Audio thumbnails for spoken content without transcription based on a maximum motif coverage criterion

Guillaume Gravier, Nathan Souviraà-Labastie, Sébastien Campion, Frédéric Bimbot

Audio thumbnails for spoken content without transcription based on a maximum motif coverage criterion

Guillaume Gravier[1], *Nathan Souviraà-Labastie*[2], *Sébastien Campion*[3], *Frédéric Bimbot*[1]

[1]CNRS, IRISA - UMR 6074 & Inria Rennes
[2]Université Rennes 1, IRISA - UMR 6074 & Inria Rennes
[3]Inria, IRISA - UMR 6074 & Inria Rennes
firstname.lastname@irisa.fr

Abstract

The paper presents a system to create audio thumbnails of spoken content, i.e., short audio summaries representative of the entire content, without resorting to a lexical representation. As an alternative to searching for relevant words and phrases in a transcript, unsupervised motif discovery is used to find short, word-like, repeating fragments at the signal level without acoustic models. The output of the word discovery algorithm is exploited via a maximum motif coverage criterion to generate a thumbnail in an extractive manner. A limited number of relevant segments are chosen within the data so as to include the maximum number of motifs while remaining short enough and intelligible. Evaluation is performed on broadcast news reports with a panel of human listeners judging the quality of the thumbnails. Results indicate that motif-based thumbnails stand between random thumbnails and ASR-based keywords, however still far behind thumbnails and keywords humanly authored.
Index Terms: spoken content processing, audio mining, motif discovery, summarization, thumbnailing

1. Introduction

Over the past few years, unsupervised word discovery methods have been proposed with the goal of finding patterns in the signal which correspond to words or short sequences of words [1, 2, 3, 4], similar to previous work in the field of music processing [5, 6]. All of these methods search for repeating acoustic patterns in speech data in a totally unsupervised manner, targeting patterns which are likely to correspond to word-like units because of their length and repetitiveness. Apart from being a fun and challenging scientific problem, unsupervised word discovery has a number of potential applications, among which transcript-free spoken content processing. Indeed, many natural language processing techniques exploit the notion of reoccurrence, i.e., repetition of a term at the lexical level, e.g., to weight terms, to find out keywords, to measure lexical cohesion. Unsupervised word-like motif discovery have made it possible to measure reoccurrence directly at the acoustic level, without resorting to a lexical representation, hence opening the door to transcript-free spoken content processing. This idea has been (briefly) explored in a few tasks such as topic segmentation [7], multiple document summarization [8], document clustering [9] and acoustic word cloud representations [10].

Following this line of thought, we propose and evaluate a transcript-free algorithm to generate audio thumbnails of spoken content, an idea initially introduced for music summarization [11, 12, 13]. Thumbnails, whether written or audio, are intended for users to quickly grasp the content of a document without having to read it or listen to it entirely. Similar to classical thumbnails, e.g., on popular web search engines, an audio thumbnail is a short audio file, typically 10 s long, providing a condensed version of the information contained in a document. The straightforward way to generate thumbnails for spoken content consists in obtaining a lexical representation via automatic speech recognition (ASR) before applying text summarization techniques [14]. Resorting to ASR is however costly from a computational standpoint and only applicable to languages for which ASR systems exist. As an alternative to ASR-based thumbnailing, word discovery and pattern reoccurrence can be exploited to directly generate an audio thumbnail.

In this paper, we introduce a criterion based on maximum motif coverage for transcript-free audio thumbnailing, elaborating on the idea that words that repeat within a spoken document are meaningful and relevant and should therefore be included in the thumbnail. The maximum motif coverage criterion hence seeks to maximize the number of motifs which have at least one occurrence included in the thumbnail, subject to length and intelligibility constraints. Via listening tests, we experimentally compare audio thumbnails based on motif discovery to various contrasts, including ASR-based keywords, on broadcast news reports with the ultimate goal of helping users navigate within audiovisual news archives as illustrated by the Texmix demo [15][1].

The paper is organized as follows. Sec. 2 briefly recalls the word-like motif discovery method used. Sec. 3 details the thumbnailing algorithm based on the notion of maximum motif coverage. Finally, human-centric evaluations are reported in Sec. 4 before a concluding discussion in Sec. 5.

