Dynamic Time Warping's new Youth

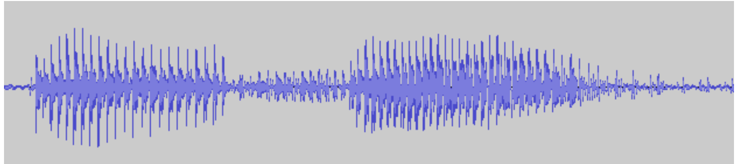
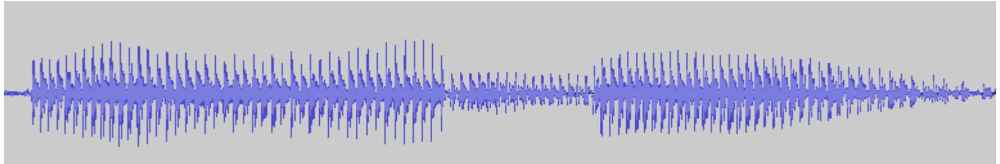
Xavier Anguera
Telefónica Research

Outline

- Was there anything before DTW?
- DTW review
- Here comes HMM
- DTW comeback
  - Algorithms
  - Speed and scalability
  - Applications
- Conclusions

Before DTW

- Speech recognition was done by means of pattern-matching input with reference patterns.
- Speaking rate variations create nonlinear fluctuations on the time axis.

- Some linear transformations were tested, but not successfully.
  - Dynamic time warping (DTW) will be able to help

Dynamic Time Warping - DTW

- DTW algorithm allows the computation of the optimal alignment between two time series X_u, Y_v ε Φ^D

$$X_U = (u_1, \ldots, u_m, \ldots, u_M)$$

$$X_V = (v_1, \ldots., v_n, .., v_N)$$

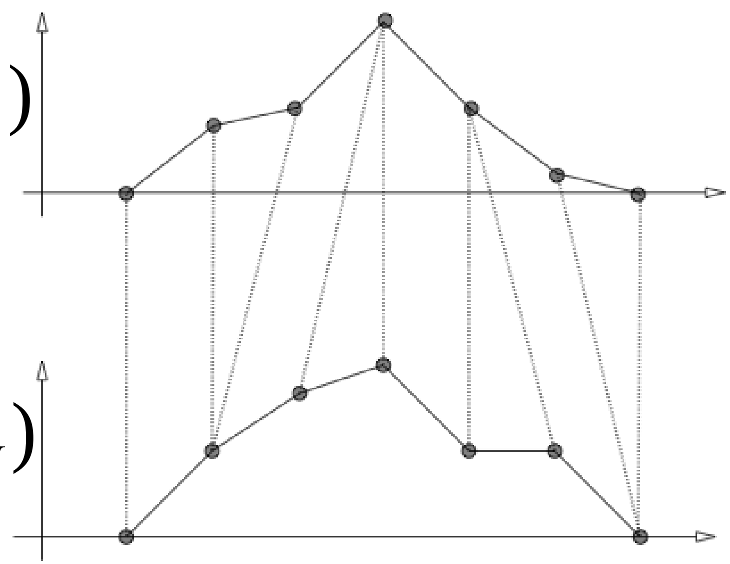


Image by Daniel Lemire

Dynamic Time Warping (II)

- The optimal alignment can be found in O(MN) complexity using dynamic programming.
- To do so, one needs to define a cost function between any two points in the series and build a distance matrix:

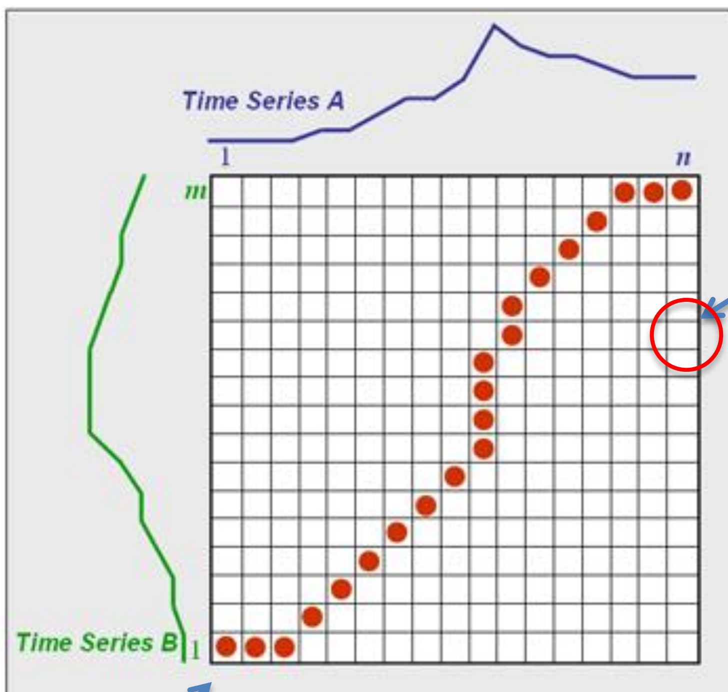


Image by Tsanko Dyustabanov

$$d : \Phi^D \times \Phi^D \rightarrow \Re \geq 0$$

Where usually:

$$d(i,j) = \left\| u_m - v_n \right\|$$

Euclidean distance

Warping function: $F = c(1),\ldots,c(K)$ where $c(i(k), j(k))$

Warping constraints

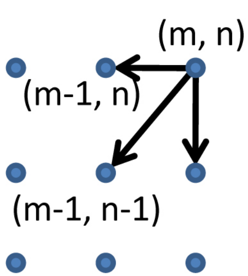
For speech signals some constraints are usually applied to the warping function F:

- Monotonicity:

$$i(k-1) \le i(k) \qquad j(k-1) \le j(k)$$

- Continuity (i.e. local constraints):

$$i(k) - i(k-1) \le 1 \qquad j(k) - j(k-1) \le 1$$



$$D(m,n) = \min \begin{cases} D(m-1,n) \\ D(m,n-1) \\ D(m-1,n-1) \end{cases} + d(u_m, v_n)$$

