euronoise

Adaptive threshold estimation for speaker verification systems

Eduardo Castilllo-Guerra[a], Roberto Díaz-Amador[b] and Cárdenas-Barreras Julian[b]

[a]University of New Brunswick, P.O. Box 4400, 15 Dineen Dr., D36 Head Hall, Fredericton, NB, Canada E3B 5A3

[b]Central University of Las Villas, Carr. a Camajuaní km 5.5, 50100 Santa Clara, Cuba

ecastill@unb.ca

This paper describes an adaptive threshold estimation mechanism for speaker authentication systems. The mechanism estimates speaker-dependent thresholds based on successful verifications considering the minimization of a cost function. Speaker authentication systems commonly use a threshold to decide whether a claimed identity matches a voice-print previously enrolled. Speaker independent threshold is a common option but it does not consider specific speaker characteristics that are relevant to achieve better system performance. Speaker dependent threshold on the contrary, uses speaker-specific data to estimate individual thresholds but the system performance can also suffer from suboptimal threshold conditioned by limited number of true scores. The algorithm reported in this paper starts with the speaker dependent threshold and use an adaptive algorithm to perform online re-estimation of the initial threshold based on speaker-dependent data. The threshold is re-estimated in each successful authentication transaction according to a custom-made confidence score. The reported technique keep the voice print up-to-date while is less sensitive to score outliers than traditional speaker dependent threshold. The algorithm provided a performance enhancement of up to 36.2% when compared to traditional speaker independent. An ad-hoc database obtained with a practical system was used involving cell and land-line utterances from male and female speakers.

1 Introduction

Voice biometric systems are a convenient and non intrusive way to authenticate remotely located users. They rely on voice prints obtained from enrolment sessions that capture unique speaker characteristics. Those unique characteristics are recognized when new utterances are authenticated. In order to determine the authenticity of a claimed identity, the score derived from the verified utterance has to be compared with a decision threshold. The threshold selection determines the way the speaker authentication system (SAS) is operated. The operating point of the SAS can be selected to minimize the decision errors such as false acceptance and false rejection errors [1][2]. It can also be selected to meet certain relationship between the false acceptance and false relation rate. In both cases the threshold must be optimized to meet the given decision criteria. The threshold selection can have two main approaches: unique to all the speakers attained to the system (speaker-independent, SI) or speaker specific (speaker dependent, SD).

The task of the SAS is reduced to accept or reject identity claims based on the utterance provided. For any utterance *A* and a claimed identity *l*, the SAS must make the decision according to the system operating criteria [3]. An optimal rule can be defined as:

$$\log\left(\frac{f_{A|l}(A\mid l)}{f_{A|\bar{l}}(A\mid \bar{l})}\right) \underset{<}{\geq} \log\left(\frac{C_{FA}\,P(\omega_{Nl})}{C_{FR}\,P(\omega_{Ml})}\right) \tag{1}$$

where $f_{A|l}(A\mid l)$ is the likelihood indicating that the given utterance was produced by speaker *l* and $f_{A|\bar{l}}(A\mid \bar{l})$ is the likelihood indicating the it was not. CFA and CFR denote the cost of false acceptance and false rejection respectively. $P(\omega_{NI})/P(w_{MI})$ is the false acceptance/false rejection probabilities ratio. The estimation of the likelihood functions is difficult in practice and lead to non-optimal decision rules.

The optimum decision rule in (1) is approximated practically with a variety of acoustic features and statistical models [2]. Despite a variety of algorithms with excellent classification capacity commonly used in SAS, the limited data used to train the models often lead to estimation errors [4][5]. However, models trained for each client are practically used to estimate the likelihood of a match between the utterance provided and the claimed identity. This likelihood approximates the left hand side (LHS) of (1) by $f_{A|\lambda_j}(A\mid\lambda_l)$. Likelihood normalization is then implemented to allow for more precise comparisons among different speakers [2]. Hence, the likelihood of the background model is used as an estimation of the denominator of LHS in (1).

