ROLE RECOGNITION IN BROADCAST NEWS USING BERNOULLI DISTRIBUTIONS

A.Vinciarelli

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

ABSTRACT

This work presents an approach for the recognition of the roles played by speakers participating in radio broadcast news (e.g. anchorman or guest). The approach includes two main stages: the first is the split of the news recordings into single speaker segments using an unsupervised approach. The second is the application of Bernoulli Distributions for role modeling and recognition. The experiments are performed over a collection of 96 news bulletins (around 19 hours of material) and show that around 80 percent of the data time is labeled correctly in terms of role.

1. INTRODUCTION

People participating in broadcast news, or more in general in radio and television programs, play a *role*, i.e. they fulfill a function that imposes constraints on timing and length of their interventions. This work presents experiments where such a role is recognized automatically. The approach proposed in this work for the above task includes two main stages: the first is the segmentation of the data into single speaker segments, the second is the actual role recognition (see Figure 1).

The speaker segmentation is performed using an *unsupervised* approach, i.e. without knowing in advance number and identity of the speakers. The reason is that there is no one-to-one correspondence between people and roles: the same role can be played by different persons and the same person can play different roles. In such a situation, the recognition of the speaker identity does not help and can be even misleading in recognizing the role. Moreover, around 50 percent of the data contains persons that talk only once (this is common in news where the journalists appear often, but guests and interviewees always change) and an identity based approach would not be helpful to recognize their roles.

The role recognition is performed by extracting, for each speaker, a vector accounting for the behavioral aspects most directly influenced by the role, i.e. timing and length of the interventions. Such vector, referred to as *behavior pattern*, is then modeled using appropriate probability density functions (Bernoulli Distributions and Gaussians). Each speaker is assigned the role maximizing the *a-posteriori* probability or the *likelihood* (the number of possible roles is six).

The experiments are performed over around 19 hours of material and the results show that around 80 percent of the data time is labeled correctly in terms of role. On the other hand, the recognition performance over the manual speaker segmentation shows that the performance can be higher then 90 percent and this represents an important margin for improvement.

Role recognition can be useful in several applications: *browsing systems* can allow users to select segments corresponding to a role of interest, *summarization systems* can select only the segments corresponding to information rich roles, *retrieval systems* can include the roles in the search clues, etc. To our knowledge, few other works have been dedicated to role recognition [1][2]. The approach proposed in [1] is based on lexical specificities related to different roles, while the technique presented in [2] is based on Social Network Analysis.

The rest of the paper is organized as follows: Section 2 presents the speaker segmentation approach, Section 3 describes the role recognition approach, Section 4 presents experiments and results and Section 5 draws some conclusions.

2. SPEAKER SEGMENTATION

The speaker segmentation technique applied in this work is fully described in [3]. 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

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

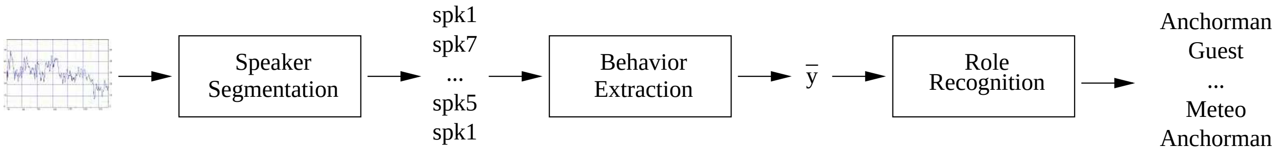

Fig. 1. General scheme. The figure shows the different steps in the recognition process.

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 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.

In some cases, different speakers are merged erroneously into a single speaker. Moreover, the system tends to oversegment and to create many short segments attributed to spurious speakers, i.e. speakers that do not actually exist in the recording. The latter problem can be addressed by observing that 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 [4]. 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 [4]. Once λ is estimated (we use 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. The $1/2$ threshold has been selected arbitrarily *a-priori* and not other values have been tried. 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.

