ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

A speaker-role based approach for detecting politicians in TV broadcast news

Delphine Charlet, Géraldine Damnati

Orange Labs, France Telecom, Lannion, France
{geraldine.damnati,delphine.charlet}@orange.com

Abstract

Politician speaker turn detection in TV Broadcast News shows is addressed in this paper. Politician speech model combines acoustical and lexical cues as well as contextual information, and does not use any specific politician model (person-independent). Politician speaker turn detection is coupled with an automatic role labeling step, which determines the contextual information and the set on which politician detector is applied. On a set of 101 TV broadcast news shows, experiments show that the politician speaker turns, which represent only 3% of the whole set of speaker turns in the corpus, are detected with a maximal F-measure of 65.6%.

Index Terms: politician speech detection, speaker role

10.21437/Interspeech.2012-405

1. Introduction

Politician speakers are of particular interest for indexing broadcast news: extracting automatically their declarations on TV enables to build, on a daily basis for instance, a kind of "politician best of". Their speaker turns may also be used as a special illustrative excerpt for automatic summary. Detecting politician speaker turn can also be a complementary step (preprocessing or confirming step) in a person identification process dealing with politicians.

Politician speech has been studied in the computational linguistics community from a lexical and structural point of view but most works aim at analyzing intra-class properties of politician speech: analyzing elocutionary features in various situations (e.g. turn taking in political interviews [1] parliamentary debates, campaign meetings…) comparing politicians one from the other, or a certain group of politicians from the other [2], or quantifying a degree of conflict in broadcast political debates [3]. In this work we are interested in retrieving politicians from TV Broadcast News (TVBN) shows, thus to detect them as a particular category distinct from the rest of the speakers involved in TVBN shows. For genericity purpose we do not consider speaker-dependent models and perform person-independent politician detection.

Politician speaker turn is characterized by several dimensions. The lexical dimension is of course characteristic of political speech in general but politicians can also be characterized by a particular elocution mode. Even though it can vary depending on the situation (discourse, debate, public allocution, interview) human listeners can quite accurately recognize politicians based on their elocution (see e.g. [4]where perceptual experiments are reported with listeners submitted to delexicalized prosodic cues), but modeling this prosodic dimension remains a challenge. Another difficulty lies in the fact that many non-politicians talk about politics in TVBN, whereas the actual politician speech is rather rare: in the TVBN corpus described in details in [5], politicians speaker turns only represent 3% of the total amount of speaker turns. In order to address this particularly unbalanced data classification task, we propose to restrict the set of candidates to the politician speaker turn detection by automatically discarding journalists speaker turns thanks to an automatic role labeling step that classifies *anchor* speaker, *reporter* and *others*.

We are interested in processing entire TVBN shows (without distinguishing pure BN speech and BC parts), in a fully automatic process, based only on the audio channel. All lexical cues are extracted from automatic transcription, even for the training phase. The only manual annotation required in our work for training was the labeling of automatically segmented politician speaker turns.
In section 2, politician speech modeling is presented. Section 3 presents the integration of this modeling with the role labeling system. Section 4 exposes a refinement in political speech modeling, and experiments are detailed in section 5.

2. Politician speech modeling

This section focuses on the core binary politician/non-politician classification task. Three information sources are explored: the way politicians talk (at the acoustical level), the words politicians use and the context in which the politicians talk. Hence, the problem is addressed in a multi-view framework, where each view corresponds to an information source. For the approaches related to the words used by the politicians or the context, automatic speech recognition transcription is used.

The same general log-likelihood ratio framework is used for each view:

X is the speaker turn to classify, obtained after automatic speech segmentation, and is represented with a sequence of T_x elements (x_1,.., x_{Tx}) (the elements depend on the explored view, and will be detailed later). The aim is to find the label among (*polit*,¬*polit*) which gives the highest probability given the sequence of elements. Assuming the independency of the elements, a length-normalized likelihood ratio score for X, for the view V_j, is computed as follows:

$$Sc_{V_j}(X) = \frac{1}{len(X)} \log \frac{P(polit|X)}{P(\neg polit|X)} = \frac{1}{len(X)} \log \frac{P(X|polit)P(polit)}{P(X|\neg polit)P(\neg polit)} \quad (1)$$

$$Sc_{V_j}(X) = \frac{1}{len(X)} \log \frac{\left(\prod_{i=1}^{T_x} P(x_i|polit)\right) P(polit)}{\left(\prod_{i=1}^{T_x} P(x_i|\neg polit)\right) P(\neg polit)}$$