Under these conditions, when an utterance is claimed to be from speaker *l*, a score still exist, according to LHS of (1) that can be compared with the specific threshold of speaker *l* as given by the right hand side (RHS) of (1). The problem is then focused on finding an optimum threshold that minimizes the cost function in the RHS of (1).

The threshold optimization is often implemented with two techniques target-impostor techniques. One approach is focused on finding a threshold that minimizes a specific cost function. The other approach determines the threshold that satisfies a specific false acceptance (FA) and false rejection (FR) rates. In this paper a cost based approach is pursued in order to minimize the decision errors involving FA and FR [7].

The main motivation of this investigation is to develop algorithms for adaptive management of speaker recognition systems. The use of SD threshold is very convenient in authentication since exploits specific speaker characteristics leading to performance enhancements. However, the number of enrolment scores is generally insufficient to provide unbiased estimations of SD threshold. This motivates the use of authentication scores to re-estimate the SD threshold but the identity of such utterances is not absolutely certain. This investigation is focused on developing a confidence index that can be used to determine those "reliable" scores at verification time. The re-estimation of SD threshold will rely only of those authentication scores that meet certain criteria preventing threshold contamination with outliers or false accepted impostor data. The convergence to stable SD threshold under this re-estimation approach must be optimized since the system could take more time to collect the minimum number of "safe" scores. Consequently, it requires optimizing the convergence process to guarantee adequate system performance at all time.

2 Thresholding

If S denotes the score of each client and t a corresponding

threshold, the probability of accepting false speakers (impostors) or rejecting true speakers associated to each client j can be stated as:

$$P_{FAj}(t) = P(s_{ij} > t) = \int_{t}^{\infty} f_{S_{ij}}(s)ds \quad (2)$$

$$P_{FRj}(t) = P(s_{Tj} > t) = \int_{-\infty}^{t} f_{S_{Tj}}(s)ds \quad (3)$$

where S_{Tj} and S_{Ij} are random variables representing scores produced by the SAS in response to a target or an impostor utterance with a claim of client *j*. Terms $f_{S_{ij}}(s)$ and $f_{S_{Tj}}(s)$ are their respective density functions. $P_{FAj}(t)$ will decrease from 1 to 0 as *t* increases from -∞ to ∞, while $P_{FRj}(t)$ increases. It is then possible to find a value of *t* that satisfies a relation such as:

$$P_{FAj}(t) = bP_{FRj}(t) - a \quad (4)$$

where the constant *a* is restricted to [0..1] and $b > 0$. A special case exists when $P_{FAj}(t) = P_{FRj}(t)$ named EER (equal error rate). If Gaussian distributions are assumed [8] for S_{Ij} and S_{Tj}, the EER relationships have a close-form expression for SD thresholds $t_{EER,j}$

$$t_{EER,j} = \frac{\sigma_{Tj}m_{Ij} + \sigma_{Ij}m_{Tj}}{\sigma_{Tj} + \sigma_{Ij}} \quad (5)$$

m and ***σ*** represent the mean and standard deviation respectively.

3 Design of confidence measurement

The confidence measurement should map the authentication scores into an index of trust. Scores closed to the SD threshold are less confident and must obtain lower scores. Authentication scores that deviate from the threshold are more confident and must received higher scores. There are three main aspects relevant to the design of this measurement: (1) the score magnitude that can be considered "reliable", (2) the desired distribution function of the index and (3) the implementation of the mapping function.

3.1 Uppers score boundary estimation

To magnitude of the score that can be considered "safe" are not known during enrolment. Only when a number of successful verifications have been accomplished a valid average magnitude can be estimated. Unfortunately, there is not absolute certainty on the identity of the verification utterances. Therefore, it is not convenient to rely on these utterances unless a confidence measure increase is available. An alternative approach is to rely on the enrolment utterances. These utterances provide higher scores than the authentication utterances as they are familiar to the models from the training session. However, if a data-based relationship can be established between the scores obtained from processing enrolment and verification utterances, the expected magnitude of what can be considered a "safe" score can be estimated at the enrolment session. Fig. 1 shows the average scores for all speakers in the database studied. It is observed a consistent 65% average ratio between both types of scores. This is observed more in details in Fig. 2.

