Can automatic speech transcripts be used for large scale TV stream description and structuring?

Camille Guinaudeau
IRISA/INRIA
Rennes, France
camille.guinaudeau@irisa.fr

Guillaume Gravier
IRISA/CNRS
Rennes, France
guillaume.gravier@irisa.fr

Pascale Sébillot
IRISA/INSA
Rennes, France
pascale.sebillot@irisa.fr

***Abstract*—The increasing quantity of TV material requires methods to help users navigate such data streams. Automatically associating a short textual description to each program in a stream, is a first stage to navigating or structuring tasks. Speech contained in TV broadcasts—accessible by means of automatic speech recognition systems in the absence of closed caption—is a highly valuable semantic clue that might be used to link existing textual description such as program guides, with video segments corresponding to program. However, high word error rates are to be expected on some programs, likely to jeopardize the usefulness of transcripts. The goal of this article is to determine to what extent automatic transcripts of TV streams, for various types of programs, can be used for structuring or navigating tasks. To this end, word-based and phonetic-based automatic association between video segments and program descriptions is used as a case study. We show that descriptions from a program guide can be associated with video segments with an accuracy of up to 65 % and provide a valuable description to validate existing program labels. Such associations constitute a first stage for structuring task as they enable video segment textual characterization.**

***Keywords*-TV stream structuring, automatic speech recognition, semantic content description**

I. Introduction

Video structuring based on image, audio and, for some specific programs such as news, on text is a widely studied domain. A first step in most structuring tasks consists in dividing the TV stream into coherent segments, *e.g.*, into topics [1], action categories in sport videos [2] or scenes [3]. Video segmentation is often based on visual features. For example, scene boundaries are often detected using a coherence measure based on color histogram. Specific visual features, such as line mark, motion or player's uniform color can be used for sport video structuring [2]. However, as video sequences are multimodal documents, some studies handle audio features [3] or associated textual information to perform video segmentation [1]. Threading can be considered as a second step in structuring tasks and consists in providing a concise and chronological view of the huge volume of information contained in a TV stream. Some studies handle visual descriptors in order to link two or more news stories. For example, [4] considers that two news stories are related if they contain a pair of similar shots (duplicate or near-duplicate keyframes). However, TV streams are also strongly qualified by the speech material contained in the soundtrack, accessible either by means of an automatic speech recognition (ASR) system or via closed-captions, the latter being absent from most (French) channels. Therefore, some works use textual clues, such as the location of words inside the text in [5] or lexical coherence in [1], to detect similar or new topics. Yet, most approaches, whether for segmentation or threading purposes, are limited to a particular type of broadcast (usually news) and, to the best of our knowledge, no studies have been performed on long TV stream, such as an entire day, containing various kinds of programs.

The goal of this article is to determine to which extent speech (transcripts) in TV streams can be used for structuring or navigation purposes, knowing that word error rates range from about 10 % on news shows to more than 60 %, for example on talk shows or series. To answer this original question, we investigate the task of associating transcripts of video segments, where segments in general correspond to programs, with textual descriptions either provided by an electronic program guide (EPG)—where descriptions vary from a precise summary of the program to the sole title—or extracted from the Internet. Indeed, associating TV segments with program descriptions can be seen as a characterization task. For example, such characterizations can be used to link segments dealing with the same topic. Alternately, they offer easily accessible user-oriented descriptions.

The association process described in this paper is mostly inspired from word-based textual information retrieval techniques and relies on the generation of shortlists of candidate segments for each description. However, particular difficulties arise from the use of automatic transcripts. First of all, transcription errors can be quite numerous, specifically on long TV streams where all sorts of programs are encountered. Moreover, a consequent number of out-of-vocabulary (OOV) words, *i.e.*, not in the vocabulary of the ASR system and therefore absent from transcripts, occurs over several days of TV, due to the high rate of proper nouns and to

Work partially funded by the Quaero project.

the diversity of topics covered. This fact is problematic for a word-level (transcript-based) association, especially since EPG/Internet descriptions often contain proper nouns likely to be OOV words. As dynamic expansion of the vocabulary of the ASR system is a delicate issue, phonetic-based spoken document retrieval techniques are also considered. The idea is to rely on comparison of phonemes[1] rather than words. Finally, EPG or Internet descriptions greatly vary in their content, some describing the entire program while others simply report on the topic (or topics) covered.

