AN OVERVIEW OF THE PICASSO PROJECT RESEARCH ACTIVITIES IN SPEAKER VERIFICATION FOR TELEPHONE APPLICATIONS

Frédéric BIMBOT[(1)], Mats BLOMBERG[(2)], Louis BOVES[(3)], Gérard CHOLLET[(4)], Cédric JABOULET[(5)], Bruno JACOB[(1)], Jamal KHARROUBI[(4)], Johan KOOLWAAIJ[(3)], Johan LINDBERG[(2)], Johnny MARIETHOZ[(6)], Chafic MOKBEL[(6)], Houda MOKBEL[(6)]

IRISA[(1)] **KTH[(2)]** **KUN[(3)]** **ENST[(4)]** **UBS-Ubilab[(5)]** **IDIAP[(6)]**

ABSTRACT

This paper presents a general overview of the current research activities in the European PICASSO project on speaker verification for telephone applications. First, the general formalism used by the project is described. Then the scientific issues under focus are discussed in detail. Finally, the paper briefly describes the *Picassoft* research platform. Along the article, entry points to more specific work also published in the Eurospeech'99 proceedings are given.

1 PRESENTATION

The 30-month European LE-Telematics project PICASSO (PIoneering Caller Authentication for Secure Service Operation) was launched in January 1998 in order to consolidate and extend the results of the previous CAVE project, on speaker verification (SV) on the telephone [1]. The partners of the PICASSO project are KPN-Telecom (NL), ENST (F), Fortis (NL), IDIAP (CH), IRISA (F), KPN-Research (NL), KTH (S), KUN (NL), Swisscom (CH), UBS-Ubilab (CH) and Vocalis (UK).

The overall objectives of the PICASSO project are to develop and test secure telematics transaction services using Speaker Verification (SV). These transactions can include actions which incur financial obligations (e.g. calling card calls, tele-shopping and other kinds of electronic commerce), which directly involve financial transactions (moving money between accounts, possibly of different owners), or which provide access to private information (e.g. a multi-media mailbox in a telecommunication service). The ultimate goal of PICASSO is to integrate speech recognition and speaker verification/identification technology to provide interfaces that are both easy to use and reasonably secure against intruders. One service based on the CAVE and PICASSO SV technology is currently in use in the Netherlands [2].

Within the PICASSO project, Work-Package WP5 is specifically dedicated to goal-oriented research on the improvement of speaker verification in the context of telecommunication transactions. In this respect, one of the outcomes of the CAVE project was to identify (and to start solving) the most significant technical issues that are still challenging for the deployment of services using SV-functionalities that are to be used by standard clients.

This paper describes the scientific issues that are addressed in the PICASSO project. It first presents the general formalism on which WP5 activities are based. Then it details the five main tasks along which the research activities are focused, namely :

- Client model estimation with scarce data
- Client / world model synchronous alignment
- Score normalization / threshold setting
- Incremental enrollment
- Password customization

Finally a section is dedicated to the description of the *Picassoft* system, a research platform shared between the project members. This paper thus serves as a general presentation of the PICASSO research activities and points to four more specific articles also published in Eurospeech'99, in which detailed results are given.

2 FORMALISM

PICASSO research activities are focused on text-dependent SV, in the sense that the verification procedure assumes that the text of the spoken utterance is known by the verification system, whether it is a fixed word (command word SV), a fixed sequence of words (e.g. a sequence of digits), a prompted sequence of words (text-prompted SV) or a particular word (or sequence of words) chosen by the user (customized password). In all these cases, the common assumption is that the system can base the verification process on a predefined speaker-dependent utterance model (in our case, a HMM) which has a left-right structure.

The verification process relies on the competition between 2 models, namely a client model (X) and a non-client model ($\overline{X}$) [3]. For a given speech utterance Y, the client model yields an estimate of the client

(1) IRISA – Sigma2, CNRS & INRIA , Campus Universitaire de Beaulieu, 35042 RENNES cedex, France.
(2) KTH – Department of Speech, Music and Hearing, Drottning Kristinas Väg 31, SE-100 44 STOCKHOLM, Sweden.
(3) KUN – Department of Language and Speech, Erasmusplein 1, NL-6525 HT NIJMEGEN, The Netherlands.
(4) ENST – Signal and Image Processing Dept, CNRS – URA 820, 46 Rue Barrault, 75634 PARIS cedex 13, France.
(5) UBS – Ubilab, Bahnhofstrasse 45, Postfach, CH-8098, ZÜRICH, Switzerland.
(6) IDIAP – Speech Group, Rue du Simplon 4, Case Postale 592, CH-1920 MARTIGNY, Switzerland.

probability density function for that utterance (likelihood), while the non-client model provides the estimate (likelihood) corresponding to the rest of the population. These two quantities will be respectively denoted as :

- client likelihood : $\hat{P}(Y|X)$
- non-client likelihood : $\hat{P}(Y|\overline{X})$

In the Picasso project, we use the *world-model* approach where the non-client model is client-independent, i.e. $\overline{X} = \Omega$.

