한국음향학회지 제34권 제2호 pp. 171~176 (2015)
The Journal of the Acoustical Society of Korea Vol.34, No.2 (2015)
http://dx.doi.org/10.7776/ASK.2015.34.2.171

pISSN : 1225-4428
eISSN : 2287-3775

Text Independent Speaker Verficiation Using Dominant State Information of HMM-UBM

HMM-UBM의 주 상태 정보를 이용한 음성 기반 문맥 독립 화자 검증

Suwon Shon, Jinsang Rho, Sung Soo Kim,* Jae-Won Lee,* and Hanseok Ko†

(손수원, 노진상, 김성수,* 이재원,* 고한석†)

Department of Electronic Engineering Korea University, *Samsung Electronics

(Received December 24, 2014; revised January 16, 2015; accepted January 29, 2015)

ABSTRACT: We present a speaker verification method by extracting i-vectors based on dominant state information of Hidden Markov Model (HMM) - Universal Background Model (UBM). Ergodic HMM is used for estimating UBM so that various characteristic of individual speaker can be effectively classified. Unlike Gaussian Mixture Model(GMM)-UBM based speaker verification system, the proposed system obtains i-vectors corresponding to each HMM state. Among them, the i-vector for feature is selected by extracting it from the specific state containing dominant state information. Relevant experiments are conducted for validating the proposed system performance using the National Institute of Standards and Technology (NIST) 2008 Speaker Recognition Evaluation (SRE) database. As a result, 12 % improvement is attained in terms of equal error rate.

Keywords: Text-independent, Speaker verification, HMM-UBM, i-vectors

PACS numbers: 43.72.Fx

초 록: 본 논문에서는 Hidden Markov Model(HMM) – Universal Background Model(UBM)의 주 상태 정보 기반의 i-vector 추출 기술을 제안한다. Ergodic HMM이 UBM을 추정하는데 쓰였으며, 이를 통해 동일 화자 음성에도 다양하게 존재하는 특성을 HMM states로 분류할 수 있다. 제안한 방법을 이용하면 HMM의 state 개수에 따라 i-vector들이 추출되는데, 주 상태 정보 방법을 통해 이들 중 하나를 선택한다. 제안한 방법을 검증하기 위해 National Institute of Standards and Technology(NIST) Speaker Recognition Evaluation(SRE) database를 이용하여 실험을 하였으며, Equal Error Rate(EER) 성능 수치에서 12 %의 성능 향상을 확인할 수 있었다.

핵심용어: 문맥독립, 화자검증, HMM-UBM, i-vectors

I. Introduction

Recently, accuracy of text-independent speaker verification has been significantly improved by the i-vector extraction paradigm with probabilistic linear discriminant analysis.[1,2] The i-vector approach in total variability space first introduced in[1] It has been considered as the state of the art in speaker verification systems. It is originated in JFA framework that consists of defining two distinct spaces: a speaker and a channel space.

A fundamental assumption of the approach is that the high dimensional GMM supervector of speech utterance can be represented by a speaker and a channel subspace based Universal Background Model (UBM). Therefore, obtaining a UBM is the first step of text-independent speaker verification. The modeling technique of UBM is based on GMM that is widely used in numerous areas of acoustic and speech processing. Although HMM is the method of choice for speech recognition,[3] for text-independent speaker verification, the model should embrace a wide range of phonetic variation rather than modeling temporal patterns of speech. Underscoring this fact, Reynolds[4] stated that there is no advantage in text-independent task using a more complex likelihood function, e.g. HMM, than

†**Corresponding author:** Hanseok Ko (hsko@korea.ac.kr)
School of Electrical Engineering, Korea University, Anam-dong, Seongbuk-Gu, Seoul 136-701, Republic of Korea
(Tel: 82-2-3290-3239, Fax: 82-2-3291-2450)

GMM.