The paper is organized as follows. In Section II, the method for word-based generation of segment shortlists for EPG/Internet descriptions is presented. Section III is dedicated to the phonetic-based shortlist generation. Experimental results on a 10 day long TV stream are presented for two tasks, segment characterization and label validation, in Section IV. A summary of the results and a presentation of future research directions conclude the paper.

II. Transcript-based shortlist generation

Automatic speech recognition systems are able to generate a textual transcription of the speech material contained in the video, where the output is a set of time-marked words with associated confidence measures which reflects for each recognized word the confidence of the ASR system in its decision. Therefore, a straightforward approach to relate segment transcripts with EPG or Internet descriptions is to rely on word-based textual information retrieval (IR) techniques to measure the similarity between a transcript and a description. Pairwise similarities are then used to determine a shortlist of relevant segment candidates for each description. We first define the pairwise distance before presenting the shortlist generation process.

A. Pairwise distance definition

The vector space model commonly used in textual IR methods consists in representing each document by a vector gathering the relevance (weight) of each index word with respect to the document. The most popular weight is the tf*idf score which takes into account the frequency of a word in a document (tf) normalized by the inverse of its frequency of occurrence over a collection of documents (idf). Document ranking according to a query—in our case, the query is either a transcript or a description—is derived from the distances between the query vector and the vectors of each document, where the distance used is usually a cosine measure [6].

In this work, the vector space model with a cosine measure is used to compute the pairwise distance between the transcript of a video segment and a description. However, to account for transcription errors, the computation of the tf*idf weights is modified for segment transcripts in order to take into account the confidence measure provided by the ASR system for each hypothesized word, aiming at reducing the weight for words with a low confidence measure [7].

B. Shortlist generation

Based on the pairwise distances, a combination of the ranking of descriptions according to segments (segment-oriented ranking) with the ranking of segments according to descriptions (description-oriented ranking) is performed to provide an association strength between each pair segment/description, as illustrated in Figure 1. The reason for considering two rankings, one with respect to segments and another with respect to descriptions, is that a good association reflects that the description best represents the (transcript of the) video segment and that the segment is the most relevant for the description. Therefore, with the goal of representing the relevance of a description with respect to a segment by its rank, a list of descriptions sorted in descending cosine similarity order is computed, for each segment. Similarly, for each description, a list of segments sorted in descending similarity order is computed.

In order to derive the segment shortlist for each description from the two ranked lists described above, one can obviously rely on the rank of the pair segment/description in either the segment-oriented ranking or in the description-oriented one. Moreover, it was found that, for either the segment-oriented ranking or the description-oriented ranking, the difference of cosine similarity score between the association considered and the next best one is also a relevant feature. Indeed, an association for which the cosine similarity difference is large with respect to the next best association is considered as more significant than one for which the score difference is small. Formally, these two considerations, rank of the association and score difference, are taken into account by defining the strength of an association between a segment s_i and a description d_j as

$$S_s(s_i,d_j) = \frac{c(s_i,d_j)\ (c(s_i,d_j) - c(s_i,d_k))}{r(d_j|s_i)} \qquad (1)$$

for the segment-oriented ranking, where $r(d_j|s_i)$ denotes the rank of description d_j in the ranked list of descriptions for segment s_i, $c(s_i,d_j)$ the cosine similarity between s_i and d_j and k is such that $r(d_k|s_i) = r(d_j|s_i) + 1$. Similarly, the description-oriented association strength is given by

