ON OPTIMUM NORMALIZATION METHOD USED FOR SPEAKER VERIFICATION

Weijie Liu, Toshihiro Isobe, and Naoki Mukawa
{wjliu, isobe, mukawa}@lit.rd.nttdata.co.jp

Laboratory for Information Technology
NTT Data Corporation, Japan

ABSTRACT

Score normalization has become necessary for speaker verification systems, but general principles leading to optimum performance are lacking. In the paper, theoretical analyses to optimum normalization are given. Under the analyses, four existing methods based on likelihood ratio, cohort, a posteriori probability and pooled cohort are investigated. Performance of these methods in verification with known imposters, robustness for different imposters and separability of the optimal threshold from the imposter model are discussed after experiments based on a database of 100 speakers.

1. INTRODUCTION

Score normalization has become a necessary step to implement a speaker verification system. In recent years, methods based on likelihood ratio[1], cohort[2,4], a posteriori probability[5] and pooled cohort[3] have been proposed. Though effectiveness of these methods has been verified with researcher-dependent experiment databases, general principles for normalization leading to optimum performance are lacking.

Since speaker verification is essentially a special case of statistical pattern recognition, the normalization can be generally explained according to some fundamental characteristics of the latter. Simply, there exist three elements of *a speaker model, an imposter model* and *a decision rule with a discrimination function and a threshold* to construct a speaker verification system. A speaker model is built for a claimed speaker based on statistical analyses to his sample data obtained from registration. An imposter model, also known as *speaker background model*[3] or *antispeaker model*[4], should give a correct description to characteristics of the imposters feigning the claimed speaker. But in real world it is by no means predictable about the imposters. So existing methods take data from speakers of a predefined set to emulate the possible imposters. As to a decision rule, a discrimination function enables a similarity measure between a measurement and a model, and a threshold is necessary to make the decision that the measurement is whether from the claimed speaker or not. When the decision rule depends not only on the speaker model but also on the imposter model, the process to make a relative similarity measure corresponds to the so-called *normalization.*

2. THEORY PREPARATION

Let X represent a d-dimensional feature vector extracted from a measurement (an utterance), $p(X|A)$ and $p(X|R)$ the likelihood functions (the probability density functions) of X for the claimed speaker and the imposters, respectively, where A means *Acceptance* and R *Rejection*. Though minimizing EER(Equal Error Rate) of FAR(False Acceptance Rate) and FRR(False Rejection Rate) has been the most widely used objective function for optimization in literatures, minimizing FAR (or FRR) with fixed FRR (or FAR) is also necessary in practice to enable different security levels of a verification system. The optimum decision rules of the normalization in the two cases are given in the following Theorem 1 and 2 (to save size, proofs of all theorems are eliminated). It is noted that the errors to be minimized in the two cases are both different from the Bayes error[6] (which is the sum of FAR and FRR) or other errors suggested for classification systems.

Theorem 1: The decision rule

$$\text{if } \frac{p(X \mid A)}{p(X \mid R)} \geq t \text{ then } X \in A(t) \text{ (Acceptance);}$$
$$\text{else } X \in R(t) \text{ (Rejection)} \qquad (1)$$

minimizes EER when t is determined by

$$P(A)\int_{X \in A(t)} p(X \mid R)dX = P(R)\int_{X \in R(t)} p(X \mid A)dX, \qquad (2)$$

where, $P(A)$ and $P(R)$ are the a priori probabilities for *Acceptance* and *Rejection* respectively, and $A(t)$ and $R(t)$ the corresponding regions in feature space. Eq.(2) is $FAR(t)=FRR(t)$. Since the dimensional of X is often over one thoustand, the integrals in (2) are difficult to calculate. In practice, the optimal t is determined from $FAR(t)=FRR(t)$, where $FAR(t)$ and $FRR(t)$ are obtained through tests on samples.

Theorem 2: The decision rule shown in Eq.(1) minimizes FAR (or FRR) for fixed FRR (or FAR) when t *is* determined by

$$P(R)\int_{X \in R(t)} p(X|A)dX = c \text{ (or } P(A)\int_{X \in A(t)} p(X \mid R)dX = c), \quad (3)$$

where, c is the predetermined value for FRR (or FAR).

It is noticed that when *p(X|A)* and *p(X|R)* are known, theoretically, an optimum normalization can be implemented. But *p(X|R)* have to be predicted, and the optimal threshold have to be calculated using a great number of samples even when *p(X|R)* gets known. Besides the likelihood function, the a posteriori probability is the alternative score measure. The following Theorem 3 shows the equivalence between the two score measures.