2. Word-like motif discovery

The discovery of repeating motifs follows the algorithm of Muscariello *et al.* [16], whose main principles are recalled hereunder. Given a set of feature vectors $\mathbf{x} = \{x_1, \ldots, x_n\}$, we seek all pairs $[a, b]$ and $[c, d]$ such that the two subsequences $\{x_a, \ldots, x_b\}$ and $\{x_c, \ldots, x_d\}$ are acoustically close enough subject to minimal constraints (minimal segment length and non overlapping segments). Segments that verify this property are considered as belonging to the same underlying *motif*.

Efficiently finding reoccurrences of a portion of $\mathbf{x}$ exploits two properties. First, we rely on the notion of *seed*, borrowed from the field of genomics, where a seed is a short segment that might be contained in a bigger motif. The foundational

[1]See http://texmix.irisa.fr/modis for examples of audio thumbnails in French broadcast news data.

idea of the seed based approach is to search for repetitions of a seed using a segmental variant of dynamic time warping (DTW) which enables locating the start and end point of subsequences close to the seed in the data, if any. A motif is hypothesized whenever such a subsequence is similar enough to the seed. In this case, the seed and its matching sequence are grown both forward and backward to find maximal length occurrences of a repeating motif by extending the optimal warping path in both directions as long as the average path distortion does not grow too high. The two extended matching segments are considered as occurrences of a motif if the average warping distortion is below a similarity threshold. Second, similar to [17], we assume that repeating words repeat at least once locally, i.e., with a minimum time between two occurrences. In other words, for a motif, there are at least two occurrences which repeat within a limited time span. This assumption enables to break down the inherently quadratic complexity of the search for repetitions, restricting the search for a reoccurrence of a seed to a short time window in the future (or, equivalently, in the past). To find long span repetitions of a motif, a library is used to store a locally reoccurring motif whenever one is discovered. The library also acts as a clustering step, grouping all occurrences of a motif within the same entry.

Combining the property of local repetition with the seed principle leads to the following algorithm sequentially processing the data to incrementally build the library which is initially empty. Starting with a short seed at $t = 0$, segmental DTW is used to search for a repetition of the seed either in the library or, if not found there, in the near future. If a close-enough repetition of the seed is found, the match is grown to find maximally repeating sequences which are considered as a reoccurrence if the DTW similarity between the extended matching sequences is below a threshold α. A reoccurrence found with a library entry indicates that a long span match of an existing motif was found. A reoccurence found in the near future means that a new locally repeating motif was discovered and must be added to the library. The algorithm proceeds by shifting the seed until the data has been entirely processed.

The crucial parameters of the motif discovery algorithm are (a) the seed length, which is directly related to the minimum length of a motif occurrence; (b) the size of the near future time window in which the seed is searched for; (c) the similarity threshold α, where the lower the threshold, the more similar the occurrences of a motif. For word discovery, a short seed length must be considered with a sufficiently loose threshold α to account for variability in speech, yet preserving precision for each motif. The size of the near future depends on the type of data considered and is only critical for computation time.

3. Thumbnailing algorithm

Elaborating on the postulate that repeated words/motifs are meaningful to spoken content understanding and should therefore be included in the thumbnail, the thumbnailing algorithm exploits motif discovery to find repeating word-like patterns. Thumbnails are generated by extracting a very limited number of segments from the original data, typically one or two, as illustrated in Fig. 1. The selection of the segments to extract is based on the maximum motif coverage principle with the goal of including at least one occurrence of as many motifs as possible, subject to constraints on the total length and on continuity.

We briefly discuss how motif discovery was used to generate a library of motifs before providing the details of the segment selection algorithm.

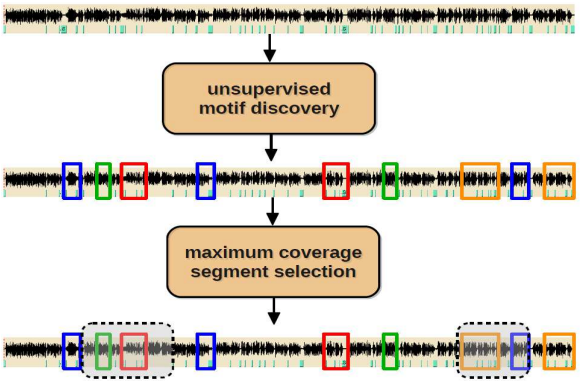

Figure 1: The different steps involved in motif discovery based thumbnail generation: motifs occurrences are indicated with colored rectangles, selected segments are highlighted in shadowed rounded boxes. Pale green regions below the waveform indicate silences.

