A STATISTICAL APPROACH TO PERFORMANCE EVALUATION OF SPEAKER RECOGNITION SYSTEMS

Guillermo Garcia, Thomas Eriksson

Communication System Group
Department of Signals and Systems
Chalmers University of Technology
412 96 Göteborg Sweden

Sung-Kyo Jung,

TECH/SSTP Lab.,
France Telecom R&D
22307 Lannion, France

ABSTRACT

In speaker recognition applications, speaker identification is the process of automatic recognizing who is speaking based on statistical information obtained from speech signals. Considering the limited number of tests in real situations during the classification phase, it is more useful to have an estimator of the probability of error for speaker recognition systems. In this work, we propose a method based on the log-likelihood of each speaker to estimate the probability of error of a speaker recognition system. We assess the performance of the estimator with experimental trials and compare with the actual number of errors. The results show that the performance of our estimator is comparable to the conventional method. The proposed method presents better reliability and fast convergence compared to the counting case. Indeed, we attain an analytical expression for the probability of error that can be used as a gradient for other optimization methods in speaker recognition applications.

Index Terms— Bayes procedures, speaker recognition , gaussian distributions, modeling, estimation.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking based on statistical information provided by speech signals. Speaker recognition can be classified into speaker identification and speaker verification [9]. This work focuses on speaker identification systems, which purpose is to determine from a set of known speakers the best match to an input utterance.

Speaker identification systems can broadly be divided into two phases: design and classification. In the design phase, a set of sample features belonging to a speaker is used to determine a model which best represents the speaker. In this study, the Expectation Maximization (EM) algorithm [13] is used to estimate Gaussian Mixture Models (GMMs) for each speaker. The EM algorithm provides Maximum Likelihood (ML) estimates for the unknown model parameters from the training samples. In the classification phase, the GMMs obtained from the design phase are used to compute the probability that unknown test samples belongs to a given speaker. In speaker identification systems, the performance of the whole system is assessed by the probability of error. It is important to measure the performance of speaker recognition systems in terms of probability of error for both system evaluation and system development. The measurement of the confidence of the probability of error has been studied in several papers e.g.,[6], [5], [2], [12], [4],[1]; wherein statistical approaches and integration of multiple sources of information are used to assess the probability of error.

Statistical estimation theory is a tool that has proven very useful to determine models in speaker recognition and other pattern classification applications [10][7]. There are several papers where algorithms for estimation of parameters are used and analyzed. The quality of the estimator can be statistically measured by computing the variance of the estimator and the convergence to the correct value. The novelty of our approach is to estimate the reliability of the classification system itself (i.e., not of the performance of the system). The advantage is that this measure is more reliable when there are limited number of tests, such as e.g., real situations.

The main contribution of this work is a theoretical study of the connection between the log-likelihood of a set of test samples from a speaker, and the classification error. We determine an analytical expression for an estimator of the classification error for speaker identification systems. Experimental results using the YOHO database [3] were performed to validate that the efficiency of the estimator over the conventional method of counting the errors.

2. STATE-OF-THE-ART SPEAKER RECOGNITION

Most state-of-the-art text-independent speaker recognition system use GMMs to represent a statistical model of each speaker. The GMM for a speaker s is defined as

$$h^{(s)}(\xi) = \sum_{k=1}^{M} w_k^{(s)} g\left(\xi, \mu_k^{(s)}, C_k^{(s)}\right) \tag{1}$$

i.e., $h^{(s)}(\xi)$ is a weighted sum of Gaussian distributions $g(\xi, \mu_k^{(s)}, C_k^{(s)})$, where μ_k is the mean and C_k is the covariance matrix of the k-th Gaussian distribution. Each speaker has a unique model, describing the particular features of his/her voice. In the classification phase, the resulting GMMs are then used to compute the log-likelihood (LL) of a set of samples from an unknown speaker, $\{\xi_t\}_{t=1}^{T}$, with respect to the actual speaker GMMs parameters,

$$LL^{(s)} = \sum_{t=1}^{T} \log h^{(s)}(\xi_t). \tag{2}$$

The speaker with the highest log-likelihood is declared as the one identified,

$$\hat{s} = \arg\max_{s, 1 \le s \le S} LL^{(s)}. \tag{3}$$

The goal of the speaker recognizer is to minimize the probability of error given by $P_e = \Pr[s \neq \hat{s}]$ [8].