$$S_d(d_j,s_i) = \frac{c(d_j,s_i)\ (c(d_j,s_i) - c(d_j,s_k))}{r(s_i|d_j)} \qquad (2)$$

where $r(s_i|d_j)$ is the rank of segment s_i in the list of ranked segments for description d_j and k is such that $r(s_k|d_j) = r(s_i|d_j) + 1$. Finally, the association strength between s_i and d_j is obtained by summing the description-oriented and segment-oriented association strengths. For each description, the shortlist is obtained by keeping the strongest associated segments, *i.e.* the associations (s_i, d_i) for all $i < l$, where

[1] The smallest segmental unit of sound employed to form meaningful contrasts between utterances.

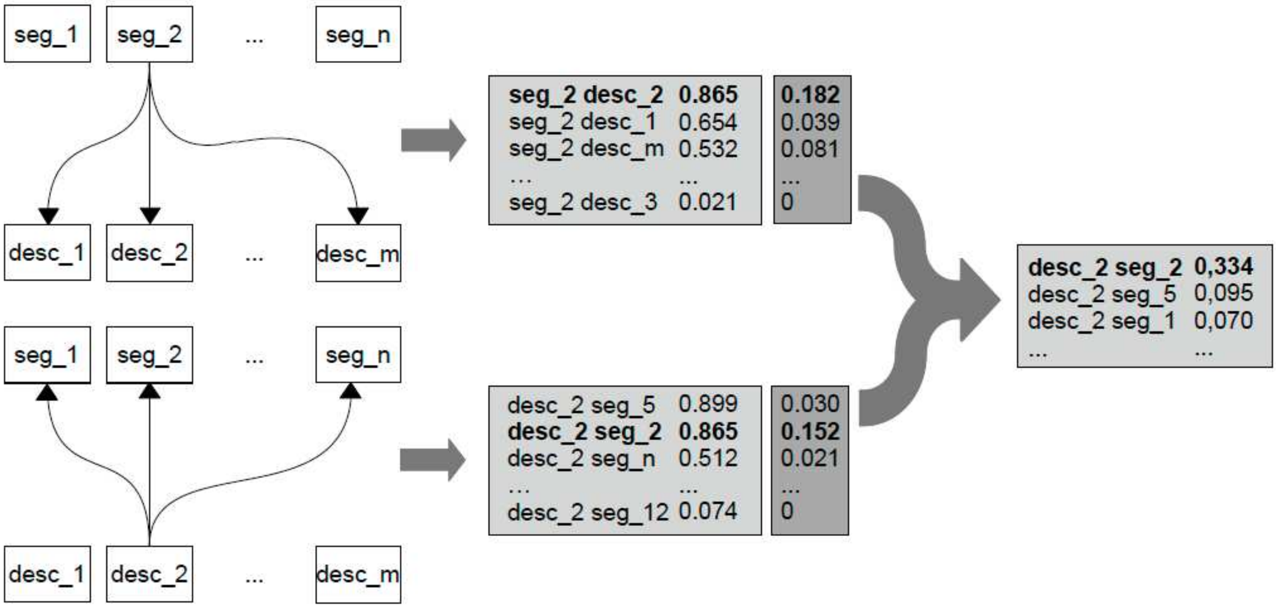

Figure 1. *Global architecture of the transcript-based shortlist generation process.*

l is such that $[S_d(s_l,d_l)+S_s(s_l,d_l)]-[S_d(s_{l+1},d_{l+1})+S_s(s_{l+1},d_{l+1})]>seuil$.

