TECHNIQUES FOR A PRIORI DECISION THRESHOLD ESTIMATION IN SPEAKER VERIFICATION

J. Lindberg[1], J. Koolwaaij[2], H.-P. Hutter[3], D. Genoud[4], J.-B. Pierrot[5], M. Blomberg[1], F. Bimbot[5,6]

lindberg@speech.kth.se koolwaaij@let.kun.nl hans-peter.hutter@ubs.com genoud@idiap.ch
pierrot@sig.enst.fr mats@speech.kth.se bimbot@sig.enst.fr
http://www.PTT-Telecom.nl/cave

ABSTRACT

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

RÉSUMÉ

Un des problèmes importants en vérification du locuteur porte sur l'estimation du seuil de décision. Dans le cadre du projet européen CAVE, plusieurs méthodes d'estimation en mode dépendant et indépendant du locuteur sont comparées. Les paramètres servant à l'évaluation de ce seuil sont estimés uniquement grâce au données de développement et sans ajout de données supplémentaires pour les clients. Les différentes approches sont testées sur la base de données en langue néerlandaise SESP.

1. INTRODUCTION

The CAVE project (CAller VErification in Banking and Telecommunications) was a 2-year project that ended in december 1997. It was 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 were Dutch PTT Telecom, KUN, KTH, ENST, UBILAB, IDIAP, VOCALIS, TELIA and Swiss Telecom PTT. In the realm of the project, 2 telephone-based system which used Speaker Verification (SV) were developed and assessed. Work Package 4 (WP4) in this project focused on the research and development aspects. The SV system used in the experiments reported here is the cave-WP4 generic SV system [7], based on the HTK software platform [2].

Laboratory evaluations of SV systems usually base their assessments on the Equal Error Rate (EER). The EER is obtained by a posteriori setting the decision threshold(s) so that false acceptance and false rejection rates become equal. The EER gives a good estimate of the modeling module of the SV system. The EER does, however, not give much information about the performance to expect in a field application. In such a case the decision threshold(s) must be estimated a priori during the enrollment phase. Bayesian theory indicate that the decision threshold(s) could be predicted for the false acceptance and false rejection costs. The mismatch between the speaker and (non-speaker) model(s) and the real data distributions requires adjustments of the threshold(s) for efficient decisions to be made.

Part of the results in this paper is also reported in [8]. In this paper a new method (SD-4) has been added and this method is compared to the results reported in [8].

2. THEORETICAL BACKGROUND

2.1 Notations

Let X denote a speaker, and χ his probabilistic model. Let $\bar{\chi}$ 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]:

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

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_χ and $P_{\bar{\chi}}$ the Probability Density Functions (PDFs) of the speaker and of the non-speaker

[1] KTH - Dept of Speech, Music and Hearing, SE-100 44 Stockholm, SWEDEN-EU
[2] KUN, Dept of Language & Speech, Erasmusplein 1, NL-6525 HT Nijmegen, THE NETHERLANDS-EU
[3] Ubilab, Union Bank of Switzerland, Bahnhofstrasse 45, CH-8021, Zürich, SWITZERLAND
[4] IDIAP, Rue du Simplon 4, Case Postale 592, CH-1920 Martigny, SWITZERLAND
[5] ENST - Dépt Signal, CNRS - URA 820, 46 Rue Barrault, 75634 Paris cedex 13, FRANCE-EU
[6] Also with IRISA / CNRS & INRIA, FRANCE-EU

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

$$PR_\chi(Y) = \frac{P_\chi(Y)}{P_{\bar{\chi}}(Y)} \underset{reject}{\overset{accept}{\gtrless}} R \quad (2)$$

where R is the Bayesian threshold:

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

2.3 Half total Error rate

Equation (3) shows that the optimal threshold only depends on the false acceptance / false rejection cost ratio and the impostor / client *a priori* probability ratio. In the particular case of equal costs of 0.5 and when clients and impostors are assumed a priori equiprobable, the choice of Θ=1 as a decision threshold should then lead to a minimum of the *Half Total Error Rate*:

