SOCIOMETRY BASED MULTIPARTY AUDIO RECORDINGS SEGMENTATION

Alessandro Vinciarelli^{1,2}

^{1}IDIAP Research Institute -CP592, 1920 Martigny (Switzerland)
^{2}Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)
email: vincia@idiap.ch

ABSTRACT

This paper shows how Social Network Analysis, the sociological domain studying the interaction between people in specific social environments, can be used to assign roles to different speakers in multiparty recordings. The experiments presented in this work focus on radio news recordings involving around 11 speakers on average. Each of them is assigned automatically a role (e.g. anchorman or guest) without using any information related to their identity or the amount of time they talk. The results (obtained over 96 recordings for a total of around 19 hours) show that more than 85% of the recording time is correctly labeled in terms of role.

1. INTRODUCTION

The problem of effectively accessing the content of spoken audio archives has been extensively studied in the last years. The most common approach is to transcribe the recordings through Automatic Speech Recognition (ASR) and then to apply techniques developed for digital texts [1]. Although successful, especially for applications like Information Retrieval, such an approach neglects the far richer information contained in speech recordings, e.g. speaker identity and emotional state, prosody, dialogue annotations, etc., that can provide useful information in applications [1].

This work focuses on the interaction between different speakers in multiparty (i.e. involving different actors) recordings. In particular, this paper shows how *Social Network Analysis* (SNA) [2], the discipline studying the interaction between people in social environments, can be used to detect the *role* (see section 2.2 for more details) of different actors in radio news recordings. The assignment of a role to different recording actors can be important in browsing (users can select specific segments based on the kind of speaker they want to listen to), thematic segmentation (in some cases, the role can be related to specific topics), summarization (speakers with certain roles can bring content more representative of the whole recording than others), etc.. To our knowledge, relatively few works have been dedicated to the study of human interaction in multimedia recordings (see [3] and references therein) and SNA has never been applied in such a context.

The rest of this paper is organized as follows: Section 2 presents the approach followed in this work, Section 3 shows experiments and results and Section 4 draws some conclusions.

2. PROCESSING APPROACH

The approach applied in this work includes two main stages. The first, named *low level processing*, includes the automatic segmentation of the recordings into single speaker segments and the application of Poisson process statistics in order to improve the segmentation quality. The second, identified as *high level processing*, involves SNA and leads to the assignment of different roles to different speakers.

2.1. Low Level Processing

The speaker segmentation technique applied in this work is fully described in [4]. The speaker sequence is modeled with a fully connected continuous density Hidden Markov Model (HMM) where each state corresponds to a single speaker. Such a model is aligned with a sequence O of observation vectors extracted from the audio data using the Viterbi algorithm. The result is the best sequence of states (i.e. the best sequence of speakers) given the model:

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

where q is a sequence of speakers 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

The author wishes to thank F.Valente and A.Peregoudov for commenting on the draft. J.F.Cosandier, G.Saudan and R.Dahler are gratefully acknowledged for providing the data. This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

1-4244-0367-7/06/$20.00 ©2006 IEEE

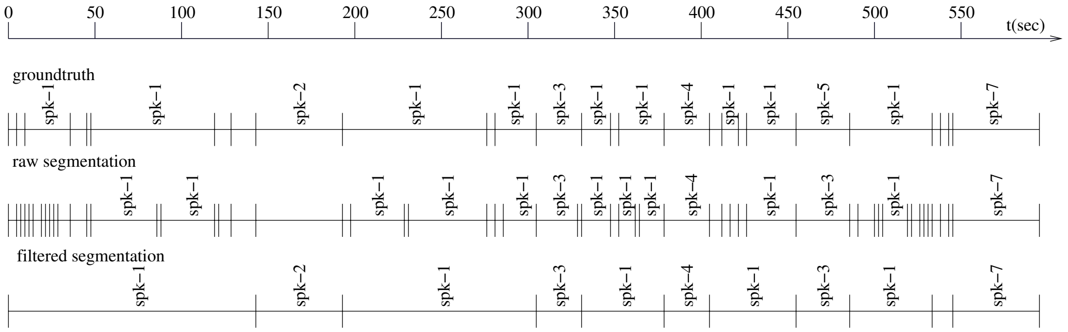

Fig. 1. Segmentation example. This figure shows an example of segmentation, the three axes correspond to groundtruth, raw and filtered segmentation respectively.

