CLIENT / WORLD MODEL SYNCHRONOUS ALIGNEMENT FOR SPEAKER VERIFICATION

Johnny MARIETHOZ, Dominique GENOUD*, Frédéric BIMBOT**, Chafic MOKBEL**
* IDIAP, rue Simplon 4, CH-1920 Martigny, Switzerland
** IRISA (CNRS & INRIA), Campus Universitaire de Beaulieu, 35042 Rennes cedex, France
marietho@idiap.ch, http://www.idiap.ch

ABSTRACT

In speaker verification, two independent stochastic models, i.e. a client model and a non-client (world) model, are generally used to verify the claimed identity using a likelihood ratio score. This paper investigates a variant of this approach based on a common hidden process for both models. In this framework, both models share the same topology, which is conditioned by the underlying phonetic structure of the utterance. Then, two different output distributions are defined corresponding to the client vs. world hypotheses. Based on this idea, a synchronous decoding algorithm and the corresponding training algorithm are derived. Our first experiments on the SESP telephone database indicate a slight improvement with respect to a baseline system using independent alignments. Moreover, synchronous alignment offers a reduced complexity during the decoding process. Interesting perspectives can be expected.

Keywords : *Stochastic Modeling, HMM, Synchronous Alignment, EM algorithm.*

1 INTRODUCTION

Many applications can use a speaker verification system to secure private information. Such systems verify the identity of a claimed client on the basis of some speech utterances. To perform the verification, client and non-client (world) models are generally computed in an enrollment phase. These models aim at discriminating between the client and impostors regarding an acoustic realization.

Speech signal conveys different information such as the pronounced words or the speaker characteristics. It is very difficult to separate these information. Thus, speaker verification systems are generally classified following their degree of dependence on the pronounced text: text-dependent, text-prompted, and text-independent systems.

Beside this classification, it is classically observed that the speaker recognition performance generally increases when introducing more knowledge about the underlying text. In this work we are particularly focusing on text-dependent speaker recognition where client and world HMMs are used to model the passwords for the client and the non-client speakers. Using two separate models does not explicitly take into account the fact that the password phonetic structure is similar for the speaker and the impostors.

This motivates the study of a *synchronous alignment* approach where the hidden process (i.e the sequence of states) is supposed identical for both client and non-client. Only the output distributions differ between the two hypotheses. The synchronous alignment approach is depicted and compared to the classical one on Figure 1.

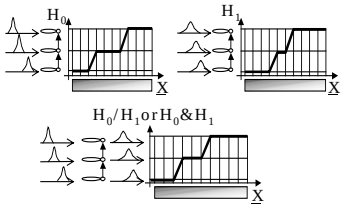

Figure 1: The synchronous alignment approach compared to the classical approach.

Besides the theoretical motivation, such structure has an important practical advantage since only one decoder is used instead of two.

In section 2, the synchronous alignment approach is detailed. Different criteria that might be used in this context are discussed. The corresponding decoding and training algorithms are also presented. In section 3, experiments conducted to validate the approach and the corresponding results are described. These experiments were conducted with a state-of-the-art system, the PICASSoft system [1] derived from the CAVE Generic System [2]. Finally, the main conclusions and the principal perspectives are drawn in section 4.

2 SYNCHRONOUS ALIGNMENT

The main idea of synchronous alignment is to make the two models share the same topology and differ in the output distributions. In order to compute the optimal path in the shared model, a global criterion is defined. Two possible criteria are proposed in this section. Specific decoding and training algorithms for both criteria are derived. The convergence properties of such algorithm are studied and the results are presented here.

2.1 Criteria for synchronous alignment

If the paths are shared between the models of the two hypotheses, a criterion must be defined in order to determine the optimal path. Two main directions can be followed :

- The criterion reflects how much the client model is more likely than the non-client model. This is a *discriminative approach* where the optimal path corresponds to the highest client likelihood and lowest anti-client likelihood.
- The criterion reflects how well the path is simultaneously good for both the client and the non-client models. A *joint likelihood* function can be used for this purpose.