Sakoe,H. and Chiba,S. (1978) "**Dynamic programming algorithm optimization for spoken word recognition**", *IEEE Trans. on Acoust., Speech, and Signal Process, ASSP-26, 43-49.*

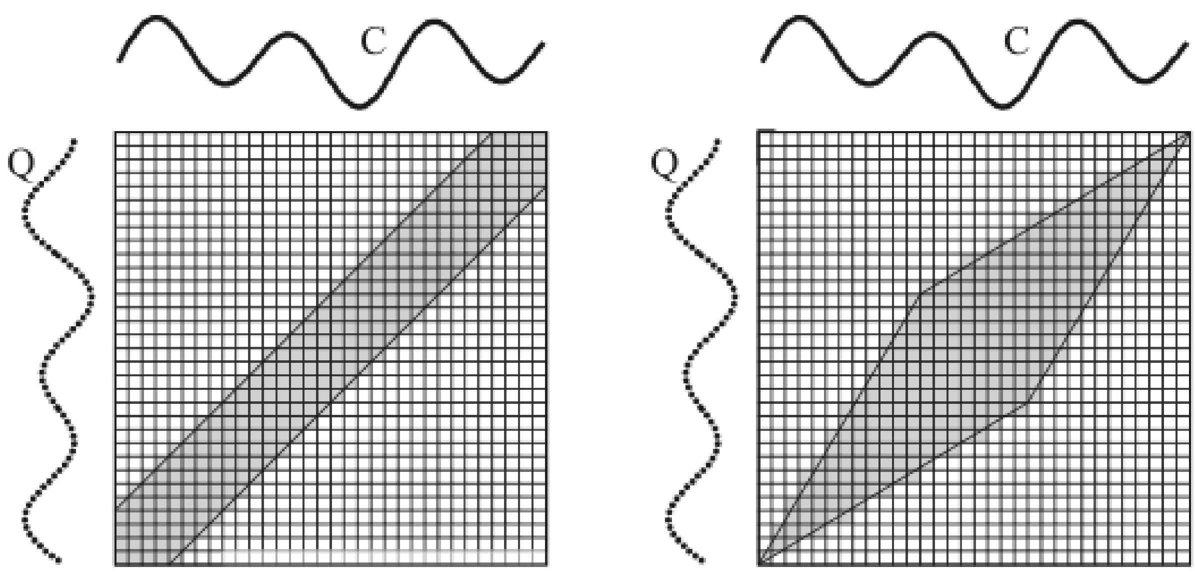
Warping constraints (II)

– Boundary condition:

$$i(1) = 1 \quad j(1) = 1 \qquad\qquad i(K) = M \quad j(K) = N$$

i.e. DTW needs prior knowledge of the start-end alignment points.

– Global constraints



Sakoe-Chiba Band   Itakura Parallelogram

Image from Keogh and Ratanamahatana

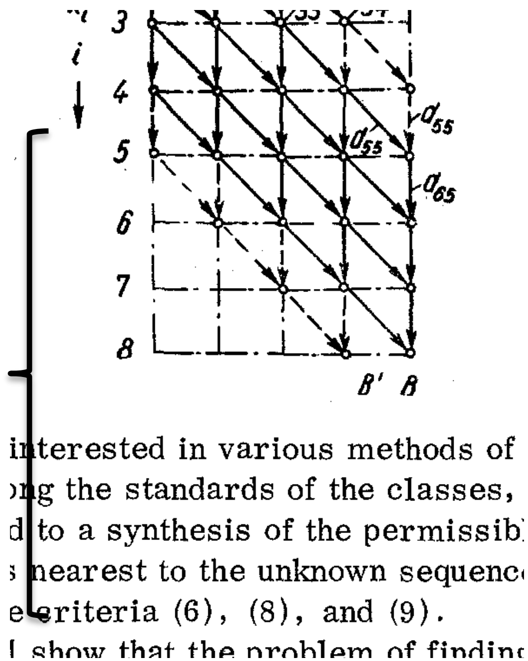
Seminal works in DTW

Hiroaki Sakoe and Seibi Chiba, “**A dynamic programming approach to continuous speech recognition**,”in 1971 Proc. 7th ICA, Paper 20 Cl3, Aug. **1971**.

Hiroaki Sakoe and Seibi Chiba, “**Dynamic Programming Algorithm Optimization for Spoken Word Recognition**”, IEEE Transactions on Audio, Speech and Signal Processing, 26(1) pp. 43-49, **1978**

Even before…

T.K.Vintsyuk, "**Speech Discrimination by Dynamic Programming**", Kibernetiks Vol. 4, No 1, pp. 81-88, **1968**.



Input query

Minimal length reference pattern

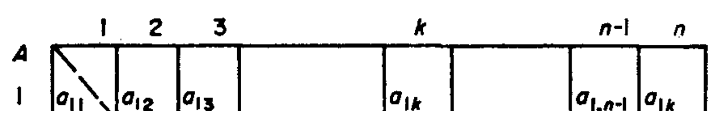
And…

V.M. Velichko and N.G. Zagoruyko, “**Automatic Recognition of 200 Words**”, Int. Journal on Man-Machine Studies, vol. 2, pp. 223-234, 1970.

matrix elements $\{a_{iK}\}$ *through* which a chosen route divided by a length of the longer word passes. In the given example with $n>m$ it will be a sum

