SEMANTIC SEGMENTATION OF RADIO PROGRAMS USING SOCIAL NETWORK ANALYSIS AND DURATION DISTRIBUTION MODELING

A.Vinciarelli[1,2], *F.Fernàndez*[3] *and S.Favre*[1,2]

[1]IDIAP Research Institute - CP592, 1920 Martigny (Switzerland)
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)
[3]Universidad Politécnica de Madrid - 28040 Madrid (Spain)
email: {vincia, sfavre}@idiap.ch, efhes@die.upm.es

ABSTRACT

This work presents and compare two approaches for the semantic segmentation of broadcast news: the first is based on Social Network Analysis, the second is based on Poisson Stochastic Processes. The experiments are performed over 27 hours of material: preliminary results are obtained by addressing the problem of splitting different episodes of the same program into two parts corresponding to a news bulletin and a talk-show respectively. The results show that the transition point between the two parts can be detected with an average error of around three minutes, i.e. roughly 5 percent of each episode duration.

1. INTRODUCTION

Radio programs are often composed of different segments following the so-called author view [1], i.e. the way authors organize the content. In general, the segments correspond to specific topics (e.g. the *stories* in the news), content categories (e.g. commercials, games and news), or other kinds of high level clues (see [2][3] for more details). The automatic detection of such segments, often called *semantic segmentation*, is useful in several applications: *browsing* systems can enable users to select the segments of interest, *retrieval* systems can use the segments as documents, i.e. as basic units of information to be retrieved in a database, etc.

This work proposes two approaches for the semantic segmentation of broadcast news: the first is based on Social Network Analysis (see Section 2) and the second, called Duration Distribution Modeling (DDM) in the following, is based on the duration of single stories (see Section 3). The experiments are performed on a collection of news programs provided by Radio Suisse Romande (RSR), the Swiss French speaking broadcasting service, for a total of 27 hours of material. Each recording is one hour long and it is composed of two segments: the first, called *news* is a bulletin and it lasts between roughly 25 and 35 minutes, the second, called *talk-show*, corresponds to the remaining part of the program. The results presented in this work are preliminary and no comparisons are performed with more established methods (e.g. [4]). On the other hand, the results are promising and encourage to continue with the application of Social Network Analysis for the segmentation of multiparty recordings into semantically coherent segments, in particular for the segmentation of news into stories.

The Social Network Analysis (SNA) based approach relies on the fact that the programs in our dataset involve two anchormen: the first one talks all along the program, while the second talks only during the first part. By identifying the two anchormen is then possible to identify the transition between first and second part. In fact, the transition can be detected as the last intervention of the second anchorman, i.e. the one that stops talking before the end of the program. Since the anchormen change at each program, no speaker recognition based approaches can be used, then it is necessary to use SNA which is identity independent and uses only *relational data* (see Section 2).

The DDM approach models the story transitions, i.e. the instants t_k where story k ends and story $k+1$ starts, as a Poisson Stochastic Process (PSP). Each PSP [5] is characterized by a parameter λ and, given a sequence of story transitions $T = (t_1, \ldots, t_{K-1})$, where K is the total number of stories in a program, the analytic expression of the likelihood $p(T|\lambda)$ is known. Since the stories of the two parts of the program are underpinned by different stochastic processes, the point where the news end and the talk-show starts can be found as the time where the parameter λ changes (see Section 3 for more details).

The rest of this paper is organized as follows: Section 2 presents the SNA based approach, Section 3 describes the PSP based technique, Section 4 shows experiments and results and Section 5 draws some conclusions.

This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2). When the work was done, F.F.Martinez was with IDIAP Research Institute.

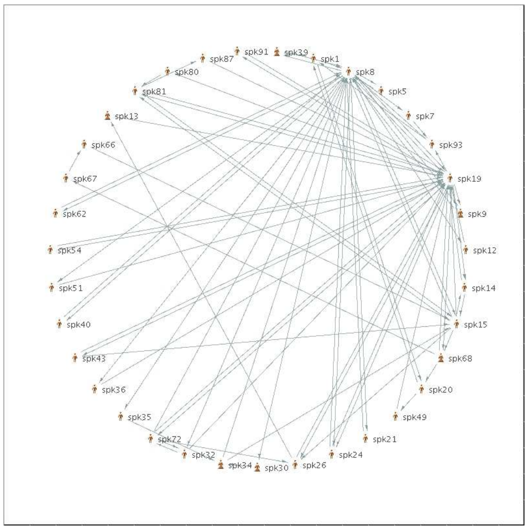

Fig. 1. Social Network. This figure shows the Social Network extracted from one of the recordings in our collection.

