Towards automatic speech summarization without transcription

G. Gravier, A. Muscariello, C. Guinaudeau, F. Bimbot

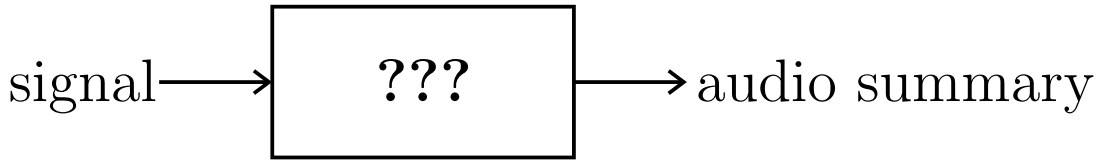
nom.prenom@irisa.fr

IRISA & INRIA Rennes

The classical approach

Use ASR as a transducer to the text domain

Direct generation of an audio summary



What can be used in speech?

Idea: find *salient words* for characterization and summarization

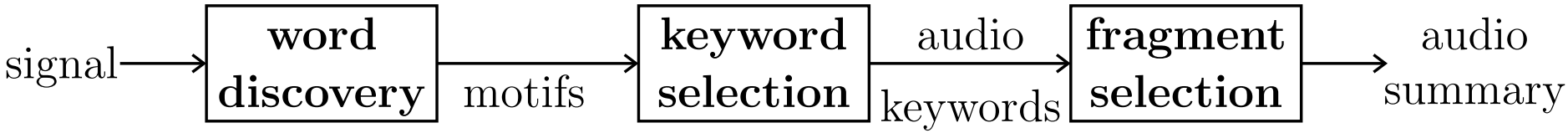
Saliency in spoken content can be measured at

1. the lexical level
   - repetition (term frequency)
   - unusual occurence (inverse document frequency)
2. the prosodic level
   - accented words or lexical stress
   - pauses

but we are facing the following difficulties

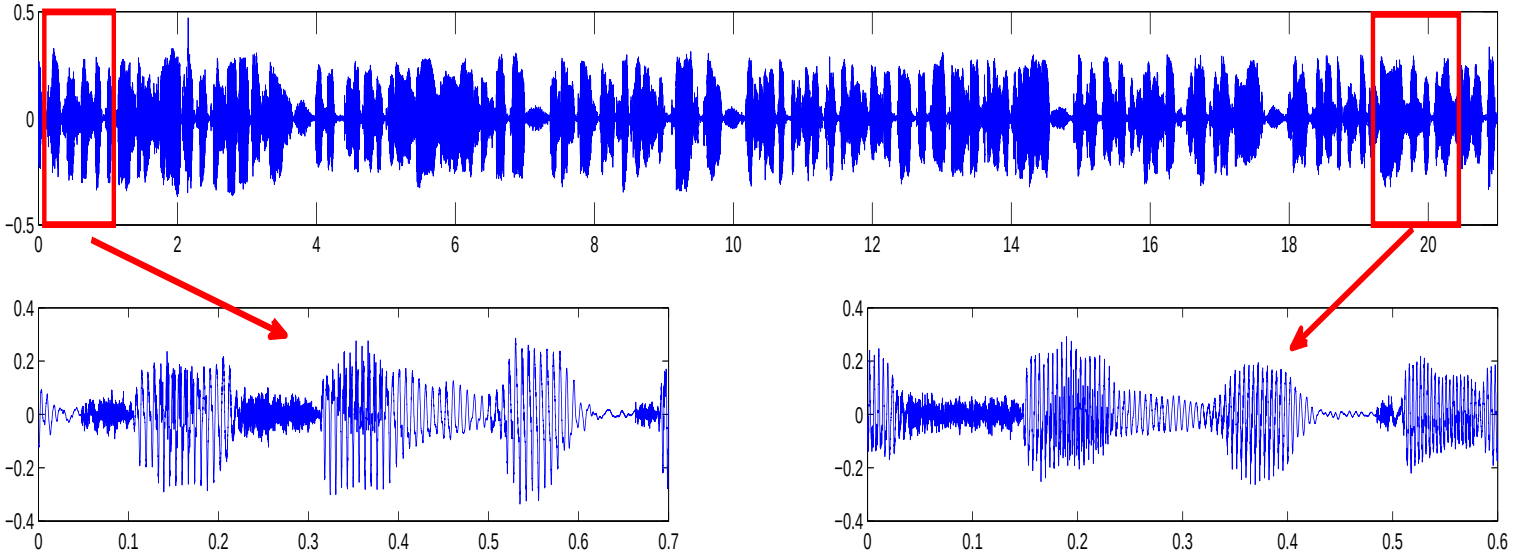
- no notion of words!
- no notion of inverse document frequency

