논문 2015-10-09

A Speaker Pruning Method for Real-Time Speaker Identification System

Min-Joung Kim, Soo-Young Suk, Jong-Hyeog Jeong*

Abstract : It has been known that GMM (Gaussian Mixture Model) based speaker identification systems using ML (Maximum Likelihood) and WMR (Weighting Model Rank) demonstrate very high performances. However, such systems are not so effective under practical environments, in terms of real time processing, because of their high calculation costs. In this paper, we propose a new speaker-pruning algorithm that effectively reduces the calculation cost. In this algorithm, we select 20% of speaker models having higher likelihood with a part of input speech and apply MWMR (Modified Weighted Model Rank) to these selected speaker models to find out identified speaker. To verify the effectiveness of the proposed algorithm, we performed speaker identification experiments using TIMIT database. The proposed method shows more than 60% improvement of reduced processing time than the conventional GMM based system with no pruning, while maintaining the recognition accuracy.

Keywords : Gaussian mixture model, Modified weighted model rank, Speaker pruning, Speaker identification

Ⅰ. Introduction

Speaker identification has been an important research topic for many years and various types of speaker models have been developed for higher identification accuracy. Among them, Continuous Hidden Markov Model (CHMM) [1, 2] has become the most popular one for construction of speaker identification system. One state CHMM, also called Gaussian Mixture Model (GMM), is widely used for speaker modeling [3-5]. K. Markov *et al.* showed that GMM can perform even better than the CHMM by using multi-states and WMR (Weighting Model Rank) also showed a better identification performance compared with ML(maximum Likelihood)[6]. Most studies mentioned above have focused only on the improvement of identification accuracy. But in real system, identification accuracy should be considered together with processing time. There have been two kinds of approaches to reduce the processing time; one is to reduce the number of frames of the input speech to be compared with models and the other is to reduce number of speakers to be compared. However, the frame reduction method could decrease the performance of system, since the important speaker's information for identification can be included in the removed frames.

In general, it is known that the WMR method has better performance than maximum likelihood method[7]. In this method, instead of likelihood, a weighting value is used to increase discrimination between speakers, when total score is calculated to decide

*Corresponding Author(jhjeong@ikw.ac.kr)
Received: 14 Oct. 2014, Revised: 14 Jan. 2015, Accepted: 15 Jan. 2015.
M.J Kim, J.H Jeong: Kyungwoon University
S.Y Suk: GITC

 2015 Apr. 10(2) 65-71
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2015.10.2.65

identified speaker. However, WMR should re-calculate the weighting value whenever the number of speaker is changed. For this reason, WMR cannot be applied to reduce the number of speaker without additional calculation burden.

Therefore, we propose a novel speaker pruning method, which can reduce the total calculation cost with maintaining identification accuracy. The proposed method takes two processing steps; speaker selection step and speaker identification step. WMR can be used in the speaker identification step. In speaker selection step, speakers having higher likelihood are previously selected as candidates of an identified speaker, where only parts of frames of input speech are used for selection of speakers. In speaker identification step, the system determines identified speaker using Modified WMR (MWMR)[8, 9], where only selected speakers are considered.

This paper is organized in the following way. Section 2 introduces conventional speaker identification such as ML, and Modified WMR. The proposed method is described in section 3. Section 4 shows some experiment results. Section 5 summarized conclusions.

II. Speaker identification methods

1. Frame level maximum likelihood method.

Given a sample of a speech utterance, speaker identification has to decide to whom of a group of N known speakers this utterance belongs.

According to the Bayes' rule[9], the identification task is to find the speaker i^* whose model λ_{i^*} maximizes a posteriori probability $P(\lambda_i|X)$ as follows:

$$P(\lambda_i|X) = \frac{P(X|\lambda_i)P(\lambda_i)}{p(X)}, \quad 1 \le i \le N, \qquad (1)$$

where, due to lack of prior knowledge, we assume equal-likely speaker models. That is, the prior probabilities $P(\lambda_i)$ are set equal:

$$P(\lambda_i) = 1/N, \quad 1 \le i \le N, \qquad (2)$$

$p(X)$ is actually the unconditional likelihood of the occurrence of the utterance X and is the same for all speakers. Therefore, $\max_i p(X|\lambda_i)$ will maximize a posteriori probability and the identification decision can be simplified to:

$$i^* = \arg\max_i p(X|\lambda_i), \qquad (3)$$

where, i^* is the identified speaker.

