Accounting for Prosodic Information to Improve ASR-Based Topic Tracking for TV Broadcast News

Camille Guinaudeau, Julia Hirschberg

Accounting for prosodic information to improve ASR-based topic tracking for TV Broadcast News

Camille Guinaudeau[1], *Julia Hirschberg*[2]

INRIA Rennes Bretagne Altantique, France[1]
Department of Computer Science, Columbia University, New York, NY, USA[2]
camille.guinaudeau@inria.fr, julia@cs.columbia.edu

Abstract

The increasing quantity of video material available on line requires improved methods to help users navigate such data, among which are topic tracking techniques. The goal of this paper is to show that prosodic information can improve an ASR-based topic tracking system for French TV Broadcast News. To this end, two kinds of prosodic information — extracted with and without a learning phase — are integrated in the system. This integration shows significant improvements in the F1-measure, by 13 and 8 points for the two techniques compared with the baseline system.

Index Terms: Topic Tracking, Prosody, TV Broadcast News, *tf-idf* criterion

1. Introduction

The increasing number of video feeds available on line requires improved methods to help users navigate within TV streams. Navigation, in this case, may mean direct access to a theme of interest, following the evolution of a particular story or seeing how a specific topic is treated by different channels. One of the main tasks necessary to support these tasks is topic tracking, to create semantic links between topic segments or stories extracted from the stream.

Topic tracking has been widely studied in recent years through the Topic Detection and Tracking (TDT) challenge which aims to develop automatic techniques for finding topically related material in streams of data (e.g., newswire and broadcast news) [1]. Most systems rely on the computation of term overlap between different segments; the more terms the segments have in common, the more likely those two segments have the same topic. As with other text-based analysis tasks, either vector space approaches [2] or statistical language models [3] have been used.

However, it is important to take into account that the original data, i.e. TV programs, is not ONLY textual. For example, [4] uses video cues — low-level (e.g., motion) and high-level (e.g., face, anchor-person) concepts — in addition to speech transcripts to detect redundancy and novelty in news stories. However, important information is also carried by the way speech is produced in the program. The fact that parts of an utterance are **accented**, or made intonationally prominent, in the signal may indicate that the information carried by this segment is new information [5] or that the speaker is emphasizing this information. Indeed, in [6], the author concludes that there exist a direct relationship between acoustic stress and information content in English and [7] shows that the use of acoustic stress information offer some slight improvement in a Spoken Document Retrieval system develop for Chinese. The goal of this paper is to demonstrate that such prosodic information can also be useful for French and that it can improve the performance of a topic tracking system based on a vector space approach. We examine how such information can be automatically extracted from the source using two different methods, with and without a learning phase, and integrated in the baseline topic tracking system.

This paper is organized as follows. First, an overview of the topic tracking system is presented. In Section 3, the computation of the vectors used for topic characterization is described. In our experiments, this computation is done using either the information retrieval criterion *tf-idf* or acoustic information. Methods used for the extraction of prosodic information are also presented in this part. Finally, in Section 4, results obtained on a corpus of over 5 hours of French TV Broadcast News are discussed.

2. Topic tracking system

The goal of the topic-tracking system we describe is to provide a navigation system for large amounts of online TV data. The approach underlying this system is to link together parts of TV programs that deal with similar topics so that users can follow the evolution of a particular story or can directly access to segments dealing with topic of interest.

Topic segments are first identified from each broadcast manually. They do not correspond exactly to reporter stories, as they contain the introduction and eventually the conclusion by the anchor speaker. The segments' extraction was done manually here, rather than automatically in order to eliminate the effects of segmentation errors on the topic tracking system's performance.

The system next computes term overlap between the automatic transcription of the speech of different segments using a vector-space approach. Each segment is described by a weighted word vector where a weight represents the extent to which the word is characteristic of the segment topic. A cosine distance is then computed between weighted vectors to evaluate their semantic proximity. A link between two segments is hypothesized where the cosine value dips below a global threshold β.