3.1. Creation of the library

Motif discovery as presented in Sec. 2 may yield significantly different results depending on the length of the seed and on the similarity threshold α which controls the amount of variability tolerated between two occurrences of a motif. Very short seeds will often result in non informative, mostly non lexical, motifs such as breath intakes. On the contrary, long seeds will result in no motifs being found. Interestingly, within the range of reasonable seed size, experience tells that different, complementary, libraries are often found when using different lengths for the seed. We exploit this property to create a library of motif which combines motifs found with a seed of 500 ms with motifs found with a longer seed of 1 s. Those values were empirically chosen and are well adapted to the broadcast news data we consider in the experimental section. The similarity threshold α is determined independently for each waveform file to process and each seed length. Setting the threshold appropriately result from a trade-off between various factors, mostly, the number of motifs within the library and the balance between precision and recall for each motif. Low thresholds yield libraries with few motifs detected with high precision and limited recall. Conversely, high thresholds will provide libraries with many entries and exhibit poor precision. To ensure that a sufficient number of motifs are found for each document processed, similarity thresholds are set depending on the file length so as to yield an average number of motif occurrences per minute. For a 0.5 s (resp. 1 s) seed, we target 20 (resp. 10) occurrences for a 2 m file. For a set of feature vectors $\mathbf{x}$, of length $L(\mathbf{x})$ minutes, this is generalized by defining the expected number of motif occurrences in $\mathbf{x}$ as

$$\widehat{K}_{\mathbf{x}} = k\left(1 + \log\left(\frac{L(\mathbf{x})}{2}\right)\right) \quad , \tag{1}$$

where k is either 10 or 20 depending on the seed length. Given $\widehat{K}_{\mathbf{x}}$, α is empirically determined by grid search so that the actual number of occurrences is as close as possible to $\widehat{K}_{\mathbf{x}}$.

3.2. Selection of the segments

The motif discovery step yields a library $M_{\mathbf{x}}$ of $N_{\mathbf{x}}$ motifs resulting from the mere concatenation of the motifs found with

seed lengths of resp. 0.5 s and 1 s. In the sequel, $M_{\mathbf{x}}(i)$ denotes the ith motif and $M_{\mathbf{x}}(i,j)$ its jth occurrence ($j \in [1, N_{\mathbf{x}}^{(i)}]$).

3.2.1. Selection criteria

Thumbnails are created by selecting a small number of *relevant* segments within $\mathbf{x}$ which are then combined to provide an audio output. In spirit, this approach is similar to selecting relevant sentences in extractive text-based summarization (see, e.g., [18] for a review on the subject). However, contrary to texts, the notion of sentence is obviously absent from audio files and one must jointly define the segments to consider and assess their relevance. The set of segments selected for thumbnailing should globally have three main properties, namely, be relevant and exhaustive with respect to the original message, be short enough and be easy to understand.

We propose to translate relevance and exhaustiveness into a motif coverage criterion which counts the number of motifs that have occurrences within the set of segments selected. Formally, if $U_{\mathbf{x}} = \{\mathbf{x}_{s_1}^{e_1} \dots \mathbf{x}_{s_n}^{e_n}\}$ denotes a collection of n segments in $\mathbf{x}$, identified by their start and end times s_i and e_i ($i \in [1, n]$), the motif coverage of $U_{\mathbf{x}}$ with respect to the library $M_{\mathbf{x}}$ is defined as

$$c(U_{\mathbf{x}}, M_{\mathbf{x}}) = \frac{\sum_{i=1}^{N_{\mathbf{x}}} \delta(M_{\mathbf{x}}(i) \cap U_{\mathbf{x}})}{N_{\mathbf{x}}} \ , \qquad (2)$$

where $\delta(M_{\mathbf{x}}(i) \cap U_{\mathbf{x}}) = 1$ if at least one occurrence $M_{\mathbf{x}}(i,j)$ is contained within a segment of $U_{\mathbf{x}}$, 0 otherwise. In plain words, coverage is the proportion of motifs in $M_{\mathbf{x}}$ which occur in $U_{\mathbf{x}}$.

Motif coverage can obviously be maximized simply by selecting one occurrence of each motif as $U_{\mathbf{x}}$. However, this trivial solution is not very satisfactory as far as understandability is concerned. Indeed, concatenating a large number of very short segments, without any context, is hard to comprehend at the audio level. It is interesting to draw a parallel with texts at this point. Playing an occurrence of each motif is similar to displaying keywords, which is different from thumbnailing where short excerpts of the text are displayed. While keywords provide an efficient way for a human reader to get a gross feeling for the content, thumbnails usually enables a better understanding. Moreover, preliminary experiments on a very limited scale with audio keywords, i.e., motifs occurrences, stressed that listening significantly differs from reading, resulting in increased difficulty to get a feeling of the content from audio keywords compared to text keywords. Hence, in addition to motif coverage, we ensure the easiness of understanding criterion by selecting a very limited number of segments, typically one or two, thus favoring long segments which include the largest possible number of motifs with a natural context.