Theorem 3: The decision rule

$$\text{if } P(A \mid X) \geq s \text{ then } X \in A(t)\ (\text{Acceptance}); \quad \text{else } X \in R(t)\ (\text{Rejection}) \tag{4}$$

is equivalent to Eq. (1) when *s* and *t* satisfy

$$s = \frac{P(A)t}{P(A)t + P(R)}, \tag{5}$$

where, *P(A|X)* is the a posteriori probability for *Acceptance*, which is calculated by

$$P(A \mid X) = \frac{P(A)p(X \mid A)}{P(A)p(X \mid A) + P(R)p(X \mid R)}. \tag{6}$$

3. METHOD INVESTIGATION

A speaker verification system is designed based on a set of training samples that are represented by *{Zij}*, where *i=1~I* means speakers and *j=1~J* utterances. *Si* represents all *J* samples of speaker *i*. When evaluating a normalization method, if the test set is the same as *{Zij}*, the test is called *a closed set test*; otherwise, *an open set test*. It is obvious that the actual situation corresponds to an open set test. But because we can never give an exact prediction for imposters, any open set test can hardly reflect all actual situations. Generally speaking, a closed set test shows the performance of a method when imposters are known, and an open set test also shows how a method is robust to different imposters. Existing methods utilize different sets of samples when building their imposter models.

3.1. Higgins Method[1]

Higgins et al. suggested to use a simple likelihood ratio to compare to a threshold in a decision rule. That is

$$LR_H = \log p(X \mid S_i) - \max_{r=1\sim I, r\neq i} \log p(X \mid S_r). \tag{7}$$

If all possible imposters are with the same likelihood function the method is optimal theoretically. Otherwise, the method emulates a serious situation for possible imposters and its imposter model is in fact not represented by a likelihood function because the integration of *max p(X|Sr) (r=1~I,r≠i)* over feature space is greater than 1.

3.2. Rosenberg Method 1[2][4]

Rosenberg et al. used the *cohort* concept in a decision rule. They defined

$$LR_R = \log p(X \mid S_i) - \log[stat_k\, p(X \mid Cohort_k^i)] \tag{8}$$

where, *Cohort* means to find a subset from *{Zij}* except for *Si*, to form the imposter model. In simple cases, a cohort is organized by individual speakers, that is, one *k* means one *Sr, r=1~I, r≠i*. *stat* means an arithmetic operator such as mean, medium, etc. When the speaker number of the *Cohort* set is *I-1* and the *stat* operator maximum, the method become the same as Higgins method. When all imposters are with the same likelihood function the method is optimum theoretically no matter what the details of *Cohort* and *stat* are. Depending on different cohort, various imposter models can be emulated.

3.3. Matsui Method[5]

Matsui et al. suggested to use the a posteriori probability in a decision rule as shown in Eq. (4), where the a posteriori probability is calculated by

$$P(S_i \mid X) = \frac{P(S_i)p(X \mid S_i)}{\sum_{r=1}^{I} P(S_r)p(X \mid S_r)} \approx \frac{p(X \mid S_i)}{p(X \mid \sum_{r=1}^{I} S_r)} \tag{9}$$

The method build a pooled model to make calculation easier, that is, samples from all speakers are taken for from one virtual speaker and are inputted to a HMM to get a unique likelihood function. According to Theorem 3, the method works as an optimum decision rule theoretically. But a premise condition is that the calculated a posteriori probability is correct, that is

$$\sum_{r=1}^{I} P(S_r)p(X \mid S_r) = \frac{1}{I} p(X \mid \sum_{r=1}^{I} S_r) \tag{10}$$

is satisfied. Ideally, when all speakers are with the same a priori probabilities and they are independent to each other, (10) exists. But in practice it is not ensured that if a HMM holds such a linearity property.

3.4. Rosenberg Method 2[4]

In fact an imposter model can be built by pooling an arbitry subset of *{Zij}*. Referring to Matsui method, a decision rule can be implemented optimally by:

$$LH_0 = \log p(X \mid S_i) - \log p(X \mid \sum_{r=1, r\neq i}^{I} S_r) \tag{11}$$

The only difference between the method and Matsui method is that whether the samples of the claimed speaker are pooled when building the imposter model.

4. EXPERIMENTSE

Experiment database consists of utterances recorded by 100 speakers, 50 males and 50 females. Each speaker was asked to utter 4 digits that was prompted at random each time, and 70 times of the utterances are recorded. After extraction, all feature vectors are grouped into two sets: Set A---speaker 1-25 with their first 30 times utterances; Set B---speaker 1-100 with their latter 40 times utterances.