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

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 a Likelihood Ratio (LR) test:

$$LR_\chi(Y) = \frac{\hat{P}_\chi(Y)}{\hat{P}_{\bar{\chi}}(Y)} \underset{reject}{\overset{accept}{\gtrless}} \Theta_X(R) \quad (5)$$

where $\hat{P}_\chi$ and $\hat{P}_{\bar{\chi}}$ 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 LRχ(Y) is obtained as the sum of the logarithm of the frame-based likelihood ratio scores $lr_\chi(y_i)$:

$$\log LR_\chi(Y) = \sum_{i=0}^{i=n} \log lr_\chi(y_i) \quad (6)$$

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:

$$\log LR'_\chi(Y) = \frac{1}{n} \log LR_\chi(Y) \quad (7)$$

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 $\bar{X}$:

$$\begin{aligned} \log LR_\chi(Y|X) &\rightarrow G(M_X; S_X) \\ \log LR_\chi(Y|\bar{X}) &\rightarrow G(M_{\bar{X}}; S_{\bar{X}}) \end{aligned} \quad (8)$$

and similarly:

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

with the obvious relations:

$$\begin{aligned} M_X &= nm_X \qquad M_{\bar{X}} = nm_{\bar{X}} \\ S_X &= ns_X \qquad S_{\bar{X}} = ns_{\bar{X}} \end{aligned} \quad (10)$$

As opposed to the utterance log-likelihood ratio, the frame-based log-likelihood ratio does not generally follow a Gaussian distribution. But, if we 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 $\log lr_\chi(y_i|X)$ (resp. $\log lr_\chi(y_i|\bar{X})$), and if we assume that the frame-based scores are statistically independent, we have (according to the Limit Central Theorem):

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

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

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

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 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 some 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 4 methods for speaker-dependent TS. Methods SD-1, SD-2 and SD-4 were tested with the unormalised log-LR, whereas SD-3 was used with normalised scores (log-LR').

4.1 Method SD-1

SD-1 consists of estimating $\Theta_x(R)$ as a linear combination of the log-LR mean, $M_{\bar{x}}$, and variance, $S_{\bar{x}}$, following an approach similar to the one proposed by Furui [6]:

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

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

The second method relies on an estimation of $\Theta_x(R)$ using also the client score obtained with the enrollment data. In this method, $\Theta_x(R)$ is obtained as a linear combination of estimates of $M_{\bar{x}}$ and M_x :

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

where $\hat{M}_{\bar{x}}$ is obtained from pseudo-impostor data, whereas $\hat{M}_x'$ is the (biased) estimate of M_x. Parameter β is optimised on a development population.

4.3 Method SD-3

This method is explicitely based on the Gaussian model of utterance log-LR distribution, as exposed in [5]. The method uses the Gaussian model introduced in subsection 2.5. Estimates $\hat{\mu}_x'$ and $\hat{\sigma}_x'$ 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 population. Then, a speaker-independent correction *h* is applied to $\hat{\mu}_x'$ only :

$$\hat{\mu}_X = \hat{\mu}_X' - h \qquad \hat{\sigma}_X = \hat{\sigma}_X' \qquad (15)$$

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):

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

4.4 Method SD-4

The fourth SD method can be viewed as a speaker dependent adjustment of an estimated SI threshold. $\Theta_x(R)$ is obtained as a linear combination of the SI-threshold and estimates of $M_{\bar{x}}$ and M_x :

$$\Theta_X(R) = \Theta_{SI} + \gamma\left(\hat{M}_X' - \hat{M}_{\bar{X}} \right) \qquad (17)$$

where $\hat{M}_{\bar{x}}$ is obtained from the enrolment set of a development population, whereas $\hat{M}_x'$ is the (biased) estimate of M_x. Parameter γ is optimised on the development population.

