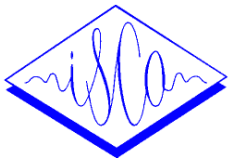

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

E-HMM approach for learning and adapting sound models for speaker indexing

Sylvain Meignier[*], *Jean-François Bonastre, Stéphane Igounet*

LIA/CERI Université d'Avignon, Agroparc,
BP 1228, 84911 Avignon Cedex 9, France.
{sylvain.meignier, jean-francois.bonastre}@lia.univ-avignon.fr,
stephane.igounet@univ-avignon.fr.

Abstract

This paper presents an iterative process for blind speaker indexing based on a HMM. This process detects and adds speakers one after the other to the evolutive HMM (E-HMM). The use of this HMM approach takes advantage of the different components of AMIRAL automatic speaker recognition system (ASR system: frontend processing, learning, loglikelihood ratio computing) from LIA. The proposed solution reduces the miss detection of short utterances by exploiting all the information (detected speakers) as soon as it is available.

The proposed system was tested on *N-speaker* segmentation task of NIST 2001 evaluation campaign. Experiments were carried out to validate the speakers detection. Moreover, these tests measure the influence of parameters used for speaker models learning.

1. Introduction

Seeking within a recording the speech sequences uttered by a given speaker is one of the main tasks of document indexing. Segmentation systems first detect breaks in audio streams and then cluster in homogeneous sound classes the segments according to those breaks.

In automatic speaker recognition, segmentation consist in finding all the speakers, as well as the beginning and the end of their contributions. The speaker segmentation problem is commonly approached by two methods.

The first method (described in [1] and [2]) is composed of two steps. The former locates the signal breaks which are caused by speakers changes. The latter determines and labels the utterances using a clustering algorithm.

The second method (as done in [3] and [1]) uses an automatic speaker recognition (ASR) system. Breaks detections and clustering tasks are carried out simultaneously. The system has to determine speakers present within a given message as well as the utterances of each of them.

[*]RAVOL project: financial support from Conseil général de la région Provence Alpes Côte d'Azur and DigiFrance.

No *a priori* information on speakers is used in these two approaches, *i.e.* the speaker models have to be built during the process. Therefore these methods are well adapted to the tasks of blind segmentation.

In this article, we propose a system adapted from the second method for blind segmentation tasks. The conversation is modeled by a Markov Model (like [3]). During the segmentation process, the Markov Model is expanded with each new detection of sound class.

The proposed system was tested on *N-speaker* segmentation task of NIST 2001 evaluation campaign [9] which uses the CALLHOME database. The experiments in this paper are done using a half of this database to select the best fitting parameters for speaker models learning. The other half remains to validate the choice of those parameters.

2. Segmentation model

2.1. Structure of the segmentation model

The signal to segment consists in a sequence of observation vectors $O = (o_1, o_2, ..., o_T)$.

The changes of sound classes are represented by a hidden Markov model (HMM). In this application the sound classes represent a speaker. Each HMM state characterize a class of sound and their transitions model the changes of classes.

The HMM λ is defined by (E, A, B) :

- Let $E = \{1, 2, ..., N\}$ be a set of states.
- Let $A = \{a_{i,j}\}$ be a set of transition probabilities between the states.
- Let B be the set of $\{b_i\}$. Let the state i be associated with a sound model C_i of the sound class $\mathcal{C}_i$. Each state i is then associated with a set b_i of emission probabilities according to C_i. Let o_t be an observation from O, $b_i(o_t)$ is the probability calculated from C_i for o_t.

The HMM is fully connected.

Transition probabilities are established according to a set of rules complying with the three following conditions:

$$\begin{cases} \forall i,\ a_{i,i} = \gamma \\ \forall (i,j),\ i \neq j,\ a_{i,j} = \frac{1-\gamma}{N-1} \\ 0 < \gamma < 1 \end{cases} \quad (1)$$

2.2. Detection of sound classes and segmentation model building

The Segmentation model is generated by an iterative process, which detected and added a new state i at each stage (i.e. i). We refer to evolutive HMM as E-HMM [4].

At the process initialization stage $i = 1$ figure 2, HMM is $\lambda^1 = (E^1, A^1, B^1)$. $E^1 = \{1\}$ is composed of a single state "1", which is associated with a sound model C_1([1]) learned from the whole signal O. At the end of the initialization process, a first trivial segmentation $s^1 = (s_1^1, ..., s_T^1) = (1, ..., 1)$ is generated. In fact, each observation o_i is simply labeled with the only sound class $\mathcal{C}_1$. This segmentation s^1 is composed of a single segment which will be challenged at the following stages.

