Boosting bonsai trees for efficient features combination : application to speaker role identification

Antoine Laurent, Nathalie Camelin, Christian Raymond

Boosting bonsai trees for efficient features combination : application to speaker role identification

Antoine Laurent[1]*, Nathalie Camelin*[2]*, Christian Raymond*[3]

[1]LIMSI-CNRS, Orsay, France
[2]LIUM, Université du Maine, Le Mans, France
[3]IRISA-INSA, Rennes, France

laurent@limsi.fr, nathalie.camelin@lium.univ-lemans.fr
christian.raymond@irisa.fr

Abstract

In this article, we tackle the problem of speaker role detection from broadcast news shows. In the literature, many proposed solutions are based on the combination of various features coming from acoustic, lexical and semantic information with a machine learning algorithm. Many previous studies mention the use of boosting over decision stumps to combine efficiently these features. In this work, we propose a modification of this state-of-the-art machine learning algorithm changing the weak learner (decision stumps) by small decision trees, denoted bonsai trees. Experiments show that using bonsai trees as weak learners for the boosting algorithm largely improves both system error rate and learning time.

1. Introduction

In this article, we focus on speaker role detection in broadcast news shows. In the literature, the problem is seen as a multiclass classification problem where each speaker of a show has to be associated with a role label. In this way, some previous studies have tackled the problem using machine learning from mainly lexical features extracted from the transcription [1, 2, 3], from acoustic / prosodic features [4, 5], or from a combination of lexical and acoustic features [6]. Those studies have highlighted the efficiency of a boosting algorithm over decision stumps to combine the various features. We propose in this work a modification of this algorithm substituting the decision stumps by small decision trees, we call *bonsai*, in order to improve the combination of these various features. Thus, we propose here a speaker role identification system capitalizing on several features coming from acoustic, lexical or semantic level of description comparable to the system of [6] with the notable exception of our classification algorithm. We show on two speaker role detection databases that our algorithm combines much more efficiently features than the original algorithm while decreasing the training time of the classifier. Speaker role detection accuracy is improved up to 4% absolute while training time could be reduced by a factor of 4.

The paper is organized in the following way: Section 2 presents our speaker role recognition system, especially the set of features involved. Section 2.4 explains the modification of the boosting algorithm we propose and the benefits we expect. Section 3 presents the comparative experiments conducted on the EPAC and REPERE databases.

2. Speaker Role Identification System

The next three sub-sections describe the feature sets, and the last one presents the characteristics of the machine learning algorithm used.

2.1. Automatic spontaneous speech characterization

A method for automatically detecting spontaneous speech in audio documents was proposed by [7]. Acoustic (vowel duration, phonemic duration, pitch...) and linguistic (number of repetitions and number of proper names, syntactic pattern size...) features were extracted by an automatic transcription system to evaluate the level of spontaneity of each speech segment.

In [8], an analysis pointed out the relationship between speech type and speaker role. For instance, a presenter tends to prepare his speech, while a guest will often see his speech identified as *highly spontaneous*. The authors directly applied a spontaneous speech detection system to the particular task of role detection and reached encouraging results with an overall classification precision of 74.4%. We reimplemented this system as the baseline for the role detection task. The specific features used in this baseline system, except the automatic transcription, are called "**SPONTA**" in the rest of this article.

2.2. Social Network Analysis (SNA)

Social Network Analysis (SNA) aims to determine each speaker position in the dialogue. The idea supported by SNA is that a specific role, an *actor*, interacts with other *actors* during *events*. These interactions may help to identify the role assigned to each speaker involved in the network. Some SNA features have been investigated in [9, 10]. The goal of this method is to determine the *centrality* of each speaker [11] compared to other speakers in a show. A speaker i is considered as interacting with a speaker j if j occurs just after i in the transcript. Drawing inspiration from [11], we propose to compute centrality according to the following equation:

$$C_i = \frac{\sum_{j=1}^{nb} \chi D_{i,j}}{\sum_{j=1}^{nb} D_{i,j}} \quad (1)$$

With $\chi = 1$ if $D_{i,j} = 1$, and $\chi = 0$ otherwise, C_i is the centrality of i, nb the number of speakers and $D_{i,j}$ the distance between speakers i and j. This distance is expressed as the number of oriented links to pass through in order to reach node j from node i. In the example of a social network in figure 1, $D_{1,2} = 1$ and $D_{1,3} = 2$. Some nodes are not connected to any

others: this is the case for the node corresponding to speaker 5 in the example. In this situation, the distance between this node and any other node is set to $nb + 1$.