As the whole system relies on vector comparisons, the quality of these comparisons is critical. The computation of vectors used for segment characterization is described in next section.

3. Vector computation

Word vector computation consists in associating scores with each word of the segment such that, the higher the score, the more characteristic the word is of the segment.

As a baseline, vector computation is done using a standard *tf-idf* approach, widely used in the Information Retrieval field and described in Section 3.1. Prosodic information can also be used as a weighting scheme to account for speaker intention or the information status of the words uttered in the segment; this weight is presented in Section 3.2. Finally, in Section 3.3, we discuss the combination of the two weighting schemes.

3.1. Tf*idf weighting

The standard *tf-idf* approach is performed using the keyword extraction tool developed by Lecorvé [8]. Given a segment s to be characterized, this tool associates a *tf-idf* score to each word of the segment, excluding stop-words such as prepositions, articles, and so on. These scores are computed based on two values: the frequency $tf(w, s)$ of the word w in the segment s, and the inverse document frequency $idf(w, \mathcal{C})$, which compares the number of documents in a reference corpus $\mathcal{C}$ to the number of documents in $\mathcal{C}$ containing w. We used 800,000 articles from the French newspaper *Le Monde*, 1987–2003, as our reference corpus[1]. All the textual resources, i.e., the transcripts and the corpus $\mathcal{C}$, are lemmatized and lemmas[2] are considered instead of words when computing *tf-idf* scores.

3.2. Weighting using prosodic information

Prosody, or intonation, is an important component of spoken communication. It may reflect various features of the utterance: the emotional state of a speaker; whether an utterance is a statement, a question, or a command; whether the speaker is emphasizing or contrasting or focusing a particular item.

We therefore examined how prosodic information could be useful in term weighting. We hypothesized that words more important to a topic might be produced with greater emphasis, in an expanded pitch range, or with greater intensity.

To explore this possibility, we first assigned an acoustic score to each lemma occurence, as described in Sections 3.2.1 and 3.2.2, where the higher the score, the more prominent the lemma. Since each lemma may have multiple occurences in a segment, we defined two strategies to deal with vocabulary repetitions : *-M-* in which the maximum acoustic value of all occurrences of the lemma in the segment is used as the acoustic score, and *-A-*, in which the average between prosodic values over all occurences of the lemma in a segment is used as the lemma's acoustic score.

The computation of acoustic scores for lemmas is done using one of two techniques: the Acoustic Information Extraction (AIE) method, described in Section 3.2.1, and the Prosodic Event Detection (PED) technique, presented in 3.2.2. We examine the usefulness of each technique to our overall goal.

3.2.1. Acoustic Information Extraction (AIE) method

In the AIE technique, two features are extracted from each non-stop word in a segment, energy (RMS) and fundamental frequency (f0). Energy represents a physical correlate of perceived loudness and f0 a physical correlate of perceived pitch. Increases in each have been shown to correlate with a word's information status. For each recording in our corpus, RMS and f0 values are extracted for each 0.01 second window of the sound file using Praat [9]. Values are then z-score normalized by speaker using the speaker diarization information generated by the ASR system. Since automatic diarization is imperfect, normalization is approximate.

Automatic transcripts are aligned with the source recordings using the temporal word boundaries provided by the ASR engine. As a lemma is usually associated with multiple values, four different strategies can be used to compute the lemma score, based on the alignment: MAX, where the maximum value is kept, AVE in which the average of the values is computed, MIN (the minimum value is kept) and SD where the standard deviation of all the acoustic values is calculated.

Each lemma u in the transcript, with the exception of stopwords, is thus associated with two scores $i(u) \in [0, 1]$ and $p(u) \in [0, 1]$ that denote energy and f0 for that lemma. A third value $b(u)$ combining both scores is also calculated by multiplying the two together.