III. Phonetic-based shortlist generation

As mentioned in the introduction, word-level comparisons between transcripts and descriptions suffer from the significant number of out-of-vocabulary words. As, by definition, such words can not be recognized, the ASR system outputs phonetically similar words, *i.e.*, a sequence of words that sounds alike the OOV words. The high OOV rate is mainly due to the presence of many unpredictable proper nouns corresponding to occasional reporters in the news (or news-like) shows, guests in talk shows, as well as fictive characters in series or movies. Moreover, if some descriptions provide a handful of details on the program contents, some are limited to the anchor and guests names, *e.g.*, in talk shows. Movie and TV series descriptions are also weakly correlated with the speech material, apart from the characters' and locations' names. To tackle this issue, a phonetic-based comparison is studied. We first briefly review phonetic-based spoken document indexing techniques before elaborating a shortlist generation function specific to proper nouns.

A. Phonetic-based spoken document retrieval method principles

Phonetic-based spoken document retrieval relies on the comparison of phoneme-level speech transcripts with a query (in general one or several words) converted into a phonetic string. Contrary to word-level retrieval approaches, the goal here is not to find an exact match for the query in the phonetic speech transcript. Indeed, apart from recognition errors, phonetic transcripts are disrupted by hesitation marks or speaker accents that influence the pronunciation of the words. Moreover, word boundaries are unknown in phoneme-level transcripts. Several methods have been designed to handle this approximate match and the lack of word segmentation. In some studies, as in [8], the phonetic transcript is split into *phonetic words* in order to use standard text-based indexing methods. However, the segmentation step is time consuming. Other approaches rely on sequences of n consecutive phonemes, known as phone(me) n-grams, as index terms [9], ignoring the sequential order in which the index terms are to be found. In this work, a segmental normalized edit distance, inspired from [10], was rather used to find the portion of the phonetic transcript that closest matches the query phonetic string according to the edit distance, thus taking into account the sequential order of the query phonemes and avoiding segmentation into phonetic words. Substitution, insertion and deletion weights in the edit distance were determined from a phonetic confusion matrix so as to take into account the typical phonetic errors made by the ASR system (*e.g.* /b/ and /d/).

B. Shortlist generation

The phonetic-based method seeks to relate descriptions containing proper nouns with the phonetic transcripts of video segments, the latter being derived from the word-level transcription using a dictionary of pronunciations[2]. Proper nouns in descriptions are automatically detected and converted to phonemes using the freely available toolkit LIA_PHON [11]. For each proper noun in a description and each segment, a score is computed based on the segmental normalized edit distance to measure how likely is the name to be in the transcript. The pairwise similarity measure between a description and a video segment is obtained by summing the segmental edit distance scores over all proper nouns in the description. Finally, for each description, a list of segments sorted in ascending order—the lower the pairwise similarity measure, the more similar the segment and the description—is obtained. As for the text-based method, an association strength between a segment s_i and a description d_j containing proper nouns is defined by taking into account the pairwise similarity measure $e(d_j, s_i)$, the rank $r(s_i|d_j)$ of the segment in the sorted list and the score difference with the next segment in the ranked list according to

$$S_p(d_j, s_i) = \frac{e(d_j, s_k) - e(d_j, s_i)}{r(s_i|d_j)\; e(d_j, s_i)} \qquad (3)$$

with k such that $r(s_k|d_j) = r(s_i|d_j) + 1$. As previously, a shortlist is generated for each description based on (3).

IV. Experimental results

The transcript-based and phonetic-based shortlists are used, either separately or jointly, to associate with each segment a unique description, the one with the highest score among those descriptions for which the shortlist contains the segment. We first present the dataset used, before discussing association results. Finally, an experiment in which the segment/description associations are used to control an existing labeling of the programs is described. This last experiment aims at providing a structured TV stream, *i.e.*, a segmented stream with validated EPG descriptions associated to each segment.