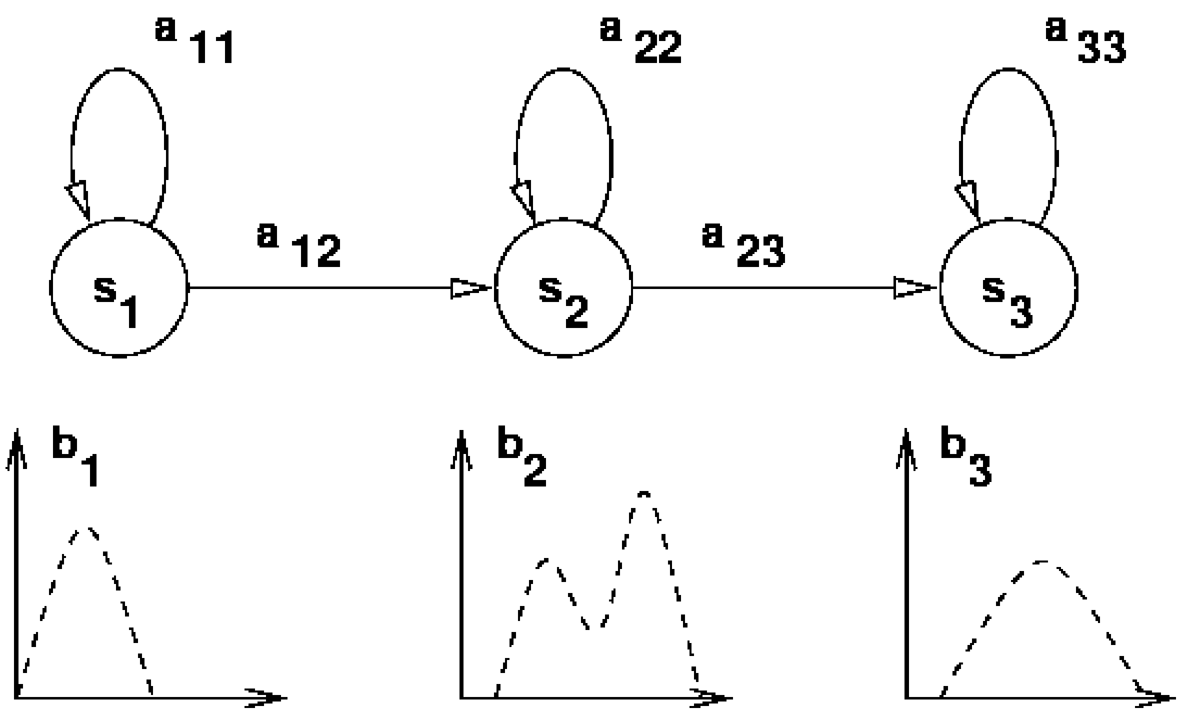
$$\frac{1}{n}(a_{11}+a_{23}+\ldots+a_{m,n-1})$$

corresponding to parts AB, CD, … QS. Path segments BC, … PQ, ST pass *between* the matrix elements and do not contribute to the common sum. The possible routes are limited: only the path segments along the horizontal from left to right, along the vertical from top to bottom, and along the diagonal from left–top to right–bottom are permitted.



Substitution by HMM's

- In the 80's, with the availability of computers with memory and possibility to better store models and statistical processing



L.R.Rabiner and B.H.Juang, "**An introduction to Hidden Markov Models**", IEEE ASSP Magazine, Vol. 3, no 1, pp. 4-16, 1986

DTW vs. ASR for Speech Recognition

Dynamic Time Warping

- Data: Some examples, no labeling needed
- Time: none for training, costly to test
- Accuracy: Mid to high

Hidden Markov Models

- Data: Lots of labeled data at phoneme level is needed
- Costly for training, light for testing
- Accuracy: high

DTW has long been abandoned in ASR for high-resourced languages, with one exception:

De Wachter et al., "**Template-based continuous speech recognition**", IEEE Trans. On Audio, Speech and Language Processing, 15(4) pp. 1377-1390

The Zero-Resource Setting

1. **No** transcribed training data
2. **No** dictionaries
3. **No** knowledge of linguistic structure

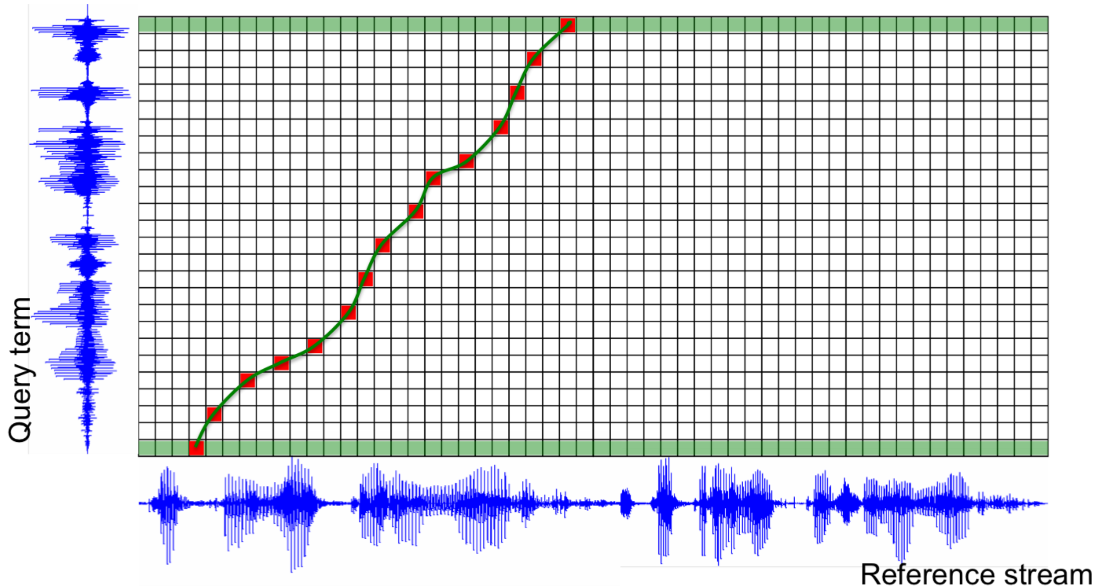
<u>But:</u> Assume you have untranscribed speech
(at least your test data)

Challenge: How can we discover linguistic structure and build useful applications and systems without much supervision?

Calling DTW back

PRO: DTW works with patterns, no need for costly transcriptions or knowledge of the language.