3.2.2. Prosodic Event Detection (PED) technique

The second technique for acoustic information extraction was developed by Andrew Rosenberg [10]. Given a (real or hypothesized) word segmentation, Rosenberg's AuToBI system detects the location and type of prosodic events in a spoken utterance from features such as energy and f0 and associates these to words (*pitch accents*) and boundary sites (*intermediate phrases, intonational phrases*) in the utterance prosodic phrases.[3] AuToBI includes classifiers trained on Standard American English but can also be trained on new data that has been labeled with prosodic events and word boundaries. For our experiment, it was trained on the C-PROM Corpus, which is a manually annotated and word-aligned corpus developed for the study of syllabic prominences in French [12]. We associated each word in the corpus with a value $d(u) \in [0, 1]$, standing for the probability for a word to be prominent in the utterance.

Table 1: Vector characterization using *tf-idf* , prosodic information and both information types

TFIDF	AIE	TFIDF+AIE	PED	TFIDF+PED
degree 1	**thing 0.99**	degree 0.93	**thing 0.41**	temperature 0.54
temperature 0.99	imbalance 0.90	temperature 0.89	temperature 0.32	degree 0.54
climatic 0.81	degree 0.87	climatic 0.80	increase 0.32	climatic 0.44
ocean 0.49	increase 0.85	ocean 0.65	ocean 0.32	ocean 0.37
planet 0.39	ocean 0.82	imbalance 0.62	degree 0.31	**thing 0.30**
imbalance 0.34	temperature 0.80	planet 0.52	climatic 0.26	increase 0.26
increase 0.16	climatic 0.79	**thing 0.55**	planet 0.19	planet 0.25
thing 0.10	planet 0.65	increase 0.50	imbalance 0.16	imbalance 0.22

3.3. Combination of *tf-idf* and prosodic information

Scores associated with each lemma in the automatic transcription of the speech contained in TV segments can also be computed by combining the *tf-idf* criterion with the acoustic information. In this case, each lemma is weighted by a new value calculated from the *tf-idf* and the acoustic score:

$$S(u) = \frac{\theta_{ir}\ tf\text{-}idf\ (u) + \theta_{ac}\ a(u)}{\theta_{ir} + \theta_{ac}}\ , \qquad (1)$$

with $a(u)$ the acoustic score obtained from the AIE and PED methods described in Section 3.2. The two factors θ_{ir} and θ_{ac} can be used to give more or less weight to the different sources of information. Table 1 provides examples of word vectors computed using *tf-idf* , the AIE method, and the PED methods alone and in combination for a segment dealing with global warming.

In the AIE column, it appears that the word *increase*, underlined, is prominent as it is associated with a high acoustic value – here, intensity. This word, which is characteristic of the topic of the segment, was not associated with a large *tf-idf* score, as its document frequency in the corpus is probably quite high.

[1] This corpus was previously used to estimate the language model probabilities and the vocabulary of the ASR system that generates automatic transcription of our data. Consequently, the vocabulary that appears in the transcripts is included in C and, as the two corpora are journalistic, the two collections are not so different.

[2] A lemma is an arbitrary canonical form grouping all inflections of a word in a grammatical category, e.g., the infinitive form for verbs, the masculine singular form for adjectives, etc.

[3] The AuToBI system is based on the ToBI standard for Standard American English ([11]). More details can be found in ([10]) and at (http://eniac.cs.qc.cuny.edu/andrew/autobi/index.html).

In this case, combinining *tf-idf* score and acoustic information (TFIDF+AIE) allows this word to be weighted more highly in the vector, reflecting the speaker's intention to put forward a term that is not distinctive from a *tf-idf* point of view. However, acoustic prominence is not always useful in re-weighting lemmas; for example, the word *thing*, in bold, is highlighted by the speaker but cannot be considered as a representative word of global warming.