Approach considered



1. discover recurring word-like audio patterns
2. select relevant motifs
3. build summary

Motif discovery illustrated



Repetitions define **iconic sounds** within the data

- unknown motifs, numbers, number of occurrences, length, etc.
- no acoustic or linguistic models trained prior to the discovery
- no additional modality of info (visual, textual, cues)

Motif discovery in equations

Motif discovery is the task of **discovering and collecting** within a document or stream χ **all possible pairs of segments** χ_a^b and χ_c^d that fulfill three requirements

$$H(\chi_a^b, \chi_c^d) < \epsilon$$ similarity condition

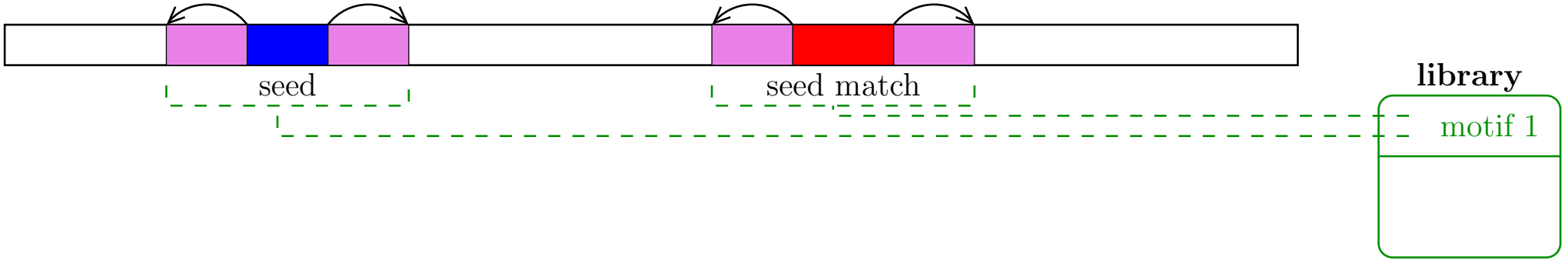
$$b - a, d - c > L_{\text{min}}$$ minimum length condition

