ESTIMATION OF FALSE ACCEPTANCE RATE IN SPEAKER VERIFICATION

Shuping Ran, William Laverty, Bruce Millar, Iain Macleod and Michael Wagner

TRUST Project
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT

In this paper, we show how the false acceptance rate depends on inter-speaker distances and propose a novel technique for the estimation of False Acceptance Rate (FAR) in speaker verification. The FAR estimate is based on the statistical technique of "bootstrapping". First, a pairwise FAR is obtained as a function of pairwise inter-speaker distances between each client and impostor pair. Second the inter-speaker distance distribution is estimated. The FAR is calculated by combining these two.

INTRODUCTION

Speaker verification systems are evaluated by estimating both the False-Rejection (or type-I) error Rate (FRR) and the False-Acceptance (or type-II) error Rate (FAR). Traditionally, these error rates are expressed as a function of a threshold which is used in a verification system to determine the acceptance or rejection of unknown speakers. The incoming speech signal from an unknown speaker is checked against his/her claimed speaker identity's model. The result is compared with the threshold to decide whether to accept or reject the speaker. The threshold can be a distortion measurement, where Vector Quantisation (VQ) is used to model the speakers, or likelihood, where Hidden Markov Models (HMM) are used. Figure 1 illustrates the FRR and FAR as a function of such a threshold.

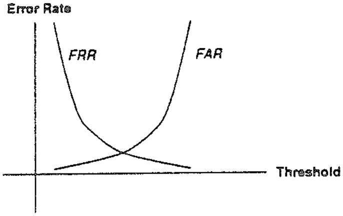

Figure 1: False rejection and false acceptance rates as a function of system threshold.

The FRR can be determined for any speaker (or "client") by analysing data collected from that speaker, ideally over a time span of weeks, months or even years. The FAR, on the other hand, can only be determined from data collected from a sample of other speakers (often referred to as "impostors"). Ideally, this sample would be representative of the population of potential impostors and should therefore normally comprise data from a large number of speakers. Because of the difficulties of collecting data covering a representative population, many previous studies have based their FAR estimation on the same "client" population (e.g. Rosenberg et. al., 1992; Matsui and Furui, 1992).

There are some shortcomings of this approach. First, the "client" population may not represent the general population or the population which intend to break into the system. The inter-speaker distance between the clients and the impostors may be very large, which will lead to a very low FAR, or may be very small, which will lead to a very high FAR. Therefore, the FAR calculated this way can not be interpreted easily in terms of the system performance in the real world, or of how the system compares with other systems which are evaluated using different populations of speakers. Second, the FAR calculated this way is an average over all the speakers disregarding the distances between clients and impostors. In reality, a pair of speakers with large inter-speaker distance will not even try to mimic each other. For example, a female speaker with a high F_0 would not try to claim to be a male speaker whose F_0 is very low, if she thought F_0 was a factor.

In this paper, we will show how the FAR depends on inter-speaker distances and we propose a novel technique for the estimation of FAR in speaker verification. This estimation provides an realistic measurement of FAR.

METHOD

The statistical properties of any decision procedure depend on the distributional properties of the data collected and ultimately on the distributional properties of the statistics used in the decision procedure. The False Acceptance Rate (FAR) and False Rejection Rate (FRR) of a speaker verification system based for example on likelihoods from a Hidden Markov Model, will depend on the distributional properties of speech. In some instances, theoretical analysis is primarily used to determine these properties. In other instances this is difficult. The ***Bootstrap Method*** (Efron, 1979) substitutes considerable amounts of computation in place of the theoretical analysis. The distribution of the data is approximated by:

- its empirical distribution as observed in the sample (Non-parametric Bootstrap method); or
- a "smooth estimate" of its distribution (such as the Kernel density estimate (Silverman, 1986)) (Smoothed non-parametric Bootstrap method); or
- a parametric estimate of the distribution of the data (Parametric Bootstrap method).

The statistical properties (in this case the FAR and FRR) of the decision procedure in question are then determined either by direct computation or by Monte Carlo simulation from this distribution. The ***Bootstrapping*** technique can be used to estimate the FRR for any client and FAR for any impostor of any client and is the basis of the procedure described below.