<u>*Acoustical level view*</u>: X is the sequence of T_x acoustical frames of the speaker turn, represented with MFCC vectors. The probability density functions are Gaussian mixtures. The length normalization *len(X)* is simply the number T_x of acoustical frames. Thus, *polit* and ¬*polit* are modeled with GMM on

MFCC. This modeling, although very simple, when applied to model *reporter* and *other*, has proven its effectiveness in [5].

Word-level view: X is the sequence of T_x transcript words of the speaker turn. The probability density functions are discrete densities based on word-frequency in each class[7]. The length normalization is given by $len(X)=1+log(T_x)$ (which experimentally gives good performances). A comparison was made between this Naive Bayes classifier and a boosting-based classifier (adaboost using bag of N-grams on the transcription and elocution features such elocution speed, number of pauses per second,…). Experiments, not reported here, showed that the approach based on likelihood ratio on word frequency performed much better on this task than the boosting-based one, because it is more robust to the low amount of training data in such unbalanced context.

Context-level view: X is the sequence of T_x transcript words for the segment of speech which is considered as the context. The probability density functions are discrete densities based on word-frequency for the context of each class, and the length normalization is given by $len(X)=1+log(T_x)$. In the next section, we will discuss on the automatic definition of the context, with an approach that integrates a role labeling step

Multi-view approach: the different likelihood-ratio scores ($Sc_{word}(X)$ for word-based view of the current speaker turn, $Sc_{GMM}(X)$ for acoustical view, $Sc_{context}(X)$ for word-based view of the context) are fused with logistic regression to obtain a final classification score:

$$Sc(X) = \frac{1}{1+\exp-\left(a_O + a_1 Sc_{word}(X) + a_2 Sc_{GMM}(X) + a_3 Sc_{Context}(X)\right)} \quad (2)$$

3. Integrating Role Labeling

In [5], 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*. *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 which mixes lexical and structural information captured using adaboost classifier on ASR transcripts and acoustical information captured with a GMM on MFCC acoustical analysis of the speech signal. This 3 class role labeling process at speaker turn level achieves 90% accuracy[5].

3.1. Context definition

Detecting *anchor* speaker enables to automatically define a context: it is assumed that each anchor speaker turn defines a new "chapter", and the context of the speaker turn to label is potentially made with all the speech sequence between 2 anchor speaker turns. This is an approximation as with such a definition several consecutive "chapters" can actually correspond to the same topic, but it is expected that within such a "chapter", there is a consistent topic. Variants on the span of the context based on this chapter definition have been explored, and experiments, not reported here, showed that best performances are obtained when using as a context the whole "chapter".

3.2. Filtering speaker turns with role labeling step

According to the level of integration of the role labeling step in the politician detection process, the set of speaker turns submitted to politician speech detector varies, as well as the availability of the context. 4 approaches are explored:

The *NoRole* Politician Detection approach considers the detection of politician speaker turns among all the speakers' turns, without any role labeling step. Thus if any role labeling step is performed, the context as defined in our paradigm is not available. As a variant, the system *NoRoleContext* Politician Detection which considers the detection among all the speaker turns, but with context available is explored.

The *Anchor* Politician Detection approach integrates only the automatic detection of the anchor speaker: the politician speech detector is only applied to the speaker turns not attributed to anchor. With the anchor detection, context is available.

The *FullRole* Politician Detection classification process fully integrates the role labeling system, and the politician speech detector is only applied on the automatically labeled *others* speaker turn. The context is defined thanks to the anchor speaker, like in the *Anchor* Politician Detection process. The whole process of the full integration of the role labeling step into the politician speech detection is presented in figure 1.