2. SOCIAL NETWORK ANALYSIS

This section presents the SNA based approach to the segmentation problem described in Section 1. SNA is the domain studying the interaction between different persons sharing a common environment and it is based on relational data, i.e. on the evidence of interaction between different individuals. Following an experimental psychology technique, we use as evidence of interaction between two individuals a_i and a_j the fact that a_i talks immediately before a_j at least once in a given program. This requires as a first step of the process the segmentation of the recordings into single speaker interventions. This converts the audio data into sequences of speaker IDs that can be used to extract the Social Networks.

The next sections explain in more detail speaker segmentation and Social Network extraction.

2.1. Unsupervised Speaker Segmentation

The speaker segmentation technique applied in this work is described in [6]. The speaker sequence is modeled with a fully connected continuous density Hidden Markov Model (HMM) where each state q corresponds to a single speaker. The audio data is first segmented into a sequence $O = \{\vec{o}_1, \ldots, \vec{o}_M\}$ of observation vectors, where M is the total number of O elements. Each $\vec{o}_i$ contains 12 *Mel Frequency Cepstrum Coefficients* (MFCC) extracted from a 30 ms long window. The MFCC are used because they have been shown to be more effective than other features in speaker recognition problems, they are then suitable to distinguish the voices of different persons.

Once O is available, the problem of speakers clustering can be thought of as finding the best sequence of states (i.e. the best sequence of speakers) given the HMM:

$$q^* = \arg\max_{q \in Q} p(O, q|\Theta) \tag{1}$$

where q is a sequence of speakers, Q is the set of all possible speakers sequences, and Θ is the parameters set of the HMM. Since the number of speakers is not known a-priori, an initial guess must be provided. In order to start with an over-segmentation, the guess must be higher than the expected number of speakers in the data. After the alignment, states that are too similar can be merged to form a single state. In other words, since the initial number of speakers is higher than the actual number of speakers, different states are attributed to the same speaker, thus it is necessary to merge them. States m and n are merged when their loglikelihood ratio satisfies the following condition:

$$\log p(O_m \cup O_n|\Theta_{m+n}) \geq \log p(O_m|\Theta_m)p(O_n|\Theta_n) \tag{2}$$

where O_t are the audio vectors attributed to state t, Θ_t is the parameter set of state t and Θ_{m+n} is the parameter set of a mixture of Gaussians trained with the Expectation Maximization algorithm over $O_m \cup O_n$. When two states are merged, the number of parameters in Θ_{m+n} is the sum of the parameters in Θ_m and Θ_n. In this way the number of parameters in the HMM remains constant and there is empirical evidence [6] that, by iterating alignment and merging steps, the likelihood increases up to a certain point and then it starts to decrease when states corresponding to different speakers are merged. This provides a good stopping criterion for the iterative process.

2.2. Social Network Extraction

The result of the speaker clustering process is that the audio data are converted into a sequence of speaker ID codes a_i, with $i \in \{1, \ldots G\}$ (G is the total number of detected speakers in the speaker clustering process described in the previous section).

We use as interaction evidence between two individuals a_i and a_j the fact that a_i talks immediately before a_j at least once. The use of the ordering includes the temporal information involved in the sequence resulting from the speaker clustering process. This allows to build the so-called *sociomatrix* X, i.e. a matrix where the element x_{ij} is the number of times speaker a_i talks immediately before speaker a_j. For each sociomatrix there is an associated directed graph where each node corresponds to a speaker and each edge corresponds to the interaction between the connected speakers: such a graph is called *Social Network* (SN) and it is shown, for one of the recordings in our data set, in Figure 1. Sociomatrices and SNs encode the so-called *relationl data*, i.e. the interaction patterns involving the speakers participating in each recording.

In the case of this work, the most important information is the speakers *centrality* [7], i.e. the inverse of the average geodesic distance between a given individual and the others (the geodesic distance between two nodes is the number of edges to be traversed to go from a node to the other):

$$C(a_i) = \frac{G-1}{\sum_{j \neq i} d(a_i, a_j)}, \quad (3)$$

where $d(a_i, a_j)$ is the geodesic distance between a_i and a_j and G is the total number of speakers. The reason for the name centrality is that such index is a measure of how much individuals are close to the others on average and then of how much they are central in the interaction pattern.