However, there have been some research efforts using more complex likelihood function like HMM. Using HMM in speaker verification was first introduced by Poritz.[5] Text-independent speaker recognition based on HMM was done by Naftali *et al*.[6] and Tomoko.[7] They used an ergodic HMM for speaker recognition and evaluated the performance with other algorithms. Recently, BenZeghiba[8] proposed User Customized Password Speaker Verification (UCP-SV) using multiple reference and background models. Because his or her own password is chosen by customer without any lexical constraints, the UCP-SV has similar freedom of input speech utterance with text-independent speaker recognition. BenZeghiba's team used a combined HMM/GMM approach with background HMM models and has shown better performance than that of the GMM-UBM baseline. R. Gajsek *et al*.[9] proposed a speaker state recognition system using a HMM-UBM based method. It shows superior results than the standard scheme of adapting GMM-UBM in both the emotion recognition task and the alcohol detection task.

By recognizing the effectiveness of capturing various characteristics of an individual speaker by the HMM-UBM, we propose to apply the HMM-UBM based i-vector feature for text-independent speaker verification system. Ergodic HMM is used for estimating UBM so that the individual speaker's characteristic is fully represented. Since an HMM has more than 1 state, several i-vectors can be extracted from GMM supervectors corresponding to each state. Using the dominant state information of test speech utterances with HMM-UBM, the score can be measured by distance between target and test i-vectors. The performance is evaluated with the i-vector baseline and other fusion rules for integrating i-vectors.

The outline of the paper is as follows. First, we describe the baseline i-vector for speaker verification briefly at section II. Section III and IV show the proposed i-vectors based text-independent speaker verification system and the scoring method. Section V presents experimental results and section 6 concludes this paper.

II. i-vectors for speaker verification

In classical Joint Factor Analysis (JFA)[10] a speaker utterance is represented by a supervector that consists of additive components from a speaker and a channel/session subspace. But in total variability perspective, a speaker utterance can be defined by both speaker and channel variability in a single space, i.e total variability space[1] as Eq.(1).

$$M = m + T_{\omega}, \tag{1}$$

where supervector M represents the speaker utterance, m denotes the speaker and channel independent supervector, i.e UBM supervector, T is a total variability matrix that is rectangular matrix of low rank and ω is total variability factor. We refer this factor as i-vector.

Consider a feature sequence of L frames y = $\{y_1, y_2, \cdots y_L\}$. The GMM-UBM model $\lambda = \{w_c, m_c, \Sigma_c\}$, $c = 1,...,C$ consists of C mixture components in a feature dimension F. For estimating the i-vector, the first step is extracting the zeroth and centered first order Baum-Welch statistics using the UBM as

$$N_c = \sum_{t=1}^{L} P(c \mid y_t, \lambda), \tag{2}$$

$$\widetilde{F}_c = \sum_{t=1}^{L} P(c \mid y_t, \lambda)(y_t - m_c), \tag{3}$$

where $c = 1,\cdots,C$ Gaussian component index, $P(c|y_t, \lambda)$ is the posteriori probability of mixture component c generating the vector y_t on UBM λ. m_c is the mean of the UBM mixture component c. Using these statistics and total variability matrix T, we can get the i-vector as follows

$$\omega = (I + T^t \Sigma^{-1} N T)^{-1} T^t \Sigma^{-1} \widetilde{F}, \tag{4}$$

where N is a diagonal matrix of dimension $CF \times CF$ whose

diagonal block is N_cI, $c = 1,...,C$ and Σ is a diagonal covariance matrix of dimesion $CF \times CF$ estimated during a factor analysis training.[11]

III. i-vector extraction based on HMM-UBM

The speaker and channel independent supervector m is estimated by GMM-UBM essentially by means of i-vector extraction. We propose to change "GMM-UBM" to "HMM-UBM" for speaker and channel independent supervector. An ergodic HMM model is employed such that all possible transitions between states are allowed for estimating the UBM as in Fig. 1 shown below as an example when there are two states. An ergodic HMM automatically forms broad phonetic classes corresponding to each state.[5,7] Hence, speech segment can be classified into one of the broad phonetic categories corresponding to the HMM states.

