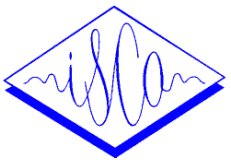

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A MAP APPROACH, WITH SYNCHRONOUS DECODING AND UNIT-BASED NORMALIZATION FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Johnny MARIETHOZ [1], *Johan LINDBERG* [2], *Frédéric BIMBOT* [3].

[1] IDIAP - BP 592, CH-1920 Martigny, Switzerland - mariethoz@idiap.ch
[2] KTH - Drottning Kristinas Väg 31, SE-100 44 Stockholm, Sweden - lindberg@speech.kth.se
[3] IRISA - Campus Beaulieu, 35042 Rennes, France - bimbot@irisa.fr

ABSTRACT

This paper presents an overview of some of the research tracks in text-dependent speaker verification followed by the Picasso Project (Research Work Package). We focus successively on the training algorithm (based on a MAP approach), the state-sequence decoding procedure (by Synchronous Alignment) and the score normalization (unit-based z-normalization). On short utterances of 1 to 5 syllables (Polyvar command words), we obtain an error rate of 5.6 % with 2 training sessions, which can be brought down to 3.3 % with 3 additional sessions in incremental training mode.

1. INTRODUCTION

Speaker Verification (SV) technology is subject to an increasing interest from telecommunication service providers, as a means of transparent and user-friendly user authentication. However, strong constraints impact the application profile, in particular, the robustness of the technology and its ability to be almost immediately operational for new users.

As a consequence, the SV products must be able to yield satisfactory performance even with limited quantities of enrolment data (typically, 2 sessions maximum) and, if possible, despite of adverse conditions of use. The work reported here mainly focuses on the first issue. It gives an overview of some of the research tracks followed by the research Workpackage of the Picasso project [B+99].

We first recall the bases of probabilistic speaker verification. Then we describe briefly the Picassoft software platform in its baseline configuration and the PolyVar / Suisse Romand database which we used in our experiments. We then present our MAP approach for adapting the world model to the client model in the initial enrollment phase and for incrementally updating the client model with additional speech material (collected for instance during system use). We also report on experiments using client/world synchronous alignment and on results obtained with various unit-based generalizations of the z-normalization technique.

2. GENERALITIES

The approach used in this paper is based on probabilistic speaker verification theory. Specifically, for an utterance Y pronounced by a speaker claiming to have the identity of X, the logarithm of the likelihood ratio is computed as :

$$s_X(Y) = \log\left(\frac{\hat{\mathcal{P}}(Y|X)}{\hat{\mathcal{P}}(Y|\bar{X})}\right)$$

where $\hat{\mathcal{P}}(Y|X)$ is the likelihood of utterance Y under the assumption that it has been uttered by speaker X, while $\hat{\mathcal{P}}(Y|\bar{X})$ is the likelihood of Y under the assumption that it has been uttered by someone else.

The probabilistic model for X (i.e., the *client model*) is estimated from training data uttered by X, whereas the model for $\bar{X}$ (i.e., the *non-client model*) is estimated from speech uttered by other speakers. In this work, the non-client model is identical for all clients (*world model* approach) and is denoted Ω.

This work is carried out in the context of text-dependent applications. All clients are assumed to have a common password. We use Left-Right Hidden Markov Models (HMMs) with Gaussian Mixture Models (GMMs) state probability density function (PDF). The client HMMs and the world model have the same topology.

Following conventional Bayesian decision theory, the log likelihood ratio $s_X(Y)$ is compared to a threshold Θ : in order to yield a decision a claimed identity. In general, the threshold Θ is calculated in order to optimize a cost function expressing a compromise between the two kinds of errors : false acceptance (an impostor is accepted) and false rejection (an genuine client is rejected).

In this work, we consider the cost function that gives equal weight to the false acceptance rate and the false rejection rate. This is equivalent to optimising the Half Total Error Rate (E), namely

$$E = \frac{1}{2}(\varepsilon_{FA} + \varepsilon_{FR})$$

where ε_{FA} and ε_{FR} are respectively the false acceptance and false rejection error rates.