First, we need to find the distribution of inter-speaker distance (δ) in the test population and estimate the distribution function ($f(\delta)$). Second, we need to calculate the FAR for each client-impostor pair and estimate the false acceptance rate as a function of inter-speaker distance ($FAR(\delta)$). The overall false acceptance is then calculated as the integral of the product of $FAR(\delta) * f(\delta)$ over δ. We illustrate these ideas by some experiments described below.

EXPERIMENT

An accurate estimation of the overall FAR, should be obtained if a database of a representative impostor population was available. A representative population is one with a free range of inter-speaker distances. Because of the limitation of resources for collecting a large database to achieve representative status, we start with a small database to evaluate the approach.

Acoustic data from 24 Australian English speakers (12 male, 12 female) were used. The utterances, typical of speaker commands to a computer, vary in length from 0.3 to 2.5 seconds. Data from two recording sessions recorded approximately one week apart were used. There were five repetitions of each of 30 utterances per speaker in each session.

These data were digitised at 20000 samples per second and down-sampled to 10000 samples per second, after bandlimiting to 60 – 4800Hz. 20 mel-frequency cepstral coefficients (MFCCs) were computed for each frame of 25.6 ms with 15.6 ms overlap between adjacent frames (Millar et. al., 1994).

The following describes the experimental procedure.

- Each speaker was modeled by using a Continuous Ergodic Hidden Markov Model (CEHMM) of eight states and eight mixtures per state as described in Ran et. al. (1994). The models were trained using five repetitions of 30 utterances from the first session.

- Each of the 24 speakers was chosen one at a time, to be the client and the remaining 23 speakers formed the impostors. The test data comprised 150 utterances (5 repetitions of 30 utterances) from the second session from each of the 24 speakers. Average likelihoods of the test data from each impostor against client's model were calculated. This process was repeated with each of the 24 speakers acting as client.
- The average likehoods of each client speaker's data testing against their own models were also obtained.
- For demonstration purposes, the threshold which gave an averaged equal error rate (10% in this case) of FAR and FRR for verification was chosen.
- The FAR of each pair of speakers was calculated, that is the FAR of each of the 24 speakers to be accepted as each of the other 23 speakers. Therefore, there were $24 * 23 = 552$ FARs (referred to as "pairwise FAR" below).
- The inter-speaker distance between each pair of speakers in the database was calculated (referred to as "pairwise inter-speaker distance" below). This calculation involves the distance between two speakers' HMM models. This distance was obtained according to the following formula:

$$\delta_{A,B} = \{[(lnLikelihood(A's\ data|A's\ model) - lnLikelihood(B's\ data|A's\ model)]+$$

$$[lnLikelihood(B's\ data|B's\ model) - lnLikelihood(A's\ data|B's\ model)]\}/2$$

where $\delta_{A,B}$ represents the inter-speaker distance between speaker A and speaker B.

This is a slightly modified version of the distance measure of Juang and Rabiner (1984). They used data randomly generated from the models instead of real data.

RESULTS

Figure 2 shows pairwise FAR vs. pairwise inter-speaker distance (δ). The $FAR(\delta)$ function was estimated from this data by using the "*Gompertz model for growth*" (Draper and Smith, 1981).

$$FAR(\delta) = 1 - \alpha exp(-\beta e^{-k\delta})$$

The parameters α, β and k were estimated using the *Statistica* software package. The resulting estimated function was:

$$FAR(\delta) = 1 - 1.001198exp(-3.82984e^{-0.399747*\delta})$$

Figure 3 presents the estimated FAR function (the continuous curve) and the data represented by using the median (represented by boxes). The middle line inside each box represents the median, the bottom line represents the 25th percentile of the population and the top line represents the 75th percentile. It should be noted that $FAR(\delta)$ is a function of inter-speaker distance (δ) specifically in this case, for the threshold which gave an averaged equal error rate. Accuracy in modelling of this curve could be improved by considering other nonlinear functions and by including the data of the client versus himself/herself.

