

A COMPARISON OF A PRIORI THRESHOLD SETTING PROCEDURES FOR SPEAKER VERIFICATION IN THE CAVE PROJECT

J.-B. Pierrot¹, J. Lindberg², J. Koolwaaij³, H.-P. Hutter⁴, D. Genoud⁵, M. Blomberg², F. Bimbot^{1,6}

(1) ENST / CNRS (2) KTH (3) KUN (4) Ubilab-UBS (5) IDIAP

pierrot@sig.enst.fr lindberg@speech.kth.se koolwaaij@let.kun.nl hans-peter.hutter@ubs.com
genoud@idiap.ch mats@speech.kth.se bimbot@sig.enst.fr

<http://www.PTT-Telecom.nl/cave>

ABSTRACT

The issue of *a priori* threshold setting in speaker verification is a key problem for field applications. In the context of the CAVE project, we compared several methods for estimating speaker-independent and speaker-dependent decision thresholds. Relevant parameters are estimated from development data only, i.e. without resorting to additional client data. The various approaches are tested on the Dutch SESP database.

1. INTRODUCTION

The CAVE project (Caller VERification in Banking and Telecommunications) is a 2-year project supported by the Language Engineering Sector of the Telematics Applications Programme of the European Union, and for the Swiss partners by the Office Fédéral de l'Education et de la Science (Bundesamt für Bildung und Wissenschaft). The partners are Dutch PTT Telecom, KUN, KTH, ENST, UBI-LAB, IDIAP, VOCALIS, TELIA and Swiss Telecom PTT. It started on December 1st, 1995. The technical objectives of the CAVE project are to design, implement and assess 2 telephone-based systems which use Speaker Verification (SV) technology. Work Package 4 (WP4) in this project focuses on the research and development aspects. The speaker verification system used in the experiments reported here is the Generic CAVE-WP4 SV system [1], based on the HTK software platform [2].

Laboratory evaluations of SV systems are usually based on the Equal Error Rate (EER), obtained by *a posteriori* setting the decision threshold(s) so as to equalise the false rejection and acceptance rates. Indeed, the EER gives a

good idea of the quality of the modeling module in a SV system. However, in the context of field applications, a specific procedure must be implemented in order to set the decision threshold *a priori*, namely during the enrollment procedure. Whereas Bayesian theory indicates that the decision threshold could be readily predicted for the false rejection and false acceptance costs, the mismatch between the speaker (and non-speaker) model(s) and the real data distributions requires the adjustment of the threshold for an efficient decision.

This paper reports on a series of comparative experiments on a priori Threshold Setting (TS) carried out by WP4. We first recall the main theoretical aspects involved in TS. Then, we express several TS procedures under a common formalism. Finally, we compare their efficiency on a task of speaker verification on a realistic telephone speech database (the SESP database).

2. THEORETICAL BACKGROUND

2.1. Notations

Let X denote a speaker, and \mathcal{X} his probabilistic model. Let \bar{X} denote the *non-speaker* model for speaker X , i.e the model of the rest of the population. Let Y be a speech utterance claimed as being from speaker X .

If we denote as \hat{X} (resp. $\hat{\bar{X}}$) the acceptance (resp. rejection) decision of the system, and p_X (resp. $p_{\bar{X}}$) the *a priori* probability of the claimed speaker to be (resp. not to be) speaker X , the total cost function of the system is [3] :

$$C = C_{(\hat{X}|\bar{X})} \cdot p_{\bar{X}} \cdot P(\hat{X}|\bar{X}) + C_{(\hat{\bar{X}}|X)} \cdot p_X \cdot P(\hat{\bar{X}}|X) \quad (1)$$

where $P(\hat{X}|\bar{X})$ and $P(\hat{\bar{X}}|X)$ denote respectively the probability of a false acceptance and of a false rejection, while $C_{(\hat{X}|\bar{X})}$ and $C_{(\hat{\bar{X}}|X)}$ represent the corresponding costs (assuming a null cost for a true acceptance and a true rejection).