Attempt of utterance Y against identity X is scored using the log likelihood ratio :

$$s_X(Y) = \log\left[\frac{\hat{P}(Y|X)}{\hat{P}(Y|\overline{X})}\right]$$

Decision is taken by comparing the log-likelihood ratio score to a client-dependent (and sometimes utterance-dependent) threshold $\theta_X(R,Y)$:

$$s_X(Y) \underset{reject}{\overset{accept}{\gtrless}} \theta_X(R,Y)$$

where R denotes the Bayesian threshold, i.e. the optimal decision threshold if the likelihood ratio was computed from the exact client and non-client probability density functions :

$$R = \frac{p}{1-p}\frac{C_{FA}}{C_{FR}}$$

with p denoting the a priori probability that the claimant is an impostor and C_{FA} (resp C_{FR}) denoting the cost of a false acceptance (resp. false rejection).

The log likelihood ratio can be submitted to various normalization operations (for instance length-norm, z-norm, h-norm, etc...), in which case the decision threshold may be chosen identical for all decisions.

3 RESEARCH ISSUES

The deployment of speaker verification technology within applications dedicated to the general public necessitates significant adjustments or add-ons to standard algorithms and procedures. One major difficulty to overcome is the lack of training material, as it is not realistic to require a large number of enrollment sessions before a system can become operational for a particular user. Practically, 2 to 5 repetitions in one single session is the amount of speech material that must be dealt with in our application context. The consequences of this lack of coverage of the client training data manifest themselves at several levels : difficulty to estimate reliably a client-model using the standard EM algorithm (based on ML optimization), poor consistency between the decoding process in the client and world-model, limited efficiency of the optimal Bayesian decision threshold, need for adaptation scheme in order to track voice drift over time. Moreover, in the case of (fully) user-customizable passwords, it is not feasible to collect corresponding non-client data and this requires a particular strategy for world-model estimation.

3.1 Client model estimation with scarce data

One of the major outcomes of the CAVE project research activities was to evidence the inefficiency of the ML criterion for training a client model with limited enrollment material. More specifically, variance parameters turn out to be impossible to estimate reliably with the typical amount of enrollment material available.

In the CAVE project, the solution adopted was the *adaptive variance flooring* approach where the variance σ^X_{ijk} of gaussian mixture k in state j for coefficient i in the client model X is prevented to go below a certain proportion γ of the overall variance σ^Ω_i for this coefficient (computed from the world-model data).

Recent experiments indicate that even simpler approaches for variance estimation can be used without significant impact on the performance. In particular, a much easier procedure of *variance scaling*, which consists in setting :

$$\sigma^X_{ijk} = \alpha\,\sigma^\Omega_{ijk}$$

yields equivalent results than variance flooring, without requiring any iterative estimation of the variance parameters. A comprehensive comparison of several alternative approaches to client model variance estimation with scarce data is proposed in [4]. Note that in most of our experiments, appropriate values for γ or α have always been in the range of 1.0 (or slightly less).

Another way to address data scarcity is to use an *adaptation scheme* for training the client model from the world model. The corresponding formalism is identical to the one mentioned in section 3.4 below (MAP approach for incremental enrollment). The technique is currently under development and its efficiency on our reference tasks will be compared to the one of variance flooring and scaling.

3.2 Client / world model synchronous alignment

The use of a HMM for modeling the probability density function of a speech utterance Y assumes that there exists a hidden process underlying the generation of that speech utterance. However, in conventional HMM-based SV approaches (using Viterbi decoding) this assumption is not fully exploited, since the sequence of states in the client and the world models are not constrained to be consistent with each other. Moreover, in case of scarce enrollment data, the sequence of states decoded in the client model is relatively prone to irrelevant state assignment, in particular when the utterance Y has a low likelihood for model X.

The scheme of client / world model synchronous alignment has been designed in order to enforce consistency between the state sequence decoding within the client and the world models, i.e. assuming a common hidden process for both models.

