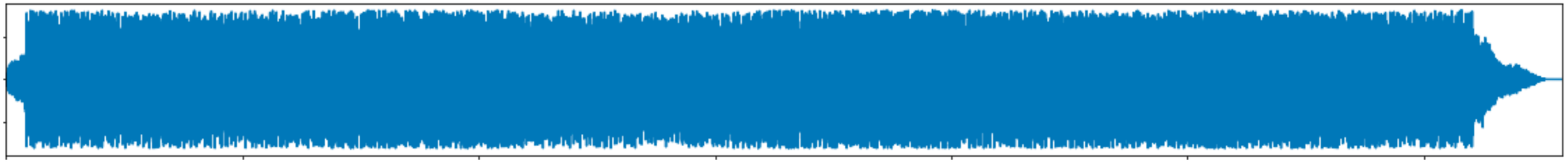
TL;DR - Duration model is very useful for improving boundary detection, for expressive models >2-gram (incl. LSTM)

Music boundary detection
based on a hybrid deep model of
novelty, homogeneity, repetition and **duration**

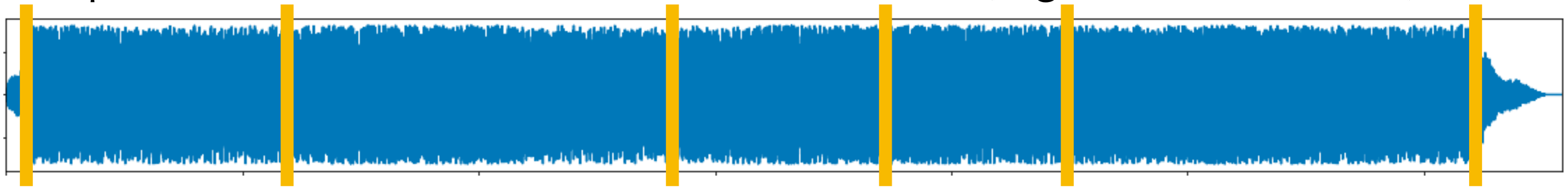
Akira Maezawa
Yamaha Corporation

Introduction
Problem setting

Input - Music audio signal



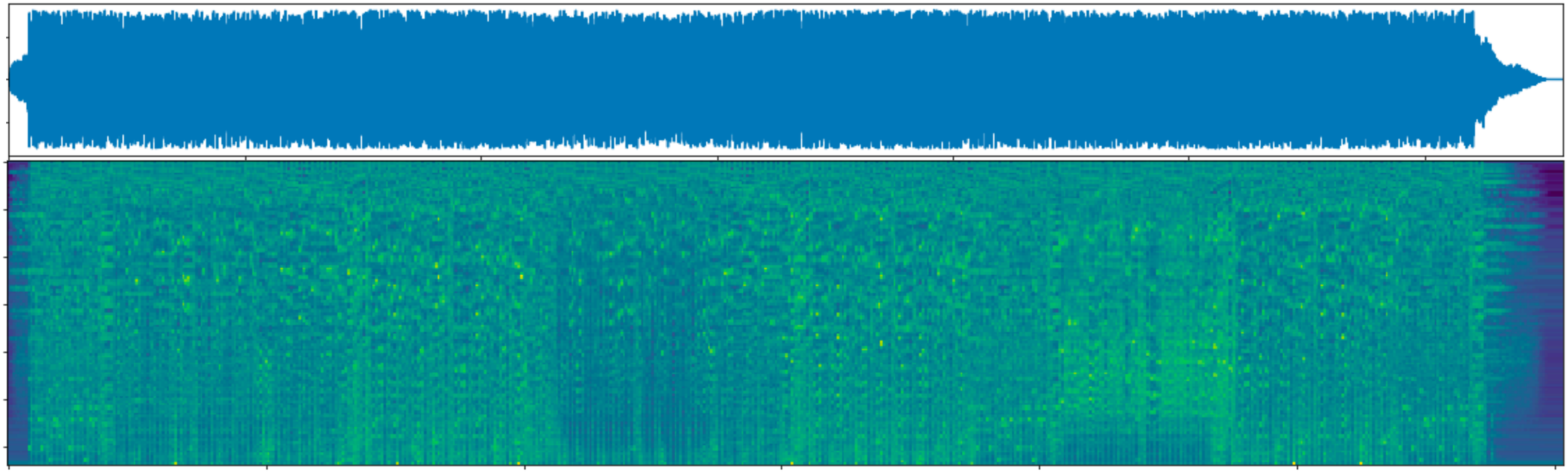
Output - Locations of structural boundaries (e.g., Verse, Chorus)