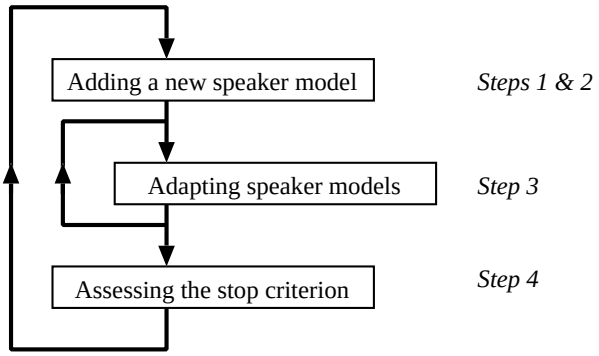

Figure 1: Diagram of the segmentation process.

The process (e.g. stage 2 & 3 figure 2) is divided in 4 steps (figure 1) for each stage i ($i > 1$) :

Step 1 A new state i is added to the set E^{i-1} ($E^i = E^{i-1} \cup i$). Transition probabilities are adapted to take into account the new number of states. Then, the new HMM $\lambda^i = (E^i, A^i, B^{i-1})$ is obtained.

Step 2 The sound model C_i is estimated from a subset of observation[2] $(o_r, o_{r+1}, ..., o_{r+t})$. r is selected such as :

$$\begin{cases} r = \underset{j \in L}{\text{ArgMax}} \left(\prod_{k=j}^{j+t} b_1(o_k) \right) \\ L = \{ j \in \{1, ..., T\} | s_j^{i-1} = s_{j+1}^{i-1} = ... = s_{j+t}^{i-1} = 1 \} \end{cases} \quad (2)$$

then, $r \in L$ is the rank of the subset $(o_r, o_{r+1}, ..., o_{r+t})$ maximizing the probabilities product for the sound class $\mathcal{C}_1$.

Moreover, the segmentation s^i is computed: the subset $(o_r, o_{r+1}, ..., o_{r+t})$ is relabeled to the sound class $\mathcal{C}_i$.

$$\begin{cases} s_j^i = s_j^{i-1}\ \forall j \notin \{r, ..., r+t\} \\ s_r^i = s_{r+1}^i = ... = s_{r+t}^i = i \end{cases} \quad (3)$$

Step 3 In this step, the process iteratively adapts the parameters of HMM λ^i :

(a) For each k in $\{1, ..., i\}$, the sound model C_k is adapted according to data which were affected to it in the segmentation s^i.

(b) The set B^i of emission probabilities are recomputed.

(c) Viterbi algorithm is applied to obtain a new version of segmentation s^i according to the HMM.

The Viterbi path $P(s^i|A^i, B^i, O)$ is computed.

$$P(s^i|A^i, B^i, O) = b_{s_1^i}(o_1) \times \prod_{j=2}^{T} (a_{s_{j-1}^i, s_j^i} \times b_{s_j^i}(o_j)) \quad (4)$$

If a gain is observed between two loops in 3, the process returns to (a)

Step 4 Lastly, the stop criterion is assessed: if

$$P(s^i|A^i, B^i, O) > P(s^{i-1}|A^i, B^{i-1}, O) \quad (5)$$

then a new stage starts back to step "1".

Note: the probability of s^{i-1} is reestimated with the transition A^i of the model λ^i, because topologies of segmentation models λ^i and λ^{i-1} must be comparable.

3. Automatic Speaker recognition System

The sound models and emission probabilities are calculated by the AMIRAL ASR system developed at LIA [5]. Emission probabilities are computed on fixed-length blocks of 0.3 second. Each emission probability is normalized by the world model.

[1] C_1 is the sound model for the not yet detected speakers (i.e. all).

[2] In this work, each subset has a 3 sec. duration (i.e. t is fixed).