Table 1 also illustrates the information provided by the PED technique, alone or associated with *tf-idf* score. Some similarity can be found between information provided by this technique and the AIE technique, since both measure acoustic prominence in different ways. Just as for the AIE technique, the PED score gives the word *increase* a high acoustic prominence score; and, similarly, the word *thing* also scores high by the PED technique. Not every word that scores high on one measure scores high on the other. For example, the word *imbalance*, which is quite characteristic of the segment topic, receives a high score by the AIE method but has a low score according to the PED technique. Therefore, its influence in the vector is lowered when the PED score is combine with the *tf-idf* score. It is not clear why a word can be associated with a high score by one method and a low one by the other but we believe that the difference is related to the learning phase of the PED technique.

4. Experiments

We evaluated the use of prosodic information from both our AIE and PED approaches to improve our ASR-based topic tracking system using 177 topic segments extracted from 8 TV news programs ($\approx$ 1/2 hour each) broadcasted in February and March 2007 on the French television channel France 2.

The ASR system used to transcribe the corpus is a radio broadcast news transcription system, exhibiting error rates ranging from 20% on broadcast news up to 70% on movies, talk shows and debates [13]. This system implements a multiple pass strategy, progressively narrowing the search space in order to use more complex language and acoustic models. In the final steps, a 4-gram LM over a vocabulary of 65,000 words is used with context-dependent phone models to generate a list of 1,000 transcription hypotheses. The language model probabilities were estimated on 500 million words of text from French newspapers and interpolated with LM probabilities estimated over 2 million words corresponding to reference transcripts of radio broadcast news shows.

A topic tracking reference was created manually by considering a semantic link between two segments if they dealt with similar or closely related stories (205 links). For example, segments that discuss the French presidential electoral campaign of two different candidates are considered as related while segments dealing with politics in a more general aspect are not. Recall and precision metrics are employed to evaluate the number and proportion of relevant links returned by our topic tracking algorithm. The F1-measure values used for comparison are obtained for an optimal value of β, the one leading to the number of links closest to that of the topic-tracking reference. In this section, we report results of using *tf-idf* and acoustic information alone and then in combination.

Word vectors used to characterize topically coherent segments can contain a varying number of words. Rather than selecting a threshold in terms of word score values we select a fixed number of the highest scoring words for each vector. For word vectors weighted with the *tf-idf* criterion, we found that, the greater the number of words used in the vector, the better the topic tracking results. When 100 words are used in each vector, the F1-measure value is 35.24, even though the difference does not represent a statistically significant improvement over a strategy using 60 words. As this difference is more important for acoustic information, as we will see later in this section, in the rest of this article the number of words used in vectors is 100 if no additionnal information is given.

Table 2: Topic tracking results with segment characterization using AIE information only (F1-measure)

		Intensity	Pitch	Both
MAX	*-M-*	37.03	**38.80**	**38.43**
	-A-	**37.09**	37.36	38.14
AVE	*-M-*	30.86	28.08	31.13
	-A-	30.99	23.05	28.33
MIN	*-M-*	30.54	*7.5*	16.62
	-A-	24.81	*5.79*	13.71
SD	*-M-*	*16.81*	37.40	34.97
	-A-	*16.71*	35.09	34.97

Results for topic tracking using word vectors weighted with acoustic scores are presented in Table 2. This table shows that the strategy used for the alignment between acoustic values and the automatic transcription has a real impact on the topic tracking system performance. The MAX strategy, where the maximum value is kept when multiple values are associated with a lemma, gives results that are significantly better (by a t-test) than the three other strategies, with an F1-measure of ≈ 37. The AVE strategy, in which the average of the values is computed; MIN technique, where the minimum value is used; and SD method, using a computation of the standard deviation of all the acoustic values, all exhibit lower performance, with the worst result given by the MIN strategy (an F1 of 5.79). We also examine the effect of our two vocabulary repetition strategies, *-M-*, in which the maximum acoustic value of all occurrences of the lemma in the segment is used as the acoustic score, and *-A-*, in which the average between prosodic values over all occurences of the lemma in a segment is used as the lemma's acoustic score. Table 2 shows that the two do not give significantly different results, although the general trend is that *-M-* is better than *-A-*. As the MAX/*-M-* strategy gives, even if not the best, at least good results for the three kind of acoustic information introduced, it is the one we use to combine with the *tf-idf* score. From the table, it can also be seen that the AIE score yields substantially similar results to the *tf-idf* score: a t-test shows no statistically significant difference between results from *tf-idf* criterion alone (F1-measure equals to 35.24) and acoustic information alone (F1=38.80). If, this is true when 100 words are used in each vector, from 70 words, the use of the *tf-idf* criterion alone give significantly better results than the exploitation of acoustic information. That can be explained by the fact that the proportion of non pertinent words is more important in the first, and best, words in acoustic vectors than in *tf-idf* ones (cf. Table 1), which is smoothed when more words are used in the vector.