A. Data

The corpus used consists of a 10 days of continuous TV programs (May 11-20, 2005) extracted from a French nation TV channel (France 2). Program descriptions were taken from an on-line EPG. In total, 650 segments were obtained by automatic alignment between the video stream and the program guide [12]. As a result, segments may not exactly correspond to a single program, segmentation errors causing some programs to be split into several segments or, on the contrary, to group programs in a single segment. Such errors clearly emphasize the fact that it is not possible to directly rely on any information contained in the EPG (genre, description, *etc.*) for segment characterization. For evaluation purposes, program titles were manually assigned to each of the 650 segments where, for segments gathering various programs, a label is considered as correct if matching one of the programs actually corresponding to the segment.

Segments were automatically transcribed with a radio broadcast news transcription system, exhibiting error rates ranging from around 20 for broadcast news up to 70 % for movies or talk shows. The automatic speech transcription system used in the experiments was designed for radio broadcast news transcription. Hidden Markov acoustic models were trained using approximately 75 hours of speech material. A 4-gram language model was obtained from about 350 million words mostly coming from a French-speaking newspaper. A word error rate of about 20 % was achieved with this system on the ESTER 1 radio broadcast news transcription benchmark [13]. Even though better transcripts could be obtained using a state-of-the-art ASR system dedicated to TV shows, error rates would still range from 10 % on clean (e.g. news) data to 70 % on the most difficult shows. Transcripts lengths vary from 7 to 27,150 words, with an average of 2,643.

As mentioned previously, descriptions were obtained from an on-line program guide. Hence, some short programs, such as fillers—used to deal with broadcasting schedule adjustments—or weather forecast reports, do not have any descriptions in the EPG. We thus consider two different corpora: one containing all the programs, the other one limited to programs longer than 600 words. Between 24 and 28 program descriptions are available for each day, varying both in length (average: 50 words) and precision (from a title to a precise description of the content). Finally, 63% of the program descriptions contain at least one proper noun with an average of 7.5 per description.

B. Results of the association method

The transcript-based and phonetic-based association methods can be applied in four different ways, depending on the corpus (all/long programs) and on whether to consider time information or not. Indeed, contrary to descriptions coming from the Internet, those from the EPG provide the broadcasting schedule. As video segments are linked to a time slot, the chronological order of EPG descriptions can be used. Thus, two search methods are proposed: one with no time information in which associations are performed on segment and description lists for an entire day; one which uses a time anchored system in which the association process is performed within overlapping two-hours time slots. The recall for the two methods is presented in Table I and detailed results for each method are respectively given in Tables II and III.

Table I shows that the use of the time anchored system leads to better recall for both methods. Moreover, it can be noted that the transcript-based method can associate descriptions with segments with an average recall of 0.45 (average of the recall for the method applied on all transcripts and

[2] Actually, ASR relies on a dictionary containing the phonetic description of each word in the vocabulary and is therefore able to tell what pronunciation was uttered.

Table I
Recall for transcript and phonetic-based methods

	without Time Anchorage System		with Time Anchorage System	
	all transcripts	> 600 words transcripts	all transcripts	> 600 words transcripts
transcript based method	0.54	0.64	0.58	0.86
phonetic based method	0.20	0.35	0.32	0.54

Table II
Precision for the transcript-based method.

	without Time Anchorage System		with Time Anchorage System	
day number	all transcripts	> 600 words transcripts	all transcripts	> 600 words transcripts
11th	0.43	**0.60**	0.60	**0.63**
12th	0.36	**0.56**	**0.73**	0.72
13th	0.32	**0.45**	0.55	**0.59**
14th	0.44	**0.50**	**0.55**	0.52
15th	0.40	**0.46**	0.56	**0.57**
16th	0.35	**0.46**	0.68	**0.73**
17th	0.37	**0.41**	0.56	**0.60**
18th	0.22	**0.25**	0.53	**0.56**
19th	0.41	**0.55**	0.58	**0.67**
20th	0.37	**0.52**	0.61	**0.66**

without the time anchorage system) whereas the phonetic method attaches a description to segments with a recall of only 0.19. This difference is obviously due to the fact that 37% of the descriptions do not contain any proper nouns.

