Impostor Detection in Speaker Recognition Using Confusion-Based Confidence Measures

Kyuhong Kim, Hoirin Kim, and Minsoo Hahn

ABSTRACT—In this letter, we introduce confusion-based confidence measures for detecting an impostor in speaker recognition, which does not require an alternative hypothesis. Most traditional speaker verification methods are based on a hypothesis test, and their performance depends on the robustness of an alternative hypothesis. Compared with the conventional Gaussian mixture model–universal background model (GMM-UBM) scheme, our confusion-based measures show better performance in noise-corrupted speech. The additional computational requirements for our methods are negligible when used to detect or reject impostors.

Keywords—Speaker recognition, speaker verification, open-set speaker identification, confidence measure.

I. Introduction

In speaker recognition, detecting imposters is regarded as speaker verification. The posterior probability for speaker verification is represented by

$$\mathrm{P}(\lambda_c \mid X) = \frac{\mathrm{P}(X \mid \lambda_c)\mathrm{P}(\lambda_c)}{\sum_{\text{all } i} \mathrm{P}(X \mid \lambda_i)\mathrm{P}(\lambda_i)}, \tag{1}$$

where X is a feature vector sequence and λ_c is a claimed speaker model. The claimed speaker may be obtained from the result of speaker identification. By giving all speakers equal prior probability, the posterior probability can be rewritten as

$$\mathrm{P}(\lambda_c \mid X) = \frac{\mathrm{P}(X \mid \lambda_c)}{\sum_{\text{all } i} \mathrm{P}(X \mid \lambda_i)} \approx \frac{\mathrm{P}(X \mid \lambda_c)}{\mathrm{P}(X \mid \lambda_{UBM})}, \tag{2}$$

where λ_{UBM} is a universal background model (UBM) [1] which represents the speaker-independent distribution of feature X. After all, in the hypothesis test, the alternative hypothesis is UBM and the likelihood ratio is compared with a predefined threshold to decide if the claimed speaker is accepted or rejected as follows:

$$\frac{\mathrm{P}(X \mid \lambda_c)}{\mathrm{P}(X \mid \lambda_{UBM})} \begin{cases} \geq \gamma & \text{accept} \\ < \gamma & \text{reject}. \end{cases} \tag{3}$$

In real situations, we should always consider the problem of noise. Speaker recognition performance is known to be degraded under mismatched conditions. To cope with environmental mismatch problems, many approaches have been studied with the categories of speech enhancement, feature compensation, model adaptation, and so on. In the confidence score for impostor detection in (3), noise corruption affects the likelihoods given by both the claimed speaker model and the UBM. If the noise corruption decreases the numerator in (3) while increasing the denominator, the likelihood ratio will be drastically decreased, and vice-versa.

To cope with this noise problem, we take advantage of confusability between the claimed speaker model and its nearest neighbor models with different criteria. We assume that the most likely model is not easily changed with slight noise corruption. This assumption is also used in [2] with application to utterance verification. Similarly, the neighbor models of the claimed speaker model are not easily changed by slight noise corruption.

In this letter we propose novel confidence scores using speaker confusability for robust impostor detection as a post-processing of open-set speaker identification.

Manuscript received Nov. 23, 2005; revised Oct. 11, 2006.

Kyuhong Kim (phone: +82 31 280 1727, email: kkyung.kim@samsung.com) was with the School of Engineering, ICU and is currently with the Computing Technology Lab., Samsung Advanced Institute of Technology, Yongin, Korea.

Hoirin Kim (email: hrkim@icu.ac.kr) and Minsoo Hahn (email: mshahn@icu.ac.kr) are with the School of Engineering, ICU, Daejeon, Korea.

II. Confusion-Based Confidence Scores for Speaker Verification

1. Speaker Confusion Rate (SCR)

It is well known that the Gaussian mixture model (GMM) is a special case of the hidden Markov model (HMM). Each speaker HMM has a single state with multiple mixtures and the self-state transition probability is one. Thus, we can rewrite the Viterbi search algorithm under GMM-based constraints. The search score $\delta_t(j)$ and path information $\psi_t(j)$ can be represented as

$$\begin{aligned}\delta_t(j) &= \max_{1\le i\le N_{speaker}} (\delta_{t-1}(i)a_{ij}\,\mathrm{P}(x_t \mid \lambda_j)) \\ &= \delta_{t-1}(j)\,\mathrm{P}(x_t \mid \lambda_j),\end{aligned} \tag{4}$$

$$\psi_t(j) = \arg\max_{1\le i\le N_{state}}(\delta_{t-1}(i)a_{ij}) = j\ , \tag{5}$$

where the search scores are recursively calculated in all speaker models and frames. In [3], a confusion-based confidence measure for utterance verification was introduced, where the momentary best-state sequence is traced during a Viterbi search, and then it is compared with a null hypothesis to calculate the confusion-based confidence score. In this letter, we use confusability in speaker verification. The momentary best-speaker sequence in the Viterbi search is traced by the following criterion:

$$m_t = \arg\max_{1\le j\le N_{speaker}} \delta_t(j). \tag{6}$$

To quantify the confusability between the claimed speaker and the momentary best speakers, we calculate the coincidence rate, or speaker confusion rate (SCR) as

$$\mathrm{SCR}(S,M) = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \mathrm{d}(s_t, m_t)\ , \tag{7}$$

$$\mathrm{d}(s_t, m_t) = \begin{cases} 0, & \text{if} \quad s_t = m_t \\ 1, & \text{otherwise}. \end{cases} \tag{8}$$