2.2. PDF Ratio and Bayesian Threshold

If we now denote by P_X and $P_{\bar{X}}$ the Probability Density Functions (PDFs) of the speaker and of the non-speaker

¹ENST - Dépt Signal, CNRS - URA 820, 46 Rue Barrault, 75634 Paris cedex 13, FRANCE-EU

²KTH, Department of Speech, Music and Hearing, Drottning Kristinas Väg 31, S-100 44 Stockholm, SWEDEN-EU

³KUN, Dept of Language & Speech, Erasmusplein 1, NL-6525 HT Nijmegen, THE NETHERLANDS-EU

⁴Ubilab, Union Bank of Switzerland, Bahnhofstrasse 45, CH-8021, Zürich, SWITZERLAND

⁵IDIAP, Rue du Simplon 4, Case Postale 592, CH-1920 Martigny, SWITZERLAND

⁶Also with IRISA / CNRS & INRIA, France

distributions, the minimisation of C in equation (1) is obtained by implementing the PDF Ratio (PR) test [4] :

$$PR_X(Y) = \frac{P_X(Y)}{P_{\bar{X}}(Y)} \begin{array}{l} \text{accept} \\ > \\ < \\ \text{reject} \end{array} R \quad (2)$$

where R is the Bayesian threshold :

$$R = \frac{C_{(\hat{X}|\bar{X})} p_{\bar{X}}}{C_{(\hat{X}|X)} p_X} \quad (3)$$

2.3. Half Total Error Rate

As can be seen from equation (3), the optimal threshold should only depend on the false acceptance / rejection cost ratio and the impostor / client *a priori* probability ratio. In the particular case when the costs $C_{(\hat{X}|\bar{X})}$ and $C_{(\hat{X}|X)}$ are equal to 0.5, and when clients and impostors are assumed a priori equiprobable, the choice of $\Theta = 1$ as a decision threshold should then lead to a minimum of the *Half Total Error Rate* :

$$HTER = \frac{1}{2} [P(\hat{X}|\bar{X}) + P(\hat{X}|X)] \quad (4)$$

2.4. Likelihood Ratio and Threshold Adjustment

In practice, however, the PR in equation (2) is calculated from likelihood functions, i.e estimations of the PDFs, which do not match the exact speaker and non-speaker distributions. As a consequence, it is usually necessary to adjust the threshold of the PR test accordingly, in order to correct for the improper fit between the model and the data [5].

Thus, the PR test becomes an LR (Likelihood Ratio) test :

$$LR_X(Y) = \frac{\hat{P}_X(Y)}{\hat{P}_{\bar{X}}(Y)} \begin{array}{l} \text{accept} \\ > \\ < \\ \text{reject} \end{array} \Theta_X(R) \quad (5)$$

where \hat{P}_X and $\hat{P}_{\bar{X}}$ denote the respective *model* likelihood functions for the speaker and the non-speaker, and $\Theta_X(R)$ is a speaker- (and cost-) dependent threshold.

2.5. Gaussian log-LR model

In most cases, the logarithm of $LR_X(Y)$ is obtained as the sum of the logarithm of the frame-based likelihood ratio scores $lr_X(y_i)$:

$$\log LR_X(Y) = \sum_{i=1}^{i=n} \log lr_X(y_i) \quad (6)$$

where y_i denotes the i^{th} frame in utterance Y , of total length n . In some variants, the average log-LR is used instead of the log-LR :

$$\log LR'_X(Y) = \frac{1}{n} \log LR_X(Y) \quad (7)$$

We will refer to these two quantities as unnormalised and normalised LR, respectively.

If n is large enough, the utterance log-likelihood ratio can be assumed to follow a Gaussian distribution. This distribution is different depending on whether the speech utterance Y was pronounced by speaker X or by an impostor \bar{X} :

$$\begin{aligned} \log LR_X(Y|X) &\longrightarrow \mathcal{G}(M_X; S_X) \\ \log LR_X(Y|\bar{X}) &\longrightarrow \mathcal{G}(M_{\bar{X}}; S_{\bar{X}}) \end{aligned} \quad (8)$$

and similarly :

$$\begin{aligned} \log LR'_X(Y|X) &\longrightarrow \mathcal{G}(m_X; s_X) \\ \log LR'_X(Y|\bar{X}) &\longrightarrow \mathcal{G}(m_{\bar{X}}; s_{\bar{X}}) \end{aligned} \quad (9)$$

with the obvious relations :

$$\begin{aligned} M_X &= n m_X & M_{\bar{X}} &= n m_{\bar{X}} \\ S_X &= n s_X & S_{\bar{X}} &= n s_{\bar{X}} \end{aligned} \quad (10)$$

