IDIAP RESEARCH REPORT

Sociometry Based Multiparty Audio Recordings Summarization

Alessandro Vinciarelli [a] [b]

IDIAP–RR 06-27

May 2006

paru dans
Proceedings of International Conference on Pattern Recognition (ICPR 2006)

[a] IDIAP - alessandro.vinciarelli@idiap.ch
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)

Rapport de recherche de l'IDIAP 06-27

Sociometry Based Multiparty Audio Recordings Summarization

Alessandro Vinciarelli

May 2006

paru dans
Proceedings of International Conference on Pattern Recognition (ICPR 2006)

his paper shows how Social Network Analysis, the study of relational data in specific social environments, can be used to summarize multiparty radio news recordings. A social network is extracted from each recording and it is analyzed in order to detect the role of each speaker (e.g. anchorman, guest, etc.). The role is then used as a criterion to select the segments that are more representative of the recording content. The results show that the length of the recordings can be reduced by more than 90 percent while still preserving most of the information about their content.

1 Introduction

The automatic summarization of a document is typically performed by selecting the segments that are more representative of its content. A summarization system is thus characterized by two key elements, the first is the way the documents are split into meaningful segments, the second is the way the segments are selected to build the summary. In most cases, summarization is applied to digital texts that can be segmented, thanks to the punctuation, into sentences. Afterwards, these can be selected through statistical analysis of the terms they contain or structural properties of the text (e.g. by selecting elements like the abstract in a paper).

In the case of spoken documents, it is possible to apply a similar approach after the recordings have been converted into text through an Automatic Speech Recognition (ASR) system. On the other hand, speech recognizers do not provide punctuation and it is necessary to rely on alternative information (e.g. the distribution of the silences) that cannot lead to a perfect segmentation into sentences or similar meaningful units. Moreover, the transcriptions can be affected by high Word Error Rates (up to 30% for spontaneous speech [2]) and this can affect the selection methods mentioned above.

This work shows how the problem of summarizing spoken documents can be addressed, in some cases, by using alternative information that can be extracted more reliably and can thus lead to higher quality summaries. The approach we propose can be applied to *multiparty* recordings, i.e. recordings involving several persons, and it is based on Speaker Clustering [1] and Social Network Analysis [5], i.e. the sociological domain studying the way people interact in specific social environments. Speaker Clustering (SC) is used to segment the recordings into speaker turns, i.e. segments where only one person talks, that can be used as a basic unit in the summarization process. Social Network Analysis (SNA) is used to detect speakers *centrality* and role (see section 3 for more details), two criteria that can be used to select the turns appearing in the summary. To our knowledge, this is the first time that Social Networks are applied in a summarization problem. The main limit of this technique is that it can be applied only on multiparty recordings where the speakers play a *role*. This is typically the case in data created in a *production* environment (e.g. radio and television), but it rarely happens for data collected in more spontaneous environments such as meeting recordings and and home videos.

Our experiments are performed over a corpus of 96 news bulletins (for a total of around 19 hours of material) involving on average 11 speakers. The results show that it is possible to achieve an average compression rate, i.e. the ratio between the lengths of the summary and of the original recording, of around 8% preserving most of the information about the actual content of the bulletins.

The rest of this paper is organized as follows : Section 2 describes our SC approach, Section 3 introduces SNA, Section 4 presents experiments and results and Section 5 draws some conclusions.

2 Speaker Clustering

The SC approach applied in this work is fully described in [1]. The algorithm is unsupervised and it is not necessary to know in advance the number of speakers appearing in the recording. On the other hand, it is necessary to make an initial guess that must be higher than the expected number of speakers, it is thus necessary to have a reasonable idea of the number of people talking in each recording.

The SC system creates a fully connected continuous density HMM [2] with one state per speaker, then aligns such a model with the sequence of feature vectors O extracted from the data using the

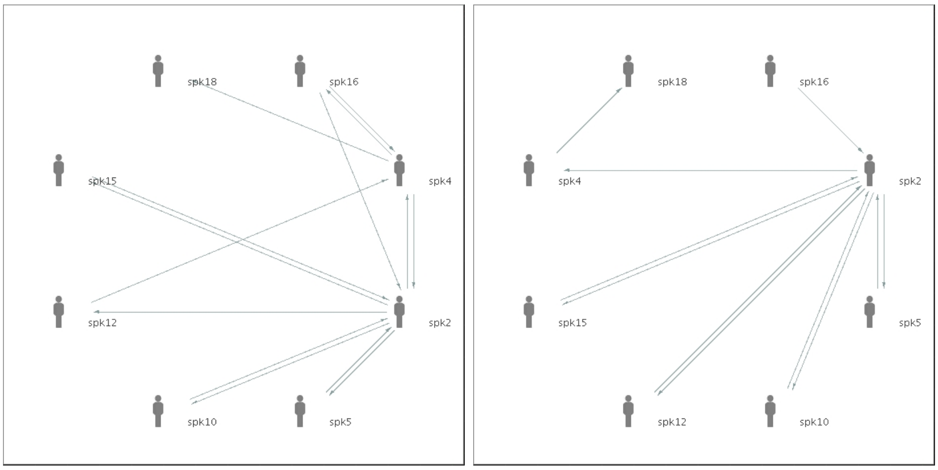

Fig. 1 – Social Networks. The two networks are extracted from groundtruth (upper network) and automatic speaker segmentation (lower network) respectively.