For minimize the text dependent variations in the test utterance, the likelihood normalization by the background speaker is important [4, 7]. The normalized likelihood can be written as:

$$p_{norm}(x_t|\lambda_i) = \frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)}, \qquad (4)$$

where, $p(x_t|\lambda_b)$ is likelihood of $b$background model at t frame, $p_{norm}(x_t|\lambda_t)$ is normalized likelihood of i model at t frame.

Using Eq. (4), normalized likelihood is accumulated over all vectors x_t, $t = 1,2,\cdots,T$ for each speaker model i. The accumulated value can be used as a new score $Sc_i(X|\lambda_i)$, to identify the speaker.

$$Sc_i(X|\lambda_i) = \frac{1}{T}\sum_{t=1}^{T} \log p_{norm}(x_t|\lambda_i). \qquad (5)$$

It should be noted that this ML method provides high accuracy only when the number of comparing speakers is small and they have fairly different vocal tract features.

2. Modified weighting model rank (MWMR) method

It is known that the Weighting Model Rank (WMR) proposed by K. Markov *et al.* has better performance than maximum likelihood method. However, it could decrease the system performance, since higher ranked frame likelihood is substituted with a big weight even when its frame likelihood is very small. If we

Table 1. N-best list of speaker models and its likelihood and weights

rank r	Likelihood	Weighting $w(r)$	Speaker model
1	p_l^t	$w(1)$	Model λ_l (max. likelihood)
2	p_j^t	$w(2)$	Model λ_j
…	…	…	…
N	p_l^t	$w(N)$	Model λ_p (min. likelihood)

consider both relative and absolute value of the likelihood at the frame level, the frames including less information of speaker's vocal tract could have smaller values of values(weights), and cause these frames have a less effect to discriminate the speaker [8] [9]. We call this method as Modified WMR (MWMR) method.

The speaker identification by MWMR method is divided into the following 3 steps:

Step 1: For each test vector x_t, $t=1,2,\cdots,T$, the likelihood of each speaker at frame level is calculated and sorted with descending order. This is, speaker model which has the biggest likelihood is placed at top, and speaker model which has the smallest likelihood is placed at the bottom.

Step 2: For each model λ_i, find its rank r, i.e. its place in the N-best list, and assign the corresponding weigh $w(r_\lambda)$ for $r_\lambda = 1,\cdots,N$ to model as follows:

$$w(r_\lambda) = \exp(\alpha - \beta r_\lambda), \quad r_\lambda = 1,\cdots,N \tag{6}$$

where, α and β are weighting factor, and are calculated using probability density function of speaker model's rank. Consequently, if speaker's number is changed, they have to be re-calculated. The weighting value could be calculated with Eq. (6) which shows the best performance in [7]. Table 1 shows relation among rank, speaker model, and weighting value.

Step 3: Total score $Sc(X|\lambda_i)$ could be calculated by multiplying the frame likelihood $p(x_t|\lambda_i)$ by corresponding its each weight $w_t(r_{\lambda_i})$ at each model λ_i:

$$Sc(X|\lambda_i) = \sum_{t=1}^{T} p(x_t|\lambda_i) w_t(r_{\lambda_i}) \tag{7}$$

where, $w_t(r_{\lambda_i})$ and $p(x_t|\lambda_i)$ is a weighting value and a frame likelihood of model i, respectively.

In our previous works [8, 9], we showed that MWMR method has better performance for speaker identification system than WMR method.

Ⅲ. Speaker pruning method

In conventional speaker identification methods, each speaker's likelihood is calculated and accumulated over all frames. In this case, as the number of test frames and the number of speakers increase, the calculation cost increases. Therefore, the reduction of calculation cost should be considered for realization of a large scale of practical system. To reduce the computational cost, it is considered to use a part of input frames in speaker selection step or to use a small number of speakers in identification step.

In this section, we describe a novel speaker pruning method, which can reduce the total calculation cost with maintaining identification accuracy. The speaker selection and speaker identification procedures are described, in details, as follows.

Speaker selection procedure

1. Calculate the frame likelihood with a part of input frames over the whole speaker models.
2. Sort each speaker models in descending order, according to its accumulated frame

likelihood.