Scores for acoustic information derived from the PED technique are lower than scores using the AIE technique, with F1-measure values of 28.34 and 23.71 for the *-M-* and *-A-* strategies, respectively. This indicates that the use of PED information is less useful than AIE information for segment topic characterization from prosodic information alone. We hypothesize that this lower performance may be due to the difference in training corpus for the PED approach (the C-PROM Corpus) vs. our news corpus. The AIE values were obtained directly from our corpus, as no learning phase is needed for this extraction method.

The *tf-idf* criterion and acoustic information can also be used together to improve topic-tracking over each information

Table 3: Topic tracking results with segment characterization using *tf-idf* criterion and prosodic scores (F1-measure)

$\theta_{ir} - \theta_{ac}$	AIE			PED
	Intensity	Pitch	Both	
1-5	37.34	39.79	38.56	40.98
1-4	38.12	41.16	39.06	41.16
1-3	39.37	41.99	39.46	42.10
1-2	41.05	44.56	42.10	42.39
1-1	46.14	**48.40**	**47.59**	**43.00**
2-1	**46.75**	44.55	45.84	41.47
3-1	44.01	42.39	41.60	37.59
4-1	42.20	40.52	40.54	36.85
5-1	39.24	39.38	39.73	35.92

source alone. In this case, the weight of *tf-idf* against acoustic information can be varied using the θ_{ir} and θ_{ac} factors (1). Table 3 presents F1-measure values obtained for topic tracking when word vectors are weighted using both cues in various ratios. From this table, it can be observed that, for the use of AIE information, the combination of both clues gives satistically better results than the one obtained when the two clues are used separately. As explained in Section 3.3 some prominent words which were not considered as important by the *tf-idf* criterion have their score boosted by acoustic information. When combined with the *tf-idf* criterion, both the AIE and the PED scores can improve the topic tracking quality, although the AIE approach provides the best combined result. This result is achieved when the $\theta_{ir} - \theta_{ac}$ weights are equal, as shown in Table 3.

Figure 1 compares precision and recall for both combined *tf-idf* and prosody methods with the simple *tf-idf* baseline.

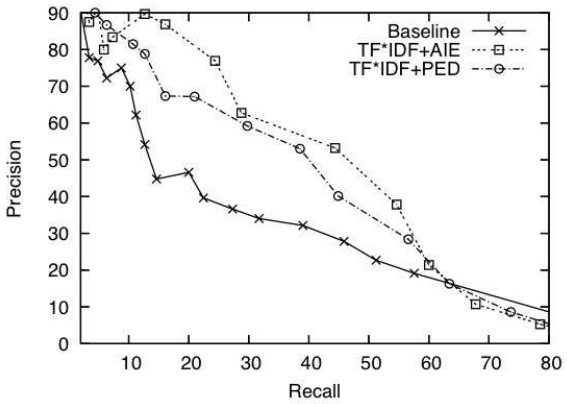

Figure 1: Recall/precision curves for topic tracking based on a combination of the *tf-idf* criterion and acoustic features.