Concerning the transcript-based method results, given in Table II, it can be pointed out that the time anchor system returns a higher precision (an absolute increase of 0.23 for all transcripts and of 0.15 for transcripts of more than 600 words). Moreover, this method gives better results on long TV segments—in general corresponding to the longer and more informative descriptions—for a large majority of days.

Table III
Precision for the phonetic-based method.

	without Time Anchorage System		with Time Anchorage System	
day number	all transcripts	> 600 words transcripts	all transcripts	> 600 words transcripts
11th	0.25	**0.30**	**0.65**	0.64
12th	0.46	0.46	0.55	0.55
13th	0.40	**0.44**	**0.75**	0.73
14th	0.44	**0.50**	0.33	**0.38**
15th	0.42	0.42	0.61	0.61
16th	0.46	0.46	0.55	**0.63**
17th	**0.08**	0.00	**0.42**	0.37
18th	0.40	**0.44**	**0.55**	0.52
19th	**0.25**	0.16	0.31	**0.36**
20th	0.41	0.41	0.57	**0.66**

The phonetic-based method results reported in Table III are not as clear as the transcript-based ones. Indeed, if the precision is better when time anchorage is used, no improvement can be noticed when the method is applied on the longest TV segments. This can be explained by the fact that the descriptions linked to long segments do not always contain a lot of proper nouns. Moreover, as the phonetic method associates a score with each proper noun—even if the latter does not appear in the phonetic transcript—, a high percentage of bad associations is returned.

Results for the phonetic method are sometimes of a poor quality. However, some of the good associations returned by this method are not discovered by the textual method (the number of good associations returned by the phonetic method but not by the textual one ranges from 9 to 24). A combination of the two techniques was attempted according to the following heuristic: if both methods disagree, the description associated with the highest score is chosen. However, this combination does not improve the rate of good associations returned since the number of associations improved is less than that deteriorated. A closer examination of the deteriorations and improvements—in terms of segments lengths, proper nouns proportion or transcripts and program descriptions quality—did not exhibited one or more parameters explaining this result. Nevertheless, some proper nouns are attached to a description with a high score as far as they are compared with sub-sequences of the phonetic transcript that do not correspond to a proper noun but which are phonetically similar. Such cases can explain the deterioration of good textual associations by wrong phonetic ones.

Overall, when ignoring time anchorage, the sole transcript-based method results in an association for the long segments with a recall of 0.64 and a precision of 0.47. Using time information, recall and precision improve respectively to 0.86 and 0.63.

C. *Validation of an existing labeling*

In [12], the segmentation task is followed by a labeling task which attaches a name, *i.e.*, a program title, to each video segment. This labeling task, performed by dynamic alignment, does not use any semantic information. The goal of this validation application is to check and, when necessary, to modify the segment labels in order to decrease the error rate of the automatic labelling task. For this purpose, the associations previously returned using the time anchored system are compared to the labels associated by [12]. If the associated description and the label are equal, the labeling is considered to be correct. Otherwise, the broadcasting times designated respectively by the description and the label are compared with the starting time of the segment. If the description starting time better matches the segment starting time, the label is replaced by the description title. However, if the two broadcasting times are too far apart from the

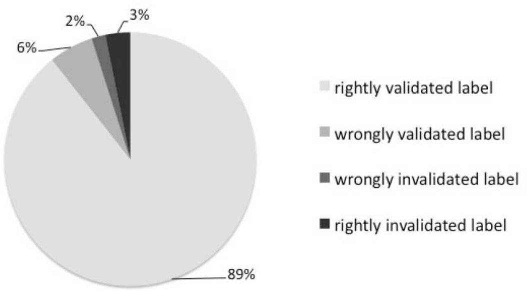

Figure 2. *Validation of an existing labeling.*

actual segment starting time (more than half an hour), the label is invalidated and no label is proposed in substitution. Results for the control of the existing labels are presented in Figure 2. The modification of the labels decrease the labeling error rate by 0.2 %, as the number of changes that improve the labeling is almost equal to the number of modifications that damage it. The cases of deterioration are always related to over-segmentation problems where the starting time of a segment does not correspond to any description.