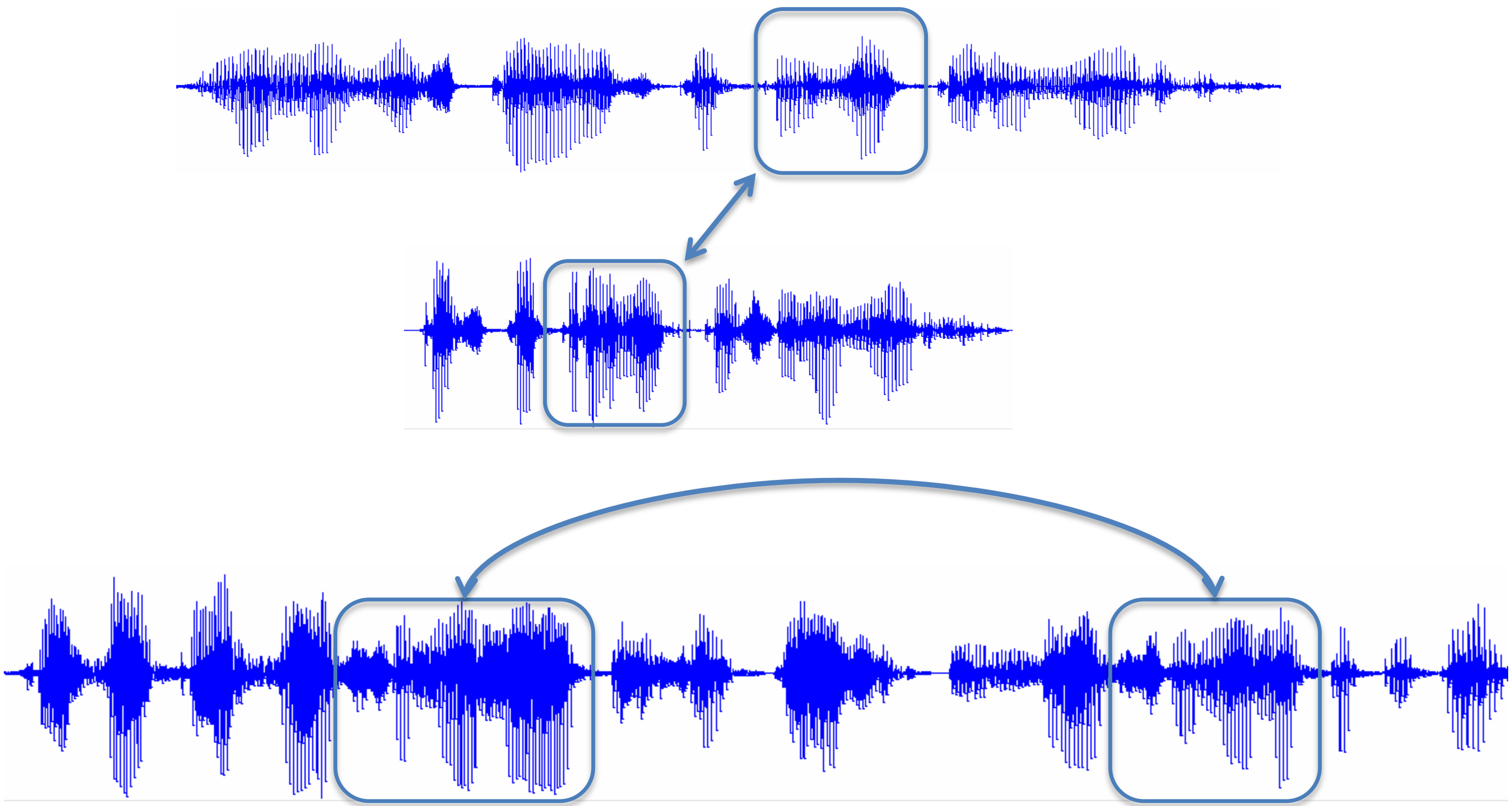
CON: DTW compares patterns given a known start-end position -> needs that at least one of the patterns be well bounded



From DTW to subsequence matching

Given several instances of an acoustic sequence,

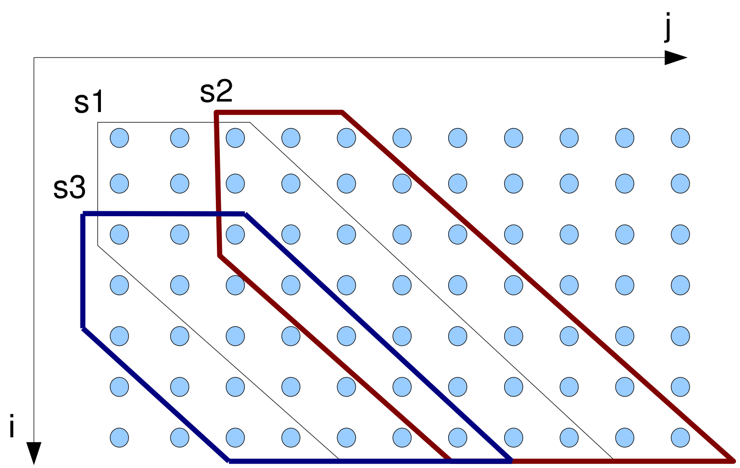
Can we modify DTW to find them?