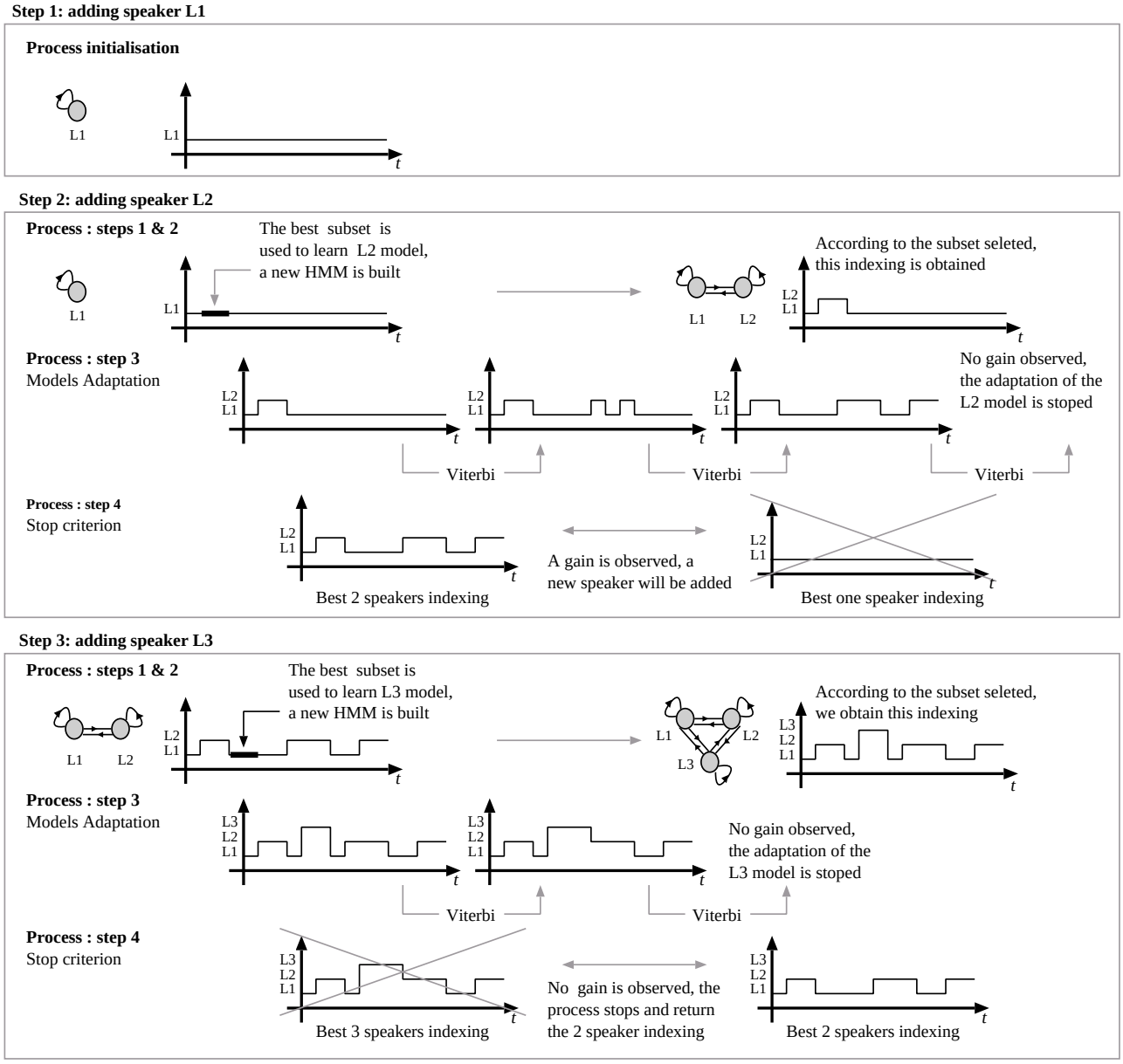

Figure 2: Example of segmentation for a 2 speakers test.

Acoustic parameterization (16 cepstral coefficients and 16 Δ-cepstral coefficients) is carried out using the SPRO module developed by the ELISA consortium [3] [10].

The sound classes are modeled by gaussian mixture models (GMM) with 128 components and diagonal covariance matrices [7], adapted from a background model.

The sound model C is first estimated over a subsequence of 3 seconds (sec. 2.2 - Step 2). Then, the sound model C is adapted from the segments which are labeled by the sound class $\mathcal{C}$ (sec. 2.2 - Step 3a).

The adaptation scheme for training speaker models is based on the *maximum a posteriori* method (MAP). For each Gaussian g, Mean μ_g of sound model C is a linear combination between the estimated $\widehat{\mu}_g$ and the corresponding mean μ_g^W in background model W. Mean $\widehat{\mu}_g$ is estimated according to the data of sound class $\mathcal{C}$:

$$\begin{cases} \mu_g = \alpha\mu_g^W + (1-\alpha)\widehat{\mu}_g \\ \alpha > 0 \end{cases} \quad (6)$$

Neither the weights, nor covariance matrices are

[3] ELISA consortium is composed of European research laboratories which work on a common platform. Members of ELISA for the participation to NIST 2001 is : ENST (France), IRISA (France), LIA (France), RMA (Belgium).

adapted. The sound model C uses the weights and covariance matrices of the background model.

4. Experiments

The proposed approach was experimented on the *N-speakers* segmentation task during NIST 2001 evaluation campaign [9]. The results are shown in sec. 4.5. Moreover, development experiments on learning method are reported in sec. 4.4.