them. States m and n are merged when their loglikelihod ratio satisifies the following condition:

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

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 $O_m \cup O_n$. After merging the states, the resulting model is aligned again with the data and the whole process is reiterated until the likelihood expressed in Equation 1 reaches its maximum.

The result of the segmentation can be observed in Figure 1. In some cases, different speakers are merged erroneously into a single speaker (e.g. speakers *spk-3* and *spk-5* in the groundtruth are given the same label in the raw segmentation). Moreover, the system tends to oversegment and to create many short segments (left blank in the figure) attributed to spurious speakers, i.e. speakers that do not actually exist in the recording. The first problem can be solved through SNA (see next subsection), the second can be solved by analyzing the distribution of the segment durations. Speaker changes are events randomly distributed in time and, like many other phenomena in nature and technology, they can be modeled with a Poisson stochastic process [5]. The probability of a segment having a duration shorter than τ can thus be expressed as:

$$p(\tau \leq t) = 1 - e^{-\lambda t}. \quad (3)$$

The λ parameter can be estimated using the groundtruth data and it represents the inverse of the average segment duration [5]. Once λ is estimated (we used a leave one out approach) it is possible to consider as spurious all segments with a duration τ such that (the threshold is selected a-priori and no other values have been tried):

$$\frac{p(\tau \leq t)}{1 - p(\tau \leq t)} \geq \frac{1}{2}. \quad (4)$$

In other words, all the segments that are likely to be produced by a Poisson stochastic process different from the one underlying the groundtruth data are considered spurious and removed. When several spurious segments follow each other, they are grouped and attributed to the most represented speaker (in terms of time) among them. When a spurious segment is isolated, it is attributed to the neighboring segment with the highest probability in Equation 3.

The result of the above algorithm can be seen in the rightmost column of Figure 1. The oversegmentation is definitely reduced and the short segments are typically merged to longer segments. In some cases, such an effect makes sense because spurious speakers are typically determined by background noises or music that induce false speaker changes. In some other cases, the short segments correspond to actual turns and their elimination is an error. However our experiments show that the application of such a processing step improves the results (see Section 3).

2.2. Social Network Analysis

Social Network Analysis [2] is the domain that involves methods and techniques to analyze *relational data*, i.e. the way people interact in a specific social environment. Given a set $A = \{a_1, \ldots, a_g\}$ of actors and a relationship $R : A \times A \rightarrow V$ (V is the set of the values that the R can take), relational data can be represented through a matrix X where $x_{ij} = a_i \rightarrow a_j$ (where $a_i \rightarrow a_j$ is the value of the relationship R between a_i and a_j). For each matrix it is possible to draw a graph, the so-called *social network*, where each node represents a person and two nodes are connected or not depending on the value of R. The definition of R depends on the specific problem under examination. In our experiments, each speaker is assumed to be an actor and $a_i \rightarrow a_j$ has value 1 when speaker i talks immediately before speaker j at least once, while it has value 0 otherwise.

Figure 2 shows the network resulting from the groundtruth of the recording used in Figure 1. Different actors appear to have different roles in the interaction pattern. Speakers labeled as *spk2* and *spk4* (the speaker segmentation is an unsupervised algorithm and the speakers are thus labeled with a