3. ESTIMATION OF THE PROBABILITY OF ERROR (P_E)

In this section, we describe the conventional method and our proposed approach to estimate the probability of error for speaker recognition systems. Assuming that a stochastic model (GMM) and a group of tests[1]are available for each speaker, we can compute the log-likelihood (LL) of these tests with respect to the actual GMMs, such that the number of log-likelihood values obtained for each test is equal to the total number of speakers.
Let $\{x_i(n)\}_{n=1}^N$ be a set of N log-likelihood values for the i-th real speaker and $\{y_i(n)\}_{n=1}^N$ be a set of the N log-likelihood of the maximum of other speakers i.e.,

$$y_i(n) = \max(x_1(n), x_2(n)..., x_{i-1}(n), x_{i+1}(n), ...x_S(n)). \quad (4)$$

Letting $\{z_i(n)\}_{n=1}^N$ be a set of the difference between the log-likelihood of the i-th real speaker and the log-likelihood of the maximum of other speakers,

$$z_i(n) = x_i(n) - \max(x_1(n), x_2(n)..., x_{i-1}(n), x_{i+1}, ...x_S(n)). \quad (5)$$

Substituting $y_i(n)$ for the $\max(x_1(n), x_2(n)..., x_{i-1}(n), x_{i+1}, ...x_S(n))$ in (5), we attain

$$z_i(n) = x_i(n) - y_i(n). \quad (6)$$

3.1. Conventional Method

The conventional estimator for the probability of error for the i-th speaker can be described as

$$P_{e_i} = \frac{1}{N}\sum_{n=1}^{N} \Phi\left(y_i(n) - x_i(n)\right), \quad (7)$$

where Φ is the unit step function which detects when $(y_i(n) - x_i(n)) \geq 0$ and N is the number of tests available. The conventional estimator will make an erroneous decision when $y_i(n)$ is larger than $x_i(n)$. The probability of error for the whole system can be defined as

$$P_e = \frac{1}{S}\sum_{i=1}^{S} P_{e_i} = \quad \tfrac{1}{S}\tfrac{1}{N}\sum_{i=1}^{S}\sum_{n=1}^{N} \Phi\left(y_i(n) - x_i(n)\right), \quad (8)$$

where S is the total number of speakers in the system.
It is obvious from (8) that some of the inherent information in $x_i(n)$ and $y_i(n)$ is discarded, since the only thing that affects P_e is the sign of $y_i(n) - x_i(n)$; all the information about the magnitude of the difference is discarded. Therefore, we propose an improved estimator of P_e in the next subsection.

3.2. Proposed method

In our proposed method, we use a statistical approach. Letting X_i and Y_i be random variables, and Z_i be the difference between X_i and Y_i, we can define the cumulative distribution function (cdf) of Z_i as

$$F_{Z_i}(z) = \Pr\left[X_i - Y_i \leq z\right], \quad (9)$$

[1]one test is a set of samples from a unknown speaker $\{\xi_t\}_{t=1}^T$. Each test corresponds to a short spoken sentence (1-2s) by the speaker.

where $F_{Z_i}(z)$ is the cdf for every z in the range from $-\infty$ to ∞. Since we are only interested on the occurrence of errors in the system (i.e., $z \leq 0$), we can define the probability of error for the i-th speaker as

$$P_{e_i} = F_{Z_i}(0). \quad (10)$$

The distribution function in (9) can be represented as a two-dimensional integral [11]. Substituting the limits obtained from (9), we attain

$$F_{Z_i}(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z+y} f_{X_iY_i}(x,y)\partial x\partial y, \quad (11)$$

where $f_{X_iY_i}(x,y)$ is a jointly probability distribution function (pdf). Rewriting $f_{X_iY_i}(x,y)$ in terms of conditional density functions, we obtain

$$F_{Z_i}(z) \quad = \quad \int_{-\infty}^{\infty} f_{Y_i}(y)\int_{-\infty}^{z+y} f_{X_i|Y_i}(x|y)\partial x\partial y \quad (12)$$

i.e., $f_{X_i|Y_i}(x|y)$ is the conditional pdf of X_i given a value of Y_i.
Rewriting the inner integral in (12) in terms of cdf, we get

$$F_{Z_i}(z) = \int_{-\infty}^{\infty} f_{Y_i}(y)F_{X_i|Y_i}(z+y|y)\partial y \quad (13)$$

where $F_{X_i|Y_i}(x|y)$ is a conditional cumulative distribution of X_i given a value of Y_i. From (10) and (13), we attain that the probability of error for the i-th speaker is

$$\hat{P}_{e_i} = F_{Z_i}(0) = \int_{-\infty}^{\infty} f_{Y_i}(y)F_{X_i|Y_i}(y|y)\partial y. \quad (14)$$