We compare the HTER obtained in two cases : on the one hand, the HTER obtained when the threshold Θ_{opt} is set so as to optimize the cost function *a posteriori* on the test set (E_{opt}) and, on the other hand, the HTER obtained when the threshold Θ_{dev} is optimized *a priori* on a separate development population of clients (E_{dev}). Note that $E_{opt} \leq E_{dev}$.

3. BASELINE SYSTEM

The baseline system is *Picassoft*, an HTK-based software platform shared by the Picasso partners for efficient concerted research.

The client and world-model HMMs have 2 states per phoneme with 3 mixtures per state. Acoustic features are 12 LPCC cepstral coefficients with log-energy, together with their first and second derivatives.

The training of the baseline system is based on a Maximum Likelihood (ML) criterion. However, only the means and weights are learned ; the world-model variances are copied and kept constant during the client model training. For these experiments we process all the training material at once (batch training).

Results are given on a subset of the PolyVar database (39 clients) with a vocabulary of 17 Swiss French command words (from 1 to 5 syllables only). For each client, 2 to 5 sessions are reserved for training. The test set is composed of approximately 15 genuine access and 40 impostor attempts per word and per client. Thus, there are approximately 36000 tests in total.

As explained above, the performances are evaluated using both HTERs E_{opt} and E_{dev}. For the latter, the test population is split into two subsets : one half for estimating Θ_{dev} and the other half for calculating the corresponding HTER. The role of the two subsets is then swapped and the error rate E_{dev} is obtained by averaging the two HTERs on the 2 subsets.

Table 1 gives the HTERs and the corresponding thresholds obtained with the baseline system, with 2, 3, 4 and 5 training sessions.

		Θ	HTER [%]
2s	opt	-2.01	6.63
	dev	-2.01 | -2.00	6.65
3s	opt	-1.14	5.04
	dev	-1.16 | -1.13	5.05
4s	opt	-0.57	4.03
	dev	-0.57 | -0.72	4.07
5s	opt	-0.34	3.31
	dev	-0.33 | -0.24	3.38

Table 1: Baseline Picassoft (ML) - HTER as a function of the number of training sessions.

4. ADAPTATION SCHEME

For increasing the robustness of the client model training, especially when the number of training session is small, we have used a MAP scheme [GL94] for :

- training the 2-session client model by adapting the world-model rather than using a ML criterion (similar to [Rey97]).
- performing incremental enrollment with each new session (3 to 5). As opposed to batch training, the speech data from all previous sessions are not reused for reestimating the model for the current session (as done in [F+00]). The client model is adapted incrementally with new speech material.

In both cases, the new mean μ_{MAP} of a given mixture for a given client is given by:

$$\mu_{MAP} = \gamma\,\mu_0 \; + \; (1-\gamma)m$$

where μ_0 is the initial value of the mean and m the client mean estimate computed as the mean value of the speech frames for which the considered mixture has the maximum a posteriori probability. Parameter γ is the relative weight given to the prior knowledge versus the new data. In our experiments, the optimal value of γ was found to be $\gamma = 0.25$.

Our results, reported in Table 2, show that doing the 2-session training by our MAP adaptation of the world-model yields, in batch mode, a significant performance improvement over the baseline system. On the other hand, the incremental enrollment degrades slightly (but reasonably) the performance as compared to MAP batch training.

5. SYNCHRONOUS ALIGNMENT

Previous work carried out in the Picasso project showed some advantage, in the ML framework, in using a technique called *Synchronous Alignment* (SA) for

		MAP batch		MAP incremental	
		Θ	HTER	Θ	HTER
2s	opt	-0.44	5.69	-0.44	5.69
	dev	-0.33 | -0.50	5.89	-0.33 | -0.50	5.89
3s	opt	-0.07	4.45	-0.20	4.68
	dev	-0.09 | -0.06	4.48	-0.21 | -0.32	4.72
4s	opt	0.13	3.60	-0.09	3.81
	dev	0.12 | 0.17	3.71	-0.10 | 0.00	3.91
5s	opt	0.28	3.19	-0.03	3.26
	dev	0.27 | 0.23	3.26	-0.01 | -0.03	3.28

Table 2: MAP / Batch vs Incremental training.