Let $\underline{X}$ denote the sequence of input feature vectors of length *T* corresponding to the utterance pronounced by the speaker to be verified, λ denote the underlying common model structure, θ_{client} and θ_{world} denote the client and world parameters respectively and, finally *S* denote a possible path in the model. The discriminative criterion is based on a weighted likelihood ratio. The optimal path can be found following this equation :

$$\textbf{(1)}\quad \hat{S} = \arg\max_{S} \left(\frac{p(\underline{X}, S/\theta_{client}, \lambda)^{\alpha}}{p(\underline{X}, S/\theta_{world}, \lambda)^{\beta}} \right) \quad with \quad \alpha + \beta = 1 \quad for \quad 0 \le \alpha \le 1$$

where α and β are weighting factors.
For the joint likelihood criterion the optimal path must be found in order to satisfy :

$$\textbf{(2)}\quad \hat{S} = \arg\max_{S} p(\underline{X}, S/\theta_{client}, \lambda)^{\alpha}\ p(\underline{X}, S/\theta_{world}, \lambda)^{\beta} \quad with \quad \alpha + \beta = 1 \quad for \quad 0 \le \alpha \le 1$$

where α and β are weighting factors.
In the rest of the paper we refer to α as the "*sync factor*".

2.2 Decoding

A decoding scheme has been developed to compute the optimal path following the two criteria described in the previous subsection 2.1. A variant of the Viterbi algorithm must be developed in order to maximize the argument in Equation (1) or (2).

2.2.1 Discriminative criterion

For the discriminative criterion, the argument to maximize in (1) can be written :

$$\textbf{(3)}\quad \frac{p(\underline{X}, S/\theta_{client}, \lambda)^{\alpha}}{p(\underline{X}, S/\theta_{world}, \lambda)^{\beta}} = \prod_{t=1}^{T} a_{S_{t-1}S_t}^{\alpha-\beta} \cdot \frac{b_{client,S_t}(\underline{X}_t)^{\alpha}}{b_{world,S_t}(\underline{X}_t)^{\beta}} \quad with \quad \alpha + \beta = 1 \quad for \quad 0 \le \alpha \le 1$$

where $a_{S_{t-1}S_t}$ represents the transition probability and is supposed to be identical for the speaker and the world parameters and, $b_{client,S_t}()$ and $b_{world,S_t}()$ are the client's and respectively the world's output distributions relative to the state S_t.
By replacing Eq. (3) into Eq.(1), it appears that the Viterbi algorithm can be used for decoding by :

- taking the transition probas at the power α - β,
- replacing for each frame the log-likelihood of an output distribution by the difference between the weighted log-likelihoods of the client and the world output distributions.

Since the discriminative criterion is mainly based on the idea that the predominant information in the measured features is relative to the speaker, a problem exists when decoding with a silence. These parts of the signal do not include any information about any speaker and the discriminative criterion is not justified. Thus we propose to first decode the signal on the world model and remove the portions corresponding to the silence. Only the speech portions of the signal are decoded using the discriminative synchronous alignment algorithm. In our experiments this procedure will be refered to as *without silence* as opposed to the standard procedure.

2.2.2 Joint likelihood criterion

For the joint client/non-client likelihood decoding, the argument to maximize in Eq.(2) can be written :

$$\textbf{(4)}\quad p(\underline{X}, S/\theta_{client}, \lambda)^{\alpha} \cdot p(\underline{X}, S/\theta_{world}, \lambda)^{\beta} = \prod_{t=1}^{T} a_{S_{t-1}S_t} . b_{client,S_t}(\underline{X}_t)^{\alpha} . b_{world,S_t}(\underline{X}_t)^{\beta} \quad with \alpha + \beta = 1 and\ 0 \le \alpha \le 1$$

where the notations are the same as for the Eq. (3).
Here too, replacing Eq. (4) in Eq. (2), shows that the Viterbi algorithm can be used for decoding in the joint likelihood synchronous alignment approach. The only modification consists in replacing, for each frame, the log-likelihood of an output distribution by a linear combination of two log-likelihoods. These two log-likelihoods correspond to the client and the world output distributions. In the joint optimization scheme, the transition probabilities remain unchanged.
In summary, for both discriminative and joint likelihood criteria, the decoding can be performed using a variant of the classical Viterbi algorithm. This offers an important practical advantage with respect to the classical method : decoding in a unique model can be performed.