Approaches

[Paulus+2010]

Repetition **Homogeneity** **Novelty** **Duration**

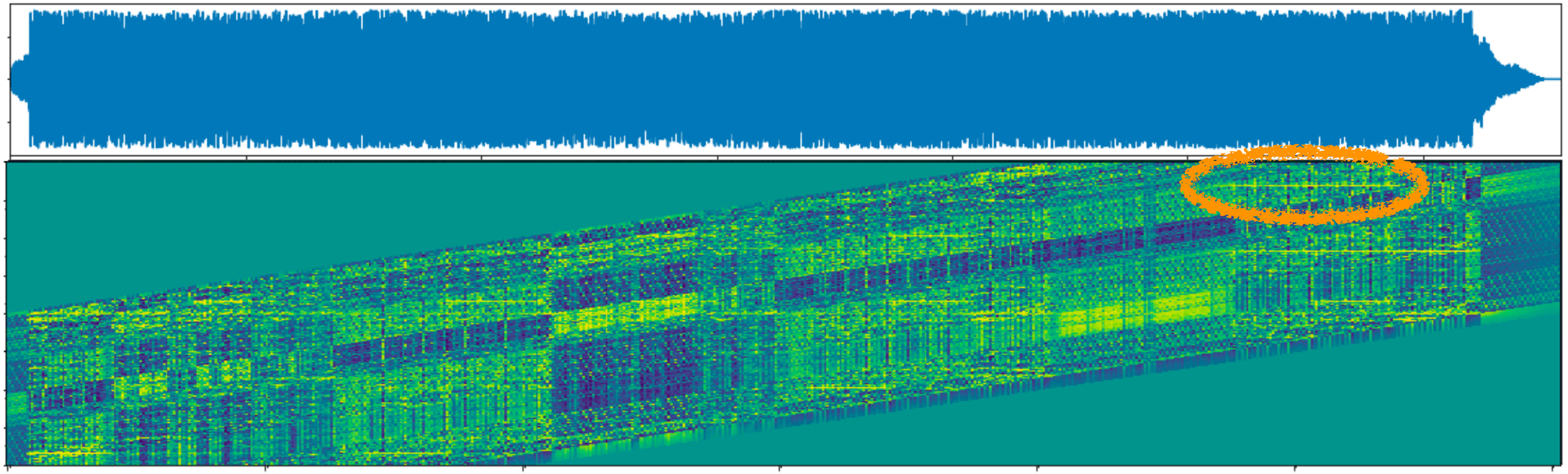


MSLS (Mel-scale log spectrogram)

Approaches

[Paulus+2010]
[Goto2003]