3. ROLE RECOGNITION

Given a recording, each speaker a_i is represented with a behavior pattern $\vec{y}_i$ (see below for more details). The role recognition problem can be thought of as the identification of the role $\hat{r}$ leading to the *Maximum A-Posteriori* (MAP) probability:

$$\hat{r} = \arg\max_{r \in \mathcal{R}} p(r|\vec{y}_i) \quad (5)$$

where $\mathcal{R}$ is the set of the predefined roles. By applying the Bayes Theorem and by observing that $p(\vec{y}_i)$ is constant for a given speaker, the above equation can be rewritten as follows:

$$\hat{r} = \arg\max_{r \in \mathcal{R}} p(\vec{y}_i|r)p(r), \quad (6)$$

where $\hat{r}$ is said the MAP role. The first term of the product in Equation 6 is the likelihood of the behavior pattern, while the second term is the *a-priori* probability of observing role r. When $p(r)$ cannot be estimated, it is assumed to be uniform and Equation 6 reduces to:

$$\hat{r} = \arg\max_{r \in \mathcal{R}} p(\vec{y}_i|r). \quad (7)$$

In this case, $\hat{r}$ is said to be the *Maximum-Likelihood* (ML) role. The experiments of this work will show result obtained by applying Equation 6 and Equation 7.

So far, we considered the case of a single speaker, but news bulletins involve G individuals and each one of them must be given a role. In this work, we make the simplifying assumption that the roles of different speakers are statistically independent, then the problem of assigning G roles to G individuals can be expressed as follows:

$$\vec{r} = \arg\max_{\vec{r} \in \mathcal{R}^G} \prod_{k=1}^{G} p(r_k|\vec{y}_k) \quad (8)$$

where $\vec{r} = (r_1, \ldots, r_G)$ is the vector such that component r_i is the role of speaker a_i. The maximization of the right hand side of Equation 8 can be achieved by maximizing separately each term of the product. This results into applying G times the approach described above for a single speaker.

The rest of this section shows how $\vec{y}$ is defined and how the probabilities $p(\vec{y}|r)$ and $p(r)$ are estimated.

Role	AM	SA	GT	AB	MT	IP
Fraction (%)	41.2	5.5	34.8	7.1	6.3	4.0

Table 1. Corpus Characteristics. This table reports the percentage of corpus time that each role accounts for.

3.1. Behavior Extraction and Modeling

In general terms, the behavior is the collection of the activities performed by a human being. The factors influencing the behavior are multiple (e.g. attitudes, emotions, values, authority, etc.), and the role is one of them. In the case of broadcast news, mainly two behavior aspects are influenced by the role, i.e. *timing* and *length* of the interventions, and the behavior pattern extraction focuses on them.

Each bulletin is split into D non-overlapping windows spanning the whole recording length. Speakers are said to be *present* in the i^{th} window if they talk during at least a fraction of it, and *absent* otherwise. This leads, for a given speaker, to a D-dimensional binary vector $\vec{x} = (x_1, \ldots, x_D)$ where each component accounts for a window: when the speaker is present in window i, then $x_i = 1$, otherwise $x_i = 0$. The binary vector accounts for the timing of the speaker interventions, in fact it provides a rough representation of the times at which the speakers talk.

The second important behavior pattern is the intervention length. Each spaker talks for a *fraction* τ of the total duration of the bulletin, where $\tau > 0$, $\sum_{k=1}^{G} \tau_k = 1$, and G is the total number of speakers in a given bulletin.

The pattern resulting from the above behavior aspects has $D + 1$ dimensions and can be written as $\vec{y} = (\tau, \vec{x})$, i.e. $\vec{y} = (\tau, x_1, \ldots, x_D)$. The probability $p(\vec{y}|r)$ is expressed as a product of two terms:

$$p(\vec{y}|r) = p(\vec{x}|r)p(\tau|r), \tag{9}$$

where we make the assumption that $\vec{x}$ and τ are statistically independent for a given role r. The term $p(\vec{x}|r)$ is modeled using *Bernoulli Distributions* and the term $p(\tau|r)$ is modeled using Gaussians.