The SCR is compared with a threshold in order to decide whether the claimed speaker is to be accepted or not. The physical meaning of SCR is how much more confusion there is between the claimed speaker and other speakers. Thus, if the SCR is smaller than the threshold, the claimed speaker is accepted by the following decision rule:

$$\mathrm{SCR}(S,M)\begin{cases} \ge \gamma & \text{reject} \\ < \gamma & \text{accept}. \end{cases} \tag{9}$$

As the most likely speaker models based on the search score are traced frame by frame, they are compared with the claimed speaker model. Since the SCR approach can reflect different frame sizes on calculating likelihood scores, it can be more noise-robust than the conventional GMM-UBM. However, towards the end of the speaker search, the score is the product of most frames. Thus, one of the drawbacks of SCR is the multiplication of frame likelihood errors while searching.

2. Framewise Speaker Confusion Rate (FSCR)

In text-independent speaker identification, the accumulated likelihood score $\delta_t(j)$ is always affected by noise corruption, which increases as frame processing proceeds. To lessen the effect of noise on frame likelihood errors, the decision criterion in (6) is modified frame by frame as

$$\hat{m}_t = \arg\max_{1\le j\le N_{speaker}} \mathrm{P}(x_t \mid \lambda_j)\ . \tag{10}$$

In a manner similar to the evaluation of the SCR, the confidence score, or framewise speaker confusion rate (FSCR) is determined by

$$\mathrm{FSCR}(S,\hat{M}) = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \mathrm{d}(s_t, \hat{m}_t). \tag{11}$$

Because the FSCR approach prevents errors from being multiplied over all the frame-level likelihoods, it can be more noise robust than SCR.

3. Accumulated Speaker Confusion Rate (ASCR)

A segment-level likelihood can be defined by $\mathrm{P}(x_n, x_{n+1}, \cdots, x_{n+m-1} \mid \lambda_j)$, where m is the length of a speech segment. Then the segment-level decision rule is defined as

$$s_n^{(m)} = \arg\max_{1\le j\le N_{speaker}} \mathrm{P}(x_n, x_{n+1}, \cdots, x_{n+m-1} \mid \lambda_j), \tag{12}$$

where n = 1, 2,⋯, N–m+1. Then, the segment-level SCR is defined by

$$\mathrm{SCR}^{(m)}(X \mid \lambda_s) = \frac{1}{N-m+1}\sum_{n=1}^{N-m+1} \mathrm{d}(s_n^{(m)}, s). \tag{13}$$

If m is equal to 1, the $\mathrm{SCR}^{(m=1)}$ is identical to the FSCR. We can average all the possible segment-level SCRs to evaluate the accumulated SCR as

$$\mathrm{ASCR}(X \mid \lambda_c) = \sum_{m=1}^{\mathrm{N}} w_m\, \mathrm{SCR}^{(m)}(X \mid \lambda_c)\ , \tag{14}$$

where the segmental weights have the following constraints,

$$\sum_{m=1}^{N} w_m = 1 \ , \ 0 \le w_m \le 1. \qquad (15)$$

Because the ASCR is the weighted sum of all the possible segmental confidence scores, it is a generalized confusion-based confidence encompassing SCR and FSCR. Thus, the SCR and FSCR are special cases of ASCR.

III. Experiments

1. Databases

For our experiments, we used the Korean Speaker Recognition Database distributed by the Electronics and Telecommunications Research Institute (ETRI). The database for our experiments consisted of 80,000 utterances of 50 speakers. We set up our baseline which had 40 enrolled speakers, and the utterances of the remaining 10 speakers were used for impostors. A total of 64,000 utterances were used for training speaker-dependent models and a speaker-independent model, and 16,000 utterances were used for performance evaluations. For noise conditions, we considered two kinds of noise conditions: sporadic noise and car noise conditions. As a sporadic noise source, a clapping sound was collected using a condenser microphone with 16 kHz and added to clean speech. Car noise was collected from a KIA Sephia at a speed of 100 km/h with all of the windows closed; the car noise was then added to clean speech with the signal-to-noise power ratios (SNRs) of 5 dB, 10 dB, and 20 dB.

