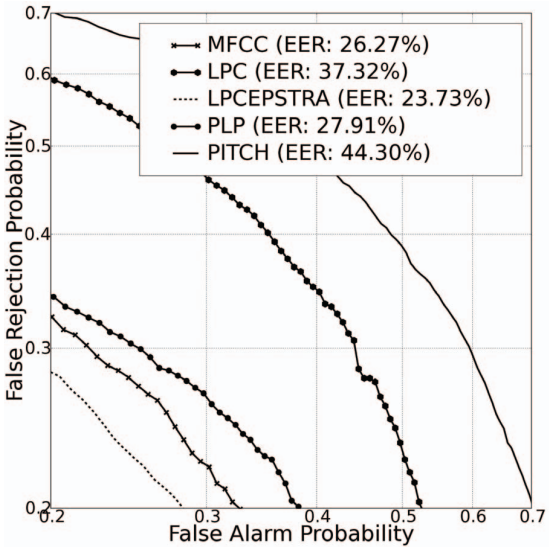

Fig. 3. DET curves of MFCC, LPC, LPCEPSTRA, PLP, and pitch.

3.2. Experimental Results

3.2.1. Verification

Fig. 3 illustrates the DET curves of MFCC, LPC, LPCEPSTRA, PLP, and pitch. The LPCEPSTRA performed the best followed by the PLP, LPCEPSTRA, LPC, and pitch. Pitch, which has an EER of slightly under 50%, performed

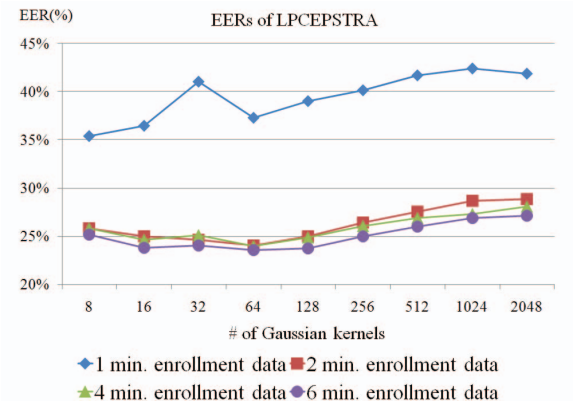

Fig. 4. EERs of LPCEPSTRA with respect to the amount of enrollment data and the number of kernels

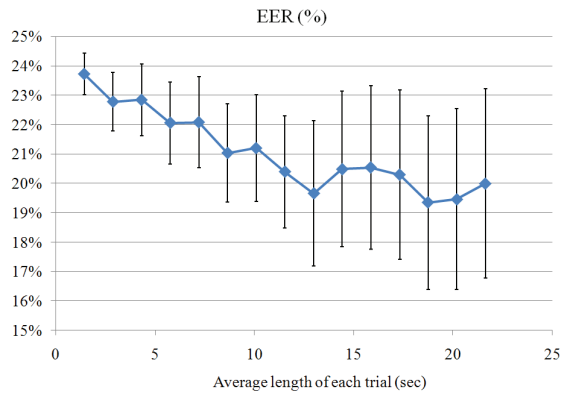

Fig. 5. EERs of LPCEPSTRA with respect to the average length of each trial: the vertical bar denotes the 60% confidence interval of EER.

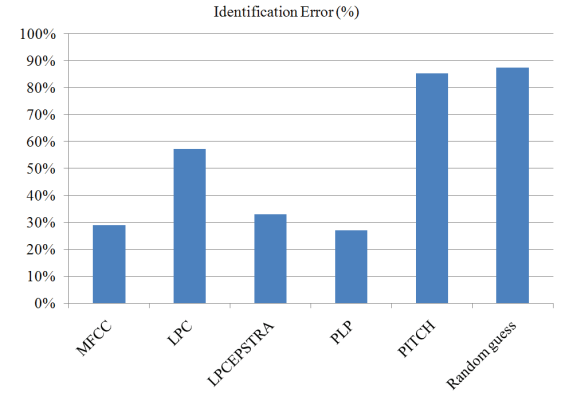

Fig. 6. Identification errors of MFCC, LPC, LPCEPSTRA, PLP, and pitch

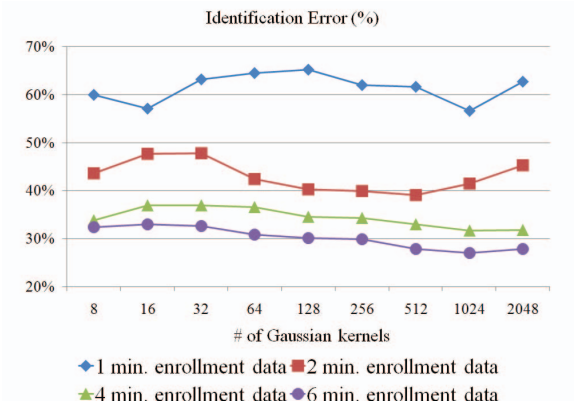

Fig. 7. Identification error of LPCEPSTRA with respect to the amount of enrollment data and the number of kernels

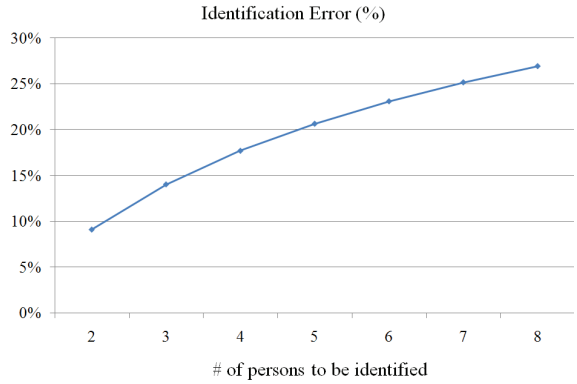

Fig. 8. Identification error of LPCEPSTRA with respect to the number of targets.