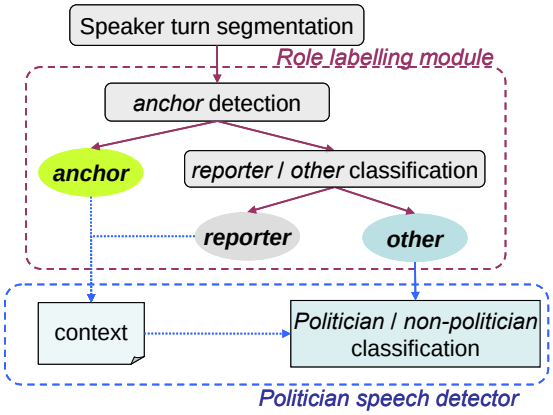

Figure 1: *FullRole Politician Detection system's architecture*

4. Refining word-level competitive models for political topics

With the context defined as the chapter between 2 anchor speaker turns, all the speakers within a same chapter shared the same context. Thus, all the speakers who talk in a chapter where a politician talks have the same context score as the politician. Moreover, as these speakers are very likely to talk about politics, they also get a good word score, because in the word-frequency based approach, words about politics are much more frequent for politicians than for non-politicians on average. Thus these speakers can generate false detection. In order to better reject these non-politicians sharing the same context as the politicians, we propose to model them explicitly, by training specific word-frequency based model for them. This class is named *politcontext*. The training corpus for this class is made with all the non politician speaker turns that are in the same chapter as the manually labeled politician speaker

This specific modeling can also be suitable for all the speakers who talk about politics, even in chapters where there is no politician speaker turns. This is the reason why we propose to apply this competitive model as an alternative competitive model for all speaker turns, and not to restrict its use to the speaker turns for which the context score is high. Then, two alternative models are available for non-politician: ¬*polit* is kept as a general competitive model, and *politcontext* is added as a supplementary competitive model. In the new likelihood ratio score, for a given test, only the competitive model giving the best likelihood is kept. The proposed likelihood-ratio score based on word-frequency is then:

$$Sc_{word}(X) = \frac{1}{len(X)} \log \frac{\left(\prod_{i=1}^{T_X} P(x_i | polit) \right) P(polit)}{L_{Alter}(X)} \quad (4)$$

where

$$L_{Alter}(X) = \max \left(\begin{array}{l} \left(\prod_{i=1}^{T_X} P(x_i | politcontext) \right) P(politcontext), \\ \left(\prod_{i=1}^{T_X} P(x_i | \neg polit) \right) P(\neg polit) \end{array} \right)$$

5. Experiments

5.1. Data collection & Corpus Analysis

A first corpus of 24 TVBN shows collected from 7 French TV channels between October 2008 and January 2009, has been manually annotated (complete transcription, speaker names and roles). As this corpus only contains 91 politician speaker turns, it was necessary to collect more data. The target was to collect enough TVBN shows to get several hundreds of politician speaker turns. Considering the low frequency of politician speech in TVBN, it was necessary to collect TVBN during several months. We collected the daily TVBN evening shows from 4 major French generalist TV channels, during spring 2011. The resulting set of 192 collected TVBN shows, totalizing about 94 hours of contents, was divided into 2 temporally separated sets: 91 shows to augment training data and 101 shows for the TEST corpus. In order to avoid an expensive and time consuming manual annotation of this corpus, we have performed a semi-automatic annotation in the following way: TVBN shows were processed with automatic speech transcription tool, which not only provides speech transcription but also speaker segmentation. Then, using the *Transcriber* interface, we quickly browsed through the automatic transcription, and manually labeled the automatic speaker turns which correspond to politicians with their names. The concept of politician is restricted to French politicians, known at a national level.

From a global training corpus totalizing about 54 hours of contents, the training set for modeling *politician* is made of 519 politician's speaker turns, from 142 different politicians (23 women, 119 men) with an average turn length of 15.4s, which makes a total amount of 2 hours and 13 min of speech.

Hence, the rareness of the politician speaker turns in TVBN leads to the necessity of a large corpus, which in turn leads to light annotations. Thus, to train models in the various approaches we investigate, training corpora have been built mixing manual annotations (on politician speaker turns) and automatic role labeling.