Figure 4 shows the inter-speaker distance distribution. The function $f(\delta)$ was estimated by (i) a Kernel density estimate (Silverman, 1986), (ii) a mixture of two normal distributions and (iii) a mixture of four normal distributions. Figure 5 shows the estimated distribution of speaker distances.

For the overall FAR calculation, we combined $FAR(\delta, \theta)$ and $f(\delta)$ (θ: threshold) as shown in Figure 6. The FAR for a given distance is the shared area below both curves from zero up to a given distance. That is

$$FAR = \int_0^{dist} FA(\delta)f(\delta)d\delta$$

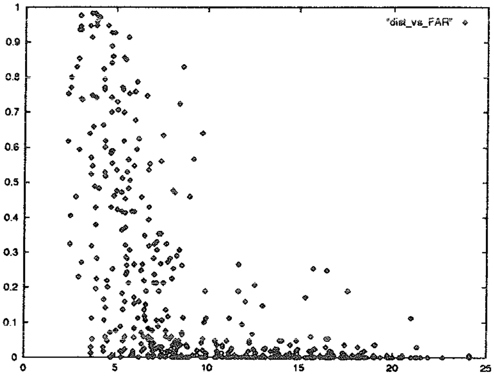

Figure 2: Pairwise false acceptance rate vs. inter-speaker distance.

where $dist$ represents a given distance; for the overall FAR it should be ∞. δ represents the variable distance.

The overall FAR for the entire population, using Simpson's rule for numerical integration of above formula, was calculated to be 17.64% for the kernel estimate of $f(\delta)$ and 32.06% for the first component of the mixture of two normals as a model of impostor distance. The later would be more realistic, assuming that only impostors in the first distribution would be likely to try to impersonate the clients.

DISCUSSION AND CONCLUSION

The overall FAR calculated using the proposed method (32.06%) is higher than the FAR calculated using the other method (10.0%). This was expected, because the FAR calculated using the proposed method covers whole population with inter-speaker distance between 0 and 25, and it covers a continuous inter-speaker distance space; whereas the FAR calculated by using the other method is an average of the FARs from samples of the test population and is discrete.

It is clear from the above description that the FAR depends on the maximum inter-speaker distance between client and impostor at which it may be expected that impostors would try to break into the system. Impostor with smaller distances from the actual clients are more likely to try to break in; the speakers with greater distances are less likely to try to break in.

It is clear that $FAR(\delta)$ function also depends on the threshold chosen for the system. We could have a family of $FAR(\delta)$ functions arising from giving different thresholds (θ) as shown in Figure 7. We therefore denote the false acceptance rate as a function: $FAR(\delta, \theta)$.

In order to evaluate a system for cross comparison purposes, data from a representative population is needed for inter-speaker distance distribution estimation. In practice, data can be collected from the populations with the high likehood of potential impostors to estimate the $f(\delta)$ and $FAR(\delta, \theta)$ functions . The overall FAR and the FAR for a given distance can then be calculated. The calculated FAR covers the whole population whose distance is less than the given inter-speaker distance, rather than averaged over the sample population which is often obtained by using other methods.

ACKNOWLEDGEMENT

This research has been carried out on behalf of the Harry Triguboff AM Research Syndicate.

REFERENCES

Draper, N. R. and Smith, H. (1981), ***Applied Regression Analysis***, John Wiley & Sons, New York.

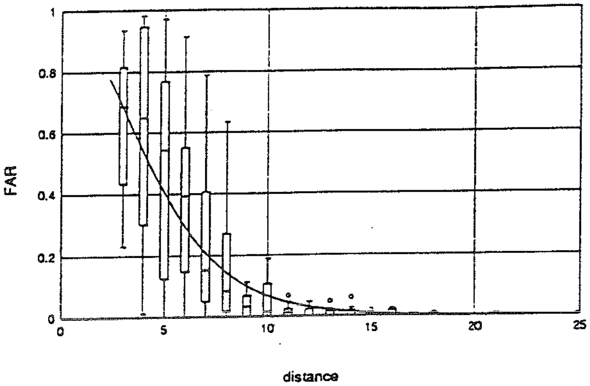

Figure 3: Statistical summary of pairwise false acceptance rate and estimated FAR function vs. inter-speaker distance.