The Bernoulli Distributions (BD) are probability density functions defined over the space of binary vectors. The expression of a BD is as follows:

$$p(\vec{x}|\vec{\mu}_r) = \prod_{i=1}^{D} \mu_{ri}^{x_i}(1 - \mu_{ri})^{1-x_i} \tag{10}$$

where $\vec{\mu}_r = (\mu_{r1}, \ldots, \mu_{rD})$. Given a training set $\mathcal{X} = \{\vec{x}^{(i)}\}$, with $i = 1, \ldots, N_r$, the loglikelihood $\log p(\mathcal{X}|\vec{\mu}_r)$ is:

$$\log p(\mathcal{X}|\vec{\mu}_r) = \sum_{n=1}^{N_r}\sum_{i=1}^{D}[x_i^{(n)} \log \mu_{ri} + (1 - x_i^{(n)}) \log(1 - \mu_{ri})], \tag{11}$$

where N_r is the number of speakers playing the role r in the training set. The Maximum-Likelihood estimates of the parameters can be obtained by maximizing the above expression and the result is:

$$\mu_{rk} = \frac{1}{N_r}\sum_{n=1}^{N_r} x_k^{(n)}, \tag{12}$$

i.e. the parameter μ_{rk} is simply the average of the k^{th} components of the $\vec{x}$ vectors in the training set.

The probabilities $p(\tau|r)$ are estimated using Gaussian distributions $\mathcal{N}(\tau|\mu_r, \sigma_r)$. The ML estimation of the parameters can be obtained, like in the BD case, by maximizing the log-likelihood of the model over a training set:

$$\mu_r = \tfrac{1}{N_r}\textstyle\sum_{n=1}^{N_r} \tau^{(n)} \quad \sigma_r = \tfrac{1}{N_r}\textstyle\sum_{n=1}^{N_r}[\tau^{(n)} - \mu_r]^2 \tag{13}$$

where N_r is the number of speakers playing role r in the training set. The parameters μ_r and σ_r^2 are also known as the *sample mean* and the *sample variance* respectively.

The last term to estimate is the *a-priori* probability $p(r)$ of observing role r (see Equation 6). In this work, $p(r)$ is estimated with the fraction of the data the role r accounts for (see Table 1).

4. EXPERIMENTS AND RESULTS

This section presents the experiments and results obtained in this work. Section 4.1 describes data and roles and Section 4.2 shows the role recognition performance.

4.1. Data and Roles

The experiments of this work have been performed over a collection of 96 news bulletins provided by Radio Suisse Romande, the French speaking Swiss national broadcasting service. The corpus includes all bulletins broadcast during February 2005 and can be considered as a representative sample of this specific kind of emissions. The average bulletin length is 11 minutes and 50 seconds and the single durations range between 9 and 15 minutes roughly. The average number of speakers participating in the bulletins is 11 and 99% of the material corresponds to people talking, while only 1% of the data contains background noise, jingles, music, etc.

Each individual plays one among six predefined roles: the *anchorman* (AM), i.e. the person managing the bulletin, the *second anchorman* (SA), i.e. the person supporting the AM, the *guest* (GT), i.e. the persons invited to present a single and specific issue, the *interview participant* (IP), i.e. the persons having an exchange in an interview, the *abstract* (AB), i.e. the person reading a short summary at the beginning of the bulletin, and the *meteo* (MT), i.e. the person presenting the whether forecast at the end of each bulletin. Table 1 shows the percentage of the corpus each role accounts for.