3. Select some high ranked speaker models and prune the others.

Speaker identification procedure

1. Calculate the weighted score (multiplication of frame likelihood and weighting value) of each frame with each selected speaker's models over all frames of test speech and sum the scores. This becomes a total weighted score for each speaker model.
2. Compare the total weighted scores and determine identified speaker whose score is the biggest.

The proposed method allows the reduction of the total calculation costs in comparison with the conventional methods with no pruning, since it uses only speakers selected through speaker selection procedure to determine the identification speaker. Identification accuracy can be increased if we apply MWMR method without re-calculation of weighting factor.

IV. Experiments

1. Database and analysis

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus[10] is used for experiments. The database includes 6,300 sentences (10 sentences spoken by each of 630 speakers), recorded in a quiet environment. 8 sentences (1 SA, 5 SX, and 2 SI) were used for training the speaker dependent models and 1 sentence (1 SA) is used for test. Speech analysis conditions are shown in Table 2. In experiments, we used the original data including silence frames and CMN was not applied in accordance with [11]. The conventional speaker identification systems often use 4 mixtures or 8 mixtures for GMM. However, we use 16 mixtures of GMM because 16 mixtures are known to be enough for TIMIT database. In candidate speaker selection, we use ML method to get better performance.

Table 2. Speech analysis conditions

Sampling Rate	16 kHz
Pre-emphasis coefficient	0.98
Hamming Windows	yes
Frame length	320 points
Frame Shift	160 points
Cepstrum vector dimension	10

Table 3. Calculation costs according to used input frame rate and selected speaker rate (%)

Selected speaker / Used frame	5	10	20	30	40	50
10	33,075	44,100	66,150	88,200	110,250	132,300
20	55,125	66,150	88,200	110,250	132,300	154,350
30	77,175	88,200	110,250	132,300	154,350	176,400
40	99,225	110,250	132,300	154,350	176,400	198,450
50	121,275	132,300	154,350	176,400	198,450	220,500
60	143,325	154,350	176,400	198,450	220,500	242,550
70	165,375	176,400	198,450	220,500	242,550	264,600
80	187,425	198,450	220,500	242,550	264,600	286,650
90	209,475	220,500	242,550	264,600	286,650	308,700
100	231,525	242,550	264,600	286,650	308,700	330,750

2. Experimental results

We conducted speaker identification tests with conventional ML method for reference. 90.47% of speaker identification rate is obtained with TIMIT database. Table 3 shows the calculation costs according to the percentage of the used frames and the selected speakers. Where, the calculation cost for one frame and one speaker model is presented by 1. Therefore, total calculation cost will be 220,500 (630 speakers x 350 input test frames) in case of non-speaker pruning.

Table 4 shows identification rates according to the rates of selection of speakers and the rates of used input frames, when ML method is used in speaker identification step. The speaker selection rates and input frame rates

Table 4. Identification rates according to the rates of selection of speakers and the rates of used input frame by ML method (%).

Used frame rates \ Speaker selection rates	10	20	30	40	50
10	85.55	88.88	90.00	90.15	90.31
20	90.00	90.47	90.47	90.47	90.47
30	90.47	90.47	90.47	90.47	90.47
40	90.47	90.47	90.47	90.47	90.47
50	90.47	90.47	90.47	90.47	90.47
60	90.47	90.47	90.47	90.47	90.47
70	90.47	90.47	90.47	90.47	90.47
80	90.47	90.47	90.47	90.47	90.47
90	90.47	90.47	90.47	90.47	90.47
100	90.47	90.47	90.47	90.47	90.47

Table 5. Identification rates by MWMR with respect to used frames

Used frame rates \ Selected speaker number	35
10	80.15
20	88.88
30	92.06
40	91.90
50	92.06
60	91.90
70	92.06
80	91.90
90	91.90
100	92.06

were varied from 10% to 50% and from 10% to 100%, respectively.

From Table 4, two cases (one for 20% of input frames with 20% of speakers and the other for 30% input frames with 10% speakers) show the same identification rate as we obtained from the conventional ML method. This means that 60% of calculation cost can be reduced without decreasing speaker identification accuracy.

To evaluate our proposed method, we applied MWMR method in speaker identification procedure. In this test, we used a weighting factor obtained from evaluating the ranks of 35 speaker's likelihood. High ranked 35 speaker models are used for identification and the others are pruned. Thus, MWMR method can be applied to identification procedure without re-calculation weighting factor, regardless of the number of speakers. Table 5 shows identification rates by MWMR with respect to used frames.