2.3 Training

For synchronous alignment, the models can be trained as classically. However, this is not fully consistent with the decoding process. Thus, a specific training algorithm has been developed. It permits to compute the client's parameters given some utterances of the password from the client. The parameters relative to the non-client or the world are supposed to be known and are not changed during the enrollment. An adaptation of the Viterbi-based variant of the "Estimation-Maximization" (EM) algorithm, is developed for this purpose.
The same criterion used during the decoding is used to train the client's parameters. Let *K* be the number of available enrollment utterances from the client. In the case of the discriminative synchronous alignment, the optimal client parameters must satisfy :

$$\textbf{(5)}\quad \hat{\theta}_{client} = \arg\max_{\theta_{client}} \prod_{k=1}^{K} \max_{S^{(k)}} \left(\frac{p(\underline{X}^{(k)}, S^{(k)}/\theta_{client}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, S^{(k)}/\theta_{world}, \lambda)^{\beta}} \right) \quad for\ \alpha + \beta = 1\ and\ \ 0 \le \alpha \le 1$$

In the case of joint likelihood synchronous alignment, the optimal client parameters must satisfy :

(6)

$$\hat{\theta}_{client} = \arg\max_{\theta_{client}} \prod_{k=1}^{K} \max_{S^{(k)}} \left(p(\underline{X}^{(k)}, S^{(k)} / \theta_{client}, \lambda)^{\alpha} \cdot p(\underline{X}^{(k)}, S^{(k)} / \theta_{world}, \lambda)^{\beta} \right)$$

$$with \quad \alpha+\beta=1 \quad and \quad 0 \le \alpha \le 1$$

2.3.1 Discriminative criterion

For iteration *n*, the optimal client's parameters at the previous iteration $\hat{\theta}_{client}^{(n-1)}$ are known. Corresponding optimal paths can be obtained using the decoding algorithm (subsection 2.2). This is the estimation stage of the EM procedure. The optimal path for the k^{th} utterance satisfies :

(7) $$\hat{S}^{(k)(n-1)} = \arg\max_{S^{(k)}} \left(\frac{p(\underline{X}^{(k)}, S^{(k)} / \hat{\theta}_{client}^{(n-1)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, S^{(k)} / \theta_{world}, \lambda)^{\beta}} \right)$$

Along the optimal path, new values of the client's parameters can be obtained by maximizing the likelihood ratio. The re-estimation equations can be derived from the optimization :

(8)

$$\hat{\theta}_{client}^{(n)} = \arg\max_{\theta_{client}} \prod_{k=1}^{K} \left(\frac{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{client}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{world}, \lambda)^{\beta}} \right)$$

$$= \arg\max_{\theta_{client}} \prod_{k=1}^{K} p(\underline{X}^{(k)} / \hat{S}^{(k)(n-1)}, \theta_{client}, \lambda)$$

Looking at Eq. (8), the maximization step of the EM procedure is equivalent to the classical one in the HMM training. Thus, in the case of discriminative training, the re-estimation equations are the same as those of classical training with the Viterbi-based EM algorithm.

2.3.2 Joint likelihood criterion

Given the estimate of the client's parameters $\hat{\theta}_{client}^{(n-1)}$ at the end of iteration $n-1$, the optimal path can be found for the training utterances. This is done using the synchronous alignment Viterbi decoding as described in subsection 2.2. The optimal path for the k^{th} utterance is the solution of :

(9) $$\hat{S}^{(k)(n-1)} = \arg\max_{S^{(k)}} \left(p(\underline{X}^{(k)}, S^{(k)} / \hat{\theta}_{client}^{(n-1)}, \lambda)^{\alpha} \cdot p(\underline{X}^{(k)}, S^{(k)} / \theta_{world}, \lambda)^{\beta} \right)$$

Given these estimated optimal paths, new estimate of the client's parameters can be obtained in the maximization step. Maximizing the joint likelihood :

(10)

$$\hat{\theta}_{client}^{(n)} = \arg\max_{\theta_{client}} \prod_{k=1}^{K} p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{client}, \lambda)^{\alpha} \cdot p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{world}, \lambda)^{\beta}$$

$$= \arg\max_{\theta_{client}} \prod_{k=1}^{K} p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{client}, \lambda)$$

As for the discriminative criterion, Eq. (10) shows that the re-estimation equations are equivalent to those of classical training with Viterbi-based EM. Thus, we can conclude that, for both criteria, training is similar to the classical training. The only difference resides in the estimation step where the optimal paths are found using the synchronous alignment Viterbi decoding algorithm.