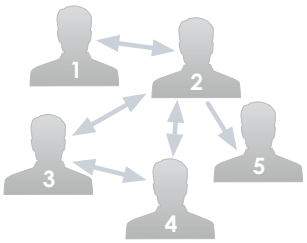


Figure 1: Example of a social network

In addition, we compute the *speaker coverage* corresponding to the elapsed time between the first and last interventions of a speaker in a show. It is normalized by the duration of the show.

In the remainder, "**SNA**" denotes the *centrality* and *speaker coverage* features.

2.3. *Ngram* patterns

One of the most obvious pieces of information to use to detect the speaker role is the transcription itself. The usual way to extract features from the transcription is to use word *Ngram* patterns. In the experiments section, the set of features extracted from the transcription is denoted as "**TEXT**".

2.4. Boosting of bonzai trees

The main idea of boosting is to combine many simple and moderately inaccurate classifiers into a single, highly accurate one. The base classifiers are trained sequentially. At each iteration of the boosting algorithm, a base classifier is trained with the data, the boosting algorithm gives more weight to samples that have been misclassified by the previous learner, forcing the next learner to focus on them. At the end the final classifier is built by a linear combination of all weak learners.

Many previous studies on speaker role identification [9, 1, 2, 7, 10] showed the efficiency of the AdaBoost.MH [12] algorithm over decision stumps (*ie* decision trees with 2 leaves) to combine the various features involved in classification based systems.

Although this algorithm has exhibited very good results, the linear combination of decision stumps may have difficulties to capture structures in the training data while boosting full decision trees may be less-efficient due to both the tree instability and data overfitting [13].

We choose to experiment the boosting of very small decision trees, named *bonsai trees*, constrained by their depth. We assess that the maximum depth of the bonsai remains very small to avoid the drawbacks of full decision trees. But we expect that a slightly more complex classifier than a decision stump will be able to capture more robust structures. Figure 2 illustrates the difference between a decision stump and a depth 2 bonsai tree.

Boosting bonsai trees in our situation, where we may want to examine millions of textual features, may benefit both the efficiency and the training time of the algorithm.

While boosting algorithm is iterative, the decision tree induction can be easily sped up with parallel processing. We expect that the use of bonsai trees as weak learners will require less iterations from the boosting algorithm than using decision stumps to reach equivalent performance. Moreover, since each bonsai tree could be induced using several parallel processes, their inductions may be not too much longer than a decision stump.

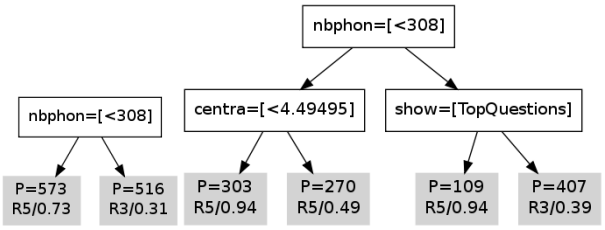
We expect 3 improvements from this modification:

1. *a performance gain* : a bonsai tree is a more complex classifier than a decision stump, thus it should be able to capture more complex structures in the data;
2. *a training time gain* : for the same reason, our boosting algorithm should produce a classifier that performs the same as the original algorithm with much fewer iterations;
3. *a fewer number of features used* : a bonsai should be able to build implicitly *Ngrams* from bag of words (*uni*-grams). If we are right, we will not have to generate explicitly *Ngrams* and evaluate them any more. The consequence would be a drastic reduction of features to be evaluated by the algorithm and therefore again we should observe a training time gain.

In the original algorithm, decision stumps are induced according to the pseudo-loss criterion proposed in [12]. We grow our bonsai by applying the same procedure recursively in each leaf of the tree. This recursion is stopped according to 2 criteria:

1. tree depth: the tree is a bonsai because it is constrained by his depth;
2. pseudo-loss gain: if a node subdivision attempt does not reduce the pseudo-loss, this node become a leaf.

Our implementation is available online [14].



(a) decision stump (b) bonsai tree width depth 2

Figure 2: Examples of bonsai trees learned on the EPAC corpus. Each node represents the test selected during the bonsai induction. Each leaf indicates the population P and the majority label along with his frequency.