From Table 5, we can find that the identification rate approaches to the maximum rate of 92.06% if we use more than 30% of frames. This indicates that our proposed method which uses about 5% of speakers and 30% of frames provides the identification rate of 2% higher than the conventional ML method, even when 65% of calculation costs are reduced.

V. Conclusion

In this paper, we proposed a new speaker-pruning algorithm that effectively reduces the calculation cost. The proposed method consists of speaker selection and speaker identification procedure. In speaker selection procedure, only a part of speaker models which have a big likelihood are selected using only a part of input frames. In speaker identification procedure, identified speaker is decided from evaluating the selected speaker models. To improve the identification performance, MWMR is applied in speaker identification. The proposed algorithm allows reducing the total calculation costs, since some speakers' models are pruned through speaker selection procedure. In the identification tests using TIMIT Acoustic-Phonetic Continuous Speech Corpus, our proposed method provided the identification rate of 2% higher than the conventional ML method, while reducing 65% of calculation costs.

References

[1] S. Furui, "Speaker-dependent feature extraction, recognition and processing techniques," Speech Communication, Vol. 10, No. 5-6 pp. 505-520, 1991.

[2] T. Matsui, S. Furui, "Comparison of text independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 157-160, 1992.

[3] D.A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech and Audio Processing, Vol. 3, No. 1, pp. 72-83, 1995.

[4] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, Vol. 17, No. 1-2, pp. 91-108, 1995.

[5] M.J. Kim, J.H. Jeong "A realization of injurious moving picture filtering system with Gaussian mixture model and frame-level likelihood estimation", Journal of Korean Institute of Intelligent System, Vol 23, No 2, pp. 184-189, 2013 (in Korean).

[6] K. Markov, S. Nakagawa, "Text-independent speaker identification on TIMIT database," Proceedings of Acoust. Soc. Jap., Vol. 1, pp. 83-84, 1995.

[7] K. Markov, S. Nakagawa, "Frame level likelihood normalization for text-independent speaker identification using Gaussian mixture Models," Proceedings of International Conference on Speech and Language Processing, pp. 1764-1767, 1996.

[8] M.J. Kim, S.J. Oh, H.Y. Jung, H Y. Chung, "Frame selection, hybrid, modified weighting model rank method for robust text-independent speaker identification," J. Acoust. Soc. Kor., Vol. 21, No. 8, pp. 735-743, 2002 (in Korean).

[9] M.J. Kim, S.J. Oh, S.Y. Suk, H.Y. Jung, H.Y. Chung, "Modified weighting model rank method for improving the performance of real-time text-independent speaker recognition system," Proceedings of Acous. Soc. Kor., pp. 107-110, 2002 (in Korean).

[10] V. Zue, S. Seneff, J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech Communication, Vol. 9, No. 4, pp. 351-356, 1990.

[11] D.A. Reynolds, "Large population speaker identification using clean and telephone speech," IEEE Signal Processing Letters, Vol. 2, No. 3, pp. 46-48, 1995.

Min-Joung Kim (김 민 정)

He received the Ph.D. degree in Information & Communication Engineering from Youngnam University, Gyeongbuk, Republic of Korea, in 2003. In September 2011, he joined Kyungwoon University, Gyeongbuk, Republic of Korea, where he is currently the assistant Professor. His research interests include Speech Recognition and Speaker Identification, Speaker Verification system.

Email: manjukz@naver.com

Su-Young Suk (석 수 영)

He received the Ph.D. degree in Information & Communication Engineering from Youngnam University, Gyeongbuk, Republic of Korea, in 2004. In 2005, he was a Postdoctoral Research Fellow in Tohoku University. In March 2013, he joined Gyeongbuk Institute of IT Convergence Industry Technology, Gyeongbuk, Republic of Korea, where he is currently the director of the headquarters. His research interests include IT System for Vehicle and Speech Recognition, Signal Processing Software.
Email: sysuk@gitc.or.kr

Jong-Hyuk Jeong (정 종 혁)

He received the Ph.D. degree in Electronics & Communications Engineering from Korea Maritime and Ocean University, Busan, Republic of Korea, in 1999. In March 2000, he joined Kyungwoon University, Gyeongbuk, Republic of Korea, where he is currently the associate Professor. His research interests include Signal Processing and Ubiquitous Sensor Network, Radio Communication System.
Email: jhjeong@ikw.ac.kr