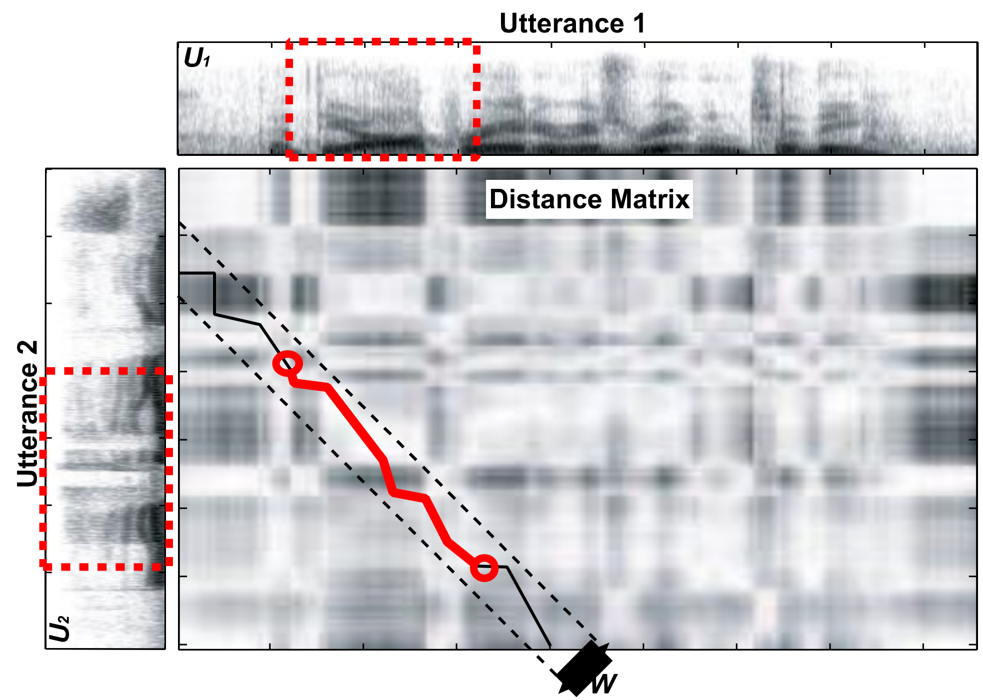


DTW-based subsequence matching

- **Segmental-DTW** by James Glass et al. at MIT
- **“image-based” DTW** by Aren Jansen et al. at John Hopkins Univ.
- **Motif Discovery** by Guillaume Gravier et al. at IRISA (Rennes).
- **DTW for music** by Meinard Müller et al. at Max Planck Institut.
- **Unbounded-DTW** by Xavier Anguera et al. at Telefonica research

Segmental DTW

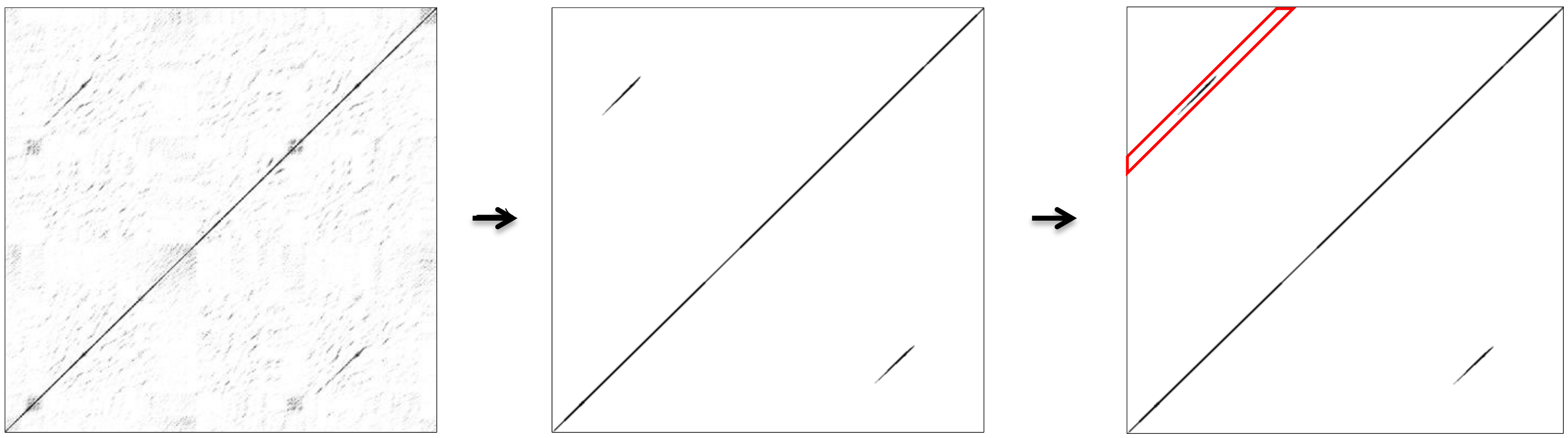




A. Park and J. Glass, "**Unsupervised Pattern Discovery in Speech**", IEEE Trans. On Audio, Speech and Language Processing, 2008