4.1. Databases

N-speakers evaluation corpus (described in [8] and [9]) is composed of 500 conversational speech tests drawn from CALLHOME corpus. Tests of varying length (< 10 minutes) are taken from 6 languages. The exact number of speakers is not provided (but is less than 10).

The NIST Corpus is divided in two parts of 250 tests named *Dev* and *Eva*.

NIST provided a separate development corpus (named *train_ch*) composed of 48 conversational speech samples extracted from CALLHOME corpus. The *train_ch* corpus permitted to learn the background model (*wld_ch*).

A second separate development data set (*train_sb*) is composed of 472 trials made of up to 100 speakers from Switchboard 2. *train_sb* permitted to learn the background model *wld_sb*.

4.2. Experiments

Experiments were carried out to estimate the influence of α parameter of MAP learning, applied on both *wld_ch* and *wld_sb* background models.

Moreover, a reference experiment based on a trivial segmentation (only one segment) named *trivial* is generated.

The results are obtained with parameters:

- The transition probabilities are estimated with $\gamma = 0.6$ (Eq. 1).
- The MAP parameters (Eq. 6) are:

  $\alpha \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$

4.3. Scoring measures

Two evaluation measures are considered:

- The mean m_{diff} of differences between the estimated number of speakers e_n and the real number of speakers r_n in the test n.

$$m_{diff} = \frac{\sum_{n=1}^{N} e_n - r_n}{N} \qquad (7)$$

- NIST speaker segmentation scoring is described in [9]. Scoring is computed on the NIST reference segments with only one speaker speaking. This score corresponds to a segmentation error.

4.4. Results

In order to compare the influence of α parameter, figures 3 and 4 are shown respectively the segmentation scores and the mean m_{diff} obtained with both background models. The best results are presented in tables 1 and 2 (respectively for *Dev* and *Eva*).

The results on *Eva* corpus is close to the results obtained on *Dev* corpus.

When the weight is close to 1, the result becomes equivalent to the trivial segmentation. The systems mainly attribute the data to only one class, besides the process does not add new speaker models (m_{diff} is near -1.5).

When the weight is 0, the adaptation learning becomes equivalent to a EM-ML training with one iteration [6], but initialized using the corresponding background model. This weight gives the best result (24.01%) for the *wld_sb* background model. Although sufficient data is provided to compute background model, the *train_sb* data is very different from the data of CALLHOME corpus (*Dev* and *Eva*).

The background model *wld_cl* take advantage of the MAP learning method. The best result (25.5%) is obtained for an $\alpha = 0.3$. Few data is used to learn this background model which is not efficient to generalize *Dev* and *Eva* data of CALLHOME corpus.

For both background models, the mean m_{diff} of differences between the estimated number of speakers and the real number of speakers is quite good (near 0.5 for best results with *wld_ch* and *wld_sb*).

Corpus	background model	α	score (%)	m_{diff}
Dev	*wld_sb*	0	24.01	0.58
Dev	*wld_ch*	0.3	25.50	0.71

Table 1: Best results for the background models

Corpus	background model	α	score (%)	m_{diff}
Eva	*wld_sb*	0	23.42	0.54
Eva	*wld_ch*	0.3	25.17	0.44

Table 2: Validation of the results obtain on Dev Corpus

4.5. NIST 2001 results

Two systems were presented to NIST 2001 *N-speaker* segmentation. They use *wld_sb* background model to

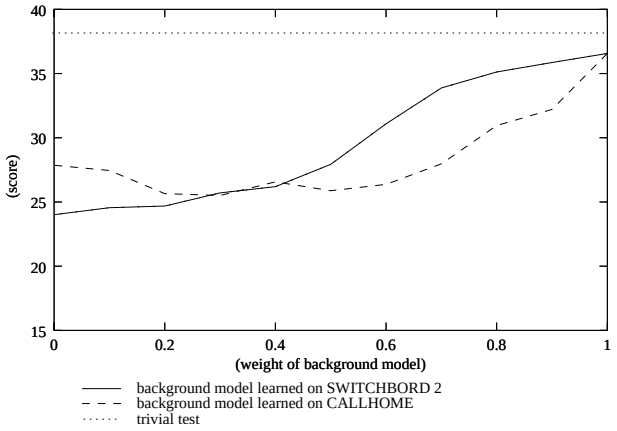

Figure 3: NIST *N-speakers* segmentation score (%): influence of α parameter.