Consider S states in an ergodic HMM. For each state, there is a GMM λ_s , $s = 1,\cdots S$ that represents the automatically classified phonetic categories as follows.

$$\lambda_s = w_c^2,\ m_c^s,\ \Sigma_c^s. \tag{5}$$

Hence, Eq.(1) can be expressed as follows:

$$M = m_s + T_s\omega_s, \tag{6}$$

where m_s represents speaker and channel independent and phonetic category dependent supervectors. T_s is the total variability matrix for phonetic category s, and ω_s is the i-vector for phonetic category s. Then, we can have

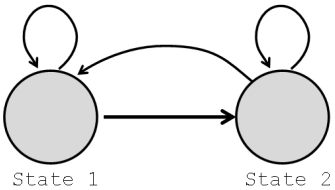

Fig. 1 Ergodic HMM model.

i-vectors $\boldsymbol{\omega}$ for each state as described below.

$$\omega = [\omega_1,\ \omega_2,...,\omega_S]. \tag{7}$$

For each state, there is UBM supervector for calculating the Baum-Welch statistics with input feature sequence y. Next, from Eq.(4) we can get the phonetic category dependent i-vector ω_s corresponding each state, i.e phonetic category.

IV. Scoring

This section describes scoring method for text-independent speaker recognizer using i-vectors. We proposed cosine distance scoring of i-vectors using dominant state information. In addition, we present general fusion rules for comparing performance with a baseline and the proposed scoring method.

4.1 Scoring using dominant state information

HMM-UBM essentially contains state transition probabilities unlike the GMM-UBM that has no state transition probability. Consider a state sequence of L frames $\mathbf{q} = \{q_1, q_2,\cdots, q_L\}$. To find the most likely state sequence $\mathbf{Q}$, it is needed to define the best score function along the path $\mathbf{q}$ at time t as

$$\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} P[q_1 q_2 \ldots q_{t-1}, q_t = i, y_1 y_2 \cdots y_t|\lambda]. \tag{8}$$

Using this best score function δ_t, we can find the most likely state sequence $\mathbf{Q}$ of input speech utterance based on Viterbi algorithm and HMM-UBM parameters.[12] In other words, the phonetic category of input speech utterance can be estimated frame by frame. Using this state sequence information, the dominant phonetic category of an input speech utterance can be determined by counting category index frame by frame. Then, the state index that appeared most is set as the dominant state information D. Finally, using this dominant state information D, the score of input

utterance can be calculated as Eq.(9).

$$score(\omega_{target,D}, \omega_D) = \frac{< wmega_{target,D}, \omega_D >}{\| \omega_{target,D} \| \ \| \omega_D}, \quad (9)$$

where $\omega_{target,D}$ is target i-vector from state D.

4.2 Feature level fusion

In feature level fusion, the normalized i-vectors for each state is simply concatenated into a vector as follows.

$$\omega = \begin{bmatrix} \overline{\omega}_1 \\ \vdots \\ \overline{\omega}_S \end{bmatrix} = \begin{bmatrix} \omega_1 / \sqrt{\omega_1^T \omega_1} \\ \vdots \\ \omega_S / \sqrt{\omega_S^T \omega_S} \end{bmatrix}, \quad (10)$$

where ω_s is i-vector from the HMM state S as Fig. 2. Both i-vector of a target and a test speech utterances should be done with Eq.(10). Using this concatenated ω, we calculate the Cosine Distance Score (CDS).

4.3 Score-level fusion

In score-level fusion, test speech utterance is scored with each phonetic category. In this paper, each i-vector calculated CDS using UBMs of each state s. So each of the i-vector extraction system acts as an independent speaker recognizer. Finally, scores were summed.

