Multi-view approach for speaker turn role labeling in TV Broadcast News shows

Géraldine Damnati, Delphine Charlet
Orange Labs, France Telecom, Lannion, France
{geraldine.damnati,delphine.charlet}@orange-ftgroup.com

1. Introduction

Speaker role recognition (SR) in TV Broadcast News is the task of assigning a role to a speaker. This task constitutes a preliminary aspect of general speaker distillation tasks aiming at structuring documents. It has been explored in the context of Broadcast News (BN), Broadcast Conversations (BC), Talk Shows, meetings or conversational phone calls analysis with various objectives, various role definitions and distributions.

This work addresses the case of TV Broadcast News (TVBN). Shows are considered as a whole including pure BN and BC. The purpose is to classify speaker turns which we believe is a relevant input for facilitating intra-document navigation or for other applications that do not require retrieving speaker clusters. The problem is an automatic segmentation process. For our labeling system classifies each speaker turn in three categories. Anchor is the main speaker reading news and introducing reports in a calm environment. Reporter is a journalist with professional elocution mode but usually in variable environment conditions. Other represents any guest speaker and concentrates all variability sources: environment, elocution mode from controlled to very spontaneous. These three categories vary from many points of view, including elocution mode, lexical field, environment: from calm for anchor to potentially very noisy for reporters and others. Intuitively, evidence for SR recognition can be found both from the acoustic and the linguistic channel. In previous work [1], we used acoustic information for anchor detection and proposed reporter/other binary classification over the linguistic channel by using Automatic Speech Recognition (ASR) output. In this work, we further exploit channel complementarities by fusing information sources at the reporter/other classification level. Furthermore, we propose a preliminary study showing that such complementarities can lead to a finer-grained characterization of other speaker turns.

2. Related work

One of the first works on this topic was published in 2000 [2] and recently the number and variety of studies around role labeling significantly increased. Approaches differ from several points of view: the amount of roles considered, the type of shows (Broadcast News, Talk Shows) and the labeling level (speech segments, speaker turns, data time or speakers). [3] provides a synthetic overview of state-of-the-art performance for several combinations of these dimensions.

3. Role labeling of speaker turns

In [1], we have proposed a multi-stage process for speaker turn role labeling. The first stage consists in determining the anchor speaker and then to classify the remaining speaker turns into reporter or other as illustrated in Figure 1.
Anchor speaker has a specific temporal distribution due to regular occurrences through TVBN shows. In previous works, [7] we have shown that anchor detection can be seen as a specific speaker clustering sub-task, for which temporal distribution information is taken into account in the choice of the relevant cluster. After processing the whole show for anchor detection, each non-anchor speaker turn is submitted to a binary reporter/other classifier. In previous work, we implemented this process by using icsiboost [8] on ASR transcripts (as briefly recalled in section 4.1). This classifier mainly relies on the lexical content of turns and on the ASR quality thanks to confidence scores. In this paper we study an alternative approach which exploits the MFCC acoustic analysis of the speech signal. We then evaluate two fusion approaches in order to take advantage of both information sources.

4. Reporter/Other classification

This section focuses on the binary reporter/other classification task, describing the two complementary classification methods and their fusion. The resulting four approaches are summarized in Figure 2 where a decision module takes as input one of the 4 different scores described in the following sections. Given that we are dealing with binary classification, scores are given as the reporter class score.

4.1. Classification based on ASR transcripts

This approach consists in using a boosting-based text classifier (icsiboost) to learn lexical cues of reporter and other speakers. As explained in more details in [1], the classification process input is built upon the ASR process output. Not only is this an interesting aspect of the approach with respect to the limited annotation effort needed to train the models, but we have shown that making use of ASR outputs can improve the classification performance. In fact, ASR transcription quality is itself a good indicator of speaker role. Reporters are professional speakers whereas other speakers are more likely to be transcribed with higher word error rates. In order to model transcription quality, ASR outputs are pre-processes on the basis of word confidence indicators: all the words whose confidence measure is below a given threshold are replaced by a generic <BAD> symbol. Bags of n-grams (from 1 to 5) over this preprocessed sequence constitute the basic textual feature set. Additionally, two numerical features are considered: the elocution speed (number of syllables per second).