2. Experimental Results

A Gaussian mixture model was used for enrolled speaker models and the universal background model trained by an EM algorithm [4]. The number of Gaussian mixtures was 512 for both speaker model and universal background model. Every utterance was pre-emphasized with a factor of 0.97, and a 20 ms Hamming window was applied with 10 ms overlapping. A speech activity detector was then used to discard silence and unvoiced frames by simply tracing frame energy because we assumed that only voiced speech frames play a major role in discriminating speakers. The feature vector for each surviving frame consisted of 12th-order static and delta mel-frequency cepstral coefficients, resulting in the final 24th-order feature vector.

To investigate the operational characteristics of the conventional and proposed methods, two error terms were used: a false alarm and a false rejection. As the threshold changed, we found the equal error rate (EER), the error rate when the false alarm is equal to the false rejection. Table 1 shows EERs of the conventional method and the proposed methods. In our experiments, the weights (w_m) in (14) were set to 1/N.

Table 1. EERs under different noise conditions (%).

Noise conditions		GMM-UBM	SCR	FSCR	ASCR
Clean		0.39	1.27	0.61	0.71
Sporadic		2.80	3.42	2.03	2.23
Car	20 dB	5.68	2.23	0.65	0.96
	10 dB	10.05	8.12	2.05	2.78
	5 dB	22.37	19.34	11.72	10.96
Multi-condition		8.26	6.88	3.41	3.53

Table 2. EERs in the multi-condition using 512, 1024, and 2048 Gaussian mixtures (%).

Methods / Mixtures	GMM-UBM	SCR	FSCR	ASCR
512	8.26	6.88	3.41	3.53
1024	6.92	6.80	2.86	3.24
2048	6.86	6.62	2.68	3.21

Table 3. Speaker identification error rate (ER) under different noise conditions.

Noise condition		GMM (% ER)	ASCR (% ER)	Improvement (%)
Clean		0.11	0.11	0
Sporadic		3.24	2.47	23.8
Car	20 dB	1.19	0.16	86.6
	10 dB	15.68	3.90	75.1
	5 dB	46.24	30.35	34.4
Multi-condition		13.29	7.40	44.3

Under clean conditions, the GMM-UBM approach is superior to the confusion-based approaches because GMM-UBM can be the best solution in an ideal environment. However, our confusion-based approaches show better performance than the GMM-UBM under different noise conditions. Because different kinds of noise are frequently involved in real environments, our confusion-based approach can be more practical.

In additional experiments, 1024 and 2048 Gaussian mixtures were involved for the performance evaluations shown in Table 2. The proposed confusion-based methods are consistently superior to the GMM-UBM.

Table 3 shows the performance of speaker identification compared with the traditional maximum likelihood-based GMM approach [5], [6] under different noise conditions, in

which 50 speakers were enrolled. From the results, we can confirm that the ASCR is a very useful method for speaker identification as well as speaker verification. In a matched clean condition, the conventional GMM-based approach can be the best solution for speaker identification. However, as shown in Tables 1 and 3, the GMM-based approaches have shown drastic performance degradations in noise conditions. Therefore, in noisy conditions, our ASCR can be a good solution for the task of detecting or rejecting impostors in open-set speaker identification.

IV. Conclusions

To detect or reject impostors in text-independent speaker recognition applications, confusion-based confidence measures are proposed. Although our measures have shown slight performance degradation in a clean condition, the degradation is negligible and we confirm that FSCR and ASCR are more practical and have better performance in noisy conditions. The computational requirement for our confusion-based approaches is heavy from the viewpoint of traditional speaker verification tasks, because it increases in proportion to the number of enrolled speakers. However, if the ASCR is used to detect or reject impostors in the application to open-set speaker identification, two advantages are obtained. First, the computational requirements are very low in the case that all the frame-level likelihood scores are already calculated in the speaker identification stage. Secondly, the ASCR, as an identification score, can be directly used for speaker verification. The GMM-UBM always requires additional calculations for alternative model likelihood, but the ASCR does not require additional computations.

One of the drawbacks in our approach is that when the number of enrolled speaker is small, our confusion-based approaches fail to represent the confusability well. To overcome this problem, our future work will include finding a virtual speaker model set in order to effectively represent the confusability.

References

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Process*, vol. 10, 2000, pp. 19-41.

[2] Y. Jeong and H.S. Kim, "Recognition Confidence Scoring Using Recognition Results from Perturbed Input Feature Vectors," *IEE Electronics Letters*, vol. 37, Aug. 2001, pp. 1143-1145.

[3] K. Kim, H. Kim, and M. Hahn, "Utterance Verification Using Search Confusion Rate and Its N-Best Approach," *ETRI J.*, vol. 27, Aug. 2005, pp. 461-464.

[4] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Statist. Soc.*, vol. 39, 1977, pp. 1-38.

[5] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Models," *Speech Communication*, vol. 17, 1995, pp. 91-108.

[6] D. Reynolds and R.C. Rose, "Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. on SAP*, vol. 3, Jan. 1995, pp. 72-83.