2.4 Convergence properties

For the Viterbi-based EM algorithm, it can be proved that, for the training utterances, the joint likelihood over the optimal path increases when the number of iterations increases. This can also be shown for the training within the synchronous alignment approach. We show this for the case of discriminative training. The proof for the joint likelihood approach is straightforward.

Eq. (8) yields :

(11)

$$\prod_{k=1}^{K} \frac{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \hat{\theta}_{client}^{(n)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{world}, \lambda)^{\beta}} \ge \prod_{k=1}^{K} \frac{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \hat{\theta}_{client}^{(n-1)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{world}, \lambda)^{\beta}}$$

$$\Rightarrow \prod_{k=1}^{K} \frac{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \hat{\theta}_{client}^{(n)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{world}, \lambda)^{\beta}} \ge \prod_{k=1}^{K} \max_{S^{(k)}} \frac{p(\underline{X}^{(k)}, S^{(k)} / \hat{\theta}_{client}^{(n-1)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, S^{(k)} / \theta_{world}, \lambda)^{\beta}}$$

Moreover :

(12) $$\prod_{k=1}^{K} \max_{S} \frac{p(\underline{X}^{(k)}, S / \hat{\theta}_{client}^{(n)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, S / \theta_{world}, \lambda)^{\beta}} \ge \prod_{k=1}^{K} \frac{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \hat{\theta}_{client}^{(n)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, \hat{S}^{(k)(n-1)} / \theta_{world}, \lambda)^{\beta}}$$

Inequalities (11) and (12) can be combined into :

(13) $$\prod_{k=1}^{K} \max_{S} \frac{p(\underline{X}^{(k)}, S / \hat{\theta}_{client}^{(n)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, S / \theta_{world}, \lambda)^{\beta}} \ge \prod_{k=1}^{K} \max_{S} \frac{p(\underline{X}^{(k)}, S / \hat{\theta}_{client}^{(n-1)}, \lambda)^{\alpha}}{p(\underline{X}^{(k)}, S / \theta_{world}, \lambda)^{\beta}}$$

Inequality (13) shows that while iterations progress, the quantity to maximize in the criterion of Eq. (5) increases. The convergence to a local maximum is thus expected.

2.5 Scoring

As mentioned previously, decision is generally taken by comparing the likelihood ratio to a predefined threshold. With discriminative synchronous alignment, the likelihood ratio can be obtained directly for $\alpha = 0.5$ (by a multiplicative constant). This is not generally the case for different values of α or for the joint likelihood criterion. Given the optimal alignment provided at the end of the decoding process, the likelihood ratio can be recomputed with different normalization methods : Sum, Mean-0, Z-norm. Please refer to [3] for more details on these normalization techniques. Results presented in this paper are obtained with the Sum normalization.

3 EXPERIMENTS AND RESULTS

Experiments were conducted within the European PICASSO project on the SESP task defined during the CAVE project [1]. The corpus of the SESP database is composed of connected Dutch digits uttered by 48 speakers (24 male and 24 female). The PICASSO reference system (Picassoft), derived from the CAVE-WP4 reference system (Genesys), is based on state-of-the-art approaches.

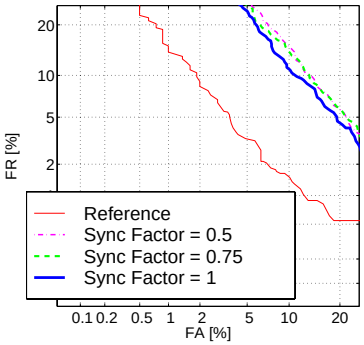

Figure 2 : Results for different sync factors (α) with the discriminative approach. Synchronous alignment is done during training. Silences removed before decoding.