Viterbi algorithm [2] and finally finds the sequence of states maximizing the likelihood of the model :

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

where q is a sequence of speakers, Q is the set of all possible speaker sequences, and Θ is the parameters set of the HMM. Since the number of speakers is overestimated (see above), the first alignment leads to an oversegmentation, i.e. segments that belong to the same speaker are attributed to different states. For this reason, two states m and n are merged when their loglikelihod ratio satisifies the following condition :

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

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 $O_m \cup O_n$. 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 final segmentation result is thus smoothed using a Poisson based filtering algorithm [4] allowing one to remove most of the spurious speaker changes due to effects like background noise or music. 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 $S^* = \{(s_j^*, \tau_j^*)\}$ (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.

3 Social Network Analysis

SNA studies the interaction between people by analyzing the so-called *relational* data, i.e. all the evidences of the fact that two or more persons interact with each other [5]. Given a news recording, it is possible to identify a set $A = \{a_1, \dots, a_g\}$ of speakers and to define a relationship $R : A \times A \rightarrow V$, where V is the set of values that R can take. This leads to a matrix X, the so-called *sociomatrix*, where $x_{ij} = R(a_i, a_j)$, where $R(a_i, a_j)$ is the value of the relationship between actors i and j.

In our experiments, $R(a_i, a_j) = 1$ when a_i talks immediately before a_j at least once in the recording and $R(a_i, a_j) = 0$ otherwise. In other words, we consider the contiguity between two speakers as the evidence of the fact that they interact. The relationship is not symmetric, thus $R(a_i, a_j) = 1 \nRightarrow$

$R(a_j, a_i) = 1$. The same condition can be expressed by saying that, in general, $X \neq X^t$ (where X^t is the transposed of X). Each sociomatrix can be represented through a graph where each node corresponds to an actor $a_i \in A$ and nodes i and j are connected or not depending on the value of x_{ij} and x_{ji}.

Figure 1 shows the social networks extracted from one of the recordings used in our experiments. The two networks have been extracted from the groundtruth (upper figure) and automatic speaker segmentation (lower figure) respectively. The groundtruth network clearly shows that speakers play different roles in the communication pattern. Speakers labeled *spk2* and *spk4* (the speaker segmentation algorithm is unsupervised and speakers are labeled with a progressive index) play a *central* role in the communication pattern, i.e. most of the paths connecting two randomly selected speakers have to pass through them. In other words, the speakers defined as *central* can reach the other actors through network paths that are shorter on average and they have a direct interaction with most of the other actors. This corresponds to the so-called *closeness centrality* (see [5] for an extensive survey on centrality definitions) and it can be measured as follows :

$$C(a_i) = \frac{g-1}{\sum_{j=1}^{g} d_{ij}} \tag{3}$$

where $C(a_i)$ is the centrality degree of actor a_i, g is the total number of actors and d_{ij} is the length of the shortest path between a_i and a_j in the network.

The centrality of the speakers is related to their role in the recording. In the case of our data, we identified five possible roles : *anchorman* (AM), i.e. a person leading and coordinating the news bulletin, *guest* (GT), i.e. the speakers invited to express their opinion or to report about a single and specific topic, *interview participant* (IP), i.e. the speakers involved in an interview, *abstract* (AB), i.e. the persons appearing at the beginning of the bulletin and providing a quick overview of the news presented in each bulletin, *meteo* (MT), the person presenting the wheather forecasts.

The AM can be detected as the speaker with the highest value of centrality :

$$AM = \arg\max_{a_j \in A} C(a_j). \tag{4}$$

The reason is that the AM must interact with most of the other speakers in order to coordinate the bulletin. In terms of content, this means that the AMs talk about most of the topics presented in each bulletin. AB and MT can be detected as the actors appearing at the beginning and the end of the bulletin respectively (they thus interact at most with one person and this increases their average distance with respect to other actors). The AB is very important in terms of summarization because it provides a description of the main news that are going to be presented in the bulletin. This corresponds to the so-called *lead summary* [3], i.e. the automatic summary obtained by selecting the initial part of a text supposed to contain a short description of the document content. All other actors fall in the GT or IP categories that can be distinguished because guests make typically a single intervention (once they start to speak they are not interrupted) while IPs have an exchange and their interventions are thus made of several turns. However, both IP and GT are characterized by the fact that they talk about a single specific topic, thus the content of their interventions is never representative of the whole bulletin content.

4 Experiments and Results

This section presents the results obtained in our experiments. 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

Role	AM	AB	GT	IP	MT
Accuracy	91.9%	97.8%	85.1%	24.3%	95.4%

Tab. 1 – This table reports the accuracy for each role.