4.1. Results of Experiments

Three experiments are executed for each of the four methods. *FAR(t) and FRR(t)* results are shown in Fig.1-4, respectively, where (a) is the results of a closed set test with Set A; (b) the results of an open set test with Set A used for training and Set B for testing; (c) the results of a closed set test with Set B. At all cases, a claimed speaker is one from the speaker 1-25. But imposters are from the remains of speaker 1-25 at (a), and from speaker 26-100 at (b) and (c).

Defects of the experiments are in speaker size and in unbalance for claimed speakers and imposters. Samples from 25+75 speakers may be too sparse to reflect a whole distribution in feature space and deficiency of samples for a claimed speaker makes the results of FRR unreliable.

4.2. Discussion

we compare the four methods in the following three aspects. (1) Results for closed set tests as shown in (a) and (c) reflect the performance of a method when the imposters are known. Also they give a measure to how much a method depends on the sample data when building models. In our case, it is found that the order is Higgins > Rosenberg1 > Matsui > Rosenberg2. Since the sample data for a claimed speaker are relatively less (or concentrated), the order means that using samples from few neighbors as imposters are effective than pooling samples of many speakers. At the case of pooling, samples from the claimed speaker himself should be included to make the imposter model not so far from the claimed speaker model. (2) At the cases of (a) and (b), models for claimed speakers and imposters are the same. So difference of the results between (a) and (b) is a measure to the robustness of a method when the actual imposters are different from the predicted ones. Since in practice (a) corresponds to the design step and (b) application step of a method, this robustness is essential to ensure a method to work at a stable state. Some quantitative descriptions are possible for comparison. From (a) we can get the threshold (score) with respect to EER (when there are more than one, their medium value are adopted), and at its two sides, the point with a distance of 5%(maximum score-minimum score) is also picked out. In detail, the scores are as follows: Higgins: -20.0±38.0; Rosenberg1: 86.5±25.8; Matsui: -8.0±26.9; Rosenberg2: 185.0±23.5. When the three points are set as thresholds, small values of FAR and FRR in (b) reflect better robustness. The results are as follows: Higgins: FAR(%)=(48.4, 13.9, 1.1), FRR(%)=(0, 0, 0); Rosenberg1: FAR(%)=(0.60, 0.05, 0), FRR(%)=(0.2, 2.2, 17.1); Matsui: FAR(%)=(0.51, 0.01, 0), FRR(%)=(0.2, 1.1, 15.7); Rosenberg2: FAR(%)=(0.74, 0.15, 0.02), FRR(%)=(1.8, 6.1, 19.4). The order from the best is Matsui > Rosenberg1 > Rosenberg2 > Higgins. (3) Though the optimal threshold is in nature dependent on the imposter model, it is hoped the dependency is weak in order to set a unique threshold reliably for different imposter models. The optimal thresholds in (a) and (c) are as follows: Higgins: -20.0/-18.8; Rosenberg1: 86.5/76.0; Matsui: -8.0/186.0; Rosenberg2: 185.0/2.2. So the order from the best is Higgins = Rosenberg1 > Matsui = Rosenberg2.

5. CONCLUSION

Four existing methods are investigated under analyses to optimum normalization. Though theoretically these methods can become optimal, their premise conditions required are hardly to be satisfied in practice. These methods are compared in the following three aspects: performance for known imposters, robustness for different imposters and separability of the optimal threshold from the imposter model.

The authors would like to thank Dr. J. Takahashi for reading the draft and giving valuable comments.

6. REFERENCES

1. **A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," Digital Signal Processing vol. 1, pp. 89-106, 1991.**
2. **A. E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F. K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," Proc. ICSLP 92, vol. 2, pp. 599-602, 1992.**
3. **E. Rosenberg and S. Parthasarathy, "Speaker Background Models for Connected Digit Password Speaker Verification," Proc. ICASSP 96, vol. 1, pp. 81-84, 1996.**
4. **C-S. Liu, H-C. Wang, and C-H. Lee, "Speaker Verification Using Normalized Log-Likelihood Score," IEEE Trans. Speech and Audio Processing, vol. 4, no. 1, pp. 57-60, Jan. 1996.**
5. **T. Matsui and S. Furui, "Similarity Normalization Method for Speaker Verification Based Speaker Recognit on a Posteriori Probability," ESCA Workshop on Automatic ion, Identification and Verification, pp. 59-62, 1994.**
6. **K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Inc., 1972.**