3. Experiments

3.1. Experimental protocol

Our role detection system is evaluated on two corpora : EPAC and REPERE, briefly described in the next sub-sections. In order to evaluate our performance, we compare with two previously published systems : the first described in [8] for the EPAC corpus and the second described in [15] for the REPERE corpus.

All results are presented in terms of precision, recall and Role Error Rate (RER). The RER corresponds to the ratio between the number of misclassified examples and the total number of examples to be classified. Evaluations are applied on the

whole corpus through a 20 cross-fold evaluation for EPAC and a 200 cross-fold evaluation for REPERE.

In all the experiments, *1g* or *2g* refers to the size of the *Ngrams* extracted (note that *2g* includes *1g*) while the d{1..4} refers to the depth of the bonsai used as weak learner in the boosting algorithm. The term ALL denotes that all features are used (*ie.* TEXT, SPONTA and SNA), otherwise the type of features used is explicitly written.

3.2. Brodcast news from EPAC Corpus

3.2.1. Corpus

The EPAC corpus was built during the EPAC project funded by the French Research Agency (ANR) from 2007 to 2010 [16]. It is composed of 100 hours of conversational speech manually transcribed. Broadcast news from 3 French radio stations (France Info, France Culture and RFI) were recorded between 2003 and 2004 . In addition to orthographic transcription, a lot of metadata have been manually annotated, including speaker role, function and job (when available). Each speaker has a main role among the following seven:

- *Auditor*: the radio listener who calls during a program;
- *Columnist*: the broadcaster or writer who reports and analyzes events in the news during a broadcast show;
- *Correspondent*: the journalist who reports on particular subjects from a remote location;
- *Guest*: person who comes to talk about his latest news;
- *Interviewee*: person who answers questions raised by the presenter;
- *Anchor*: person who has multiple functions within a show—can host a talk, may take calls from listeners, or has the responsibility to give news, or weather information. . . ;
- *Voice-over*: pre-recorded voice commonly used to introduce programs.

The orthographic transcription that we used for the training and the evaluation comes from an automatic speech recognition system [17] which yields a WER of 17.3% on the EPAC test corpus.

3.2.2. Results

We compare our system to a baseline system trained with the same configuration as used in [8] in which we include the SNA feature as they do in [6]. This system corresponds in our notation to *ALL-2g-d1*.
Experimental results point out interesting conclusions.

Firstly, Table 1 compares results from our proposed system, tuned with a depth of 3, and the baseline system, which is depth 1, trained with different feature combinations. It shows how each feature type (TEXT,SNA,SPONTA) contributes to the performance of the final system. We can observe that our approach outperforms significantly the baseline for all configurations. The results show that a decision stump is too weak to capture some relevant information from data. Figure 3 tends to confirm: indeed, while learning curves for boosting over depth 2/3/4 bonsai converge around the same plateau, boosting over decision stumps gets stuck.

Secondly, Table 3 shows that a depth 2 tree as base learner for boosting gives notably better results than a decision stump.

	d3	*d1*
TEXT -2g	68.9	64.5
+ SPONTA	75.8	71.6
+ SNA	**77.7**	**73.9**
+ SNA	75.0	70.9
SPONTA	67.9	65.2
+ SNA	74.1	70.9
SNA	60.0	54.6

Table 1: RER of the systems trained with various feature combinations and various depth : d3 corresponds to our system trained with a bonsai of depth 3 and d1 corresponds to the baseline system trained with classical decision stumps.

Increasing depth to 3 outperforms again, while depth 4 yields similar performance. Thus these results show that using trees only slightly deeper than the usual decision stumps brings notable improvements to system performance.

Ngram/tree depth	**d1**	**d2**	**d3**	**d4**
1g	72.9	76.2	77.7	78.1
2g	73.9	77.8	77.7	78.4

Table 3: RER of the system trained with every features on EPAC.