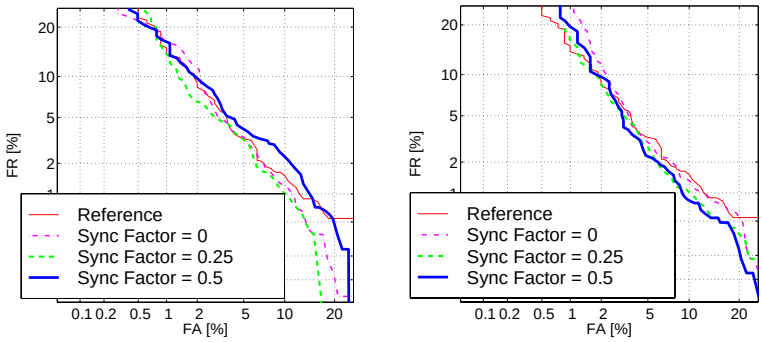

Figure 3 : Results for different sync factors (α) with the joint likelihood approach. Independent training (left) vs. synch. alignment training (right). Silences are not removed

In our experiments, speaker verification is performed in text-dependent mode. Left-right HMMs with 2 states per phoneme are used to define the client and world-models. The output distribution are modeled by 3 Gaussian mixtures per state. Feature vectors are extracted from 25ms signal frames with a 10ms shift. These vectors are formed of 12 LPCC and the Energy on log scale plus first and second derivatives. The results are presented as DET curves [4] showing false rejection rates as a function of false acceptance rates.

Figure 2 presents a summary of the results obtained with the discriminative criterion. In Figure 2, the results are provided in the case when the client models are trained with the synchronous alignment algorithm as described in the subsection 2.3 (the results obtained with independent training of the client and world models are worse).

Figure 3 presents the performance for joint likelihood synchronous decoding with different sync factors. These results are given for both independent and synchronous alignment training. Although the results are much better than those obtained with the discriminative criterion, no improvement is obtained over the baseline reference. With synchronous alignment training, the best results are obtained with a synchronous factor of 0.25. A slight improvement with respect to the reference system is noticed, although it may not be statically significant. It shows however that the synchronous alignment approach is at least promising. Moreover, the synchronous alignment approach has the advantage of a simpler decoding process.

In summary, the synchronous alignment approach provides slight improvements on the SESP database when the joint likelihood criterion is used. A more comprehensive study can be found in [5].

4 CONCLUSIONS

Two main sources of information are expressed in a speech signal: the underlying text and the speaker characteristics. In general speaker recognition performance increases when considering the underlying text during the modeling process. This was one motivation for the development of the synchronous alignment method. This method considers a speech utterance as the result of a single hidden process common to the client and world model and therefore associates two sets of output distributions to a single HMM automaton. In this framework, we studied two criteria. The discriminative criterion assumes that the speaker information is predominant in the signal. In contrast, the joint likelihood criterion searches for an optimal path that maximizes a joint likelihood of both hypotheses assuming that the underlying text is the predominant information in the signal. We also derived a decoding algorithm and a training algorithm for both criteria (these algorithms were implemented within the HTK toolkit). They have been experimented on the databases of the PICASSO project. They have been compared to state-of-the-art speaker verification techniques. The results show that equivalent results (or slight improvement) can be obtained with the joint likelihood criterion. This offers the advantage of both a cheaper decoding algorithm and a more consistent interpretation of the frame-based terms in the likelihood ratio. An other significant result of this work is that the discriminative criterion has limited performance. This tends to show that the predominant information in the speech signal is the underlying text, and opens the door for other applications of this approach, in particular in speech recognition.

5 ACKNOWLEDGEMENTS

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

6 REFERENCES

[1] F. Bimbot et al., "Robust approaches to speaker verification on the telephone : an overview of the PICASSO project activities", Eurospeech 1999.

[2] 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," Eurospeech 97, pp. 971-974, 1997.

[3] G. Gravier and G. Chollet : "Comparison of normalisation techniques for speaker verification", RLA2C, pp. 97-100, 1998.

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

[5] J. Mariéthoz, C. Mokbel : "Synchronous alignment", IDIAP Research report # 99-06, 1999.