SPEAKER VERIFICATION PERFORMANCE AND THE LENGTH OF TEST SENTENCE

Jialong He, Li Liu

Dept. of Speech & Hearing Science
Arizona State University
Tempe, AZ 85287-1908, USA

ABSTRACT

It is known that the performance of a speaker verification system improves with the length of test sentences. However, little is known about the exact relation between the performance and the test length. That makes it difficult to compare the results from various studies in which different test lengths have been used to evaluate the systems. In this paper, we have proposed a method to calculate the verification error rates at any lengths of test sentences, as long as the error rates at two different lengths are given. The accuracy of this calculation method is demonstrated with a speaker verification experiment and with the results reported in literature. Good agreement is shown between the calculated values and that measured through experiments.

1. INTRODUCTION

The performance of a speaker verification system depends on many factors like the speaker model used, the available data, and the way how it is trained and tested. Usually, it is not easy to compare the performance of different systems. Even for systems that are based on the same kind of speaker models and have been evaluated with the same speech database, the reported error rates could be quite different. For example, comparing the results reported in [1] and [2], both of them used the Gaussian mixture speaker model (GMM) and the two systems were evaluated with the same YOHO database. However, different verification performances were reported. This can be attributed to the fact that in Reynolds' experiment he concatenated four digit strings as a single test sentence, while Liu *et al*. used single string as a test sentence.

To deal with such kind of comparison issue, we propose the following method which can be used to estimate the verification performance of a system at any test lengths if its performance at two different lengths is known. The accuracy of this estimation method is demonstrated by a speaker verification experiment and with the results taken from the literature.

2. ESTIMATION PROCEDURE

In this section, a set of calculation equations will be derived. This calculation is based on two assumptions. First, we assume that the sentence scores have a normal distribution. In general, this assumption is reasonable insofar as there are enough data, thus the central limit theorem can be applied. The exact meaning of the scores could be different. For example, in a VQ based speaker verification system [3], the sentence score is the average distance of test vectors to their nearest codes. While in the GMM or HMM based speaker verification systems, the sentence score is the average log-likelihood or the average log-likelihood ratio [2][4]. Another assumption is that the decision threshold does not change with different lengths of test sentences.

2.1 Mean and Variance of Scores

Given a test sentence from a speaker whose identity is to be verified, a vector sequence, $X = \{x_1, x_2, \cdots, x_T\}$, can be derived. Each vector x_i is obtained from one short segment of the speech signal by using short-time analysis techniques. Suppose the frame scores are $s(x_i)$, the average of all frame scores from this sentence is used as the sentence score,

$$S(X) = \frac{1}{T}\sum_{i=1}^{T} s(x_i) \tag{1}$$

Let μ and σ^2 be the mean and the variance of the frame score,

$$E[s(x)] = \mu, \quad Var[s(x)] = \sigma^2 \tag{2}$$

Then the mean and the variance of the sentence score are given by

$$E[S(X)] = E\left[\frac{1}{T}\sum_{i=1}^{T} s(x_i)\right] = \mu \tag{3}$$

$$Var[S(X)] = E[S(X)^2] - E^2[S(X)] \quad (4)$$

Put Eq. (1) into (4), we have

$$\begin{aligned} Var[S(X)] &= E\left[\left(\frac{1}{T}\sum_{i=1}^{T} s(x_i)\right)^2\right] - \mu^2 \\ &= \frac{1}{T^2}\sum_{i=1}^{T}\left(E[s^2(x_i)] - \mu^2\right) + C(x_i, x_j) \quad (5) \\ &= \frac{\sigma^2}{T} + C(x_i, x_j) \end{aligned}$$

where $C(x_i, x_j)$ is defined as

$$C(x_i, x_j) = \frac{1}{T^2}\sum_{i=1}^{T}\sum_{\substack{j=1 \\ j\neq i}}^{T}\left(E[s(x_i)s(x_j)] - \mu^2\right) \quad (6)$$

It is easy to see that the mean value of the sentence score equals to the mean value of the frame score, but the variance of the sentence score depends on the length of vector sequences and $C(x_i, x_j)$. The value of $C(x_i, x_j)$ is related to the degree of correlation among the frame scores.