Summary	*S1*	*S2*	*S3*
Precision	97.8%	91.9%	93.4%
Recall	92.2%	95.5%	93.7%
Compression	8.4%	45.1%	38.3%

Tab. 2 – Summarization Performance. This table reports the average values of π, ρ and compression rate for the different kinds of summary.

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 3) 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. Table 1 shows the effectiveness of the system in assigning the correct role to each speaker in terms of Accuracy.

The results show that the only role for which the performance is not satisfactory is the IP. The reason is that in the intevrviews the journalists tend to make very short interventions that are often filtered by the Poisson based process (see Section 2). In this way, interviewees and journalists are merged into a single speaker that is typically interpreted as a guest. In our context this is not a major problem because the only important aspect is to capture segments that concern a single topic and can thus be selected as meaningful units in building a summary.

4.1 Summarization

The goal of a summarization system is to select segments that account for a fraction as small as possible of the documents they are extracted from while still preserving information about their content. Based on the speakers role, it is possible to obtain three kinds of summaries from our data. The first (called *S1*) can be obtained by selecting the part of the bulletin where the AB speakers are talking. They appear at the beginning of each bulletin and they provide a quick description of the news that are going to be presented. The second (called *S2*) is composed of the concatenation of all AM turns. AM speakers are active all along the bulletins and some times they present some news, while in other cases they simply introduce a guest or an interview with some details. This kind of summary can thus provide an extensive version of the abstract where all of the news presented in the bulletin are discussed in detail. The third (called *S3*) is the concatenation of GT and IP interventions. This kind of summary is supposed to contain the news that are presented in more detail (and are thus more important) in each bulletin.

Since *S1*, *S2* and *S3* represent ideal summaries that can be used as a term of comparison, it is possible to perform an *intrinsic* performance evaluation of the summarization system [3], i.e. to measure the system effectiveness through the similarity between ideal and automatic summaries. This can be done by measuring *Precision* π and *Recall* ρ achieved in selecting specific kinds of interventions. Table 2 reports the average of the π and ρ values obtained over the recording corpus used in our experiments, while Figure 2 shows the (π, ρ) plane where each point corresponds to a single bulletin (points corresponding to different summary kinds are plotted with different symbols). In most cases, both π and ρ are above 90% and the automatic summaries are thus close to the ideal ones. In the case of $S1$, the average summary length is around 60 seconds, thus a π value of 92.2% means that, on average, less than 6 seconds of the automatic $S1$ summaries do not correspond to an actual AB

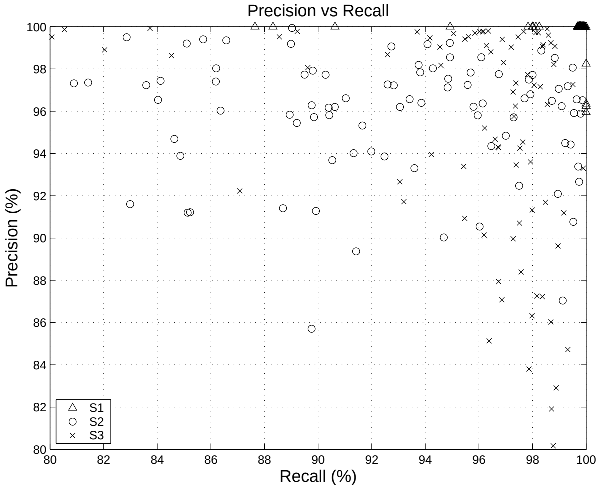

FIG. 2 – (π, ρ) plane. Each point corresponds to a single recording, $S1$, $S2$ and $S3$ summaries are represented with different symbols.

intervention. In the case of $S2$ and $S3$, the same considerations show that, on average, only 15 and 16 seconds respectively are not part of the corresponding ideal summaries.

Another important aspect of the summarization is the so-called *Compression Rate*, i.e. the ratio between the length of the summary and the length of the summarized document. Table 2 shows the compression rates achieved for the different summary kinds. In the case of $S1$, it is possible to remove, on average, more than 90% of the bulletins lengths while still preserving most of the information about their content. In the case of $S2$ and $S3$, the compression rates are higher, but the summaries are composed of several interventions and users can manually select one of them in order to achieve better compression rates. Even if it requires a manual intervention, such an approach represents a good trade-off between performance and user effort required to achieve it.

5 Conclusion

This paper has presented an SNA based approach to summarize multiparty audio recordings. Experiments performed on radio news bulletins show that different speakers can be assigned roles which can be used as a criterion to select recording segments more or less representative of the bulletin content. The results show that it is possible to obtain summaries accounting, on average, for less than 10% of the recordings from where they are extracted while still containing information about the whole recording content.

Acknowledgements. This work is supported by the Swiss National Science Foundation through the National Center for Competence in Research on Interactive Multimodal Information Management (IM2).

Références

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

[2] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, 1998.

[3] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization evaluation methods : Experiments and analysis. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, pages 60–68, 1998.

[4] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw Hill, 1991.

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