Incorporating the maximum length criterion of the thumbnail with motif coverage and easiness of understanding, the segment selection step consists in finding the set of segments maximizing the global, empirically defined, cost function

$$C(U_{\mathbf{x}}; M_{\mathbf{x}}) = c(U_{\mathbf{x}}, M_{\mathbf{x}}) \ \left(1 - \frac{l(U_{\mathbf{x}})}{25}\right)^3 \ (1 - 0.2|1-n|) \ , \qquad (3)$$

where $l(U_{\mathbf{x}}) = \sum e_i - s_i$ is the total duration in seconds. The first term of the product in $C(U_{\mathbf{x}}; M_{\mathbf{x}})$ is the coverage, the second accounts for the total duration of the audio thumbnail, with an upper limit of 25 s, while the third limits the number of segments included. Parameters were empirically fixed on a few examples.

Algorithm 1 Two-level greedy optimization algorithm for (3)

$\widehat{U}_{\mathbf{x}} \leftarrow \emptyset$
for $m = 2$ to $N_{\mathbf{x}}$ **do** ▷ for all possible subset sizes
  $U_{\mathbf{x}} \leftarrow$ MAXIMIZE$(M_{\mathbf{x}}, m)$ ▷ find best segment
  $\widehat{U}_{\mathbf{x}} \leftarrow U_{\mathbf{x}}$ **if** $C(U_{\mathbf{x}}, M_{\mathbf{x}}) > C(\widehat{U}_{\mathbf{x}}, M_{\mathbf{x}})$
  $M'_{\mathbf{x}} \leftarrow M_{\mathbf{x}} -$ MOTIF$(U_{\mathbf{x}}, M_{\mathbf{x}})$ ▷ remove motifs in $U_{\mathbf{x}}$
  for $m' = 2$ to $N_{\mathbf{x}} - m$ **do** ▷ proceed at level 2
    $U'_{\mathbf{x}} \leftarrow U_{\mathbf{x}} \cup$ MAXIMIZE$(M'_{\mathbf{x}}, m')$
    $\widehat{U}_{\mathbf{x}} \leftarrow U'_{\mathbf{x}}$ **if** $C(U'_{\mathbf{x}}, M_{\mathbf{x}}) > C(\widehat{U}_{\mathbf{x}}, M_{\mathbf{x}})$
  end for
end for
return $\widehat{U}_{\mathbf{x}}$

// find shortest segment containing exactly m motifs of M
function MAXIMIZE(M, m)
  return $\arg\max_U C(U; M)$ s.t. $\sum_i \delta(M(i) \cap U) = m$
end function

3.2.2. Algorithmic implementation

The maximization of (3) is performed in a greedy heuristic way, considering in turn various combinations of motifs from the library and limiting ourselves to a maximum of $n = 2$ segments selected. The idea is to search for combinations of motifs with consecutive occurrences for all possible sizes of such combination. The solution is initialized to the empty set. For all sizes m of the possible motif subsets, we determine the segment in $\mathbf{x}$ that maximizes (3) across all tuples of motifs of size m with consecutive occurrences. Since the first and third terms in (3) are fixed in this case, the best segment is actually the smallest segment which includes occurrences of exactly m motifs. This selection procedure is recursively pursued in search of a potential second segment after removing from the library the motifs considered at level one. In the end, we pick the configuration that maximizes the global cost function among the one or two segments solutions enumerated, as illustrated by Algorithm 1.

In practice, segment boundaries are adjusted according to silences in Algorithm 1. Silence detection is performed based on the waveform energy profile. When searching for the minimum length segment maximizing (3), boundaries of the segments are extended to the nearest silence of at least 0.2 s, including silences within the segment to avoid concatenation artifacts.

4. Human-centric evaluations

Evaluation was performed on reports from TV broadcast news which were presented to users along with a variety of audio thumbnails and of keywords. We first describe the data and setting for this evaluation before reporting results.

4.1. Data and settings

A total of 20 reports, taken from the regular 20h broadcast news show on the French channel France 2, were selected to include a diversity of topics (sports, politics, etc.). Reports have an average duration of 1 m 48 s, including the anchorperson's introduction preceding the report itself, with mostly a unique speaker during the report. The introduction was considered as being part of the report and was included in all of the experiments.