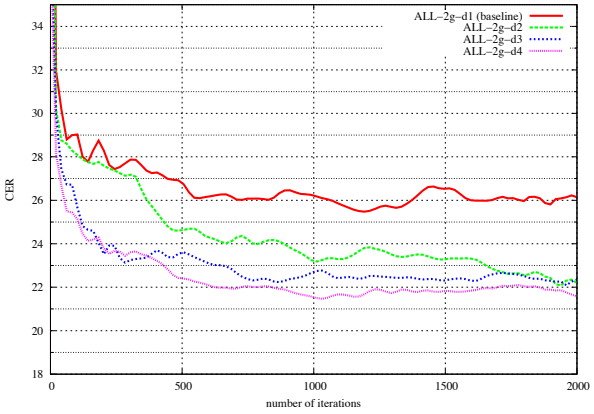


Figure 3: Results in term of 100-RER for learning curves of various depths trees wrt the number of boosting iterations.

Thirdly, Table 2 details the results for all roles obtained with our depth 3 bonsai system using *1g* and *2g* as text features. We compare theses configuration to the system proposed in [6] that actually correspond to the notation *ALL-2g-d1* in the table. We can observe that a consistent improvment is observed through all different roles.

Moreover, these results show that using a bonsai on *1g* gets better performance than decision stumps on *2g*. Using bonsai on *2g* does not bring improvement over bonsai on *1g*. This is an interesting result, since it tends to show that it is not necessary to generate explicitly *2g* or more when using bonsai trees: *Ngrams* are captured by the tree.

		ALL-2g-d1		**ALL-1g-d3**		**ALL-2g-d3**	
Role	Nb.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Auditor	262	**88.7**	92.7	86.3	93.1	87.7	**95.0**
Columnist	173	60.2	59.5	**69.9**	**63.0**	**69.9**	**63.0**
Correspondent	133	48.9	51.1	**61.3**	54.9	61.2	**59.4**
Interviewee	215	59.5	60.9	**65.3**	**67.4**	63.4	63.7
Guest	265	76.0	72.8	**76.6**	**79.2**	76.3	77.7
Anchor	231	92.5	90.5	**94.3**	**93.5**	**94.3**	**93.5**
VoiceOver	14	**88.9**	**57.1**	66.7	**57.1**	**88.9**	**57.1**
All roles	1293	73.86		**77.73**		77.65	

Table 2: Comparative results, in precision and recall, of our depth 3 bonsai system with [6], denoted *ALL-2g-d1* for various roles of EPAC. The first column indicates the number of each role and the last line shows the RER for the three systems.

System		R1		R2		R3		R4		R5	
	RER	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
(Bigot et al.)	73.1	60	18.6	52.4	23.4	62.8	47	69.9	58	75.8	95.9
(Bigot et al.) +show	86.9	82.7	71.6	85	72.3	85.6	76.2	74.3	75	89.7	95.1
ALL-1g-d3	88.6	86.1	83.0	78.6	70.2	88.5	85.9	86.7	78.0	90.5	95
ALL-1g-d3+show	91.0	88.0	83	89.7	74.5	90.7	91.8	92.9	83.5	91.8	96.3

Table 4: Comparative results of our best system, that uses a 3 depth bonsai, with the reference [15] system on the REPERE corpus. The term *show* indicates the show-dependent classifier.

3.3. Brodcast news from REPERE Corpus

3.3.1. Corpus

The REPERE corpus was built during the DEFI-REPERE french evaluation campaign from 2010 to 2014 [18]. Data from various TV shows were recorded from French TV channels : BFM and LCP. Twenty four hours have been recorded : about 14 hours of broadcast news, about 6.5 hours of debates and 3.5 hours of parliamentary sessions of the French National Assembly.

The corpus has been manually segmented according to the speaker and manually enriched with annotations relative to five speaker roles:

- R1 : the anchor;
- R2 : journalists, equivalent to the EPAC columnist;
- R3 : reporters, equivalent to the EPAC correspondent and voice-over roles;
- R4 : guests;
- R5 contains all remaining speakers with another role, like EPAC interviewee or auditors, or politicians ...

The [17] system adapted for REPERE yields a WER of 15.18% on the REPERE test corpus (phase 1).

3.3.2. Results

Similarly to the experiments done on the EPAC corpus we propose in Table 4 different results obtained for different depth of bonsai. We compare our results with the SVM based approach published in [15], denoted Bigot. They propose 2 different systems based on low level features, one using the show information (denoted Bigot+show) and the other not. It should be noted that the set of features used in their system and ours is not the same. Boosting system over decision stumps exhibits better performance than the SVM approach proposed in [15].