For modeling specifically ¬*polit*, the training corpus for ¬*polit* depends on the level of integration of the role-labeling process. The table 1 summarizes the size of the training corpus for modeling ¬*polit* according to the approach.

Politician Detection Approach	*set of speaker turns*	*# spk turns*	*Total amount of speech data*
NoRole	All non-politicians	11442	~ 46 hours
Anchor	Reporter and other non-politicians	8763	~32 hours
FullRole	Other non-politicians	4068	~13 hours

Table 1: *influence of the role labeling integration for ¬ polit training corpus .*

For training models for context-level view, the semi-supervised approach that mixes manual labeling of politician turns and automatic role labeling is also used: chapters are defined between 2 automatically detected anchor speaker turns. For training model for the word-level view for *politcontext*, we use the automatically defined chapter around a manually labeled politician speaker turn. Discarding the politician speaker turns in these chapters, the training corpus for *politcontext* is made of a set of 2270 speaker turns, for an amount of about 8 hours of speech.

The corpus for testing is composed of 101 TVBN shows containing 387 politician speech segments from 96 different politicians (19 women, 77 men), with 42 also present in the training set. As this TEST corpus is not so big, we did not want to divide it into dev and test. Then, to learn fusion coefficients in logistic regression, the TEST corpus is split into 5 equal parts and for each part, logistic regression was trained on the 4 remaining parts.

Automatic transcription is performed using the VoxSigma speech recognizer V3.5 from Vocapia Research, which is based on LIMSI technology[6]. This software provides speech recognition, with speaker segmentation. The evaluation of the transcription is performed on the first 24 shows fully annotated corpus: the word error rate (wer) is 15.9%. As automatic speaker segmentation is used, there can be errors of segmentation: we discard from the training and testing set the automatic speaker turns that contain speech from politicians and non-politicians within the same turn. For the test set, this case concerns 18 speaker turns, to be compared to the 387 pure politicians speaker turns.

5.2. Politician speech detection evaluation

According to the level of integration of the role-labeling step, the politician speech detector is not applied on the same test set.

Table 2 summarizes the set of politician and non-politician speaker turns which are submitted to the politician speech detector, for the different approaches. The table shows that errors in role labeling step lead to the rejection of a certain number of politicians speaker turns (from the total 387 politician speaker turns set, 22 politicians speaker turns are detected as anchor, and 33 as reporter).

Evaluation of the politician speech detector is performed with precision and recall rates, computed at the speaker turn

level. Evaluation is performed for the whole processing that takes into account the errors of role-labeling step, when it is used: thus, the recall rate for politicians is always computed on the initial 387 speaker turns set.

Politician detection approach	*# polit spkturns*	*# ¬polit spkturns*
NoRole	387	10203
Anchor	365	7859
FullRole	332	3619

Table 2: *size of speaker turns set submitted to politician speaker turn detector*

Figure 2 shows the performances of the system according to the level of integration of the role labeling step and with the inclusion or not of alternative *politcontext* model (referred as Alter). The system *NoRole* Politician Detection cannot use any context information and apply the politician detector to all speaker turns: the performances show that the proposed politician speech detector is clearly not suitable without any role information nor context. When context information is introduced in *NoRoleContext*, a very significant improvement is observed. When using anchor turn detection, not only to define context, but also to reduce the set of speaker turns submitted to politician detector, in *Anchor* Politician Detection, no significant improvement compared to submitting all speaker turns to the politician detection is observed. Hence, making a difference between reporters and anchor does not help in this task, the only important point is to use anchor for context definition. When restricting the set of speaker turns submitted to the politician detector to labeled *others* speaker turns in the *FullRole* Politician Detection approach, performances are much better. The *¬polit* class is better modeled when restricted to *others*.

