Evolutive HMM for Multi-Speaker Tracking System

Sylvain Meignier, Jean-François Bonastre, Corinne Fredouille, Teva Merlin

EVOLUTIVE HMM FOR MULTI-SPEAKER TRACKING SYSTEM

Sylvain Meignier, Jean-François Bonastre, Corinne Fredouille, Teva Merlin*

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

ABSTRACT

Seeking within a speech sequence the speaker utterances is one of the main tasks of indexing.

In this paper, the proposed speaker tracking system is defined in the case where all speaker identities are known beforehand. The conversation is modeled as an evolutive HMM-like model, in which speaker models computed are added one by one. A temporary indexing is proposed after each speaker adding and then challenged at the next step. This process is iterated until all the speakers are detected.

The system has been assessed using multi-speaker messages generated by concatenation of Switchboard mono-speaker segments. The obtained results show the potentiality of the proposed solution.

1. INTRODUCTION

Seeking within a recording the speech sequences uttered by a given speaker is one of the main tasks of document indexing. Speaker indexing and tracking consist in finding the number of speakers, as well as the beginning and the end of speaker contributions. In speaker indexing task, the system has no prior information about the speakers. The speaker models have to be built during the indexing process.

In this paper, we propose a speaker tracking system designed for the case where all the potential speakers are known beforehand, ie the speaker models are available. The system has to determine the subset of speakers present within a given message as well as the utterances of each of them.

The utterance sequence is modeled by a HMM (like in [1]). The originality of the proposed work consists in building the HMM using an incremental process. The speaker models are detected and added one by one to the HMM. The use of a HMM approach takes advantage of the AMIRAL speaker verification system [2]. Indeed, this system produces standardized scores in a probabilistic domain.

The proposed solution may solve the problem of multiple short interventions by exploiting all the information (detected speakers) as soon as it is available.

The system was tested on simulated multi-speaker messages computed by concatenation of mono-speaker phone conversations extracted from Switchboard database (NIST/NSA98[1]).

2. SPEAKER TRACKING SYSTEM

The conversation model is based on a HMM (section 2.2). The model states represent the speakers of the message and the transitions model the speaker turns. Each state is associated with a speaker model (section 2.1).

At the beginning of the iterative process (section 2.3), the HMM is initialized with two default states, which represent a generic speech model and a generic noise model. The following steps detect the speakers one by one. At this time, the Viterbi algorithm is applied, giving a temporary indexing which will be challenged at the next step. The process is iterated until all the speakers are detected.

2.1. Speaker verification system

2.1.1. Front-end processing

This system relies on the standard ELISA[2] consortium parameterization module. The speech signal is represented, every 10ms, by 16 cepstrum coefficients derived from filter bank analysis. Cepstral Mean Normalization (CMN) is applied to minimize channel-induced perturbations.

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

[1] http://www.nist.gov/speech/spkrec98.html

[2] The ELISA consortium is composed of European research laboratories working on a shared reference platform for the evaluation of speaker recognition systems. These labs are: ENST (France), EPFL (Switzerland), IDIAP (Switzerland), IRISA (France), LIA (France), RIMO — Rice (USA) and Mons (Belgium) —, RMA (Belgium), VUTBR (Czech Republic).

2.1.2. Speaker modeling

Speaker model training relies on the EM (Expectation-Maximization [3]) algorithm to estimate Gaussian Mixture Models (GMM [4]). In this paper, the Gaussian mixtures are made of 16 components, summarized by full covariance matrices.

2.1.3. Block-segmental approach and normalization

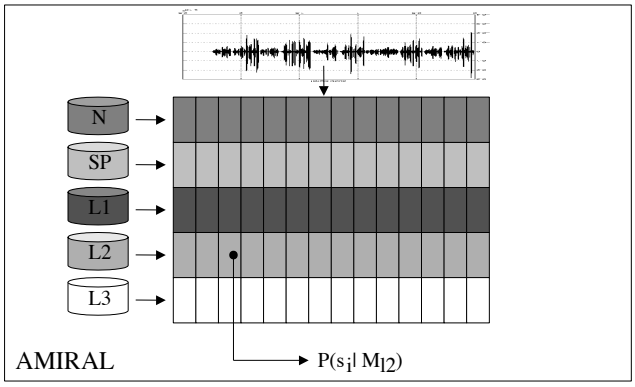

Figure 1: *Use of AMIRAL to compute the emission probabilities*

A specific aspect of the AMIRAL system is to consider speech signal at a segmental level. During test, the speech signal is split into short temporal blocks of fixed length (0.3 second) on which a similarity measure is computed.