Role	all	AM	SA	GT	AB	MT	IP
ML	92.2	97.9	76.0	98.5	99.8	97.9	0.0
MAP	92.7	97.9	76.0	99.7	99.8	96.8	0.0

Table 2. Upper bound performance. The table shows the results obtained over the *manual* speaker segmentation.

Role	all	AM	SA	GT	AB	MT	IP
ML	80.9	94.9	0.5	94.5	58.9	77.7	0.0
MAP	81.1	94.9	1.0	95.8	58.9	73.4	0.0

Table 3. Recognition Results. The second column reports the overall accuracy, i.e. the percentage of data time correctly labeled in terms of time. The results of this table have been obtained over the automatic segmentation.

4.2. Role Recognition Results

The first step in applying the recognition approach presented in Section 3 is the selection of the hyperparameter D, i.e. the number of non-overlapping windows spanning each bulletin. Since the longest bulletin is around 15 minutes long, D has been set *a-priori* to 15. In this way, the windows are never longer than one minute, a time comparable with the average intervention length in the corpus. The hyperparameter D is likely to affect the recognition performance, but no other values than 15 have been tested in this work.

The training of the models is performed using a *leave-one-out* approach: the models are trained over all bulletins except one, the so-called *left-out*, which is used as a test set. All bulletins in the corpus are used as left-out so that the whole corpus can be used as test set without the risk of overfitting [5].

The first set of experiments has been performed over the manual speaker segmentations. This is expected to provide an upper bound of the role recognition performance and to show if the models actually capture the characteristics of the roles. The performance is measured in terms of accuracy, i.e. of the data time percentage correctly labeled in terms of role. The overall performance is 92.2 percent for the ML approach and 92.7 percent for the MAP approach. Table 2 shows the results role by role. The only role which is not recognized at all is the IP, the reason is that it is confused with the GT, a role with a similar $\vec{x}$ distribution, but a much higher $p(\tau|r)$ component.

The second set of experiments has been performed on the automatic speaker segmentations obtained with the system described in Section 2. The results are reported in Table 3 and show that there is an accuracy difference of around 10 percent with respect to the results obtained over the manual segmentations. However, the difference is not the same for all roles: in some cases (AM and GT) the degradation is slight and the segmentation errors seem to have a limited impact. In other cases (AB and MT), the degradation is more evident and it is due essentially to the music used as background in correspondence of certain roles. Finally, the performance for SA and IP is definitely unsatisfactory. Fortunately, the latter roles account for a small percentage of the total amount of data.

Most of the errors are concentrated at the transition between one speaker and the following one. In fact there is a delay between the actual transition and the time where the system actually detects the transition. The average recording time is around 12 minutes (720 seconds) and 20 percent of errors means roughly 140 seconds. Since there are on average 30 interventions, such an error amounts to around 5 seconds per transition, a delay that can be easily managed by users.

The difference between MAP and ML is relatively low, i.e. the *a-priori* probability of a given role seems not to play a major role. The main reasons is, in our opinion, that most of the role relevant information is conveyed by the behavior patterns thus the contribution of the *a-priori* probability is not determinant.

5. CONCLUSIONS

This work has presented an approach for the role recognition in broadcast news. The approach is based on simple Machine Learning techniques including Bernoulli Distributions and single Gaussians. The results show that certain roles (i.e. AM and GT) are recognized with high accuracy, while others are recognized poorly or even systematically missed (i.e. SA and IP), however, these latter account for a small fraction of the data and their impact on the overall performance is acceptable.

6. REFERENCES

[1] R. Barzilay, J. Collins, M. Hirschberg, and S. Whittaker, “The rules behind the roles: identifying speaker role in radio broadcasts,” in *Proceedings of AAAI/IAAI*, 2000, pp. 679–684.

[2] A. Vinciarelli, “Sociometry based multiparty audio recordings segmentation,” in *Proceedings of IEEE Conferences on Multimedia and Expo*, 2005, pp. 1801–1804.

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

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

[5] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer Verlag, 2006.