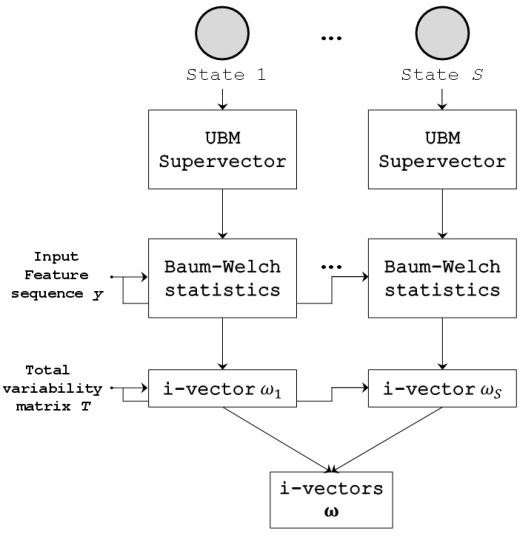

Fig. 2. i-vectors extraction based on HMM-UBM flow chart.

V. Experiments

5.1 Experimental condition

Experiments were conducted on the core task of NIST 2008 Speaker Recognition Evaluation (SRE), namely short2-short3, with common condition 7 (telephone-telephone-English) for female only. Both GMM-UBM and HMM-UBM was trained on telephone utterances selected from NIST2005 and 2006 SRE database with female gender dependent 1024 Gaussian mixture and diagonal-covariance matrices. For HMM-UBM, we use ergodic HMM that consist of 2 states. Our experiment operated on 13-dimensional MFCC feature extracted using 25-ms Hamming window. We apply feature warping to 13 dimensional feature vector on 3 seconds length sliding window.[13] Then delta and acceleration coefficients were appended to produce total 39-dimentioanl feature vectors. i-vector was extracted using 400 dimensional subspace matrix T that was trained with same database as UBM. All i-vector are normalized by whitening and scaling the length of each i-vector to a unit length.[14] The dimensionality of i-vectors is further reduced to 250 by Linear Discriminant Anlysis (LDA). LDA transform matrix were estimated using same data bases as training UBM for each states s of HMM-UBM.

5.2 Experimental result

The result of the baseline, proposed and other fusion rule method are shown as table 1. The three indices, Equal Error Rate (EER), Detection Cost Function (DCF) 08 and DCF10, of performance were measured. The DCF008 and DCF10 are defined in the NIST SRE plan of 2008 and 2010. The baseline system evaluated with GMM-UBM based i-vector extraction scheme. All three proposed, DSI, feature fusion and score fusion system evaluated with HMM-UBM based i-vector extraction scheme via three scoring ways as described at chapter 4. DSI means system scoring using dominant state information.

Table 1. Performance evaluation result of speaker verification systems.

System name		EER (%)	DCF08	DCF10
baseline		5.70	0.2740	0.7873
Proposed	DSI	4.98	0.2389	0.6357
	Feature fusion	5.11	0.2493	0.6671
	Score fusion	5.27	0.2541	0.7028

From the result, it is apparent that the proposed systems shows the better performance than baseline. In other words, speech segment can be classified into one of the broad phonetic categories corresponding to the HMM states and shows better performance than using only one phonetic category such as GMM-UBM based baseline system. Especially, the proposed system with DSI shows best performance than other system. It shows improved performance in all three evaluation metrics than other systems.

VI. Conclusion

In this paper, we proposed i-vectors extraction based on dominant state information of HMM-UBM. For validating the performance of proposed algorithms, the experiments were conducted using NIST 2008 SRE database. Through the experimental evaluation, we found the performance improvement in EER and DCF when using proposed i-vectors extraction. Scoring methods were used in three ways. Among them, using dominant state information of HMM-UBM shows best performance in speaker verification. From this result, we developed more robust speaker verification system with utilizing HMM-UBM and its dominant state information.