Normalization is applied on the similarity measures in order to cope with variability problems such as message content, noise and degradation due to signal recordings and transmission channels, and mismatch across training and testing conditions (different lines, different handset types...).
Another purpose of the normalization step is to provide probabilities as the output of the speaker verification system *(figure 1)*.

The normalization method combines two techniques classically used for speaker verification.
First, a classical world model-based likelihood ratio is computed for each block.
Then, a Maximum A Posteriori (MAP) normalization is applied to the similarity ratios. Therefore, the normalized similarity measure for a block refers to the a posteriori probability of recognizing the target speaker. The MAP normalization function has to be learned using a separate tuning data set. However, the preliminary world model-based normalization allows to reduce the amount of data and tuning conditions which are usually required for this learning phase.

2.2. HMM-based conversation model

In the indexing system proposed, a HMM is used to model the conversation. It is defined by:

- A set of states representing the speakers, the "Speech" model and the "Noise" model. To each state is associated a set of emission probabilities computed by the verification system (using the corresponding model).
- A set of weighted transitions between states, representing the speaker changes.

An ergodic graph is chosen, since the system does not have a priori knowledge about the duration of utterances.

HMM transition probabilities are not learned, but they are chosen according to fixed rules expressed by inter-state weight matrix *(Table 1)*. The weights verify three conditions:

- All the probabilities of staying in the same state (s_i) are equal.
- The probabilities ($P(s_i \rightarrow s_j)$) between speaker models are equal.
- Let $P(s_i \rightarrow s_j)$ be the transition probabilities between states. Then $P(s_i \rightarrow s_j) < P(s_i \rightarrow s_i)$.

Models	Speech	Noise	Speaker I	Speaker J
Speech	5	1	5	5
Noise	5	1	5	5
Speaker I	5	1	10	1
Speaker J	5	1	10	1

Table 1: *Weight matrix: X model (row) to Y model (column), $I \neq J$*

2.3. Automatic detection of speakers

The speaker number detection algorithm consists of two parts. First, the HMM is initialized with the two default states (speech and noise models). Viterbi algorithm gives a "Speech"/"Noise" segmentation *(figure 2)*.

Then, the speakers are detected one by one during an iterative process, which proposes a temporary indexing at each step *(figure 3)*. The iteration steps are:

1. The best speaker model is chosen (section 2.4).

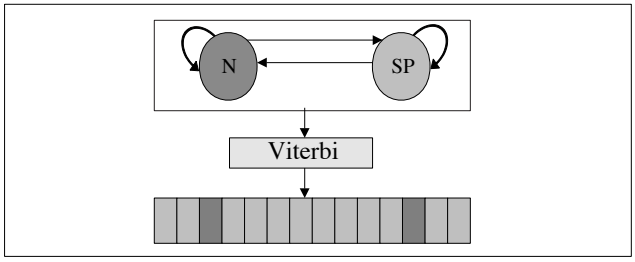

Figure 2: *Iterative process initialization*

2. Then, a state is added to represent the new speaker in the HMM. Transition weights are adapted according to the new model to take into consideration the new number of states.

3. Viterbi algorithm is applied to obtain the best indexing according to the HMM.

4. Lastly, the stop criterion is tested: is the last proposed indexing better than the previous one ? if so a new iteration begins.

NB: In practice, the system checks a second stop criterion: it is still possible to create a new speaker model ? (do segments labelled as "Speech" exist in the proposed indexing ?)

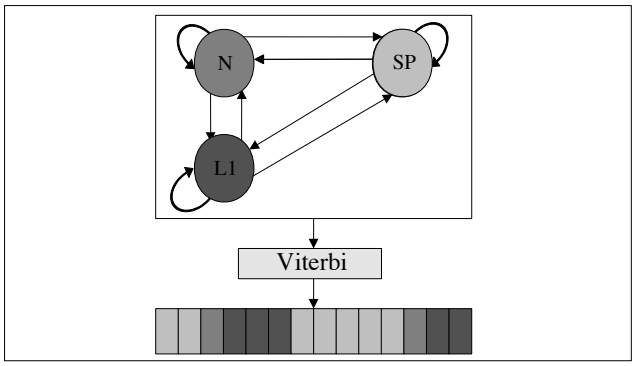

Figure 3: *Iterative process: adding the first detected speaker model*

2.4. Choice of a new speaker model

The model selection method is based on the SWGM algorithm [5] applied on the blocks labelled as "Speech". The blocks are first sorted (best scores first). Then, the algorithm selects the optimal set of blocks on which to take the decision. SWGM tends to find a compromise between the set size and the mean score of the block set.