The integral in (14) is an expectation, and we can write it as

$$\hat{P}_{e_i} = E_{Y_i}[F_{X_i|Y_i}(Y_i|Y_i)], \quad (15)$$

where E_{Y_i} is an expectation over all values of $\mathbf{y}_i$. The probability of error for the i-th speaker can be approximated by the sample mean

$$\hat{P}_{e_i} = \frac{1}{N}\sum_{n=1}^{N} F_{X_i|Y_i}(y_i(n)|y_i(n)), \quad (16)$$

in which as in (8), N is the number of tests available. The probability of error for the whole system can be derived in a similar way as the conventional method,

$$\hat{P}_e \quad = \quad \frac{1}{S}\sum_{i=1}^{S}\hat{P}_{e_i}, \quad (17)$$

$$\quad = \quad \frac{1}{S}\frac{1}{N}\sum_{i=1}^{S}\sum_{n=1}^{N} F_{X_i|Y_i}(y_i(n)|y_i(n)) \quad (18)$$

where S is the total number of speakers. We will approximate $F_{X_i|Y_i}(x|y)$ as the cdf of a conditional normal distribution with mean $\mu_{X_i|\mathbf{Y}_i}(y_i)$ and variance $\sigma^2_{X_i|Y_i}$, which are defined as

$$\mu_{X_i|Y_i}(y_i) \quad = \quad \mu_{X_i} + \frac{C_{X_i,Y_i}}{\sigma^2_{Y_i}}\left(y_i(n) - \mu_{Y_i}\right), \quad (19)$$

$$\sigma^2_{X_i|Y_i} \quad = \quad \sigma^2_{X_i} - \frac{C^2_{X_i,Y_i}}{\sigma^2_{Y_i}} \quad (20)$$

$$\quad = \quad \sigma^2_{X_i}(1 - \rho^2_{X_iY_i}). \quad (21)$$

where

$$\mu_{X_i} = \frac{1}{N}\sum_{n=1}^{N} x_i(n), \quad \mu_{Y_i} = \frac{1}{N}\sum_{n=1}^{N} y_i(n), \tag{22}$$

$$\sigma^2_{X_i} = \frac{1}{N-1}\sum_{n}^{N-1}\left(x_i(n) - \mu_{X_i}\right)^2, \tag{23}$$

$$\sigma^2_{Y_i} = \frac{1}{N-1}\sum_{n}^{N-1}\left(y_i(n) - \mu_{Y_i}\right)^2, \tag{24}$$

$$C_{X_i,Y_i} = \frac{1}{N-1}\sum_{n}^{N-1}\left(x_i(n) - \mu_{X_i}\right)\left(y_i(n) - \mu_{Y_i}\right). \tag{25}$$

i.e., μ_{X_i} and μ_{Y_i} are the sample means, $\sigma^2_{X_i}$ and $\sigma^2_{Y_i}$ are the variances and C_{X_i,Y_i} is the covariance of a set of the log-likelihood values for the i-th real speaker (X_i), and a set of the log-likelihood of the maximum of other speakers (Y_i). Finally in (21), $\rho_{X_iY_i}$ is the correlation coefficient defined as $\rho_{X_iY_i} = \dfrac{\mathrm{C}_{X_i,Y_i}}{\sigma_{X_i}\sigma_{Y_i}}$ which satisfies the condition $|\rho_{X_iY_i}| < 1$.
Then, $F_{X_i|Y_i}(x|y)$ is described as

$$F_{X_i|Y_i}(x|y) = \frac{1}{\sqrt{2\pi\sigma^2_{X_i|Y_i}}}\int_{-\infty}^{x}\exp\left(-\frac{\left(r-\mu_{X_i|Y_i}(y_i)\right)^2}{2\sigma^2_{X_i|Y_i}}\right)\partial r, \tag{26}$$

$$= Q\left(\frac{x-\mu_{X_i|Y_i}(y_i)}{\sigma_{X_i|Y_i}}\right). \tag{27}$$

Substituting (27) into (18), give us the estimator for the probability of error

$$\hat{P}_e \approx \frac{1}{S}\frac{1}{N}\sum_{i=1}^{S}\sum_{n=1}^{N} Q\left(\frac{y_i(n)-\mu_{X_i|Y_i}(y_i(n))}{\sigma_{X_i|Y_i}}\right) \tag{28}$$

Comparing (28) with (8), we notice certain similarities between the estimators. In both cases, the probability of error is defined as an average over the number of speakers and the number of tests of a function of $y_i(n)$. However in our estimator, the statistics of $y_i(n)$ and $x_i(n)$ represented as $\mu_{X_i|Y_i}(y_i)$ and $\sigma_{X_i|Y_i}$, provide inherent information that is not used in the conventional method.