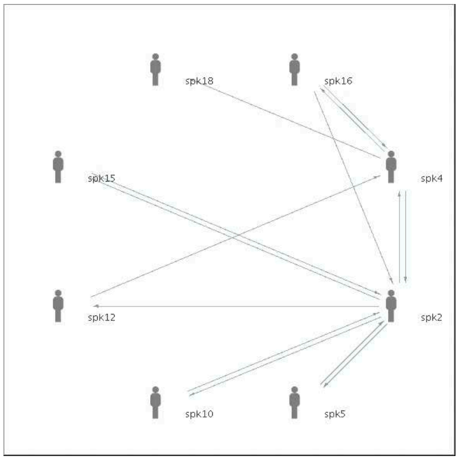

Fig. 2. Social network. This figure shows the Social Network corresponding to the groundtruth speaker segmentation of Figure 1

progressive identifier) appear to have a *central* position in the network. By central it is meant that most of the paths leading from one speaker to another have to pass through them. The other speakers occupy a marginal position and interact only with the central speakers. Given a specific social environment, the position of an actor in the network is typically associated to a specific social role, thus SNA can be used to detect the role of each speaker. The main advantages in using SNA for audio recordings are essentially two. The first is that the role can be attributed independently of the identity of the speakers, thus roles can be correctly attributed even when they are played by different actors in different recordings. The second is that the recording length is not taken into account, thus high duration variability can be tolerated.

Actor position relevant information can be extraced from the matrix of the *geodesic distances* D where the element d_{ij} is the distance between actor i and actor j. The geodesic distance between two nodes is the length (i.e. the number of edges to be traversed) of the shortest path connecting them (in a path each edge can be traversed only once). Since central actors have many connections, a good centrality measure $C(i)$ is given by the inverse of the average distance from the other actors:

$$C(i) = \frac{g-1}{\sum_{j=1}^{g} d_{ij}} \tag{5}$$

where g is the total number of actors.

In our experiments, D conveys two main kinds of information. The first is the value of C that allows one to group users based on their role. In fact speakers with the same C are likely to play the same social role. The second is the presence of paths connecting several speakers. Since in our experiments the relationship, thus the presence of an edge, depends on the fact that the speakers talk after each other, a path identifies a sequence of speaker turns that potentially represents a meaningful higher level unit.

In our experiments we identified five roles that can be assigned using the above information. The first is the *anchorman* (AM), i.e. the role of the speakers coordinating the newscast (in the case of our data there are two AM per bulletin). The second is the *guest* (GT), i.e. speakers invited to express their opinion or to report about a specific topic (there are three to five GT per bulletin). The third is the *interview participant* (IP), i.e. actors that are involved in an interview. The fourth is the *abstract* (AB), i.e. the persons providing the list of the news that will be presented in each bulletin, and the fifth is the *meteo* (MT), i.e. the role of the person reading the wheather forecast. The AM, AB and MT roles are played by different persons in different recordings (there are around 20 different persons that alternatively appear in different newscast as AM, AB or MT). Different persons play the GT or IP role in the same bulletin and, with few exceptions, they change at each newscast. In most cases, guests and interview participants are invited only once since they are related to very specific topics.

Each role can be detected using the social network information. AM can be identified as the speakers with the highest C value (see Equation 5), GT are speakers that can be found in the shortest paths starting and ending with the AM, while IP are the nodes along any longer path starting and ending with the AM. For what concerns AB and MT, they can be easily detected as the speakers that talk first (no path leads to them) and last (no path starts from them).

3. EXPERIMENTS AND RESULTS

This section describes the experiments performed in this work. The data we used have been collected from Radio Suisse Romande (the Swiss national broadcasting service) during February 2005. For each working day (monday to friday) we collected five recordings corresponding to news bulletins diffused at different times. Since in february there are 20 working days and four recordings were lost because of technical problems, the resulting number of recordings is 96 for a total of 18 hours and 56 minutes. The average recording duration is 11 minutes and 50 seconds (the minimum is 9 minutes and 4 seconds and the maximum is 14 minutes and 28 seconds). The average number of speakers per recording is 11.0 and the average number of interactions is 29.0.

Different roles (see Section 2.2) account for different fractions of the total corpus time, AM corresponds to 46.7% of the material, AB to 7.1%, GT to 34.8%, IP to 4.0%, meteo to 6.3% and the remaining 1.0% includes essentially music, jingles, noise and anything else cannot be attributed to one of the above roles.