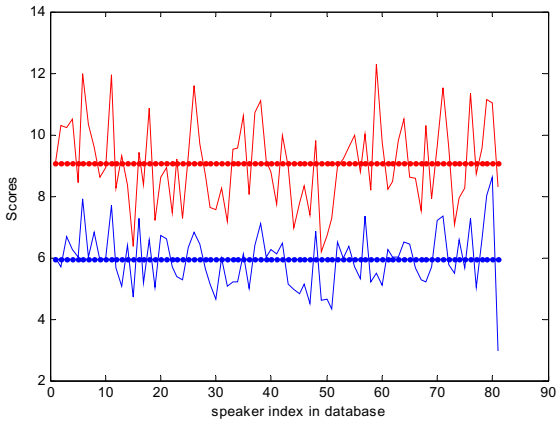

Fig. 1 Average enrolment (top) and verification (botom) scores. Average is shown for each case.

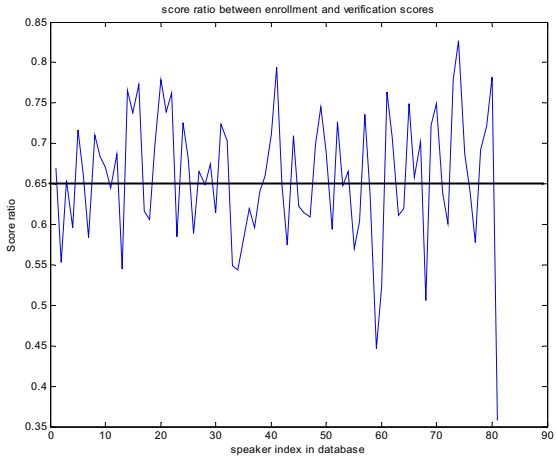

Fig. 2 Score ratio between enrolment and verification.

Based on these observations, the upper boundary needed to develop the confidence score can be obtained as the 65% of the average score obtained from enrolment ($\lambda_{REF} = 0.65 * \lambda_{ENROLMENT}$).

3.2 Score distribution

The average distribution function was derived from the database studied. All FA scores were analyzed and the zones where they were located above the SD threshold were documented. 69% of the scores associated with the FA cases were located in the 0-0.5 interval above the selected threshold for each speaker, 23% were located in the 0.5-1 range, 6 % lied in the 1-2 range and 2% were above 2. This a non-linear score distribution. Intervals closer to the speaker threshold are less confident and while the scores deviate from the speaker threshold the function exponentially increase the confidence. The confidence index should be 50% for a threshold equal to the SD threshold and change nonlinearly as scores deviate more

from the SD threshold. A sigmoid function was appropriate to model the score distribution defined as:

$$y = 1/(1 + Ae^{-Bx}) \quad (6)$$

A allows shifting the function in the *x* axis; *B* changes the slope and width of the transition band and *x* control the range mapped. Positive increments of *x* were used to match the distribution of the score into confidence interval between 50 – 100 %. This is because of the confidence index is only given when the resulting score is greater than the threshold of the verified speaker. Fig. 3 shows plots of the sigmoid function with different values of B with A and x constant. B = 2 provides the best modeling option. Smaller B values provided pessimistic confidence indexes and higher B values provided optimistic scores.

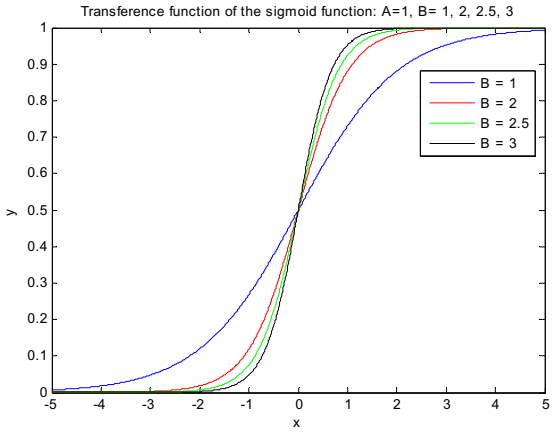

Fig. 3 function of the sigmoid function: A=1, x = 5, B= 1, 2, 2.5 and 3.