4. EXPERIMENTAL EVALUATION

In this section, we compare the performance of the proposed estimator of P_e with the conventional method in a speaker recognition system.

4.1. Database Description

The experiments were conducted using the YOHO database [3]. The first session of each of the 137 speakers in the enrollment sessions has been used for training. Each speech file, after removing silence at the beginning and end, was segmented into frames of 25 ms length with an overlap of 10 ms. Each frame was pre-emphasized and Hamming windowed. Then, a 12-th order MFCC were created and used to train an 8-mixture GMM. The evaluation database was created from the 3 remaining enrollment sessions and the verify session in a similar way to the one used for training, obtaining 112 tests ($N = 112$) for each speaker. From the computation of the log-likelihood of each test, we obtain 137 log-likelihood values, one for each speaker. We assume that there is only one log-likelihood value that corresponds to the i-th real speaker and one maximum log-likelihood value selected from the other 136 other speakers.

4.2. Experiments Implementation

In this section, we describe the implementation of the estimator for the probability of error. Table 1 shows the algorithm used.

Step 1	Using the evaluation database, Compute μ_{X_i} and μ_{Y_i} from (22). Compute $\sigma^2_{X_i}$ and $\sigma^2_{Y_i}$ from (24). Compute C_{X_i,Y_i} from (25).
Step 2	Using the values in step 1, Compute $\mu_{X_i\|Y_i}$ from (19). Compute $\sigma^2_{X_i\|Y_i}$ from (21).
Step 3	Using the evaluation database and the values in step 2, Compute $\hat{P}_e = \frac{1}{S}\frac{1}{N}\sum_{i=1}^{S}\sum_{n=1}^{N} Q\left(\frac{y_i(n)-\mu_{X_i\|Y_i}(y_i(n))}{\sigma_{X_i\|Y_i}}\right)$.

Table 1. Algorithm used for estimation of P_e.

4.3. Experimental Results

The performance of the estimator was compared against conventional. To obtain a more reliable performance of the estimator in a statistical way, we sort our sequence of 112 tests in 50,000 random ways. Figure 1 shows a comparison between the standard deviation of our estimator and the conventional method over 50,000 realizations of our sequence of tests. We can observe that the standard deviation of the estimator is lower compared to the standard deviation of the conventional method regardless of the number of tests, indicating that the stability of the estimator is better than the conventional case. Figure 2 shows the estimated probability of error for the conventional method as a function of the number of tests. The different realizations shown are drawn from our sequence of 112 tests. Figure 3 shows the same approach, but for our estimator. Comparing both figures, we observe that our estimator converges faster and has a lower variance for small number of tests than the conventional method. Figure 4 shows a comparison between our estimator and the conventional method for a single realization. It shows that our method only requires 10 tests to estimate with high accuracy the probability of error for a sequence of 112 tests.

5. CONCLUSIONS

Being able to assess the performance (i.e., the probability of error) over a long term is of great importance for the speaker identification system optimization. Because of the limited number of tests available in real situations, measuring the performance is not a trivial task. The main contribution of this paper is an estimator based on the log-likelihood values from a set of speaker test samples. Our estimator was able to overcome the limitation of the number of tests available, it achieves convergence on a small number of tests (5 tests) and has

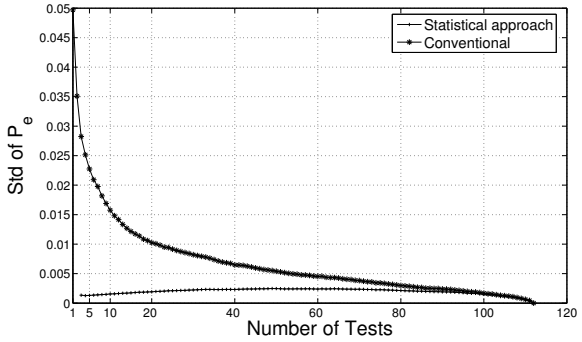

Fig. 1. Comparison of standard deviations between our estimator for the probability of error and the conventional method.