In practice, the classifier is trained and applied on shorter segments delimited by pauses. Posterior probabilities of each class given each segments are weighted by segments duration and interpolated. Let \( X \) be a speaker turn composed of \( N(X) \) speech segments \( s_n \) of duration \( d(s_n) \). Let \( P_{\text{ASR}}(\text{reporter}|s_n) \) be the posterior probability of the class reporter provided by icsiboost trained from ASR-based features for segment \( s_n \): \n
\[
S_{\text{ASR BOOST}}(X) = \sum_{n=1}^{N(X)} \frac{P_{\text{ASR}}(\text{reporter}|s_n) \times d(s_n)}{\sum_{n=1}^{N(X)} d(s_n)}
\]

4.2. GMM decoding on frame level acoustic analysis

Related works report classifiers based on acoustical features computed on the entire speaker turn [3]. Hence, for a given speaker turn, a single vector of acoustical features is computed and submitted to a classifier to determine the role of the speaker. In this work, we evaluate the hypothesis that all the speakers which play the same role share some acoustical properties which can be captured at the frame-level (short term analysis window). Thus, a GMM-classifier based on a frame-level likelihood computation is used here for role classification. This approach uses GMM based on MFCC features at the frame-level. Each role is then modeled by a GMM, trained with Expectation-Maximization algorithm. The decision score for a given speaker turn \( X \) of \( T_x \) frames relies on a time-normalized log-likelihood ratio:

\[
S_{\text{GMM}}(X) = \frac{1}{T_x} \log \frac{P(\text{reporter}|X)}{P(\text{other}|X)} = \frac{1}{T_x} \log \frac{P(X|\text{reporter})P(\text{reporter})}{P(X|\text{other})P(\text{other})}
\]

where \( P(X|\text{reporter}) \) and \( P(X|\text{other}) \) are the likelihoods computed by each role GMM at a frame-level and accumulated over the entire speaker turn. In practice, \( a \) priori probabilities \( P(\text{reporter}) \) and \( P(\text{other}) \) are considered equal.

4.3. Early fusion through feature enhancement

The first fusion approach consists in taking advantage of the ability of icsiboost to perform global optimization from both textual and continuous numerical features. The \( S_{\text{GMM}} \) score obtained on a given speaker turn is provided for the classification process of each segment as an additional feature. Note that experiments consisting in computing the \( S_{\text{GMM}} \) score at the segment level yielded equivalent results for early fusion but we keep the turn level computation for the sake of comparison between the two fusion methods. The approach is represented in Figure 2 by the "early fusion" arrow. Following the same two-step approach as the one described in 4.1, the segment posterior probabilities are combined by weighted interpolation and the resulting turn level score \( S_{\text{EARLY}} \) is provided to the decision module. Let \( P_{\text{ASR,GMM}}(\text{reporter}|s_n) \) be the posterior probability of reporter provided by icsiboost trained from the enhanced feature set for segment \( s_n \): \n
\[
S_{\text{EARLY}}(X) = \sum_{n=1}^{N(X)} d(s_n) \times P_{\text{ASR,GMM}}(\text{reporter}|s_n)
\]

4.4. Late fusion through score combination

The ASR-based classifier score \( S_{\text{ASR BOOST}} \) and the GMM-MFCC classifier score \( S_{\text{GMM}} \) are fused with logistic regression in order to obtain a final classification score:

\[
S_{\text{LATE}}(X) = \frac{1}{1 + e^{-a_0 + a_1 S_{\text{ASR BOOST}}(X) + a_2 S_{\text{GMM}}(X)}}
\]

Logistic regression is trained on a development set as a reporter detector, thus, the fused score should tend to 1 for a reporter speaker turn, and to 0 for an other speaker turn.

---

![Figure 2: reporter/other classification approaches](image)
5. Experiments

5.1. Data description

The corpus is composed of 38 TVBN shows collected from 7 French TV channels between October 2008 and January 2009, with a variable length from 10 to more than 40 minutes and a number of different speakers ranging from 10 to 80. On the overall, the corpus corresponds to 14.5 hours of speech, for a total amount of 158k words uttered by 1400 speakers. 24 shows are used for training while the TEST corpus is composed of the 14 remaining shows. TVBN shows are from the main French generalist channels and BC portions (interviews, live reports…) are kept in the evaluation process. Apart from the anchor speakers, 70% of them can be identified from the audio. From 337 such non-anonymous speakers in the TEST corpus, 38 also occur in the training corpus, representing 15.5% of reporter and other speaker turns.

Automatic transcription is performed using the VoxSigma speech recognizer V3.4 from Vocapia Research, which is based on LIMSI technology [9]. The word error rate (wer) on the TEST corpus is 21.9% ranging from 13.4% on planned speech to 39.7% wer on spontaneous speech. We use the automatic word transcription as well as word level confidence measures (posterior probabilities). Automatic speaker turn segmentation is also provided by the VoxSigma recognizer and shorter speech segments are further extracted by splitting speaker turns on detected pauses.