The variable *x* allows controlling dynamically the score range mapped. This range is determined for each speaker as a function of the reference score λ_{REF}, the verification score (λ_{CS}) and the threshold and (λ_{SD}). This approach covers the total expected range of variation between the scores and the SD threshold. It ensures a speaker dependent function that is tailored for each speaker and is adaptively updated from the speaker dependent threshold. The range is obtained as:

$$x = \frac{\lambda_{CS} - t_{SD}}{\lambda_{REF} - t_{SD}} \quad (7)$$

Table I shows examples of applying equations (6) and (7) with $x(t_{SD},\lambda_{REF},\lambda_{CS}) = x(3.07,5.84,\lambda_{CS})$ and $f(x,A,B) = f(x,1,B)$. From this table it can be noticed that the mapping function with $x \leq 2$ produce very optimistic confidence scores in the less reliable range. It does not provide a great separation between scores that deviates from the less reliable range. This separation is required to differentiate those utterances with higher than normal scores that are suitable to re-estimate the SD threshold. Based on this criteria B = 2 provided the best compromise between low confidence scores around the speaker threshold and high confidence for scores more separated from the threshold.

3.3 Evaluation of mapping function

The performance of the function with the testing utterances from the database is described in Table II (only a small fraction of the database is shown). Any negative value of x or confidence score below 50% corresponds to a rejected utterance included in this table just for comparison purposes but would not occur in normal processing.

$x(t_{SD},\lambda_{REF},\lambda_{CS})$	$\lambda_{CS} = t_{SD} + k$						
y(x,A,B)	k=0	k=0.5	k=1	k=2	k=3	k=4	k=5
$x(3.07,5.84,\lambda_{CS})$	0	0.180	0.361	0.721	1.081	1.442	1.803
y(x, 1, 1.5)%	50	56.7	63.2	74.7	83.5	89.7	93.7
$x(3.07,5.84,\lambda_{CS})$	0	0.180	0.361	0.721	1.081	1.442	1.803
y(x, 1, 2)%	50	58.9	67.3	80.9	89.7	94.7	97.4
$x(3.07,5.84,\lambda_{CS})$	0	0.180	0.361	0.721	1.081	1.442	1.803
y(x, 1, 2.5)%	50	61.1	71.1	85.8	93.7	97.4	98.9

Table 1 Performance of the mapping function

Notice that for all true speakers shown the confidence score is ranging between 75-97% while the confidence of impostors is well below 50% (0-29%). Based on this performance, the newly developed adaptive confidence score provides an accurate measurement that changes adaptively while the speaker dependent threshold is adapted. This provides an enhanced confidence score that contributes to reliably detect utterances that are good for updating the speaker dependent threshold. The confidence measurement also benefits from any improvement made in the SD threshold, forming a closed supervision mechanism that continuously tunes unique speaker characteristics.

SPK idx	x (Conf) utt. 1	x (Conf) utt. 2	x (Conf) utt. 3	x (Conf) utt 4	x (Conf) utt 5	Avg Impost
1.	1.0819 (0.8967)	1.2124 (0.9187)	1.0805 (0.8967)	1.2402 (0.9228)	0.70772 (0.8046)	-0.75887 (0.1996)
2.	1.0204 (0.8850)	0.25238 (0.6236)	0.93307 (0.8660)	0.8329 (0.8410)	0.59992 (0.7685)	-0.64679 (0.2286)
3.	1.0587 (0.8926)	0.85372 (0.8465)	1.159 (0.9104)	0.98727 (0.8781)	1.0052 (0.8819)	-0.78744 (0.1913)
4.	0.84942 (0.8454)	0.86551 (0.8495)	0.84833 (0.8451)	0.73843 (0.8141)	0.89951 (0.8580)	-0.82165 (0.1749)

Table 2 Example of the performance of confidence function

Scores with confidence values ≥75% exhibit a separation from the decision threshold in the non-mapped scale of 60% or more. This suggests that these scores are adequate to choose the threshold for SD re-estimation.