In the case that the frame scores are uncorrelated, we have

$$E[s(x_i)s(x_j)] = E[s(x_i)]E[s(x_j)] \quad (7)$$

then $C(x_i, x_j) = 0$, in consequence, $Var[S(X)] = \sigma^2 / T$. That is, the variance of $S(X)$ is inversely proportional to the length of vector sequences.

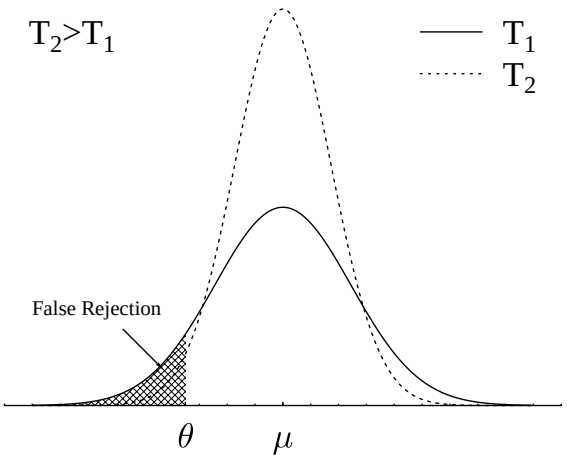

Figure 1: the variance of sentence score decreases with the length of test sentences.

Figure 1 illustrates how the distribution of sentence scores shrinks with the length of test sentences. Obviously, for the fixed decision threshold the verification error becomes smaller for longer test sentences.

Another extreme case is that the frame scores are completely correlated. In such situation,

$$E[s(x_i)s(x_j)] = E[s^2(x_i)] \quad (8)$$

by putting Eq. (8) into (5), we have

$$Var[S(X)] = \sigma^2 \quad (9)$$

It is seen that the variance of the sentence score equals to the variance of the frame score. Since both the mean and the variance of $S(X)$ does not change with the length of test sentences, the verification performance can not be improved by adding more completely correlated test vectors.

In practical situations, the frame scores are neither completely correlated nor ideally uncorrelated. Therefore, the speaker verification performance usually improves with the length of test sentences, but the improving rate is slower than the ideal uncorrelated situation.

To estimate the error rate for a given length, we need to know the variance of $S(X)$ at that length. Because it is difficult to obtain the value of $C(x_i, x_j)$, we have to find an approximation to $Var[S(X)]$.

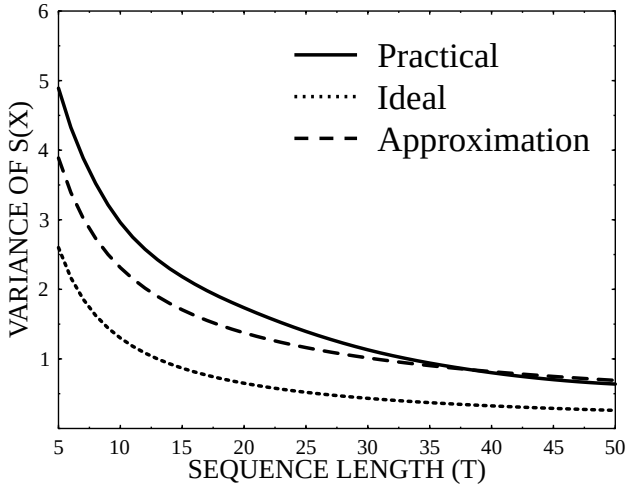

Figure 2: Variance of the sentence score with the length of test sentences in frames. The dashed curve is an approximation to the practical curve.

As an example, Figure 2 shows how the variance of the sentence score decreases with the length of vector sequences. The solid line was obtained with the speech data from the YOHO database. For comparison, the ideal uncorrelated case is also shown in dotted line.