5. DATABASE

All our experiments on TS were carried out on the realistic telephone speech database SESP [1], collected by KPN Research. 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. During each call, the speaker was asked to utter a number of items, including a speaker-dependent sequence of 14 digits (twice) and a few other sequences of 14 digits, corresponding to 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 rest of the 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 q=3 Gaussian densities per state, and p=3 states per phoneme q=2 Gaussian densities per state. In these experiments, both the client and world model, have the same topology. These configurations were chosen as they were those that worked best in terms of Equal Error Rate, in previous experiments on SESP [1].

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

TS method	Eval. data	dev. data	p = 2, q = 3			p = 3, q = 2		
a posteriori (sp.-dep. thresholds)			EER			EER		
EER	SESP-a SESP-b	- -	*0.57* *0.46*			*0.99* *0.63*		
EER-N	SESP-a SESP-b	- -	*0.57* *0.26*			*0.99* *0.89*		
a priori			FR	FA	HTER	FR	FA	HTER
$\Theta = 1$	SESP-a SESP-b	- -	12.11 8.21	0.25 0.00	6.18 4.10	13.25 9.76	0.25 0.00	6.75 4.88
SI	SESP-a SESP-b	SESP-b SESP-a	0.86 1.72	4.60 1.73	2.73 1.72	1.85 1.47	4.01 1.61	2.93 1.54
SI-N	SESP-a SESP-b	SESP-b SESP-a	1.63 2.25	4.95 1.96	3.29 2.11	2.73 2.12	2.15 1.61	2.44 1.87
SD-1	SESP-a SESP-b	SESP-b SESP-a	4.08 1.05	2.26 3.69	3.17 2.37	3.25 1.44	3.59 2.98	3.42 2.21
SD-2	SESP-a SESP-b	SESP-b SESP-a	2.83 1.28	1.82 1.12	2.32 1.20	2.72 1.02	2.52 1.80	2.62 1.41
SD-3	SESP-a SESP-b	SESP-b SESP-a	4.86 1.65	1.66 2.44	3.26 2.05	2.80 1.76	1.89 3.11	2.35 2.43
SD-4	SESP-a SESP-b	SESP-b SESP-a	0.38 1.08	4.18 1.61	2.28 1.35	0.58 1.47	2.28 1.61	1.43 1.54

Table 1: Equal-Error Rates and comparative results for several a priori threshold setting methods, on the SESP-a and SESP-b databases.

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, followed by those obtained with the various speaker dependent TS methods presented above.

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.

Loosely speaking, the HTER is about 3 to 5 times larger than 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 and SD-4, which seem to perform consistently better. This may come from the fact that these methods only use 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 methods for Threshold Setting methods for other cost functions than the HTER.

8. REFERENCES

[1] Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J., and Pierrot J.-B., "Speaker Verification in the Telephone Network: Research activities in the CAVE Project," *Proc. EUROSPEECH'97*, Rhodes, Greece, 1997, Vol. 2, pp971-974.

[2] Young S., Jansen J., Odell J., Ollason D., Woodland P., "The HTK BOOK, HTK 2.0 Manual", 1995.

[3] Duda R.O., Hart P.E., "Pattern Classification and Scene Analysis", John Wiley & Sons, 1973.

[4] Scharf L.L., "Statistical Signal Processing. Detection, Estimation and Time Analysis", Addison-Wesley Publishing Company, 1991

[5] Bimbot F., Genoud D., "Likelihood ratio adjustment for the compensation of model mismatch in speaker verification", *Proc. EUROSPEECH'97*, Rhodes, Greece, 1997, Vol. 2, pp1387-1390.

[6] Furui S., "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. on ASSP, vol 29, no 2, pp. 254-272, 1981.

[7] Jaboulet C., Koolwaaij J., Lindberg J., Pierrot J.-B., Bimbot F., "The cave-WP4 generic speaker verification system", *Proc. RLA2C*, Avignon, France, 1998

[8] Pierrot J.-B., Lindberg J., Koolwaaij J., Hutter H.-P., Genoud D., Blomberg M., Bimbot F., "A comparison of a priori threshold setting procedures for speaker verification in the cave project", *Proc. ICASSP*, Seattle, USA, 1998