3.1. Automatic Role Assignment

The first processing step is the segmentation of the recordings into speakers. The result is a sequence $S = \{(s_i, \tau_i)\}$, where

$i = 1, \dots, M$, of pairs including a speaker label s_i and a duration τ_i. If $S^* = \{(s_i^*, \tau_i^*)\}$, where $i = 1, \dots, N$, is the groundtruth, the segmentation performance can be measured through the fraction of recording time such that $s = s^*$. In the following, Such a measure is called *Accuracy* α and it can be interpreted as the probability that the speaker assigned by the system to a certain time segment actually corresponds to the real speaker.

The average α before the Poisson based filtering process (see Section 2) is 85.0%. The errors are mainly due to background noise and music that tend to determine many short turns attributed to spurious speakers (see central column in Figure 1). The main resulting problem is that the average number of interactions is increased from 29.0 (see above) to 52.1, i.e. around 45% of the interactions is likely to be the effect of an error. This is problematic because it heavily affects the Social Network and it can lead to wrong role assignment.

For the above reason, it is necessary to apply a Poisson based filtering that reduces the number of spurious segments and thus decreases the spurious interactions potentially affecting the role assignment process. After the filtering, the average α is decreased to 80.6%, but the average number of interactions is 16.9. This corresponds to around 40% of interactions lost with respect to the real amount of speaker turns, but most of the interaction losts are quick exchanges between the two AM. This has thus a limited impact on the role assignment performance, because the role is rather detected through the interaction the two AM have with the other speakers. The result of the filtering is a sequence $F = \{(f_i, \tau_i)\}\}$ with $i = 1, \dots, K$.

The result of the role assignment process produces a sequence $R = \{(r_i, \tau_i)\}$, with $i = 1, \dots L$, where L corresponds to M when the SNA is applied to the result of the segmentation and to K when it is applied to the result of the filtering. The process performance can be measured again with an *Accuracy* value accounting for the percentage of recording time labeled with the correct role.

The average α is 68.3% after the raw segmentation and 86.8% after the filtering. The application of the Poisson based duration analysis improves the Accuracy by 27.1% even if the performance of the speaker segmentation is decreased from 85.0% to 80.5%. The reason is that what is actually important in the role assignment process is to preserve the interaction related information rather than the correct labeling of the speakers. Although the number of interactions is heavily affected both before and after the filtering, this last seems to be able to keep essentially correct interactions while getting rid of spurious ones. Table 1 reports the results per role before and after the filtering. The performance is high on AM, AB and MT, while it is low on GT and IP. On the other hand, it happens in most cases that the two roles are confused with each other. In other words, when they are misclassified GT and IP tend to be labeled as IP and GT respectively. However, for many applications this is not a major problem, e.g. in thematic segmentation both guest interventions and interviews typically concern a single topic, thus a segment can be considered topic coherent independently of the fact of being labeled as GT or IP.

Role	α bf	α af
AM	88.3%	91.9%
AB	33.2%	97.8%
GT	44.2%	85.1%
IP	71.2%	24.3%
MT	98.7%	95.4%

Table 1. Results per role. Average α for before (*bf* column) and after (*af* column) the filtering process.

4. CONCLUSION

This paper shows that it is possible to apply Social Network Analysis to multiparty audio recordings in order to assign different roles to different speakers. To our knowledge, such a task has never been addressed before in the literature and SNA has never been used to process multimedia recordings. SNA enables to assign roles to different speakers independently of their identity (it is thus possible to process recording involving different actors) as well as the amount of time they talk (it is thus possible to tolerate time variability in the data). In our opinion, the role of the speakers, whenever it is possible to identify clear roles like in the case of the news bulletins, can be an important metadata and it can help in applications like thematic segmentation, browsing, etc.. As a future work, we plan to apply the same kind of analysis to other kinds of multiparty recording in order to further validate our approach.

5. REFERENCES

[1] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 61–69, 2005.

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

[3] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, 2005.

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

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