For the reporter/other classification approach described in 4.2, a 256-components GMM is trained for both roles, using 36 dimensional acoustical vectors (12 MFCC + delta + delta-delta). "reporter" GMM is trained with 2h55min of speech and "other" GMM with 1h55min of speech. When used as an additional feature for early fusion, GMM models are trained by leaving-one-out, considering the current training TVBN show as a held-out show. Finally, the late fusion regression coefficients are also learnt by leaving-one-out.

In order to evaluate our approaches, automatically segmented speaker turns must be assigned a reference label. To this end, we assign the role of the speaker with the maximum temporal overlap. Several problems can arise with this approximation. There can be segmentation errors (several speakers in one turn) or speech activity detection errors (single speaker with low overlap or no speaker at all). We consider that a speaker turn has a good matching with reference label when it corresponds to only one speaker with more than 85% temporal speech coverage. 77% of derived turns reference label have a good matching (95.7% of anchor, 87.4% of reporter and 63.3% of other). In order to have a good evaluation, we have chosen to keep all speech turns, should this approximation. There can be segmentation errors (several speakers in one turn) or speech activity detection errors (single speaker with low overlap or no speaker at all). We consider that a speaker turn has a good matching with reference label when it corresponds to only one speaker with more than 85% temporal speech coverage. 77% of derived turns reference label have a good matching (95.7% of anchor, 87.4% of reporter and 63.3% of other). In order to have a good evaluation, we have chosen to keep all speech turns, should they have an approximate reference label. The TEST corpus with derived reference role annotation is depicted in Table 1.

5.2. Reporter/other classification evaluation

Classification performances depend on the threshold applied on the output classification score in the decision module of Figure 2. For a given threshold value, two types of errors can be computed (the rate of reporter turn misclassified as other turns, and the rate of other turns misclassified as reporter turns). Figure 3 shows the evolution of those two errors depending on the threshold for the 4 classification approaches studied in this paper. Additionally, in order to summarize performances in one numerical value, focus is made on one particular point: the Equal Error Rate (EER), where the reporter→other confusion rate is equal to the other→reporter confusion rate.

In order to evaluate the various approaches proposed in section 4, we focus on those speaker turns that are not detected as anchor by the initial speaker clustering process. The anchor detection method recalled in section 3 yields a high precision rate of 94.0% however a few other turns are detected as anchor. Thus the following curves reflect the behavior of the approaches on 1082 turns (601 reporter and 481 other).

Figure 3: other→reporter confusions for individual classification ScASR_BOOST, ScGMM and their fusion.

Curves in Figure 3 show the good complementarity of ASR-based and GMM approaches, as both fusion methods lead to a significant error rate reduction compared to the initial ASR approach (37% relative reduction of the EER for late fusion). Late fusion should be preferred if high recall values are targeted for the class other (top left of the curve).

5.3. Overall speaker turn labeling evaluation

In order to have an overview of the performances of the complete 3 class anchor/reporter/other classification task, we evaluate the overall classification accuracy (i.e. the number of speaker turns which are assigned the correct role over the total number of speaker turns) obtained when setting thresholds to 0.5. In [1], the icsiboost classifier was trained on a subset of 15 TVBN shows out of the 24 available shows. Classification accuracy for ScASR_BOOST on TEST was 85.7%. In these experiments, free parameters have been tuned by performing leaving-one-out over the 24 shows of the training corpus, leading to 86.8% accuracy for the same method.

Table 2: Overall performances on the TEST corpus.

<table>
<thead>
<tr>
<th>classification</th>
<th>accuracy</th>
<th>anchor f-measure</th>
<th>reporter f-measure</th>
<th>other f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScASR_BOOST</td>
<td>86.8%</td>
<td>93.1</td>
<td>87.2</td>
<td>81.8</td>
</tr>
<tr>
<td>ScEARLY</td>
<td>89.6%</td>
<td>93.1</td>
<td>90.0</td>
<td>87.2</td>
</tr>
<tr>
<td>ScLATE</td>
<td>90.1%</td>
<td>93.1</td>
<td>90.5</td>
<td>87.7</td>
</tr>
</tbody>
</table>

At this particular operating point, both fusion methods improve classification accuracy, with a slight advantage for late fusion, as observed on Figure 2 DET curves. The largest improvement is observed for the category other whose f-measure is increased by 5.9% absolute with late fusion. The overall classification accuracy reaches 90.1% on automatically segmented and automatically transcribed speaker turns.
5.4. Detailed analysis of other speaker turns