For each report, a thumbnail based on motif discovery was created according to Algorithm 1 along with two contrasts,

namely a random thumbnail and an oracle thumbnail. Random thumbnails are made by selecting at random between single or multiple segments. In the first case, an initial utterance (as provided by the ASR partitioning system) is chosen at random, selecting contiguous utterances to ensure at least a 10 s thumbnail. In the second case, a set of non contiguous utterances are chosen at random until a minimum duration of 10 s is reached. Oracle thumbnails simply consist in taking the introduction of the report by the anchorperson, authored by journalists as a teasing summary. In addition to audio thumbnails, keywords were also used as a different type of thumbnail, presenting a list of keywords to the user instead of an audio summary. For each report, a set of 10 keywords was selected, either by a human listener or automatically from the ASR transcript. Note that in the first case, keywords were chosen with no constraints, in particular no requirement to select keywords in a thesaurus or in the transcripts, with listeners often knowing the context of the news. ASR-based keywords were selected using the standard tf-idf weighting after lemmatization, considering only adjectives, nouns and verbs. Transcripts exhibit a word error rate of about 18 % [19] and inverse document frequencies were computed on a large collection of newspaper articles.

For evaluation, subjects were presented with the original report with either one of the audio thumbnails or keyword lists. They were asked to judge whether the thumbnail or keywords summary was characteristic of the report and provide a score on a scale from 1 to 7, the lowest score meaning "not at all" and the highest "that's exactly it". Evaluation was implemented via a dedicated web page that subjects could access any time they wanted, each time with a randomly chosen pair among the 100 possible tuples (20 reports, 5 thumbnail types). Subjects were recruited mostly among researchers and students in computer science, with some researchers in the speech communication field. A few votes from non technical people were also registered. All had limited knowledge of what they were evaluating and the contrasts were never described to the subjects who obviously didn't know which type of thumbnails they were rating.

4.2. Results

About 100 subjects participated in the evaluation[2], one third female and two third male. In total, 515 pairs report/thumbnail were evaluated, with about 100 responses for each type of thumbnails. The average number of votes per subject (excluding anonymous votes) is 4.2 but we observed significant variations across subjects with a standard deviation of 6.1: Around 50 % of the subjects only evaluated one or two pairs while the top 5 contributors account for about 30 % of the votes. This is clearly not an optimal situation but we believe it is the price to pay to ensure a sufficient number of responses by imposing almost no constraint for participation to the evaluation process.

Results are reported in Tab. 1, averaged across all votes for a given type of thumbnails. We report the empirical mean of the scores, the empirical standard deviation and the coefficient of variation, along with the number of votes on which these statistics were established. The coefficient of variation is a measure of how spread the votes are, independent of the value of the mean, thus providing a better measure of dispersion than the standard deviation.

Obviously, human-based approaches still outperform by far automatic ones to provide a short audio or keyword summary of multimedia spoken data, with significant differences in the average score to the benefit of human keywords and oracle thumbnails. Also, comparing these two strategies, i.e., audio thumbnail and keywords, in the human generated case, keywords appeared as more relevant to characterize a report than audio thumbnails. We believe that this is partly attributable to the fact that reading is faster and easier than listening to quickly comprehend the content. However, it is interesting to note that the average score for oracle audio thumbnails is reasonably high, indicating that in the absence of ASR transcription, audio thumbnails provide an interesting option.

[2]To keep things as light as possible, subjects were not required to identify themselves but were rather invited to enter a pseudo: Some did not and voted anonymously.

thumbnail type	l	#vote	μ	σ	$100\frac{\sigma}{\mu}$
human keywords	—	117	5.58	1.06	18.9
ASR keywords	—	102	3.07	1.26	41.1
oracle thumbnail	12.9	98	4.78	1.41	29.5
motif thumbnail	14.9	99	2.92	1.50	51.5
random thumbnail	11.7	99	2.52	1.46	58.0

Table 1: Average duration (l, in seconds), number of votes and mean (μ), standard deviation (σ) and coefficient of variation ($100\ \sigma/\mu$) of the score for each type of thumbnail

Focusing on the comparison of the machine-generated audio thumbnails, motif-based thumbnails stand between random ones and oracle ones, with a large difference for the latter. Variations in the scores across subjects is also much higher for the machine-generated thumbnails than for the oracle ones, as indicated by the variation coefficients. Two factors can be invoked to explain this observation. On the one hand, automatically created thumbnails vary in quality from one report to another, depending on a variety of factors (number of actual motifs, quality of the motif discovery, randomness in the random case, etc.). On the other hand, facing poorer global quality, judgments greatly vary according to the degree of expectation of the subjects. Finally, the difference between random thumbnails and motif thumbnails is only marginally significant (p=0.057 according to the unpaired Student test). Interestingly, the difference in scores between motif thumbnails and ASR keywords is not significant (p=0.44). While word-discovery based thumbnails are only marginally better than random selection, they provide information comparable to ASR based keywords without the burden of transcription.