		MAP batch		MAP incremental	
		Θ	HTER	Θ	HTER
2s	opt	-0.24	5.73	-0.24	5.73
	dev	-0.24 | -0.34	5.79	-0.24 | -0.34	5.79
3s	opt	0.00	4.56	-0.12	4.67
	dev	-0.02 | 0.02	4.63	-0.07 | -0.11	4.72
4s	opt	0.21	3.82	-0.06	3.82
	dev	0.14 | 0.22	3.90	-0.06 | -0.05	3.84
5s	opt	0.31	3.49	0.09	3.27
	dev	0.27 | 0.31	3.55	0.08 | 0.12	3.31

Table 3: MAP + Synchronous Alignment ($\alpha = 0.5$)

training the client model parameter and for computing the likelihood ratio score [M+99]. The hypothesis behind the SA technique is that, for a speech utterance, the hidden process (i.e the state sequence) in the client (speaker-dependent) and world (speaker-independent) models should be identical.

Under the SA approach, the training and the scoring are thus carried out under the hypothesis of a common sequence of states $\hat{S}$ in the client and world-model :

$$\hat{S} = \arg\max_{S} \alpha \log \hat{\mathcal{P}}(Y|X,S) + (1-\alpha) \log \hat{\mathcal{P}}(Y|\bar{X},S)$$

where α denotes the weight given to the client model in the optimal state sequence decoding. We have recently integrated the SA procedure within the MAP scheme, which is a straightforward generalization of the technique formerly developed in the ML case. Note that, in incremental training mode, SA impacts the way the data are assigned to each state before adaptation.

Table 3 show that, in our experiments, the SA approach used in conjunction with MAP training, causes some minor degradations of the HTER in batch mode and yields equivalent performance as asynchronous alignment, in incremental mode. In these experiments, α was chosen equal to 0.5, as the globally optimal value on the test set in MAP batch mode.

6. SCORE NORMALIZATION

Likelihood ratio normalization has proven to be a factor of noticeable improvement in speaker verification, especially the z-norm technique [LP88]. The z-norm approach assumes the gaussianity of the impostor score distribution and uses an external impostor population to center and normalize this distribution for each client. The adjusted log-likelihood ratio is then computed as :

$$\tilde{s}_X(Y) = \frac{s_X(Y) - \mu_X}{\sigma_X}$$

where μ_X and σ_X are respectively the mean and standard deviation of the impostor score, as estimated on the external impostor population. This normalization procedure usually shows experimentally a beneficial effect on the accuracy of the decision based on a client-independent threshold.

In the conventional z-norm technique μ_X and σ_X are estimated and applied at the whole utterance level. However, the SA approach makes it possible to generalize it at the word, phoneme and even state level, because the likelihood ratio obtained in SA mode can be rewritten as the sum of frame-based likelihood ratios, which can themselves be grouped by states, phonemes or words :

$$s_X^{\mathrm{SA}}(Y) = \frac{1}{N} \sum_t \log \frac{\hat{\mathcal{P}}(y_t|X,S_t)}{\hat{\mathcal{P}}(y_t|\bar{X},S_t)}$$

$$= \sum_{y_t \in U_k} \beta_k \sum_t \log \frac{\hat{\mathcal{P}}(y_t|X,U_k)}{\hat{\mathcal{P}}(y_t|\bar{X},U_k)} = \sum_k \beta_k s_X^{\mathrm{SA}}(Y,U_k)$$

where the utterance Y is assumed to be composed of N frames y_t, where S_t denotes the state to which frame y_t is assigned in the SA procedure and where U_k denotes a set of states (i.e word, phoneme or single state) to which we suppose N_k frames where assigned ($\beta_k = N_k/N$ and $\sum \beta_k = 1$).