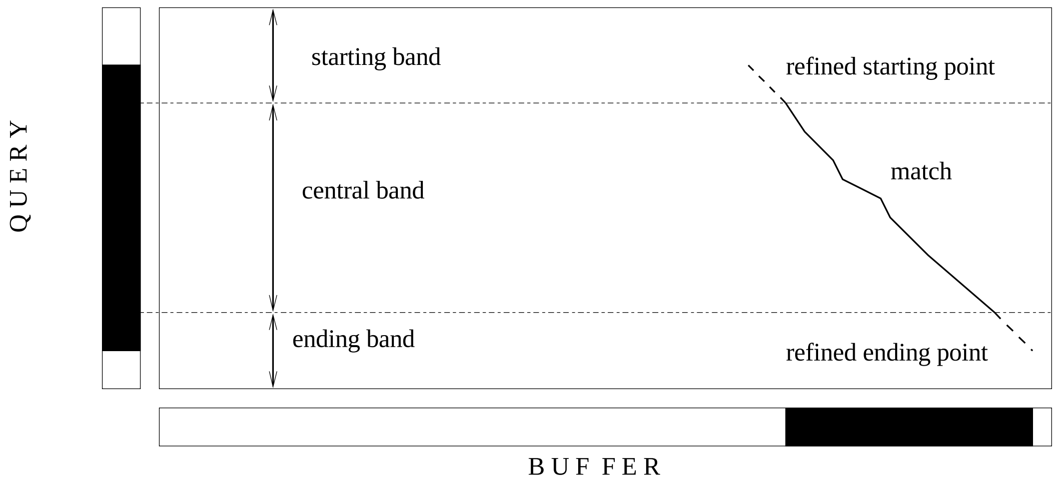
"Image-based" DTW

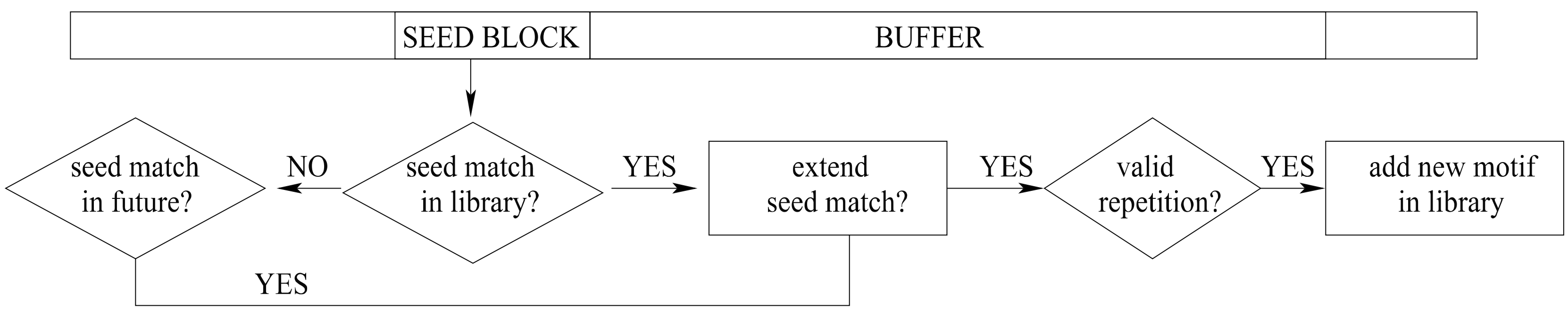
Image processing can be used to speedup the discovery of possible matches in the similarity matrix