Acknowledgement

This research was funded and supported by Samsung Electronics Co., Ltd.

References

1. N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. on Audio, Speech, and Lang. Process. **19**, 788-798 (2011).
2. P. Kenny, "Bayesian speaker verification with heavy tailed priors," Odyssey Speker and Language Recognition Workshop, Brno, Czech Republic, (2010).
3. T. H. Kwon and H. S. Ko, "Performance improvement in speech recognition by weighting HMM likelihood" (in Korean), J. Acoust. Soc. Kr. **22**, 145-152, 2003.
4. D. a. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing **10**, 19-41 (2000).
5. A. Poritz, "Linear predictive hidden Markov models and the speech signal," ICASSP, 1291-1294 (1982).
6. N. Z. Tishby, "On the application of mixture AR hidden markov models to text independent speaker recognition," IEEE Transactions on Signal Processing **39**, 563-570 (1991).
7. T. Matsui, and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's," IEEE Transactions on Speech and Audio Processing **2**, 1992-1995 (1994).
8. M. F. BenZeghiba, and H. Bourlard, "User-customized password speaker verification using multiple reference and background models," Speech Communication **48**, 1200-1213 (2006).
9. R. Gajšek, F. Mihelič, and S. Dobrišek, "Speaker state recognition using an HMM-based feature extraction method," Computer Speech & Language **27**, 135-150 (2013).
10. P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, (Report) CRIM-06/08-13, 1-17 (2005).
11. P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," IEEE Transactions on Speech and Audio Processing **13**, 345-354 (2005).
12. L. R. Rabiner, "A tutorial on hidden markov-models and selected applications in speech recognition," Proceedings of the Ieee **77**, 257-286 (1989).
13. J. Pelecanos, and S. Sridharan, "Feature warping for robust speaker verification," Interspeech, 213-218 (2001).
14. D. Garcia-romero, and C. Y. Espy-wilson, "Analysis of i-vector length normalization in speaker recognition systems.," Interspeech, 249-252 (2011).

Profile

▸ Suwon Shon (손수원)

He received his B.S. degree and M.S. degree in Electrical Engineering from Korea University, Seoul, Korea, in 2010 and 2012 respectively. Since 2010, he has been a Ph.D. student at Korea University. His research interests are multi-channel acoustic processing and speech, speaker recognition.

▸ Jinsang Rho (노진상)

He received his B.S. degree in Electrical Engineering from Korea University, Seoul, Korea, in 2011 respectively. Since 2011, he has been a Ph.D. student at Korea University. His research interests are speaker recognition and speech dereveberation.

▸ Sung Soo Kim (김성수)

Sung Soo Kim was born in Bucheon, Korea, in 1982. He received the B.S. degree in electrical, electronics and radiowave engineering from Korea University, Seoul, Korea, in 2008, and the M.S. degree in electrical engineering and computer science from Seoul National University (SNU), Seoul, Korea, in 2010. Since 2010, he has been with the Samsung Electronics, where he is currently an engineer of multimedia solution team. His research interests include speech recognition and machine learning.

▸ Jae-Won Lee (이재원)

Jae-Won Lee received the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science & Technology (KAIST), Daejon, Korea, in 1993 and 1999, respectively. In 1996, he joined the Samsung Electronics as a research staff member. He is currently a principal engineer of multimedia solution team. His research interests include speech recognition, natural language processing, and machine learning.

▸ Hanseok Ko (고한석)

He received his B.S. degree from Carnegie Mellon University, in 1982, his M.S. degree from Johns Hopkins University, in 1988, and his Ph.D. from the CUA, in 1992, all in the field of electrical engineering. At the onset of his career, he was with the WOL, Maryland, where his work involved signal and image processing. In March of 1995, he joined the faculty of the Department of Electronics and Computer Engineering at Korea University, where he is currently a professor. His professional interests include speech/image signal processing for pattern recognition, multi-modal analysis, and intelligent data fusion.