4 Adaptive threshold

Based on the confidence index, the limited set of true scores available to re-estimate the SD threshold can be increased with the detected reliable authentication scores. It can be achieved with a minimized risk of contaminating the scores with impostor data. The re-estimation has the limitation that while the number of true scores is not sufficient to converge to an optimum threshold the performance of the system will fluctuate as the SD thresholds converge. This limitation can be minimized by using an adaptive algorithm for the threshold re-estimation. The adaptation algorithm starts

with the SI threshold computed at design time and continuously adapted with the reliable verification scores.

It is convenient now to estimate the number of successful reliable verification scores needed by the SD threshold to converge to optimum values. From a previous work reported in [2], true-speaker scores were simulated based on statistics obtained from the studied database (simulation required due to insufficient number of true verifications scores in the database). This experiment consisted on increasing the number of true scores for each speaker in the database (only Pin data was used) from five to ∞ while keeping the number of impostors at 50. It was concluded that 50% of the total possible improvement using SD threshold was achieved with 10 true score varied. A total of 74% was obtained for 20 true scores and 92% for 50 [2]. This shows that to get all the benefit of the SD threshold, more than 50 reliable verification scores are necessary. The system may require certain amount of time to reach this number, especially when only very reliable scores are used for re-estimation. Before achieving the optimum threshold, the performance of the system can fluctuate creating undesired effects. This confirms the necessity to develop an adaptive threshold that can adapt the threshold while controlling the performance of the system while the threshold converges to optimum values.

A basic Least Square Algorithm (LMS) was implemented to adapt the threshold from the SI threshold. The adaptation step (μ) was changed from 0 to 1 where 0 corresponded to the basic SI threshold and 1 corresponded to the standard SD threshold.

5 Experiment

An experiment was conducted to compare the performance of a speaker verification task using SI, SD and SD adaptive thresholds. In both cases the equal error rate given by the FA and FR rate was minimized.

In this work, a real set of scores was used, based on an ad-hoc database consisting of spoken digits from several male and female speakers uttering their pin and telephone numbers through telephone lines. A HMM-based text-dependent SAS was used to estimate the scores. Background models were created with a subset from the ad-hoc database for a group of predefined digits to normalize speaker model scores. Each digit was modeled by a left to right continuous-density HMM with 11 states and 8 Gaussian mixtures per state. Speaker models were created by adaptation from the background model, using three iterations of the MAP adaptation algorithm. During verification a speaker score was obtained as the average of the normalized log-likelihood ratios of the digits.

In this experiment the models were trained with four utterances (balanced pin and telephone) and tested with 6 utterances (also with balance composition). A total of 50 impostors were used for all speakers in the database. The initial threshold used was the SI threshold computed to achieve a minimum the ERR over all speaker in the database. Different adaptation steps were tested in order to evaluate the benefits of the adapted SD threshold (SD_{ADAPT}). Fig. 4 shows the performance improvement of SDadapt over SI threshold.

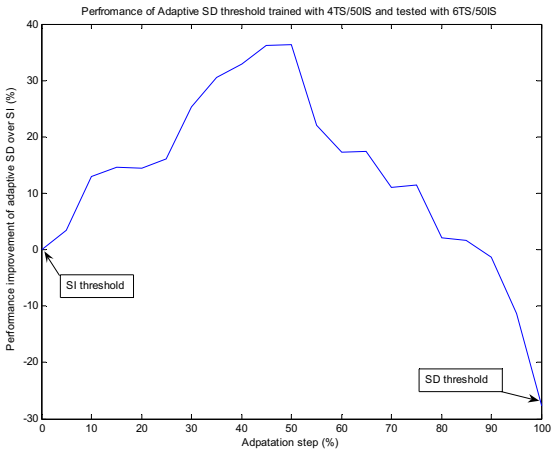

Fig. 4 Performance for different adaptation steps

Adaptation below 0.8 provided a performance enhancement with respect to SI threshold with μ = 0.45 providing the highest system performance enhancement (36.19%). Adaptation steps greater than 0.8 decreased the SI performance. This is attributed to the limited number of scores available which resulted in a biased SD threshold.