A. Jansen et al., "**Towards spoken term discovery at scale with zero resources**", Interspeech 2010

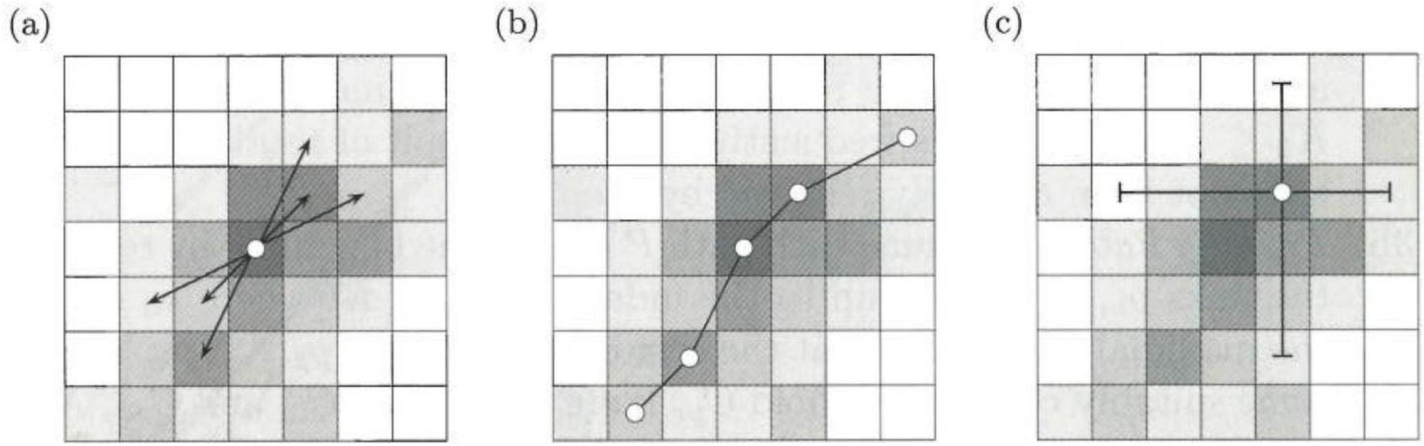
Automatic motif discovery

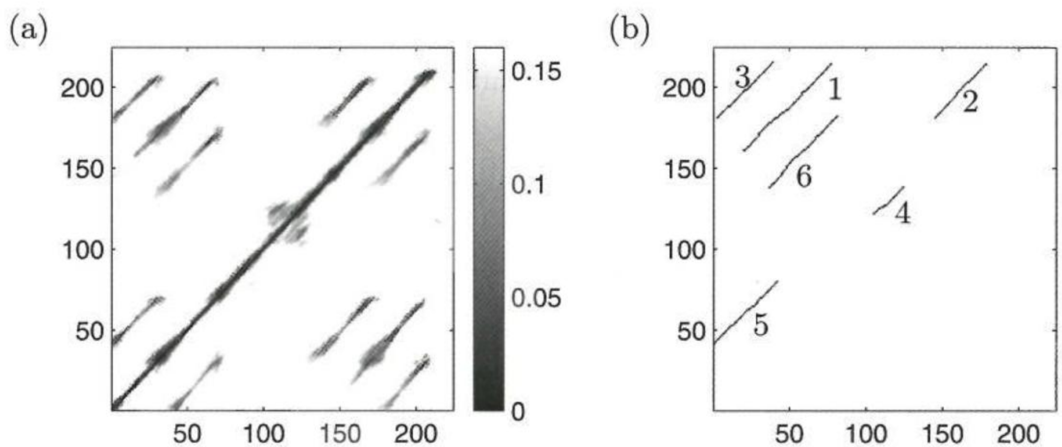




A. Muscariello et al., “**Towards Robust Word Discovery by Self-Similarity Matrix Comparison**”, Proc. ICASSP 2011

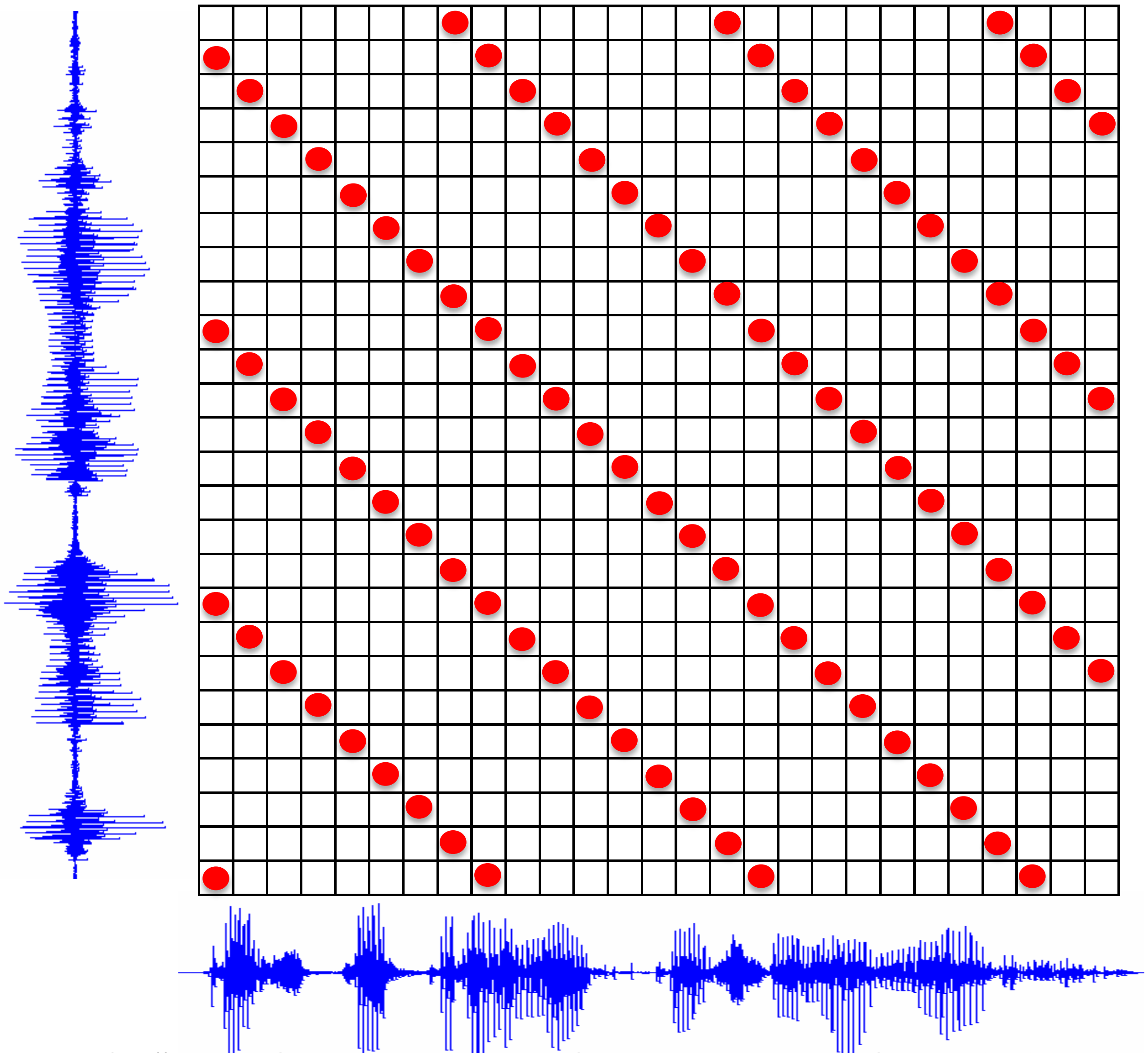
Music structure analysis





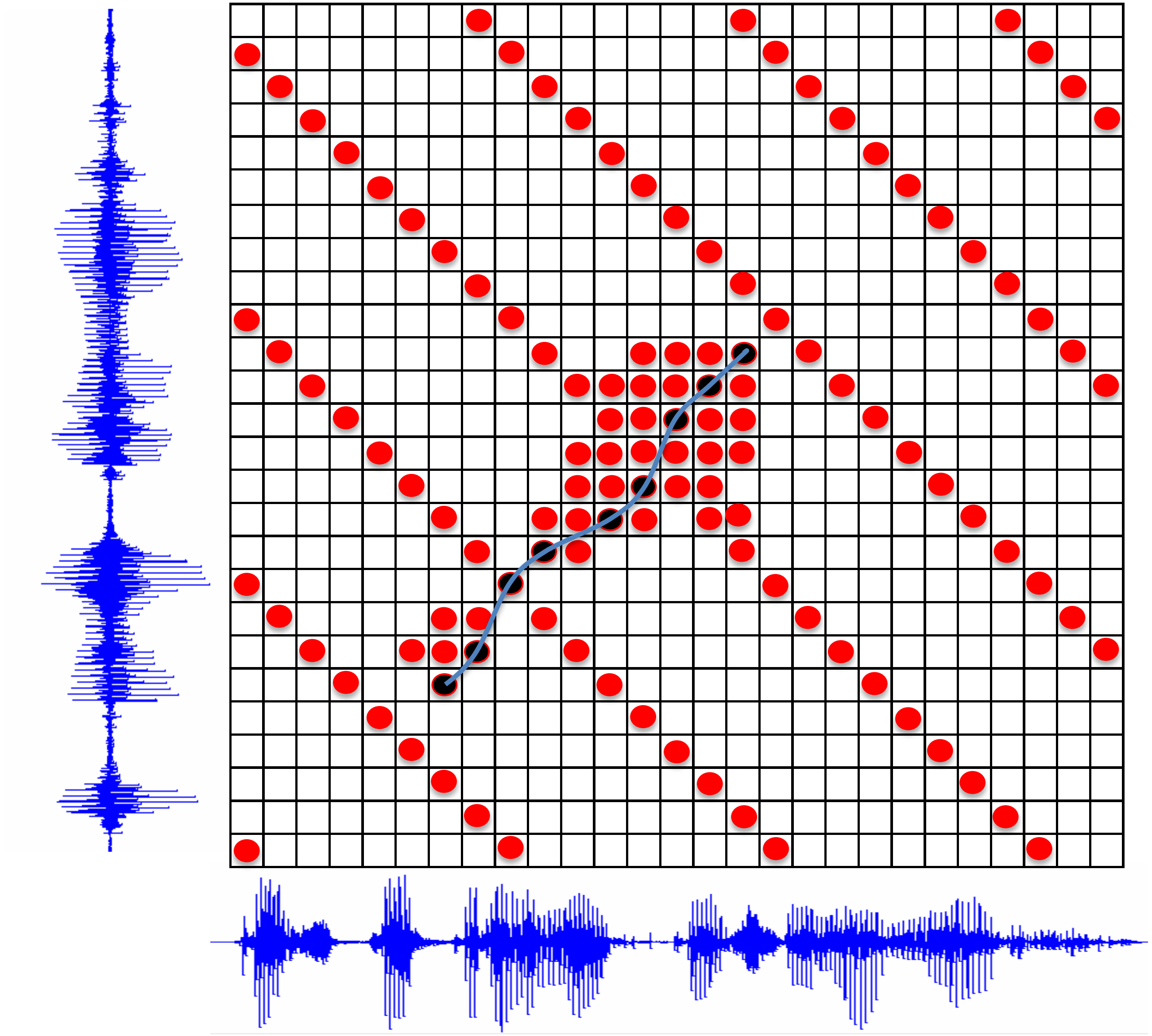
Meinard Müller, "**Information Retrieval for Music and Motion**", Springer-Verlag, ISBN 978-3-540-74047-6, pp. 147-150, 2010