Fig. 9. Identification error with respect to the average length of each trial: the vertical bar denotes the 90% confidence interval of the identification error.

the worst. This indicates that pitch information from humming is not conducive to human verification. Fig. 4 illustrates the EERs of LPCEPSTRA with respect to the amount of enrollment data and the number of Gaussian kernels in the

GMM. The vertical bar denotes the 60% confidence interval of EER, where the false alarm and the miss probabilities are considered as a binomial statistic:for details, please refer to [10]. The EER decreases as the amount of enrollment data increases, and the minimum EER of 23.50% was achieved when the amount of enrollment data and the number of Gaussian kernel are 6 min. and 64, respectively. Our conjecture for this is that the amount of data is insufficient to train the UBM with larger number of kernels than 64. Fig. 5 illustrates the EERs of LPCEPSTRA with respect to the average length of each trial. Longer trials are produced by concatenating short trials whose average length is 1.11s. Consequently, the number of trials decreases as the average length increases, and the confidence interval, which is inversely proportional to the square root of the number of trials, increases as the average length increases. From Fig. 5, it can be stated that the EER of 22.13% at 11.54s is better than the EER of 23.01% at 1.45s with 60% confidence. For larger confidence than 60%, it is required to perform experiments with larger database.

3.2.2. Identification

Fig. 6 illustrates the identification errors of MFCC, LPC, LPCEPSTRA, PLP, and pitch. Unlike the verification experiments, the PLP performed the best. The identification error of pitch is similar to that of random guess, which indicates that pitch is not conducive to human identification. Fig. 7 illustrates the identification error with respect to the amount of enrollment data and the number of Gaussian kernels. As in the case of verification, the identification error decreases as the amount of enrollment data increases, and the minimum error of 26.94% was achieved when the number of Gaussian kernels is 1 024. Fig. 8 illustrates the identification error with respect to the number of targets. As the number of targets to be identified increases, the identification error increases. The identification error is less than 20% for 5 targets which is a normal family size. Fig. 9 illustrates the identification error with respect to the average length of each trial. The confidence interval is computed by considering the identification error as a binomial statistic. Longer trials are produced in the same manner as in the verification experiment. The identification error decreases as the average length of each trial increases. With 90% confidence, the error rate of 19.51% at 14.43s is different from that of 26.94% at 1.45s. For higher confidence, further experiments are required.

4. CONCLUSION

This paper considered humming-based human verification and identification systems. The verification and identification performances of the system considered were evaluated with MFCC, LPC, LPCEPSTRA, PLP and pitch. Our experimental results indicate that pitch, which is conducive to humming-based music retrieval and speaker verification, is not conducive to human verification and identification. The reason for this is that pitch is highly dependent on the melody of humming and less on the target. In our experiments, LPCEPSTRA and PLP performed the best in verification and identification, respectively. Our future work will focus on analyzing humming of patients with speech disorder for real applications.

5. REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems*, vol. 14, no. 1, pp. 4–20, 2004.

[2] N. Suzuki, Y. Takeuchi, K. Ishii, and M. Okada, "Effects of echoic mimicry using hummed sounds on human-computer interaction," *Speech Communication*, vol. 40, no. 4, pp. 559–573, 2003.

[3] K. Andersson and L. Schalén, "Etiology and treatment of psychogenic voice disorder: Results of a follow-up study of thirty patients," *Journal of Voice*, vol. 12, no. 1, pp. 96–106, 1998.

[4] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: musical information retrieval in an audio database," *in Proc. the 3rd ACM international conference on Multimedia*, pp. 231–236, 1995.

[5] J. Shifrin, B. Pardo, C. Meek, and W. Birmingham, "HMM-based musical query retrieval," *Ann Arbor*, vol. 1001, pp. 48109–2110.

[6] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J .Godfrey, "Modeling prosodic dynamics for speaker recognition," *Proc. ICASSP*, vol. 4, 2003.

[7] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[8] Jean-Luc Gauvain and Chin-Hui Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[9] Jian lai Zhou, Ye Tian, Yu Shi, Chao Huang, and Eric Chang, "Tone Articulation Modeling For Mandarin Spontaneous Speech Recognition," in *Proc. ICASSP*, 2004, vol. 1, pp. 997–1000.

[10] David A. van Leeuwen, Alvin F. Martin, Mark A. Przybocki, and Jos S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Computer Speech and Language*, vol. 20, pp. 128–158, 2006.