V. Conclusion and future work

The speech material contained in TV programs, accessible by means of ASR systems, can be useful for characterizing TV programs. Indeed, on the longest TV segments and using the time anchored system, our system associates program descriptions with TV segments with a precision of 0.63 when the textual method is used and of 0.55 with the phonetic method. However, descriptions automatically associated to TV segments for characterization purposes are not acurate enough when the association process is not constrained by either the segment length or time anchorage. Hence, the association process should be improved in order to be used in practical settings, *e.g.*, to account for the genre of programs which strongly impacts the EPG description. Nevertheless, these results show that speech transcripts can be used in order to associate a textual description with a TV program even though word error rates can be quite high.

A first way to improve the results is to find how to properly combine the textual and phonetic methods. Since the correct associations returned by the two techniques are not the same, a good combination could lead to a better precision (an absolute increase included between 0.03 and 0.07 depending on the implementation). Since segment lengths or transcripts quality do not seem to have an impact on the combination results, the most immediate problem to fix is the matching between proper noun phonetic strings with sub-sequences of phonetic transcript that do not correspond to a proper noun, *e.g.* by detecting named entities in the transcripts in spite of transcription errors. Finally, as textual transcripts of TV stream give some promising results in our association system, it seems interesting to use them for thematic segmentation purposes. Indeed, as the TV stream structuring task can consist in linking TV programs that deal with a same topic, partitioning the TV stream into broadcasts or parts of broadcasts that addressed a unique subject is a first stage to reach this goal.

References

[1] I. Ide, H. Mo, N. Katayama, and S. Satoh, “Topic threading for structuring a large-scale news video archive,” in *3rd International Conference on Image and Video Retrieval*, 2004.

[2] Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi, “Automatic parsing of TV soccer programs,” in *2nd International Conference on Multimedia Computing and Systems*, 1995.

[3] Z. Liu, Y. Wang, and T. Chen, “Audio feature extraction and analysis for scene segmentation and classification,” *Journal of VLSI Signal Processing Systems*, vol. 20, no. 1/2, pp. 61–79, 1998.

[4] X. Wu, C.-W. Ngo, and Q. Li, “Threading and autodocumenting news videos,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, p. 59, 2006.

[5] F. Fukumoto and Y. Suzuki, “Event tracking based on domain dependency,” in *23rd annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.

[6] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[7] G. Lecorvé, G. Gravier, and P. Sébillot, “An unsupervied Web-based topic language model adaptation method,” in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2008.

[8] P. S. Cardillo, M. Clements, and M. S. Miller, “Phonetic searching *vs.* LVCSR: how to find what you really want in audio archives,” *International Journal of Speech Technology*, vol. 5, no. 1, pp. 9–22, 2002.

[9] K. Ng and V. M. Zue, “Subword-based approaches for spoken document retrieval,” *Speech Communication*, vol. 32, no. 3, 2000.

[10] A. Muscariello, G. Gravier, and F. Bimbot, “Variability tolerant audio motif discovery,” in *15th International Multimedia Model Conference*, 2009.

[11] F. Béchet, “Lia_phon : un système complet de phonétisation de textes,” *Traitement automatique des langues*, vol. 42, no. 1, pp. 47–67, 2001.

[12] X. Naturel, G. Gravier, and P. Gros, “Fast structuring of large television streams using program guides,” in *4th International Workshop on Adaptive Multimedia Retrieval*, 2006.

[13] S. Galliano, E. Geoffrois, D. Mostefa, J.-F. Bonastre, K. Choukri, and G. Gravier, “The ESTER phase II evaluation campaign for the rich transcription of french broadcast news,” in *Interspeech*, 2005.