As opposed to the utterance log-likelihood ratio, the frame-based log-likelihood ratio does not generally follow a Gaussian distribution. But, if we now denote as μ_X and σ_X (resp. $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$) the mean and variance of the distribution of the frame-based client (resp. impostor) log-likelihood ratio $\log lr_X(y_i|X)$ (resp. $\log lr_X(y_i|\bar{X})$), and if we assume that the frame-based scores are statistically independent, we have (according to the Limit Central Theorem) :

$$\begin{aligned} m_X &= \mu_X & m_{\bar{X}} &= \mu_{\bar{X}} \\ s_X &= \sigma_X/\sqrt{n} & s_{\bar{X}} &= \sigma_{\bar{X}}/\sqrt{n} \end{aligned} \quad (11)$$

Under the assumption that the client and impostor log-LR follow Gaussian distributions, the optimal decision threshold can be obtained as :

$$\Theta_X(R) = \arg_t \left[\frac{\mathcal{G}(M_X; S_X)(t)}{\mathcal{G}(M_{\bar{X}}; S_{\bar{X}})(t)} = R \right] \quad (12)$$

and similarly for log-LR'.

In practice, it is feasible to obtain reasonable estimates of $M_{\bar{X}}$ and $S_{\bar{X}}$, from scores yielded by a population of *pseudo*-impostors. Conversely, in real applications, M_X and S_X have to be estimated from the enrollment data themselves and are therefore strongly biased, especially in the case when very few enrollment data are available.

3. SPEAKER-INDEPENDENT (SI) THRESHOLD

A classical method for adjusting the threshold $\Theta_X(R)$ in equation (5) consists in estimating a speaker-independent threshold so as to optimise the cost function of equation (1). In practice, this optimisation is carried out on a development data set, composed of enrollment and test data for a population of speakers which is distinct from (but representative of) the actual client population. In our experiments, we have tested the SI method both with unnormalised and normalised LR. We denote these two approaches as SI and SI-N, respectively.

The SI and SI-N methods do not make any particular assumption as regards the shape of the log-LR distribution. However, the fact that the threshold is speaker-independent relies on the hypothesis that the mismatch between the likelihood function and the actual client PDF translates into a client-independent shift between the log-PR and the log-LR. This is obviously a very simplistic hypothesis as part of the model mismatch is certainly variable across speakers.

4. SPEAKER-DEPENDENT (SD) THRESHOLD

Conversely, the estimation of a speaker-dependent threshold accounting for the variability in modeling accuracy can be hindered by the lack of proper data for estimating that threshold. Indeed, in the context of practical applications, enrollment material is so limited that it is not reasonable to reserve any of it for threshold setting. The speaker-dependent threshold must be derived from the same client data as those used for training the client model (and from some pseudo-impostor data).

In the next sections, we present 3 methods for speaker-dependent TS. The first method (SD-1) consists of estimating $\Theta_X(R)$ as a function of the log-LR mean and variance only, following an approach similar to the one proposed by Furui [6]. The second method (SD-2) relies on an estimation of $\Theta_X(R)$ using also the client score obtained on the enrollment data. The third method (SD-3) is based on the Gaussian model introduced in subsection 2.5. Methods SD-1 and SD-2 were tested with the unnormalised log-LR, whereas SD-3 was used with the normalised one (log-LR').

4.1. Method SD-1

In this method, $\Theta_X(R)$ is obtained as a linear combination of estimates of $M_{\bar{X}}$ and $S_{\bar{X}}$ only :

$$\Theta_X(R) = \hat{M}_{\bar{X}} + \alpha \hat{S}_{\bar{X}} \quad (13)$$

where $\hat{M}_{\bar{X}}$ and $\hat{S}_{\bar{X}}$ are obtained from pseudo-impostor data, whereas α is optimised on a development population.

4.2. Method SD-2

In this method, $\Theta_X(R)$ is obtained as a linear combination of estimates of $M_{\bar{X}}$ and M_X :

$$\Theta_X(R) = \beta \hat{M}_{\bar{X}} + (1 - \beta) \hat{M}_X^\circ \quad (14)$$

where $\hat{M}_{\bar{X}}$ is obtained from pseudo-impostor data, whereas \hat{M}_X° is the (biased) estimate of M_X obtained on the client enrollment data. Parameter β is optimised on a development population.