5. Discussion

The human-based experimental evaluation of audio thumbnails and keywords reported in this paper clearly establishes that transcript-free thumbnailing is possible and provide a level of information similar to ASR-based keywords. But reaching the level of humanly authored thumbnails and keywords require further work. The maximum coverage criterion is based on repetitions but do not account for the notion of frequency, less that of inverse document frequency. These two notions are known to be crucial in keyword extraction and should therefore be taken into account. While accounting for frequency is fairly easy, defining the notion of inverse document frequency in the framework of unsupervised motif discovery remains a challenge. Another interesting question to address is that of contextualization. Motif-based thumbnails include motifs plus context while ASR keywords, which obtained a similar average score, only include motifs. Measuring the importance of the context in audio thumbnails thus appear as a key to a better understanding of human acceptability of summaries and thumbnails.

6. References

[1] A. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transaction on Acoustic, Speech and Language Processing*, vol. 16, no. 1, pp. 186–197, Jan. 2008.

[2] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Conf. of the Intl. Speech Communication Association (Interspeech)*, 2009, pp. 2843–2846.

[3] X. Anguera, R. Macrae, and N. Oliver, "Partial sequence matching using an unbounded dynamic time warping algorithm," in *IEEE Intl. Conf. on Acoustics Speech and Signal Processing*, 2010, pp. 3582–3585.

[4] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011, pp. 401–406.

[5] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," in *Journal of New Music Research*, vol. 32, no. 2, 2003, pp. 153–163.

[6] J. Paulus and A. Klapuri, "Music structure analysis by finding repeated parts," in *ACM Workshop on Audio and Music Computing Multimedia*, 2006, pp. 59–68.

[7] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Annual meeting of the Association for Computational Linguistics*, vol. 45, no. 1, 2007.

[8] X. Zhu, G. Penn, and F. Rudzicz, "Summarizing multiple spoken documents: Finding evidence from untranscribed audio," in *Proc. Joint Conf. of the Annual Meeting of the ACL and of the Intl. Joint Conf. on Natural Language Processing of the AFNLP*, 2009, pp. 549–557.

[9] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. Conf. on Empirical Methods in Natural Language Processing*, 2010, pp. 460–470.

[10] R. Flamary, X. Anguera, and N. Oliver, "Spoken Wordcloud: clustering recurrent patterns in speech," in *Workshop on Content-based Multimedia Indexing*, 2011.

[11] M. Cooper and J. Foote, "Summarizing popular music via structural similarity analysis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[12] W. Chai and B. Vercoe, "Structural analysis of musical signals for indexing and thumbnailing," in *IEEE Joint Conference on Digital Libraries*, 2003.

[13] G. Peeters, *Computer Music Modeling and Retrieval*. Springer Berlin Heidelberg, 2004, ch. Deriving musical structures from signal analysis for music audio summary generation: Sequence and State approach, pp. 143–166.

[14] J. Goldman, S. Renals, S. Bird, F. de Jong, M. Federico, C. Fleischhauer, M. Kornbluh, L. Lamel, D. W. Oard, C. Stewart, and R. Wright, "Accessing the spoken word," *International Journal on Digital Libraries*, vol. 5, no. 4, pp. 287–298, 2005.

[15] M. Bréhinier, S. Campion, and G. Gravier, "Texmix: An automatically generated news navigation portal," in *Intl. Conf. on Multimedia Retrieval*, 2012.

[16] A. Muscariello, G. Gravier, and F. Bimbot, "Unsupervised motif acquisition in speech via seeded discovery and template matching combination," *IEEE Trans. on Audio, Speech and Language*, vol. 20, no. 7, pp. 2031–2044, 2012.

[17] C. Herley, "ARGOS: Automatically extracting repeating objects from multimedia streams," *IEEE Transactions on Multimedia*, vol. 8, no. 1, pp. 115–129, 2006.

[18] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, 2010.

[19] 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.