While in the conventional procedure, the client and world model likelihoods are respectively computed as :

$$\hat{P}(Y|X) \approx \hat{P}(Y|X, S_X) \text{ with } S_X = Arg\max_Q \hat{P}(Y|X, Q)$$
$$\hat{P}(Y|\Omega) \approx \hat{P}(Y|X, S_\Omega) \text{ with } S_\Omega = Arg\max_Q \hat{P}(Y|\Omega, Q)$$

where S_X and S_Ω correspond to the state sequences in the client and world models respectively, the synchronous alignment scheme computes the likelihood along a jointly optimized path S, namely :

$$\hat{P}(Y|X) \approx \hat{P}(Y|X, S) \quad \text{and} \quad \hat{P}(Y|\Omega) \approx \hat{P}(Y|\Omega, S)$$
$$\text{with } S = Arg\max_Q \left\{ \hat{P}(Y|X, Q)^a \, \hat{P}(Y|\Omega, Q)^{(1-a)} \right\}$$

The optimal path S is obtained by assigning a different weight to the client (a) and world ($1-a$) models. In the case when $a = 0$, the client path is simply synchronized on the world model path. Note that a consequence of the use of the synchronous alignment scheme is that both client and world models must have the same topology (i.e. number of states), but they do not need to have the same number of mixtures per state.

The use of synchronous decoding turns out to be more consistent if similar constraints have been introduced during the training of the client model. Detailed decoding and training procedures for the synchronous alignment scheme can be found in [5].

A first set of recent experimental results (also reported in [5]) show a slight benefit in terms of EER for the synchronous alignment approach. Moreover, the approach offers a slightly reduced computational complexity and provides a simple decomposition of the utterance log likelihood ratio in terms of a sum of frame-by-frame likelihood ratios. For all these reasons, the synchronous alignment procedure appears to be a very relevant extension of the HMM scheme to likelihood ratio computation in speaker verification.

3.3 Score normalization / threshold setting

Bayesian theory offers a general framework for decision threshold setting in two-class problems such as SV. However, the risk ratio R (as defined in section 2) is the optimal threshold only in the case when the likelihood functions are equal to the true probability density functions of the 2 classes. As this is not the case in practice, the threshold has to be adjusted by taking into account several factors, in particular the claimed identity, but also the speech utterance itself.

Quite equivalently to formulating it in terms of threshold adjustment, the problem to be addressed can be expressed in terms of likelihood ratio normalization. In that case, the decision rule becomes :

$$\Psi[s_X(Y)] \underset{reject}{\overset{accept}{\gtrless}} R$$

where Ψ denotes a normalization function that is applied to the likelihood ratio in order to compensate as well as possible for the modeling inaccuracies. Most conventional normalization techniques and threshold setting procedures fall under this formalism. Function Ψ can itself depend on X, via statistics of the client and/or (pseudo-)impostor scores. It can also depend on the speech utterance Y via the sequence of decoded states, especially if a synchronous alignment approach is used.

Whereas threshold setting procedures usually assume that the cost function is already known, an other challenge is to find a normalization function that yields reasonable results for a large range of risk ratios, so that the likelihood ratio normalization can be performed independently from the respective false acceptance / rejection costs. This requires a global modeling of the client and impostor score distributions.

3.4 Incremental enrollment

To overcome the difficulties raised by the limited amount of training data collected during the (active) enrollment session(s), it is possible to extend the training material using "passive" enrollment, i.e. during the actual use of the SV system, provided that there is a way to certify (or to neutralize the risk) that the spoken material was (or not) actually uttered by the true client. Moreover, it is well-known that people's voice changes over time and this requires a process for regularly updating client models in order to track this evolution. In such a context, one option is to store the speech material in order to use it at a later stage for a complete (batch) retraining of the client model. This however requires significant storage resources. An alternative option is to use incremental training for updating progressively the speaker model.

In the PICASSO project, we focus on the Bayesian adaptation (or MAP) approach for HMM models with gaussian mixtures [6] which offers a well-defined framework for addressing incremental enrollment as an adaptation problem [7]. Moreover, a particular way of deriving the MAP priors from the initial model (i.e. the model before update) requires only a limited amount of information to be carried from one session to the next one. In practice, the past enrollment data can be summarized by the gaussian mixture (and transition) occupancy for each state in the model. These quantities are used to derive the priors for the next increment and they can be easily updated at each new session [8].

This approach is currently under evaluation within the project. Preliminary results show convergence of the training process but, as could be expected, slightly less good performance than the batch approach (i.e. using all speech material at once). Moreover, the enrollment material used in the protocol is a priori known as belonging to the actual client (supervised update) and the approach has to be evaluated in more adverse cases (unsupervised update). A more detailed description of the approach together with experimental results will be the subject of a future article.

3.5 Password customization

In real services, a very desirable feature is the possibility for the user to choose the speech utterance on which the

verification is to take place (user-customized password). Firstly, this offers better user-friendliness and it is perceived as an essential functionality by some service providers. Secondly, this warrants increased security as fraudulent access requires prior knowledge of the client's password. Customized password is therefore a means to decrease the vulnerability of SV systems against intentional imposture, in particular "technological imposture" with concatenated words against SV systems using a fixed vocabulary [9].