Unbounded-DTW



X. Anguera et al., "**Partial Sequence Matching using an Unbounded Dynamic Time Warping Algorithm**", ICASSP 2010

Unbounded-DTW



The search for speed and scalability

- Coarse-to-fine approximation of DTW
  - S. Salvador and P. Chan, "**FastDTW: Toward accurate dynamic time warping in linear time and space**". 3rd Wkshp. On Mining Temporal and Sequential Data, ACM KDD 2004
- Intelligent bounding of DTW
  - Work by Eamon Keogh (pure DTW, no subsequences)
  - Y. Zhang and J. Glass, "**A Piecewise Aggregate Approximation Lower-Bound Estimate for Posteriorgram-based Dynamic Time Warping**", Interspeech 2011
- Use of IR techniques
  - A. Jansen et al., "**Efficient Spoken Term Discovery using Randomized Algorithms**", ASRU 2011

(some) applications of the technology

- Finding structure in an unknown language -> Zero-resources approaches (JHU Workshop this summer)
  - Helping ASR by increase of training data (Jansen_2011)
- Acoustic documents comparison and topic detection
  - NLP on speech (Drezde_2010)
- Query-by-example search (Metze_2011)
- Spoken term discovery (Muscariello_2011)
- Spoken Summarization (Jansen_2010, Flamary_2011)
- Acoustic indexing enhancement via transcription propagation

Conclusion

- The old DTW is back
  - It will not reclaim ASR, but it takes on new challenges
- Lots of research to be done
  - On scalability (matching patterns is still expensive)
  - On generality
  - Robustness

References

- T.K.Vintsyuk, “Speech Discrimination by Dynamic Programming”, Kibernetiks Vol. 4, No 1, pp. 81-88, 1968.

- V.M. Velichko and N.G. Zagoruyko, “**Automatic Recognition of 200 Words**”, Int. Journal on Man-Machine Studies, vol. 2, pp. 223-234, 1970.

- H. Sakoe and S. Chiba, “**A dynamic programming approach to continuous speech recognition**,”in 1971 Proc. 7th ICA, Paper 20 CI3, Aug. 1971.

- H. Sakoe and S. Chiba, “**Dynamic Programming Algorithm Optimization for Spoken Word Recognition**”, IEEE Transactions on Audio, Speech and Signal Processing, 26(1) pp. 43-49, 1978

- S. Salvador and P. Chan, “**FastDTW: Toward accurate dynamic time warping in linear time and space**”. 3rd Wkshp. On Mining Temporal and Sequential Data, ACM KDD 2004

- De Wachter et al., “**Template-based continuous speech recognition**”, IEEE Trans. On Audio, Speech and Language Processing, 15(4) pp. 1377-1390, 2007

- A. Park and J. Glass, “**Unsupervised Pattern Discovery in Speech**”, IEEE Trans. On Audio, Speech and Language Processing, 2008

References (II)

- Meinard Müller, “**Information Retrieval for Music and Motion**”, Springer-Verlag, ISBN 978-3-540-74047-6, pp. 147-150, 2010

- X. Anguera et al., “**Partial Sequence Matching using an Unbounded Dynamic Time Warping Algorithm**”, ICASSP 2010

- A. Jansen et al., “**Towards spoken term discovery at scale with zero resources**”, Interspeech 2010

- R. Flamary et al.,“**SpokenWordCloud: Clustering Recurrent Patterns in Speech**” in Proc. CBMI 2011

- M. Drezde et al. “**NLP on Spoken Documents without ASR**”, in Proc. EMNLP 2010

- A. Muscariello et al., “**Towards Robust Word Discovery by Self-Similarity Matrix Comparison**”, Proc. ICASSP 2011

- A. Jansen et al., “**Efficient Spoken Term Discovery using Randomized Algorithms**”, ASRU 2011

- Y. Zhang and J. Glass, “**A Piecewise Aggregate Approximation Lower-Bound Estimate for Posteriorgram-based Dynamic TimeWarping**”, Interspeech 2011

- A. Jansen et al., “**Towards unsupervised training of speaker independent acoustic models**,” in Proc. Interspeech, 2011

- F. Metze, “The spoken web search task at Mediaeval 2011”, ICASSP 2011