Through a number of trials, we found the following equation can be used to approximate the variance of the sentence score at a given length of *T*,

$$\sigma_T^2 = \frac{\sigma^2}{T^\alpha} \tag{10}$$

where σ^2 is the variance of the frame score, α is a constant between 0 and 1. The choice for α depends on the speech materials. We see that α=1 corresponds to the ideal uncorrelated situation, while α=0 is another extreme case that the frame scores are completely correlated. However, we do not need to specify a value for α in advance. As to be described in the next section, its value can be determined from the known error rates. The dashed line in Figure 2 comes from the Eq. (10) with α=0.7.

2.2 Error Rate vs. Sequence Length

Because the sentence scores are assumed to have a normal distribution, from Figure 1, it is easy to see that the false rejection rate (*FR*) can be calculated with

$$FR = \frac{1}{2}\left[1 - erf\left(\frac{\mu - \theta}{\sqrt{2}\sigma_T}\right)\right] \tag{11}$$

where θ is the decision threshold and $erf(\cdot)$ is the error function defined as

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt \tag{12}$$

put Eq. (10) into (11) and define β as $\frac{\mu - \theta}{\sqrt{2}\sigma}$, we have

$$FR = \frac{1}{2}\left[1 - erf\left(T^{\alpha/2}\beta\right)\right] \tag{13}$$

There are two unknown parameters, α and β, in Eq. (13). In order to calculate the *FR* for a given length, we have to know the error rates at two different lengths. Suppose the error rates at length T_1 and T_2 are FR_1 and FR_2, respectively, we obtain the following two equations

$$\begin{aligned} T_1^{\alpha/2}\beta &= erfinv\left(1 - 2FR_1\right) \\ T_2^{\alpha/2}\beta &= erfinv\left(1 - 2FR_2\right) \end{aligned} \tag{14}$$

where $erfinv(\cdot)$ is the inverse function of $erf(\cdot)$. By solving these two equations, we have

$$\begin{aligned} \alpha &= \frac{2\log\left(erfinv\left(1-2FR_1\right) / erfinv\left(1-2FR_2\right)\right)}{\log\left(T_1 / T_2\right)} \\ \beta &= \frac{erfinv\left(1-2FR_1\right)}{T_1^{\alpha/2}} \end{aligned} \tag{15}$$

By putting α and β into Eq. (13), we can calculate the *FR* at any lengths. The calculation procedure is summarized as follows,

1. From the known error rates at two different lengths, calculate α and β using Eq. (15). The value of the inverse error function $erfinv(\cdot)$ can be obtained by table-looking or using a utility such as MATLAB.
2. The *FR* for the test sentences with *T* vectors is estimated with Eq. (13). Again, the value of the error function $erf(\cdot)$ can be found by table-looking or using a utility.

In the above discussion, we have considered only the false rejection rate. For the false acceptance rate (*FA*), if we define β as $(\theta - \mu)/\sqrt{2}\sigma$, the final calculation equations will be the same as Eq. (13). By knowing all *FR* and *FA*, it is easy to obtain the equal error rate (*EER*).

It should be noted that when deriving Eq. (13), we implicitly assumed that the known error rates are less than 50%. If the error rate is larger than 0.5, Eq. (13) should be changed to

$$FR = \frac{1}{2}\left[1 + erf\left(T^{\alpha/2}\beta\right)\right] \tag{16}$$

and the corresponding item $(1-2FR)$ in Eq. (15) should be replaced with $(2FR-1)$.

Finally, in Eq. (13), *T* is assumed to be the number of feature vectors. However, we can use any time unit to specify the length of sentences, such as second or millisecond. By putting the relation $T = \kappa\tau$ into Eq. (13) and (15), where τ is the time unit defined and *k* is a conversion coefficient, it can be shown that *k* will vanish from the final calculation formulas.