4.3. Method SD-3

This method is explicitly based on the Gaussian model of utterance the utterance log-LR distribution, as exposed in [5]. Estimates $\hat{\mu}_X^\circ$ and $\hat{\sigma}_X^\circ$ of μ_X and σ_X are initially obtained from the client enrollment data, whereas $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ are estimated from the pseudo-impostor data. Then, a speaker-independent correction h is applied to $\hat{\mu}_X^\circ$ only :

$$\begin{aligned} \hat{\mu}_X &= \hat{\mu}_X^\circ - h \\ \hat{\sigma}_X &= \hat{\sigma}_X^\circ \end{aligned} \quad (15)$$

where h is optimised on a development population. Then, estimates of m_X , s_X , $m_{\bar{X}}$ and $s_{\bar{X}}$ are obtained from $\hat{\mu}_X$, $\hat{\sigma}_X$, $\hat{\mu}_{\bar{X}}$ and $\hat{\sigma}_{\bar{X}}$, as in equation (11). Finally, $\Theta_X(R)$ is obtained as in equation (12) :

$$\Theta_X(R) = \arg_t \left[\frac{\mathcal{G}(\hat{m}_X; \hat{s}_X)(t)}{\mathcal{G}(\hat{m}_{\bar{X}}; \hat{s}_{\bar{X}})(t)} = R \right] \quad (16)$$

5. DATABASE AND PROTOCOL

All our experiments on TS were carried out on the realistic telephone speech database SESP, collected by KPN Research (the research laboratory of the Dutch Telecom). It contains telephone utterances from 21 male and 20 female speakers calling with different handsets (including some calls from mobile phones) from a wide variety of places (such as restaurants, public phones and airport departure lounges). During each call, the speaker was asked to utter a speaker-dependent sequence of 14 digits (twice) and another (session-dependent) sequence of 14 digits, corresponding to the number assigned to one of the other speakers.

Each session contains therefore 2 utterances of the client card number. For the experiments described in this paper we used 2 enrollment sessions with a low level of background noise, corresponding to 2 calls placed from 2 different handsets. Two other calls were reserved as extended enrollment material. The remaining calls were used as test material.

In our experiment on TS, we have split the SESP data into 2 sub-populations which we denote SESP-a and SESP-b. SESP-a contains 11 male and 10 female speakers while SESP-b contains 10 male and 10 female speakers. Each data set is composed of approximately 800 genuine trials and 250 impostor attempts from other clients (out of which about 75 % are same-sex attempts). We use SESP-b as pseudo-impostors and development data for SESP-a and vice-versa.

Acoustic features are 16 LPC cepstral coefficients with log-energy, together with their first and second derivatives. Cepstral mean subtraction is applied. Our tests were carried out using Left-Right HMM digit models, with 2 different topologies : $p = 2$ states per phoneme $\times q = 3$ Gaussian densities per state, and $p = 3$ states per phoneme $\times q = 2$ Gaussian densities per state. In these experiments, both the client model and the non-client model (here, a world-model) have the same topology. These configurations were chosen as they were those that we found to work best in terms of Equal Error Rate, in previous experiments on SESP [1].

In all our experiments, we aim at optimising the HTER, as defined in equation (4).

6. RESULTS

Comprehensive results are reported in Table 1. We provide separate performances for SESP-a and SESP-b. We first give Equal Error Rates for both unnormalised and normalised likelihood scores. Then, we give the performance with the fixed threshold $\Theta = 1$, followed by those obtained with the various TS methods presented above. For each method, we compare the performance obtained with an external development population to those (always better) that are reached when the development population is the same as the test population. The latter scores are given in italics.

7. COMMENTS AND CONCLUSIONS

On our task, normalisation by the utterance length seems to have little effect. But SESP utterances all have quite similar lengths. Therefore, the real impact of normalisation can not be studied accurately.