The performance of the system when using the SD threshold and the SD_{ADAPT} threshold is shown in Fig. 5 and Fig. 6. A summary of the system performance is given in Table 3.

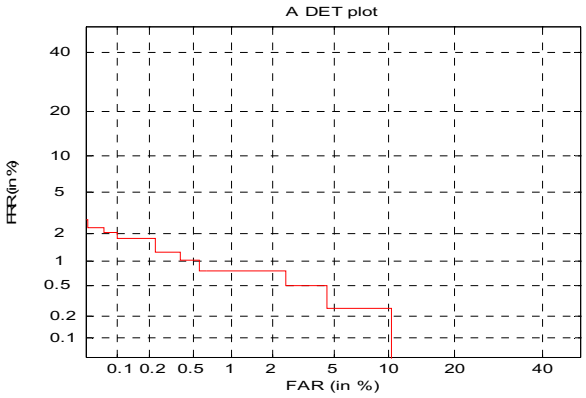

Fig. 5 System performance obtained with SD_{ADAPT} threshold

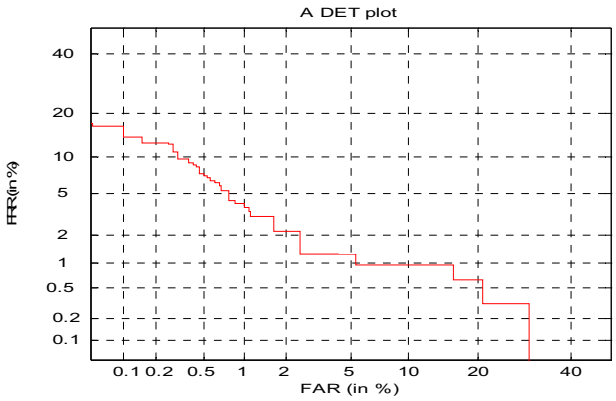

Fig. 6 System performance obtained with SD threshold

Thresholds	FAR	FRR	AER
SI	2.153	2.301	2.227
SD	0.876	4.812	2.844
SD_{ADAPT}	0.751	2.092	1.421

Table 3 System Performance for SI/SD/SD_{ADAPT} thresholds

6 Conclusion

A performance index was reported that is suitable to implement adaptive threshold re-estimation. An adaptive algorithm to achieve SD threshold was developed that outperformed traditional SI and SD thresholds. The performance advantage is more readily appreciated when the number of available true scores is limited. The combination of both techniques allows continuous adaptation of speaker dependent threshold needed to manage SAS with a minimized risk of contamination the threshold calculation with impostor data.

The score measurement could also be relevant for other tasks related with management of SAS such as template adaptation due to long-term variation of speaker and channel characteristics.

Acknowledgments

Authors appreciate NSERC and UNB for the financial contributions to this work.

References

[1] R.O. Duda, and P.E. Hart, "Pattern classification and scene analysis", *John Wiley & Sons,* 2nd Edition (2000)

[2] M. Stevenson, "Speaker dependent threshold task" *Technical report submitted to Diaphonics Inc.* (2007)

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas ''Score normalization for text-independent speaker verification systems'' *Digital Signal Processing* 10, 42–54 (2000)

[4] D. Falavigna, "Comparision of different HMM based method for speakers verifications" *EUROSPEECH*, 371-374 (1995)

[5] F. Bimbot, et al., ''A tutorial on text-independent speaker verification", *Eurasip Journal on Applied Signal Processing* 4, 430-451 (2000)

[6] N. Mirghafori, and M. Hebert, "Parameterization of the score threshold for a text-dependent adaptive speaker verification", *ICASSP* (2004)

[7] D.A. Reynolds, T.F. Quateri, R.B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing,* ICASSP 10, 19-41 (2004)

[8] J.R. Saeta, J. Hernando, "New speaker threshold estimation method in speaker verification based on weighting score" *3th International Conference on Non-Linear Speech Processing*, 34-41 (2005)