For each speaker model not already included in the HMM, the SWGM is computed. Finally, the model for which the SWGM is maximal is selected.

3. EXPERIMENTS

3.1. Data set

The method proposed in this paper has been experimented with on a data set issued from NIST/NSA 1998 speaker verification evaluation campaign. It is composed of recordings stemming from Switchboard database and built from telephone speech segments of one speaker. We used two different data subsets defined by the ELISA consortium:

- The first one is the development data set (denoted Dev) which is used to estimate the transition weight matrix. The data set is composed of 25 male speech signals.
- The second one is the validation data set (denoted Eva), with the same size and structure as the previous one, but on a different speaker population.

For each subset, two minutes of signal per speaker are used to train speaker models and 30 more seconds of signal are available for the generation of multi-speaker segments.

3.2. Multi-speaker message generation

The multi-speaker message is generated from concatenated mono-speaker segments. The method used is:

- l different speakers are selected.
- i ($i \geq l$) different segments are chosen, such that each speaker is present at least once.
- The duration d of each speech segment is selected.

l, i and d are Gaussian random numbers *(Table 2)*. The speaker selection, the appearance order of segments and the choice of segments are generated by using a uniform distribution.

The generated messages are close to the real conditions, although there is no situation with mixed speaker segments. 5000 messages are generated for each data set.

3.3. Test results

Experiments were carried out to validate:
• The conversation model as an adaptive HMM with incremental speaker model addition.

Parameter	Mean	Standard deviation
l	3	1
i	32	96
d (# of 0.3s blocks)	6	30

Table 2: *Parameters of Gaussian distributions*

- The SWGM method used to choose the speaker.

Criteria selected to measure the performance of the system are:
- Percentage of correctly indexed blocks, denoted CB.
- Percentage of badly indexed blocks, denoted BB.

A block attribution error rate (EB) is computed from this two values:

$$EB = \frac{BB}{BB + CB}$$

This two values represent the indecision rate:
- Percentage of segments labelled as "Noise" (NB).
- Percentage of segments labelled as "Speech" (SB).

The results shown in *(Table 3)* are very encouraging. The attribution errors are correct (EB is 27% on Eva) regarding the short speaker utterance duration (the average length is 1.8s.) as well as the intrinsic difficulty of Switchboard. Moreover, low indecision rates are observed (NB is 0.7% and 5.6% for SB on EVA).

Finally, the rate of speakers detected in the message is around 97% on Dev and Eva. However, the number of wrongly added speakers in the HMM are quite important (around 87% on both data sets).

Data	CB	BB	NB	SB	EB
Dev	71.6%	26.2%	0.04%	0.18%	26.8%
Eva	66.7%	27.0%	0.7%	5.6%	28.7%

Table 3: *Multi-speaker tracking proposed system: Rates computed on 2 corpra (Dev et Eva) which composed each of 5000 messages. For all blocks: CB= % Correctly indexed blocks, BB= % Wrongly indexed blocks, NB= % Noise blocks, NS= % Speech blocks, EB= Block attribution error rate,*

4. CONCLUSION

A multi-speaker tracking based on an evolutive HMM-like model is proposed. This approach, based on an iterative algorithm, detects and adds one by one the speaker models. It proposes a temporary indexing at each step using all the knowledge available at this level. This indexing is improved during the next steps.

The obtained results are encouraging, regarding the low attribution error rates. Nevertheless, too many speaker models are badly added to the HMM. This issue may come from the stop criterion as well as the (transition) weight matrix. Futher work will focus on this point, by introducing an explicit duration model into the HMM.

5. REFERENCES

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

[2] C. Fredouille, J.-F. Bonastre, T. Merlin, Similarity normalization method based on world model and a posteriori probability for speaker verification, *EUROSPEECH*, 1999.

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

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

[5] J-F. Bonastre, P. Delacourt, C. Fredouille, S. Meignier, T. Merlin, C. Wellekens, Différentes stratégies pour le suivi de locuteur, *accepted for publication, RFIA 2000*, Jan. 2000.