Repetition　Homogeneity　Novelty　Duration

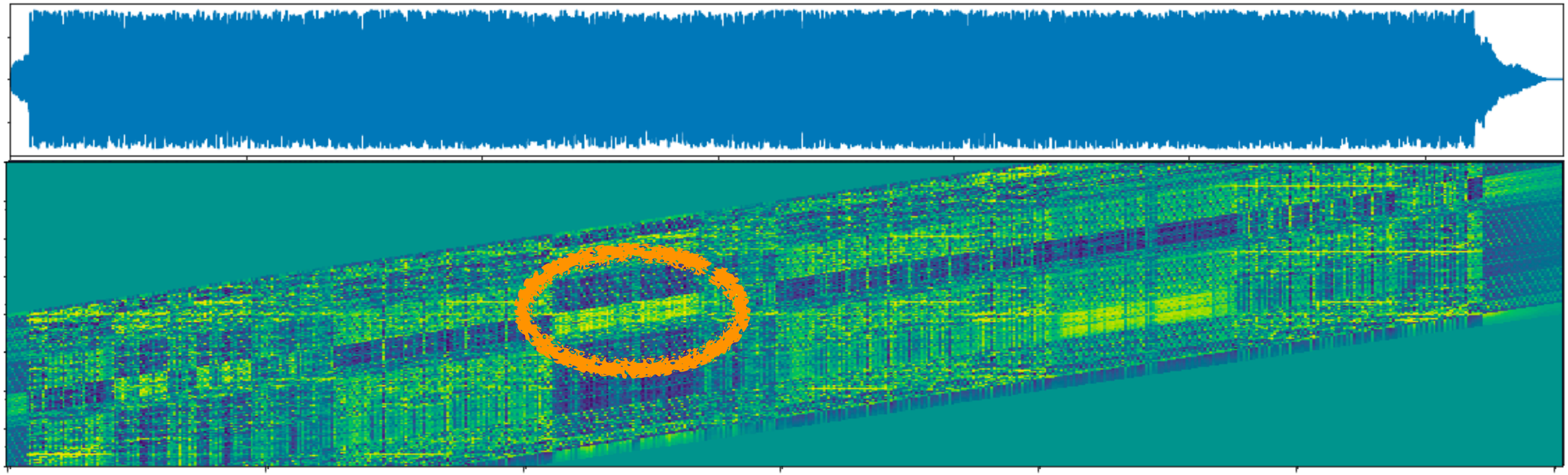


SSM (Self-similarity matrix)

Approaches

[Paulus+2010]
[Cooper+2003]

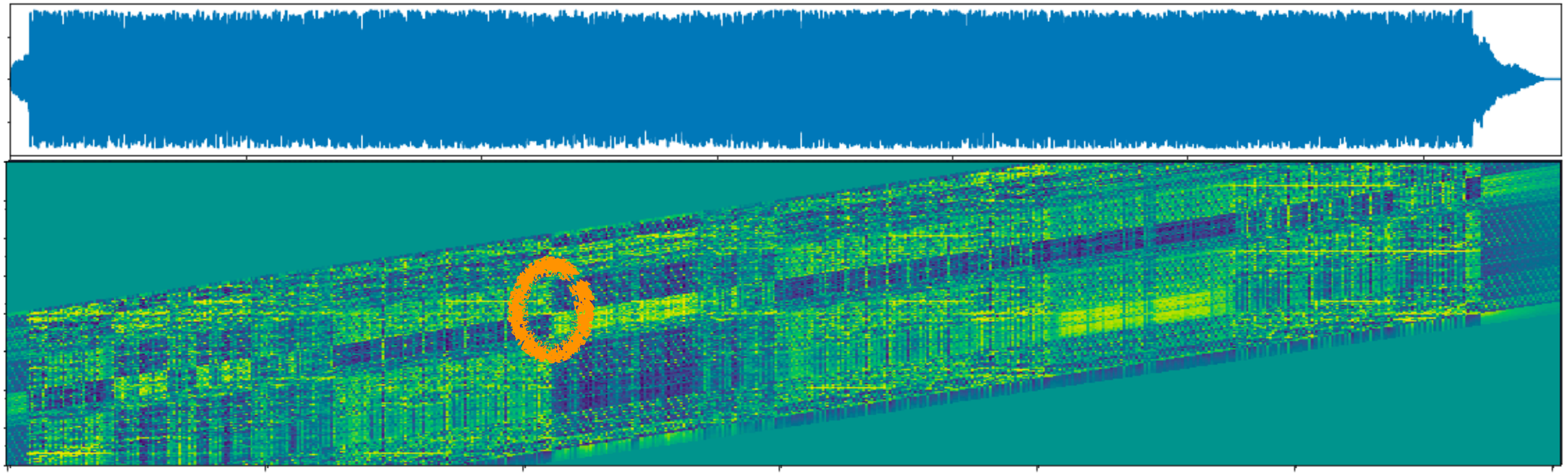
Repetition　Homogeneity　Novelty　Duration



Approaches

[Paulus+2010]
[Foote2000]