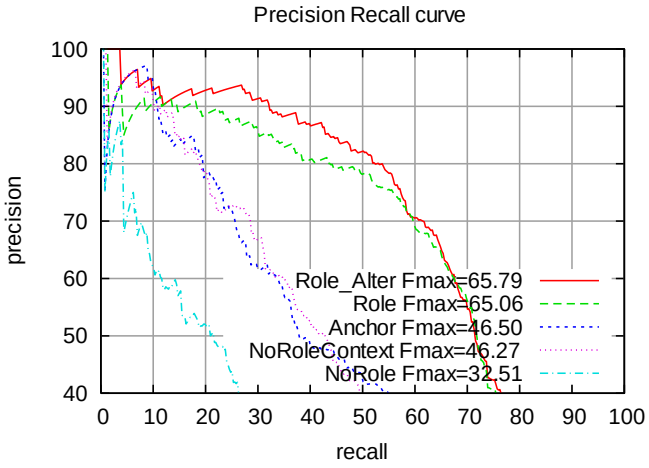

Figure 2: *Role Labeling integration for Politician Speech detection.*

Another interesting point is the influence of the inclusion of explicit alternative *politcontext* model using L_{Alter} score in $Sc_{word}(X)$. For the *FullRole* Politician Detection, using L_{Alter} improves in the high precision rate area. Without using L_{Alter}, the system discards a lot of politician speaker turns which get the same score as non-politicians talking about politics. Using the specific *politcontext* alternative model, some of these politicians speaker turns can be distinguished from the ones of non-politicians talking about politics, leading to a much higher recall rate. For a precision rate of 80%, the recall rate goes from 43.7% to 53.7% using L_{Alter}. For this operating point, 208 politicians speaker turns from 58 different politicians are detected, with 31 politicians being in the training corpus, and the other 27 ones never seen before. This confirms the interest of our person-independent politician detection approach. From a first analysis, false rejection are mainly due to bad context boundaries (error in anchor detection or the anchor talks about different subjects in the same turn), and difficult acoustic conditions (politicians speaking in the street during a demonstration) or unusual person (young woman whereas the majority of the politicians in the corpus are men over 50).

Although these performances are encouraging, further work is envisaged: we plan to model more specifically politicians according to the interaction situation (discourse, interview,…), particularly at the acoustic level. Such interaction-dependent modeling will require a supplementary annotation effort.

6. Conclusions

The task of generic politician speech detection in TVBN shows is explored. As politicians represent only 3% of the speaker turns in these shows, a large corpus of 192 TVBN shows has been collected, and the approach relies on a very lightly annotated training corpus, mixing manually labeled politician speaker turns and automatically role-labeled speaker turns. Politicians are modeled in a generic likelihood ratio framework on acoustical, lexical and contextual levels. An automatic role labeling between (*anchor*, *reporter*, *other*) enables to define the context and to refine the modeling, applying the politician speaker turn detector only to non-journalist speaker turns. A specific competitive model in the likelihood ratio framework is introduced and improves the recall in the high precision rate area. The proposed method could be generalized to the detection of other specific categories of people in TVBN, that share topics and a certain form of elocution, such as sportsmen.

7. References

[1] Adda-Decker, M., Barras, C., Adda, G., Paroubek, P., de Mareuil, P. B., Habert, B., "Annotation and analysis of overlapping speech in political interviews". In Proceedings of LREC, Marrakech, 2008.

[2] Yu, B., Kaufmann, S. and Diermeier, D. "Classifying party affiliation from political speech", Journal of Information Technology in Politics, 5(1): 33-48, 2008.

[3] Kim, S., Valente, F. and Vinciarelli, A., "automatic detection of conflicts in spoken conversations: rating and analysis of broadcast political debates", Proc. of ICASSP'12, Kyoto, 2012.

[4] Obin, N., Dellwo, V., Lacheret, A., Rodet, X, "Expectations for discourse genre identification; a prosodic study", Proc. Interspeech'10, Makuhari, 2010

[5] Damnati, G., Charlet, D. "Multi-view approach for speaker turn role labeling in TV Broadcast News shows", Proc. of Interspeech'11, Florence, 2011.

[6] Gauvain, J.L., Lamel, L. and Adda, G., "The LIMSI Broadcast News Transcription System". *Speech Communication*, 37(1-2):89-108, 2002.

[7] Frank, E., Bouckaert, R., "Naive Bayes for Text Classification with Unbalanced Classes", *Proceedings of the 10th* European Conference on Principles and Practice of Knowledge *Discovery in Databases*, Springer, Berlin, Germany, 2006, pp. 503-510