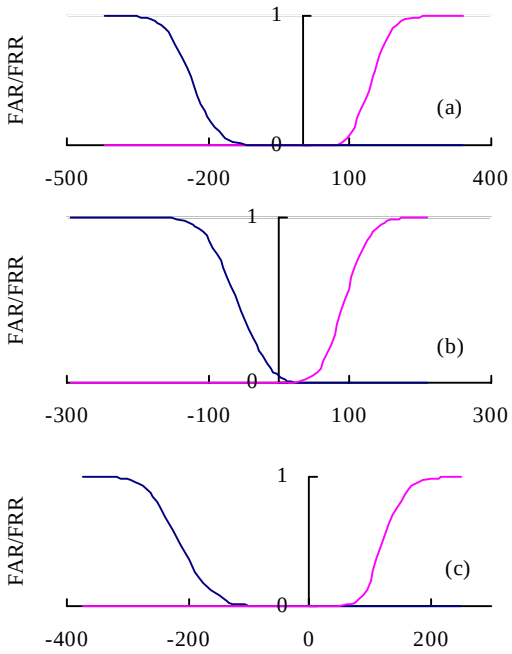

Figure 1. Results of Higgins Method

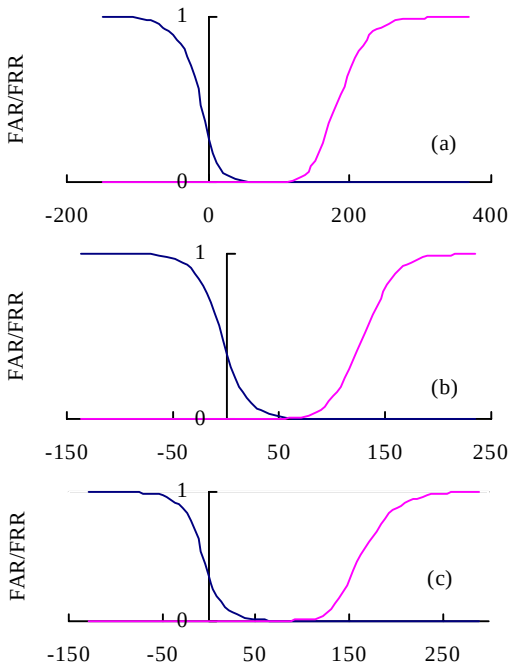

Figure 2. Results of Rosenberg Method 1

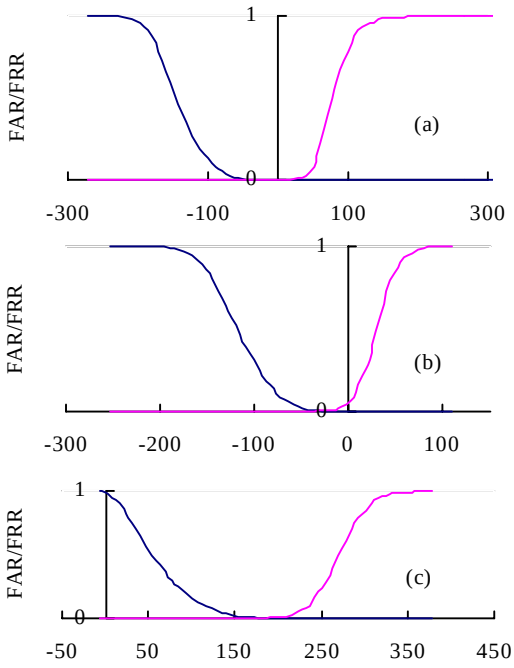

Figure 3. Results of Matsui Method

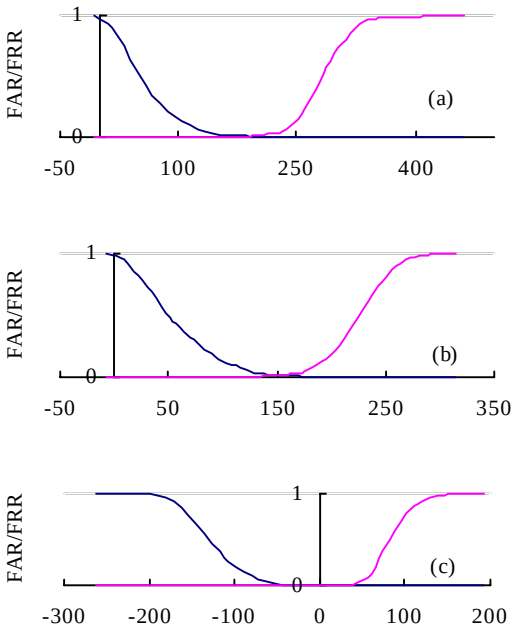

Figure 4. Results of Rosenberg Method 2