Efron, B. (1979), "Bootstrap methods: another look at the jacknife", ***Ann. Statist.***, vol. 7, pp. 1-26.

Juang, B.-H. and Rabiner, L. R. (1984), "A probabilistic distance measure for hidden Markov Models:, ***AT&T Technical Journal***, vol. 64, No. 2, pp. 391-408.

Matsui, T. and Furui, S. (1992), "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", ***ICASSP-92***, vol. 2, pp. 157-160.

Millar, B., Chen, F., Macleod, I., Ran, S., Tang, H., Wagner, M., and Zhu, X. (1994), "Overview of speaker verification studies towards technology for robust user-conscious secure transactions", ***Proc. of SST-94***.

Ran, S., Millar, B., Laverty, W., Macleod, I., Wagner, M. and Zhu, X. (1994), "Speaker recognition using continuous ergodic HMMs", ***Proc. of SST-94***.

Rosenberg, A. E., DeLong, J., Lee, C.-H., Juang, B.-H., and Soong, F. K. (1992), "The use of cohort normalized scores for speaker verification", ***ICSLP-92***, pp. 599-602.

Silverman, B.W. (1986), ***Density estimation for statistics and data analysis***, Chapman & Hall, New York.

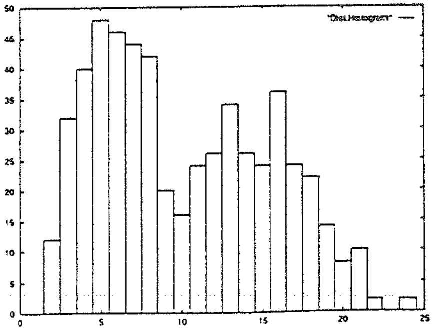

Figure 4: Inter-speaker distance distribution.

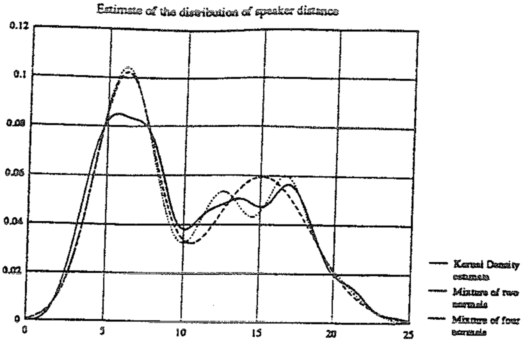

Figure 5: Estimate of inter-speaker distance distribution. Solid line: Kernel density estimate; Dotted line: mixture of four normals; Dash line: Mixture of two normals.

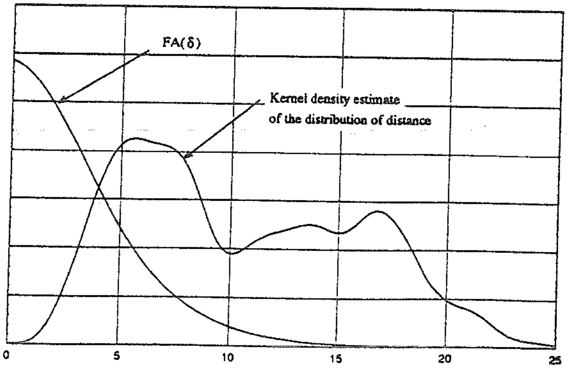

Figure 6: Estimate of $FAR(\delta)$ and Kernel density estimate of distance distribution.

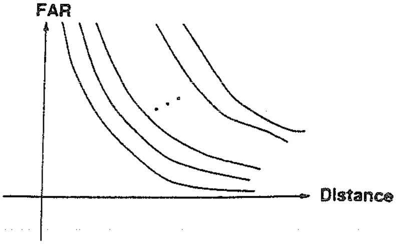

Figure 7: Different FAR functions for different thresholds.