In this context, the problem of estimating a client model is not more difficult than in the case of fixed text : it is realistic to ask the client to repeat several time his/her new password (preferably in a single session). The difficulty comes from the necessity to estimate a world-model for this particular password, i.e. a model of the way other speakers would pronounce this very password. In this case, the problem is therefore to infer a speaker-independent model from a single-speaker set of examples.

The approach that we are investigating consists in the following steps :

1. Transcribe the password into a sequence (or a graph) of speech symbols using a set of speaker-independent acoustic units,
2. Build a speaker-independent password HMM by substituting each speech symbol in the graph obtained at step 1 by its speaker-independent acoustic model ; this yields the password world-model,
3. Train a client (speaker-dependent) model from the password utterances, by standard training or by adaptation of the world-model inferred in step 2.

Three approaches are being compared, which differ according to the way they address step 1. One of them is based on a phonetic HMM for obtaining the symbolic transcription. A second approach uses a neural network. A third one resorts to ALISP units [10]. These approaches are currently under development and will be compared on the same task.

4 THE PICASSOFT RESEARCH PLATFORM

As for the CAVE project [11], a significant effort is being dedicated to the development, maintenance and improvement of a common software platform, aiming at providing to each partner algorithms corresponding to the state-of-the-art reached within the project. The most significant novelties since the CAVE platform (Genesys) are :

- the introduction of explicit experiment configurations that allow flexible combinations of different sets of populations (development, pseudo-impostor, test,...)

- the implementation of a wide variety of likelihood ratio normalization techniques, which can be gender, speaker, handset, ... dependent, so that these techniques can be extensively compared.

- the possibility of keeping track of the likelihood values at the frame level, so that several normalizations techniques can readily be applied at the frame, segment, speech unit and/or utterance levels.

- the integration of advanced variance estimation strategies (variance flooring, scaling, tying, etc...).

- the use of a larger variety of assessment tools, in particular (besides the EER), the use of DET curves [12], the computation of Decision Cost Functions and also the distribution of errors over the client population.

Like the CAVE (Genesys) platform, the Picassoft platform is centered on HTK (v2.1), but it also uses shell scripts, and Matlab v5.0 functions.

5 CONCLUSION

The results obtained so far in the Picasso project consolidate previous findings and open new tracks for improved approaches in speaker verification. Moreover, the different issues addressed are likely to meet and ultimately merge towards a more unified framework. Synchronous alignment offers a simplified log-likelihood ratio computation, which could in turn benefit from variance scaling approaches (that may simplify further the frame likelihood ratio calculation). The use of a common sequence of states is also an interesting property for developing utterance-dependent normalization schemes. Moreover, it appears clearly that adaptation techniques can be used to address the client model estimation at several steps : initial estimation by adapting the world-model and incremental enrollment by updating the current client model. The maintenance and regular upgrade of the *Picassoft* platform will allow the partners to continue these investigations in a concerted and consistent way.

6 ACKNOWLEDGEMENTS

This work was funded by the Telematics Programme of the European Commission (project LE4-8369) and by OFES (Office Fédéral de l'Education et de la Science - project 97.0494-2).

7 REFERENCES

[1] F. Bimbot et al. : An overview of the CAVE project research activities in speaker verification. Proc. RLA2C Workshop, pp. 215-220, Avignon, 1998

[2] E. Den Os et al. : Speaker verification as a user-friendly access for visually impaired people. Eurospeech'99.

[3] L. Scharf : Statistical signal processing – Detection, estimation and time analysis. Addison-Wesley Publishing Company. 1991.

[4] H. Melin, J. Lindberg : Variance flooring, scaling and tying for text-dependent speaker verification. Eurospeech'99

[5] J. Mariéthoz, D. Genoud, F. Bimbot, C. Mokbel : Client / world model synchronous alignment for speaker verification. Eurospeech'99.

[6] J-L. Gauvain, C. Lee : Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. IEEE Trans. SA, Vol. 2, n° 2, pp. 291-298, April 1994.

[7] C. Mokbel, O. Collin : Incremental enrollment of speech recognizers. ICASSP'99, Vol. 1, pp. 453-456.

[8] C. Mokbel et al : Picasso WP5 deliverable D5.1 on incremental enrollment. 1998.

[9] J. Lindberg, M. Blomberg : Vulnerabilty in speaker verification : a study of technical impostor techniques. Eurospeech'99.

[10] G. Chollet et al. : Towards ALISP : Automatic Language Independent Speech Processing. In NATO-ASI on Computational Models of Speech Pattern Processing, R. Moore Ed., 1998.

[11] C. Jaboulet, J. Koolwaaij et al. : The CAVE-WP4 generic speaker verification system, Proc. RLA2C, pp. 202-205, Avignon, 1998

[12] A. Martin, M. Przybocki : The DET curve in assessment of detection task performance, Eurospeech'97, pp. 1895-1898.