$$a < b < c < d$$ disjointness condition

$\Rightarrow$ **naive approach is intractable!**

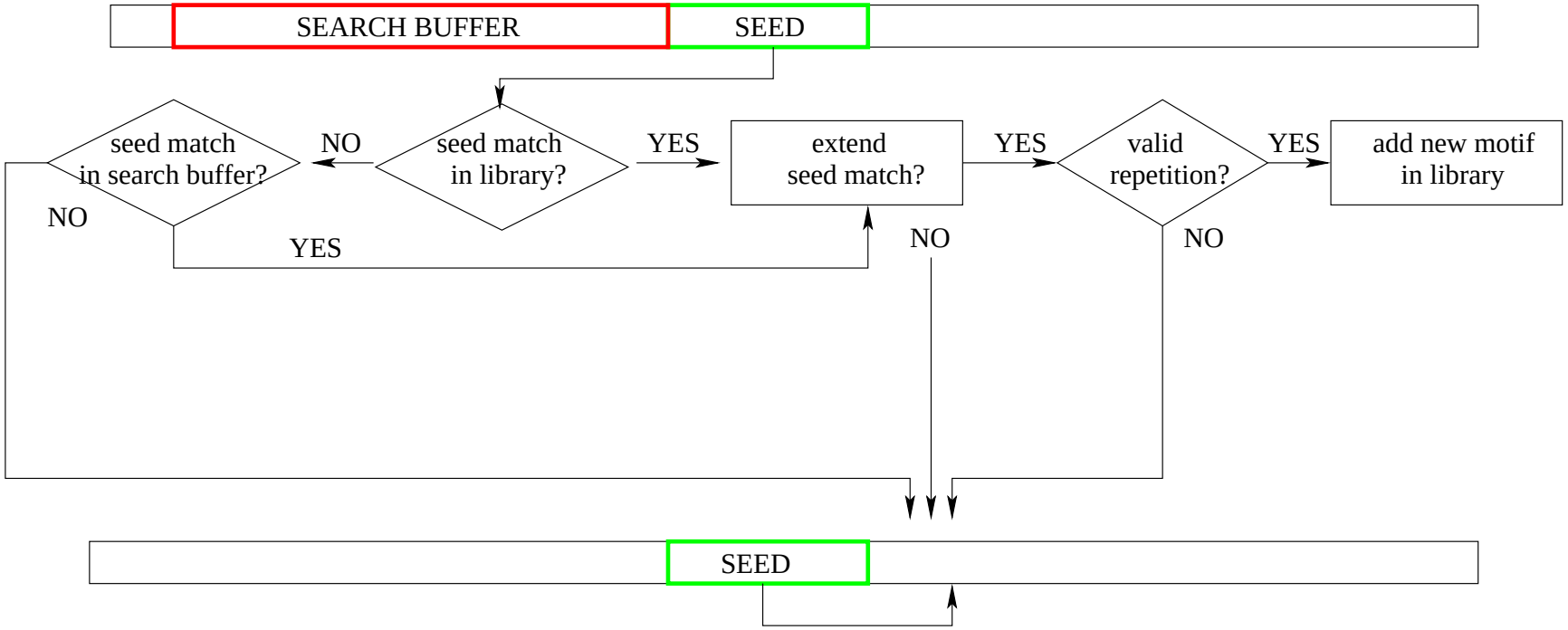
The *seed* principle

- seed = fragment of a potential motif occurence
- seed match = fragment matching the seed
- motif candidate = motif defined by two occurences resulting from the extension of a seed and its seed match
- library = library of motifs built from motif candidates verifying $H(x, y) < \epsilon$



The algorithmic of word discovery

motif discovery = build a library of motifs



Some facts:

- segmental DTW (+ self-similarity matrix comparison)
- seed length = 0.25s

Does it work?

Results on 2h of radio broadcast news [buffer size = 90s]

threshold	N	P	R	CPU time
$\epsilon = 1.2$	207	62.7%	45.1%	0:44
$\epsilon = 1.4$	777	49.8%	47.2%	1:23
$\epsilon = 1.6$	1948	39.4%	48.1%	2:38
$\epsilon = 1.8$	3866	27.5%	55.4%	3:38
$\epsilon = 2.0$	4808	17.3%	59.4%	4:18

Motifs discovered can correspond to

- words or locutions
- non speech vocal sounds
- noise

Prosody for keyword selection

Idea: extract prosodic information for motifs (or words) to select relevant ones.

- Two approaches to prosodic characterization
  1. acoustic information extraction: directly use mean/max pitch and intensity for each word
  2. prosodic event detection: apply classification methods for pitch accent detection and classification (AuToBI)
- Motifs (or words) with high pitch/intensity/accent are better

Can prosody provide meaningfull keywords?

Prosody for keyword selection

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

Showtime!

So what now?

We have

- presented a set of techniques towards the concept of transcript-free generation of audio summaries for spoken content
- shown that the goal of transcript-free summarization is reachable

But lots of things remain to be done, such as

- improve motif discovery (better distances (SSM), motif clustering, etc.)
- filter non speech motifs (supervised approach?)
- select relevant keywords and build summaries for real
- evaluate the whole thing