We point out that once again boosting bonsai outperforms significantly boosting decision stumps with observed RER of 85.1%, 88.1% and 88.6% for respectively bonsai depth 1, 2 and 3. Furthermore, as previously shown on the EPAC corpus, a consistant improvment is observed through all the different roles. We can notice that the *ALL-1g-d3* system that does not take into account the show information already outperforms the (Bigot et al.)+show. Taking into account the name of the show as a new feature of our system (*ALL-1g-d3+show*) allows an extra-benefit of 2.4%.

Finally, to be fair, we also evaluate our system without using the transcription as in [15]. The system *SPONTA+SNA-d3* obtained a RER of 90.2%, which is 7 points better than the reference system.

4. Conclusion

We have investigated and presented in this paper a modification of the boosting algorithm widely used in the literature for speaker role identification. We propose to replace decision stumps used as weak learner for the boosting algorithm by small decision trees, we called bonsais. A speaker role identification system based on this algorithm is evaluated and compared to state-of-the-art systems published in [6] and [15] on two different corpora.

Experiment results tend to confirm our 3 expectations about the use of bonsai as weak learner for the boosting algorithm: it allows significant improvement in terms of accuracy while bonsai needs only to be very small to get these improvements. It permits a strong reduction of the number of iterations of the boosting algorithm. It is also able to produce similar results using *1g* features and *2g* features, because it is able to implicitly build them: this result is quite interesting because it seems that we don't need to generate explicitly complex *Ngram* features, saving a lot of training time. Actually, taking in consideration the previous facts, we computed that building a boosting based system with depth 3 bonsais is approximately 4 times faster than building a boosting based system on decision stumps.

5. References

[1] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *AAAI*, 2000, pp. 679–684.

[2] G. Damnati and D. Charlet, "Robust speaker turn role labeling of tv broadcast news shows," in *ICASSP*, Prague, République Tchêque, 2011.

[3] Y. Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Human Language Technology Conference of the NAACL*, New York, USA, 2006, pp. 81–84.

[4] H. Salamin, S. Favre, and A. Vinciarelli, "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," in *IEEE Transactions on Multimedia*, vol. 11, 2009, pp. 1373–1380.

[5] B. Bigot, I. Ferrané, J. Pinquier, and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in *Searching Spontaneous Conversational Speech*, Firenze, Italie, 2010, pp. 5–10.

[6] R. Dufour, A. Laurent, and Y. Estève, "Combinaison d'approches pour la reconnaissance du rôle des locuteurs," in *JEP*, Grenoble, France, 2012.

[7] R. Dufour, Y. Estève, P. Deléglise, and F. Béchet, "Local and global models for spontaneous speech segment detection and characterization," in *ASRU*, Merano, Italie, 2009.

[8] R. Dufour, Y. Estève, and P. Deléglise, "Investigation of spontaneous speech characterization applied to speaker role recognition," in *Interspeech*, Florence, Italie, 2011.

[9] P. N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli, "Role recognition for meeting participants: an approach based on lexical information and social network analysis," in *ACM Multimedia Conference (MM'08)*, Vancouver, Canada, 2008, pp. 693–696.

[10] W. Wang, S. Yaman, K. Precoda, and C. Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *ICASSP*, 2011, pp. 5556–5559.

[11] A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Transaction on Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.

[12] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000, http://www.cs.princeton.edu/~schapire/boostexter.html. [Online]. Available: http://www.cs.princeton.edu/~schapire/boostexter.html

[13] J. R. Quinlan, "Bagging, boosting, and C4.5," in *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI 1996)*, vol. 1. AAAI Press, 1996, pp. 725–730. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.2457&rep=rep1&type=pdf

[14] C. Raymond, "bonzaiboost," http://bonzaiboost.gforge.inria.fr/, 2010. [Online]. Available: http://bonzaiboost.gforge.inria.fr/

[15] B. Bigot, C. Fredouille, and D. Charlet, "Speaker role recognition on tv broadcast documents," in *SLAM@INTERSPEECH*, 2013, pp. 66–71.

[16] Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news," in *LREC*, Valletta, Malte, 2010, pp. 1686–1689.

[17] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?" in *Interspeech*, Brighton, Angleterre, 2009, pp. 2123–2126.

[18] J. Kahn, O. Galibert, M. Carr, A. Giraudel, P. Joly, and L. Quintard, "The repere challenge: Finding people in a multimodal context," in *Odyssey 2012 - The Speaker and Language Recognition Workshop*, 2012.