The other category naturally concentrates the largest variety of speakers. In order to deeper analyze the various approaches behavior we have extracted four particular types of speakers:

- **Politicians** are particular in the sense that they have both a particular lexical and semantic field and a particular elocution mode, closer to professional mode.
- **Translators** speaking over foreign language speakers are affiliated to the other categories while they actually usually are reporters. What's more translation is usually delayed and prepared, resulting in clean speech where disfluencies for instance are not reproduced.
- **Identified** speakers are speakers whose name has been identified by human annotators from the audio signal only. Those speakers are usually introduced by name by journalists (similarly to the turns retrieved in [5]).
- **Anonymous** are speakers who cannot be identified from audio signal. They usually are punctually asked for a testimony and introduced by generic location such as "these inhabitants", "the victim's lawyer".

The first two sets are not completely exclusive, since there can be a translator of non French speaking politicians (12 turns in common). The last two sets are mutually exclusive and do not overlap with the first two ones. Politicians represent 8% of other speaker turns but 11.2% in terms of duration. 60% of other turns are anonymous speakers, representing half of the total duration. However some subsets are small, observing the behaviour of our approaches gives us a tendency about their appropriateness to retrieve such or such categories of speakers.

For the following evaluations, reporter set is unchanged, and other set is restricted to the different subtypes. Thus, each curve provides the contribution of the subtype to the total amount of other–reporter confusions in abscise and gives an idea of the potential performances for a show with journalists and only anonymous speakers or journalists and only politicians for instance.

![Figure 4: Contribution of politician, translators, identified and anonymous to the other–reporter confusions for individual ScASR_BOOST and ScGMM](image)

Figure 4 shows that ASR-based boosting classification method is not sensible to the different categories of other speakers while the MFCC-based GMM classification behaves differently for the various sub-categories. Due to its construction through a combination of weak classifiers (1000 iterations in our experiments), icsiboost is more robust to class heterogeneity than the GMM classifier. What's more, the elocution mode in the other class is particularly contrasted from usually planned speech for politicians or translators to very spontaneous speech for anonymous. Hence GMM classifiers naturally encounter more difficulties discriminating reporters from politicians or translators, while they perform very well to discriminate reporters from anonymous speakers.

<table>
<thead>
<tr>
<th>Category</th>
<th>turns</th>
<th>average duration</th>
<th>% total duration</th>
<th>ScASR_BOOST EER</th>
<th>ScGMM EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politicians</td>
<td>39</td>
<td>17.2s</td>
<td>11.2%</td>
<td>15.4</td>
<td>10.3</td>
</tr>
<tr>
<td>Translators</td>
<td>60</td>
<td>12.0s</td>
<td>12.1%</td>
<td>13.3</td>
<td>11.7</td>
</tr>
<tr>
<td>Identified</td>
<td>104</td>
<td>16.4s</td>
<td>28.5%</td>
<td>13.5</td>
<td>7.7</td>
</tr>
<tr>
<td>Anonymous</td>
<td>290</td>
<td>10.5s</td>
<td>50.8%</td>
<td>13.5</td>
<td>7.9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>481</td>
<td>12.4s</td>
<td>5968s</td>
<td>13.9</td>
<td>8.7</td>
</tr>
</tbody>
</table>

It is interesting to notice in Table 3 that even if GMM classifiers have more difficulties with politicians and translators (with around 20% EER), late fusion still improves performances for the four sub-categories.

Finally, this analysis suggests exploring some particular strategies towards a finer granulated characterization of other speaker turns. Future work will address this issue by further exploiting the complementarity of the two classification approaches, exploiting ASR transcriptions on one hand and acoustic analysis on the other hand.

6. Conclusions

We have extended our speaker turn role labeling approach which consists in detecting the anchor speaker in a first stage and performing binary reporter/other classification in a second stage on the remaining speaker turns. We have proposed a multi-view approach for the second stage through the fusion of classifiers exploiting lexical cues and classifiers exploiting acoustic cues. A 37% relative EER reduction is observed for reporter–other confusions when comparing the multi-view approach to the single lexical view approach. As a result, an overall three-way classification accuracy of 90.1% is obtained from automatically segmented and transcribed speaker turns on complete TV news shows including BN and BC. Additionally, we have proposed a detailed analysis along several sub-categories of other speaker turns suggesting that further improvements could be obtained thanks to a finer-grained multi-view characterization of these speakers.

7. References