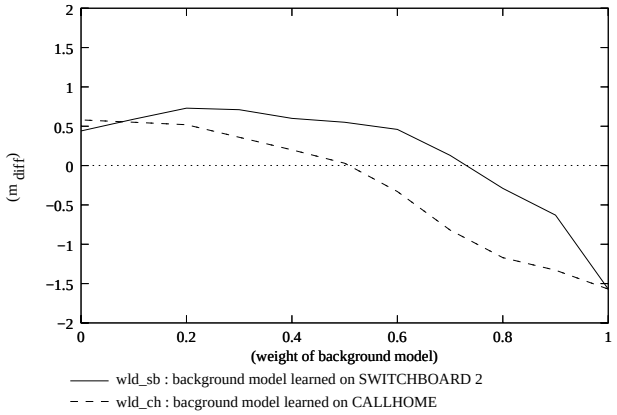

Figure 4: NIST *N-speakers* m_{diff}: influence of α parameter on MAP.

adapt speaker models with an α parameter equal to 0. The difference between the system is the value of γ parameter used to compute HMM transition probabilities. In the first system *LIA10*, γ is equal to 0.6, In the second one, $\gamma = 0.5$.

Note: α and γ parameter was estimated before NIST 2001 on *train_ch* corpus according to the evaluation rules.

Tables 3 and 4 show the results of *LIA10* and *LIA20* as well as the results of the *trivial* segmentation.

Tables 3 show the scores computed by the number of speakers present. Systems are well adapted for tests where a lot of speakers speak. The number of speaker are detected correctly. However the scores (22% and 24%) is close to the score of the *trivial* system (26%) for 2 speakers tests.

LIA10 and *LIA20* scores is almost equal for the different speaker languages. The chosen learning method is well adapted when speaker language is unknown.

5. Summary

In this article, the segmentation system uses an evolutive HMM to model the conversation and to determine automatically the sound classes present in messages. The approach is based on an iterative algorithm which detects and adds the sound models one by one. At each stage, a segmentation is proposed, according to available knowledge. This segmentation is called into question at the following iteration until the optimal segmentation is reached.

Within sight of the results, the system behaves satisfactorily. Experiments showed that MAP training is well adapted for the selected task of segmentation. As for the weight between the background and the estimated sound model, it has a influence on the segmentation error.

Further work will focus on this two points, by adapting the background model data of CALLHOME to the SWITCHBOARD background model and by introducing an explicit duration model into the HMM to improve the speaker detection.

6. References

[1] P. Delacourt, D. Kryze, C.J. Wellekens. Use of second order statistic for speaker-based segmentation, *EUROSPEECH*, 1999.

[2] H. Gish, H-H Siu, R. Rohlicek. Segregation of speakers for speech recognition and speaker identification, *ICASSP*, pages 873-876, 1991.

[3] L. Wilcox, D. Kimber, and F. Chen, Audio indexing using speaker identification, *SPIE*, pages 149-157, July, 1994.

[1] K. Sönmez, L. Heck, M. Weintraub, Speaker tracking and detection with multiple speakers, *EUROSPEECH*, 1999.

[4] S. Meignier, J.-F. Bonastre, C. Fredouille, T. Merlin, Evolutive HMM for Multi-Speaker Tracking System, *ICASSP*, june 2000.

[5] C. Fredouille, J.-F. Bonastre, T. Merlin, AMIRAL: a block-segmental multi-recognizer approach for Automatic Speaker Recognition, Digital Signal Processing, Vol.10, Num.1-3, pp.172-197 Jan.-Apr. 2000.

[6] D. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Stat. Soc.*, Vol. 39, pp 1-38, 1977.

[7] D. A. Reynolds, Speaker identification and verification using gaussian mixture speaker models, *Speech Communication*, pp 91-108, Aug. 1995.

System	*Dev+Eva* Files 500	2-spkrs 303	3-spkrs 136	4-spkrs 43	5-spkrs 10	6-spkrs 6	7-spkrs 2
trivial	38	26	39	49	50	56	61
LIA10	24	22	25	23	30	35	37
LIA20	24	24	24	22	29	38	34

Table 3: NIST 2001 % scores for *N-speaker* segmentation task by speakers number in each test

System	arabic 95	english 56	german 67	japanese 68	mandarin 118	spanish 96
trivial	40	24	30	33	42	42
LIA10	24	23	19	26	25	26
LIA20	22	26	24	26	24	25

Table 4: NIST 2001 % scores for *N-speaker* segmentation task by different languages

[8] The 2000 NIST Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2000/doc/spk-2000-plan-v1.0.htm.

[9] The NIST Year 2001 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v53.pdf.

[10] Elisa Consortium, Overview of the ELISA consortium research activities, *Odyssey*, 2001. 91-108, Aug. 1995.