In this context, the z-norm approach can be generalized by estimating and applying a different normalization for each unit U_k :

$$\tilde{s}_X^{\mathrm{SA}}(Y) = \sum_k \beta_k \tilde{s}_X^{\mathrm{SA}}(Y,U_k) = \sum_k \beta_k \frac{s_X^{\mathrm{SA}}(Y,U_k) - \mu_X(U_k)}{\sigma_X(U_k)}$$

In our experiments, we compare 3 different levels of normalization : the conventional z-norm where the normalization parameters are only client-dependent (but independent of the word and of the state sequence), the word-dependent z-norm (wz-norm) for which distinct means and variances are estimated and

used for the different words on which our tests are carried out and finally a state-sequence dependent z-norm (ssz-norm) where a distinct mean and variance is estimated and applied to each state.

We compare the performance of estimates of the means and variances obtained either on an external impostor population (33 pseudo-impostors, with one access per speaker) or on the test impostor population (i.e the other clients) : we denote these 2 configurations as *a priori* vs *a posteriori* normalization, the latter being naturally unrealistic in a real application.

		Unit-based estim.		Frame-based estim.	
norm	HTER	A post.	A prio.	A post.	A prio.
no	opt	5.73	5.73	5.73	5.73
	dev	5.79	5.79	5.79	5.79
z-	opt	4.87	5.53	4.89	5.65
	dev	4.96	5.57	4.91	5.70
wz-	opt	4.03	6.29	4.06	5.67
	dev	4.09	6.32	4.09	5.75
ssz-	opt	5.72	7.38	5.17	7.11
	dev	5.76	7.41	5.22	7.25

Table 4: Comparison of normalization schemes.

Furthermore, we compare *unit-based* estimates (obtained by averaging means and variances estimated on separate utterances) versus *frame-based* estimates (obtained by pooling together all the speech frames belonging to a same unit, before computing the mean and variances),

Table 4 summarizes the normalization results obtained in these various configurations, in the case of 2 sessions MAP batch training and Synchronous Alignment.

This latter series of results shows that a noticeable benefit can be obtained by standard (word-independent) z-normalization bringing the HTER down from 5.8 % to 5.6 %. A further gain could seemingly be obtained by wz-norm if the mean and variance parameters were estimated accurately (see *a post.* results). However, this potential advantage is not observed with *a priori* estimates, perhaps because of the limited amount of external impostor data. Moreover, a very interesting result is observed with ssz-norm : even with an ideal (a posteriori) estimation of the means and variances of the impostor scores, the state-based normalization performs very poorly, which is bound to come from the fact that the gaussianity of the scores may not be verified at the state-level.

7. CONCLUSIONS

On short utterances like the command words of the Polyvar database (i.e 1 to 5 syllables), the Picassoft platform yields 5.6 % HTER for 2 training sessions. This error rate can be brought down to 3.3 % with 3 additional sessions in incremental training mode. Beside the Incremental MAP training scheme, several innovative schemes have been developed in the context of the Picasso project, in particular Synchronous Alignment and Unit-based Normalization. In spite of the fact that these schemes look attractive, they have yielded only a marginal benefit (when any) in terms of error rate reduction, in our experiments. Further investigations, including experiments on other databases and tasks, are necessary to conclude on their exact impact in terms of performance and robustness.

ACKNOWLEDGEMENTS

This work was funded by OFES (Office Fédéral de l'Education et de la Science), project No. 97.0494-2 and by the CE (Commission Européenne) Telematics Program LE4 (project 8369).

8. REFERENCES

[B+99] F. Bimbot et al. An overview of the picasso project research activities in speaker verification for telephone applications. In *6th european conference on speech communication and technology — eurospeech'99*, volume 5, pages 1963–1966, Budapest, Hungary, September 5–10 1999.

[F+00] C. Fredouille et al. Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. In *ICASSP2000 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 5–9 2000.

[GL94] J. L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains. In *IEEE Transactions on Speech Audio Processing*, volume 2, pages 291–298, April 1994.

[LP88] Kung-Pu Li and Jack E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *ICASSP1988*, pages 595–597, 1988.

[M+99] J. Mariéthoz et al. Client / world model synchronous alignement for speaker verification. In *6th European Conference on Speech Communication and Technology — Eurospeech'99*, Budapest, Hungary, September 5–10 1999.

[Rey97] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Eurospeech 97*, volume 2, pages 963–966, 1997.