TS method	eval. data	dev. data	$p = 2, q = 3$			$p = 3, q = 2$		
<i>a posteriori</i> (sp.-dep. thresholds)			EER			EER		
EER	SESP-a	-	<i>0.57</i>			<i>0.99</i>		
	SESP-b	-	<i>0.46</i>			<i>0.63</i>		
EER-N	SESP-a	-	<i>0.57</i>			<i>0.99</i>		
	SESP-b	-	<i>0.26</i>			<i>0.89</i>		
<i>a priori</i>			FR	FA	HTER	FR	FA	HTER
$\Theta = 1$	SESP-a	-	12.11	0.25	6.18	13.25	0.25	6.75
	SESP-b	-	8.21	0.00	4.10	9.76	0.00	4.88
SI	SESP-a	SESP-b	0.86	4.60	2.73	1.85	4.01	2.93
		<i>SESP-a</i>	<i>2.22</i>	<i>2.72</i>	<i>2.47</i>	<i>2.59</i>	<i>2.41</i>	<i>2.50</i>
	SESP-b	SESP-a	1.72	1.73	1.72	1.47	1.61	1.54
		<i>SESP-b</i>	<i>0.26</i>	<i>2.26</i>	<i>1.26</i>	<i>0.97</i>	<i>1.91</i>	<i>1.44</i>
SI-N	SESP-a	SESP-b	1.63	4.95	3.29	2.73	2.15	2.44
		<i>SESP-a</i>	<i>3.06</i>	<i>1.62</i>	<i>2.34</i>	<i>3.06</i>	<i>1.44</i>	<i>2.25</i>
	SESP-b	SESP-a	2.25	1.96	2.11	2.12	1.61	1.87
		<i>SESP-b</i>	<i>0.42</i>	<i>2.66</i>	<i>1.54</i>	<i>1.59</i>	<i>1.61</i>	<i>1.60</i>
SD-1	SESP-a	SESP-b	4.08	2.26	3.17	3.25	3.59	3.42
		<i>SESP-a</i>	<i>3.35</i>	<i>2.26</i>	<i>2.80</i>	<i>3.72</i>	<i>2.43</i>	<i>3.07</i>
	SESP-b	SESP-a	1.05	3.69	2.37	1.44	2.98	2.21
		<i>SESP-b</i>	<i>1.17</i>	<i>3.00</i>	<i>2.09</i>	<i>0.79</i>	<i>3.34</i>	<i>2.07</i>
SD-2	SESP-a	SESP-b	2.83	1.82	2.32	2.72	2.52	2.62
		<i>SESP-a</i>	<i>2.83</i>	<i>1.82</i>	<i>2.32</i>	<i>2.61</i>	<i>2.52</i>	<i>2.57</i>
	SESP-b	SESP-a	1.28	1.12	1.20	1.02	1.80	1.41
		<i>SESP-b</i>	<i>1.28</i>	<i>1.12</i>	<i>1.20</i>	<i>1.28</i>	<i>1.52</i>	<i>1.40</i>
SD-3	SESP-a	SESP-b	4.86	1.66	3.26	2.80	1.89	2.35
		<i>SESP-a</i>	<i>0.97</i>	<i>2.20</i>	<i>1.58</i>	<i>1.08</i>	<i>1.92</i>	<i>1.50</i>
	SESP-b	SESP-a	1.65	2.44	2.05	1.76	3.11	2.43
		<i>SESP-b</i>	<i>0.57</i>	<i>2.71</i>	<i>1.64</i>	<i>0.57</i>	<i>2.68</i>	<i>1.63</i>

Table 1: Equal Error Rates and comparative results for several a priori Threshold Setting methods, on the SESP-a and SESP-b databases. *Scores in italics are obtained with the evaluation data used as development data.*

Loosely speaking, the HTER is about 3 to 5 times larger than the EER. This stresses once more the fact that the EER figure is a very optimistic evaluation of the actual performance of a SV system.

All methods yield similar results except method SD-2 which seems to perform consistently better. This may come from the fact that SD-2 is only using the means of the log-LR distributions, which are probably estimated more reliably than the variances, given the small amount of data and the strong bias in the client estimates.

It must also be noted that the SI methods do not perform especially worse than the SD methods, which tends to show that a large part of the model mismatch can be accounted for by a speaker-independent shift of the Bayesian threshold.

Quite important differences are observed between performances obtained on SESP-a and SESP-b, which illustrates the relatively large confidence interval that must be taken into account when interpreting these results.

Future work will consolidate these results, by extending the amount of experiments and the size of the database, and by testing the merit of the various Threshold Setting methods for other cost functions than the HTER.

8. REFERENCES

- [1] F. BIMBOT, H.-P. HUTTER, C. JABOULET, J. KOOLWAAIJ, J. LINDBERG, J.-B. PIERROT : *Speaker verification in the telephone network : research activities in the CAVE project*, Proc. Eurospeech '97, pp. 971-974. 1997.
- [2] S. YOUNG, J. JANSEN, J. ODELL, D. OLLASON, P. WOODLAND : *The HTK BOOK*, HTK 2.0 Manual. 1995.
- [3] R.O. DUDA, P.E. HART : *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.
- [4] L.L. SCHARF : *Statistical Signal Processing. Detection, Estimation and Time Series Analysis*. Addison-Wesley Publishing Company, 1991.
- [5] F. BIMBOT, D. GENOUD : *Likelihood ratio adjustment for the compensation of model mismatch in speaker verification*, Proc. Eurospeech '97, pp. 1387-1390. 1997.
- [6] S. FURUI : *Cepstral Analysis Technique for Automatic Speaker Verification*. IEEE Trans. on ASSP, vol. 29, no 2, pp. 254-272, 1981.