In the experiments of this work, we show that the two anchormen (see Section 1) are the individuals with the highest centrality. In other words, the extraction of the Social Network and the calculation of the centrality index enable one to find the anchormen a_i^* and a_j^* as follows:

$$a_i^*, a_j^* = \arg \max_{a_i, a_j \in (1,\dots,G)} C(a_i) + C(a_j). \quad (4)$$

The anchormen are detected with an accuracy of 68.8% (the accuracy is the percentage of anchormen time identified as such by the system). If $\tau(a_k)$ is the time at which the last intervention of speaker k ends, then the approach described in this section identifies the transition time t^* between news and talk-show as follows:

$$t^* = \arg \min_{a_k \in \{a_i^*, a_j^*\}} \tau(a_k), \quad (5)$$

in other words, the transition is considered to take place at the end of the last intervention of the anchorman that disappears first from the program.

3. DURATION DISTRIBUTION MODELING

This section describes the Duration Distribution Modeling. The rationale behind such approach is that our data can be considered as a sequence of stories and that the transition points between consecutive stories follow a Poisson Stochastic Process [5]. This can be seen by observing the following: given a recording m in the collection, consider the *staircase* function $f_m(t)$ which gives the number of story transitions that took place between time 0 and time t. Such function is called staircase because it increases by one each time there is a transition and then it remains stable until there is another transition. The average number of transitions $n(t)$ in the dataset at a given time t can be estimated as follows:

$$n(t) = \frac{1}{M} \sum_{m=1}^{M} f_m(t), \quad (6)$$

where M is the total number of recordings in the dataset. The function $n(t)$ is plotted in Figure 2 and it consists of two linear pieces that can be expressed as $n_1(t) \simeq \lambda_1 t$ and $n_2(t) \simeq \lambda_2 t$. This shows that the transitions actually follow a PSP and that the PSP underpinning the transitions changes at a certain point of the program. The change of slope corresponds to the transition between the news and the talk-show: the segmentation process can be thought of as finding the story in correspondence of which the underlying PSP (and the corresponding λ parameter) changes.

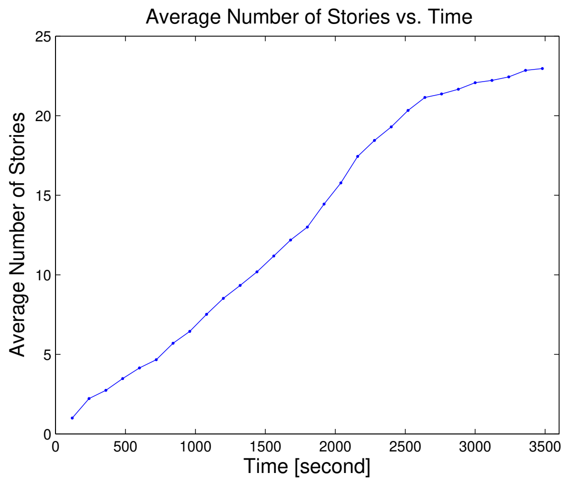

Fig. 2. Average number of stories vs time. This plot shows the average number of stories (estimated at two minutes long time steps) as a function of the time.

Since the transition points follow a PSP, the probability of a story being long τ can be written as follows [5]:

$$p(\tau|\lambda) = \lambda e^{-\lambda\tau}, \quad (7)$$

this means that the likelihood of a sequence $T = \{\tau_1, \dots, \tau_N\}$ of story durations in a given recording can be expressed as follows:

$$p(T|\lambda_1, \lambda_2) = \prod_{k=1}^{n} p(\tau_k|\lambda_1) \prod_{l=n+1}^{N} p(\tau_l|\lambda_2) \quad (8)$$

where n is the index of the story where the PSP underlying the story transitions changes, i.e. the index of the story where the news end and the talk-show starts (see Section 1). The value of n can be found by maximizing the logarithm of the likelihood:

$$n = \arg\max_m m \log \lambda_1 + (N-m) \log \lambda_2 - \lambda_1 \sum_{k=1}^{m} \tau_k - \lambda_2 \sum_{k=m+1}^{N} \tau_k. \quad (9)$$

The last problem to be solved is the estimation of the parameters λ_1 and λ_2. This is performed using a leave-one-out approach, i.e. by using all recordings except the one used for testing the algorithm. Given a set of recordings for which n is known, the λ_i values are those that maximize the likelihood of all the T sequences observed in the training set:

$$\lambda_i = \frac{N_i}{\sum_{k=1}^{N_i} \tau_k^{(i)}}, \quad (10)$$

where N_i is the total number of stories following a stochastic process with parameter λ_i and $\tau_k^{(i)}$ is the k^{th} story following the same stochastic process.