3. ACCURACY OF CALCULATION

3.1 Speaker Verification Experiment

We did a speaker verification experiment to demonstrate the accuracy of the proposed calculation method. The

speech signals were taken from the YOHO database, 40 speakers (20 males, 20 females) were used in this experiment. The speaker model was the Gaussian mixture model (GMM) with 16 mixture components. Each feature vector was composed of 16 MFCC coefficients. The analysis window size was 32 ms with 16 ms overlap. To show how the error rate changes with the length of test sentences, we first concatenated all 40 digit strings from the verification sessions into a long sentence and then cut it into segments of specified length. Each segment was used as one test sentence.

Figure 3 plots the false rejection rate as a function of the test length in frames. Since the advance step between the successive frames is 16 ms, the actual length of test sentences is 16×*T* ms. The curve with the label "measured" was obtained from the verification experiment. The calculated values were obtained by assuming the error rates at 50 and 300 frames are known. We see that the calculated values are comfortingly close to the measured ones. The maximum relative estimation error is about 3.4%. Because the estimation result for the *FA* is quite similar to that for the *FR*, we do not repeat it here.

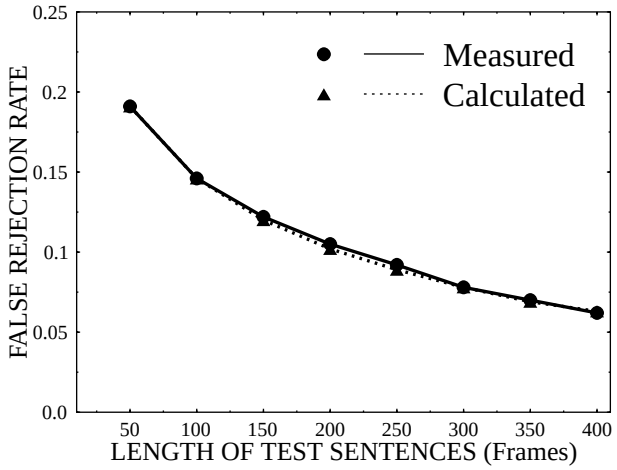

Figure 3: The false rejection rate obtained from the verification experiment and calculated by the proposed method.

3.2 Verifying Published Results

To show that the proposed method is also applicable to other speaker models as well as to other speech databases, we verify the results published in literature. The data shown in the second column of Table 1 was taken from the paper by Tishby [4]. In his experiment, he used AR hidden Markov models. The third column in Table 1 was calculated by assuming that the two points at 3 and 7 are known. We see that the calculated values are very close to that measured through experiments.

Length (digits)	Experimental EER (%)	Calculated EER (%)
1	10.5	9.3
2	5.6	5.6
3	3.8	3.8
4	2.8	2.8
5	2.0	2.1
6	1.5	1.6
7	1.3	1.3
8	1.1	1.1
9	0.9	0.9
10	0.8	0.7

Table 1: Comparing the EER reported in the paper by Tishby [4] and estimated with the proposed calculation method.

4. SUMMARY

In this paper, we have proposed a method to estimate the verification performance (false rejection rate, false acceptance rate and equal error rate) of a speaker verification system with the length of test sentences. By using this method, we can derive the verification error rate at any lengths if the error rates at two different lengths are known. On the other hand, we can also know how long the test sentences should be for a system to reach a specified error rate. Because the correlation among the frame scores has been taken into account, the calculated values are very close to that directly measured through experiments. Therefore, we suggest that when assessing a verification system the performance at two different lengths should be tested in order to characterize how the performance changes with the test length.

5. REFERENCES

[1] Liu L., He J. and Palm G. "A comparison of human and machine in speaker recognition," *Proceedings of Eurospeech Conference*, Vol. 5, pp. 2327-2330, Rhodes, Greece, 1997.

[2] Reynolds D., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, Vol. 17, pp. 91-108, 1995.

[3] Rosenberg A. E. and Soong F. K., "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Computer Speech and Language*, Vol. 22, pp. 143-157, 1987.

[4] Tishby N. Z., "On the application of mixture AR hidden Markov models to text independent speaker recognition," IEEE Trans. Signal Processing, Vol. 39, pp. 563-570, 1991.