5. Conclusion and future work

In this paper we have shown that prosodic information can improve the performance of an ASR-based topic tracking system in tracking related segments in a French newscast. This improvement is realized by an increase in F1 of +13 and +8 for two approaches to modeling prosodic information, one using a direct modeling technique (AIE) and the other using a prosodic event detection approach (PED). We have also demonstrated that the joint use of the two criteria, *tf-idf* and acoustic information, improves over either method used alone. Finally, we have found that the method used for acoustic information extraction has an impact on the quality of topic tracking, with direct modeling performing better than prosodic event detection. This difference is probably related to the fact that the event detector was trained on a different type of corpus than our test corpus.

In order to validate this hypothesis, it would be interesting to train the event detector on a news corpus and check whether the performance is higher with a better trained PED system. Moreover, as a learning phase does not seem to be necessary to achieve good performance with the AIE technique, accounting for prosodic information by using intensity or pitch values can easily be done on another kind of TV program. Therefore, we plan to work on a larger corpus composed of different kinds of TV programs. Preliminary experiments, run on a corpus composed of 16 hours of reports on current affairs, show improvements in the F1-measure by +5, which suggests that the improvement is not dependent on the type of the corpus studied. Finally, we also envisage a study of the integration of other cues to improve the topic tracking quality which is degraded by transcription errors and extensive use of synonyms. For example, confidence measures — provided by the automatic speech recognition system and corresponding to the probability of a word being correctly transcribed — can be used in order to counteract transcription errors and semantic relations can be integrated as additional information to take into account the semantic links that exist between words.

6. References

[1] C. L. Wayne, "Topic Detection and Tracking (TDT)," in *DARPA Broadcast News Workshop*, 1998.

[2] I. Ide, H. Mo, N. Katayama, and S. Satoh, "Topic Threading for Structuring a Large-Scale News Video Archive," in *3rd ACM Intl. Conf. on Image and Video Retrieval*, 2004.

[3] P. V. Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron, "Text Segmentation and Topic Tracking on Broadcast News via a Hidden Markov Model Approach," in *5th Intl. Conf. on Spoken Language Processing*, 1998.

[4] X. Wu, "Threading Stories and Generating Topic Structures in News Videos across Different Sources," in *13th ACM Intl. Conf. on Multimedia*, 2005.

[5] J. Hirschberg, "Communication and Prosody: Functional aspects of Prosody," *Speech Communication*, vol. 36, no. 1-2, pp. 31 – 43, 2002.

[6] F. Crestani, "Towards the use of prosodic information for spoken document retrieval," in *24th Intl. Conf. on Research and development in information retrieval, SIGIR '01*, 2001.

[7] B. Chen, H. min Wang, and L. shan Lee, "Improved spoken document retrieval by exploring extra acoustic and linguistic cues," in *7th European Conf. on Speech Communication Association, Eurospeech'01*, 2001.

[8] G. Lecorvé, G. Gravier, and P. Sébillot, "An unsupervied Web-based topic language model adaptation method," in *33rd Intl. Conf. on Acoustics, Speech and Signal Processing*, 2008.

[9] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9-10, pp. 341 – 345, 2002, http://www.fon.hum.uva.nl/praat/.

[10] A. Rosenberg, "AuToBI - A Tool for Automatic ToBI Annotation," in *11th Intl. Conf. on Speech Communication Association, Interspeech'10*, 2010.

[11] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "TOBI: A Standard for Labeling English Prosody," in *2nd Intl. Conf. on Spoken Language Processing*, 1992.

[12] M. Avanzi, A. Simon, J.-P. Goldman, and A. Auchlin, "C-PROM. an annotated corpus for French prominence studies," in *Prosodic Prominence: Perceptual and Automatic Identification, Proceedings of Speech Prosody 2010 Satellite Workshop*, 2010, http://sites.google.com/site/corpusprom/.

[13] S. Huet, G. Gravier, and P. Sébillot, "Morpho-syntactic post-processing of N-best lists for improved french automatic speech recognition," *Computer Speech and Language*, no. 24, pp. 663–684, 2010.