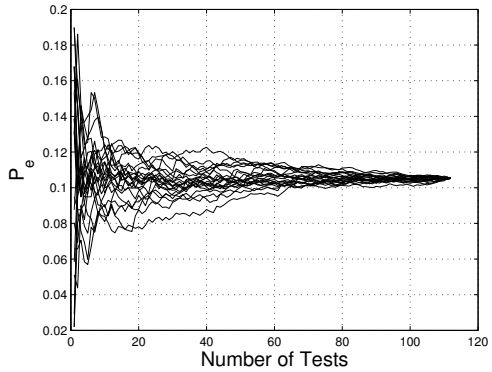

Fig. 2. Probability of error as a function of the number of tests for different realizations using the conventional method.

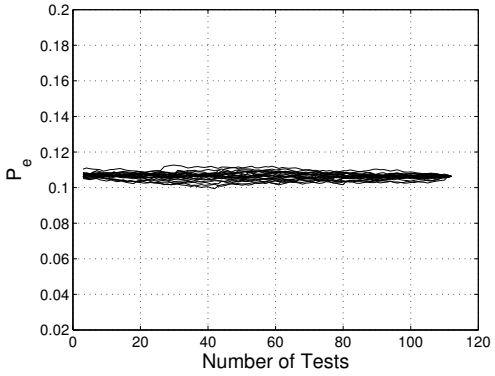

Fig. 3. Probability of error as a function of the number of tests for different realizations using our statistical approach.

lower variance regardless of the number tests compared to the conventional method. Moreover, the analytical expression obtained can foster new paths for developing optimization methods to improve the probability of error and better feature extractors for speaker recognition.

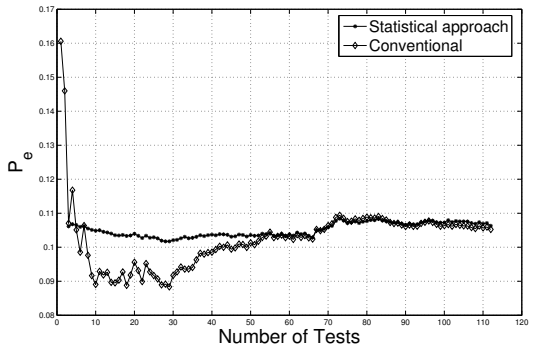

Fig. 4. Comparison between our estimator for probability of error and the conventional method for a single realization.

6. REFERENCES

[1] F. Botti, A. Alexander, and A. Drygajlo, "An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data," *citeseer.ist.psu.edu/botti04interpretation.html*, 2004.

[2] N. Brummer and J. d. Preez, "Application-independent Evaluation of Speaker Detection," in *Proceedings CSL*, vol. 20, 2006, pp. 230–275.

[3] J. Campbell, "Testing with YOHO CD-ROM voice verification corpus," in *Proceedings ICASSP*, 1995, pp. 341–344.

[4] W. Campbell, D. Reynolds, J. P. Campbell, and K. J. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *Proceedings ICASSP*, vol. 1, 2005, pp. 717–720.

[5] C. Cortes and M. Mohri, "AUC Optimization vs. Error Rate Minimization," in *Proceedings NIP*, 2004.

[6] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "SHEEPS, GOATS,LAMBS and WOLVES: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," in *Procceedings ICSLP*, 1998.

[7] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience Publication, 2001.

[8] T. Eriksson, S. Kim, H.-G. Kang, and C. Lee, "An information-theoretic perspective on feature selection in speaker recognition," *IEEE Signal Processing Letters*, vol. 12, no. 7, pp. 500 – 503, 2005.

[9] S. Furui, "Recent advances in speaker recognition," in *Proceedings ICASSP*, vol. 1, 1989, pp. 429–440.

[10] S. M. Kay, *Fundamentals of Statistical Signal Processing, Estimation Theory*, 2nd ed., ser. Signal Processing. Prentice Hall, 1993.

[11] A. Papoulis and S. U. Pillai, *Probability , Random Variables and Stochastic Processes*, 4th ed. Mac Graw Hill, 2002.

[12] J. Richiardi, P. Prodanov, and A. Drygajlo, "Speaker verification with confidence and reliability measures," in *Proceedings ICASSP*, vol. 1, 2006, pp. 641–644.

[13] Y. Zhang, M. Alder, and R. Togneri, "Using Gaussian mixture modeling in speech recognition," in *Proceedings ICASSP*, vol. 1, 1994, pp. 613–616.