4. EXPERIMENTS AND RESULTS

This section presents the experiments performed over a collection of 27 recordings provided by Radio Suisse Romande, the French speaking Swiss broadcasting service. Each recording is one hour long and it is composed of two parts: the first is referred to as *news* and the second is referred to as *talk-show*. The goal of the experiments is to detect automatically the transition point between news and talk-show by using the techniques presented in Section 2 and Section 3. The transition point is variable and ranges between 35 and 45 minutes. In other words, there is no strong a-priori constraint enabling one to detect the transition point in a short neighborhood centered around a predefined time.

The performance is measured in terms of *accuracy* α, i.e. in terms of the percentage of time where the semantic class (news or talk-show) is assigned correctly. Since each recording contains only two segments, $100 - \alpha$ expresses the distance (in terms of percentage with respect to the total duration of the recording) between the actual transition point and the tansition point detected automatically. In other words, if the accuracy in a recording is 95%, then the difference between the real transition and the detected transition accounts for 5% of the total duration of the recording.

The results of the experiments ar as follows: the α is **94.5%** for the SNA method and **99.8%** for the DDM method. The method based on the story transitions performs better than the other, but such a performance is overestimated. In fact, the results are obtained over a manual segmentation, i.e. the story transitions have been detected by a human assessor. The process is then not fully automatic.

On the contrary, the results obtained using the SNA based approach are realistic because the process does not involve any manual intervention. The speaker segmentation (see Section 2) is automatic as well as the analysis of the resulting Social Network. The average distance between the actual transition and the transition detected automatically is around 3 minutes. This means that a potential user does not need to listen to more than 6 minutes (3 minutes before and 3 minutes after the detected point) in order to find the actual transition between news and talk-show. This reduces by roughly 40% the variability range observed in our data (the transition point is between $\simeq$ 35 and $\simeq$ 45 minutes), then it decreases the amount of time needed for an operator to find the real transition point.

5. CONCLUSIONS

This paper has presented preliminary experiments involving two approaches for the semantic segmentation of broadcast news: the first is based on Social Network Analysis and the second on the application of Poisson Stochastic Processes to story transitions. The first approach has been evaluated in realistic conditions, i.e. by performing automatically all the steps of the processing, and the results show that, on average, roughly 95% of the recording time is labeled correctly in terms of semantic class. This is equivalent to say that the transition point between the two parts of our recordings is found with an average error of around three minutes.

The second approach has been tested by performing manually one step of the process (the story segmentation) and the corresponding results are thus significantly overestimated. On the other hand they provide an upper bound and they show that the story transitions capture the information necessary to perform the segmentation. Although taylored to our specific kinds of data, the two approaches presented in this work can be extended, with some modifications, to other kinds of data. However, the task addressed in this work is relatively simple and the results must be considered preliminary.

The results presented in this work concern a task which is rather specific of the data at disposition. However, this is just a preliminary work and the final goal is a more general task, i.e. the segmentation of the recordings into stories by combining the two approaches presented in this work. This is the subject of future work.

6. REFERENCES

[1] C.G.M. Snoek and M. Worring, "Multimodal video indexing: a review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[2] Z. Xiong, X.S. Zhou, Y. Rui, and T.S. Huang, "Semantic retrieval of video," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 18–27, 2006.

[3] M. Chen, S.C. Chen, M.L. Shyu, and K. Wickramaratna, "Semantic event detection via multimodal data mining," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 38–46, 2006.

[4] M.A. Hearst, "TextTiling: segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[5] A. Papoulis, *Probability, Random Variables, ans Stochastic Processes*, McGraw Hill, 1991.

[6] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, 2003.

[7] S. Wasserman and K. Faust, *Social Network Analysis*, Cambridge University Press, 1994.