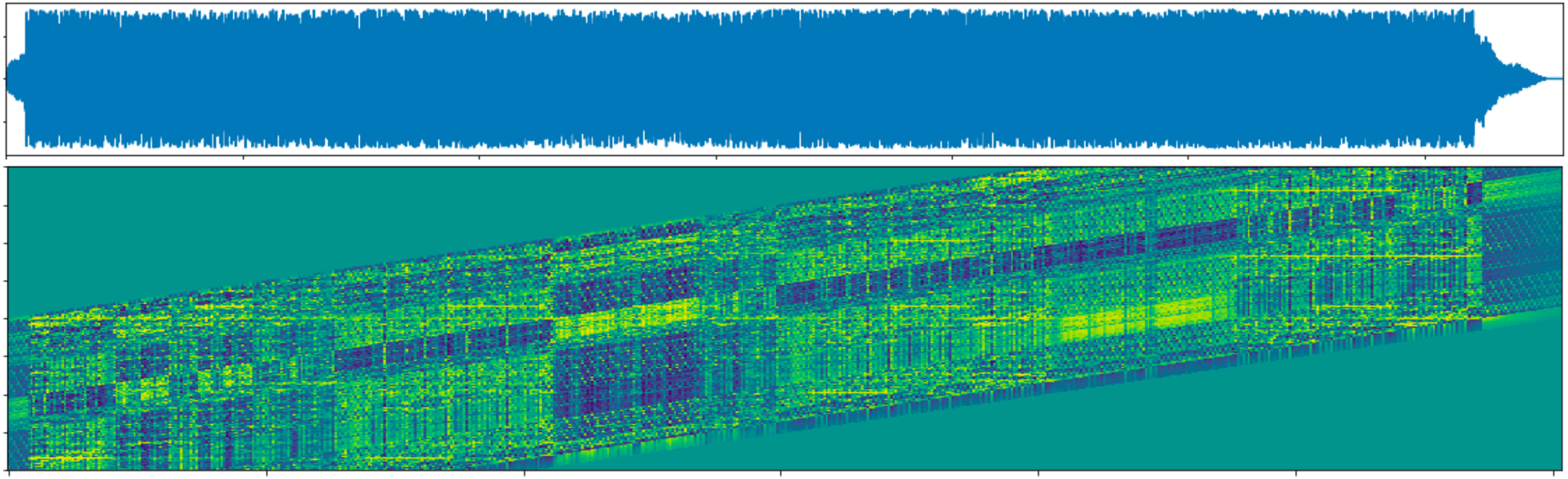
Repetition　Homogeneity　Novelty　Duration



Approaches

[Paulus+2010]
[Cheng+2018]
[Paulus+2009]

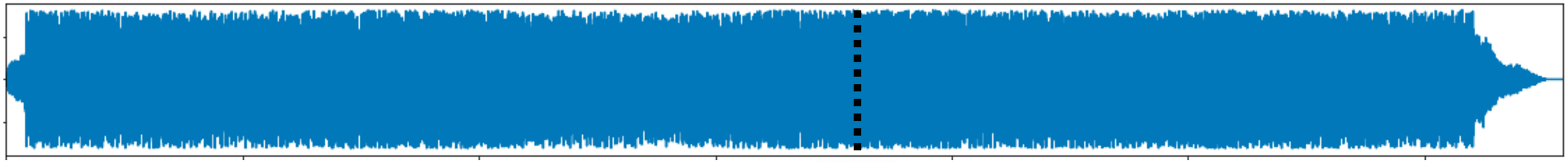
Repetition　Homogeneity　Novelty　Duration

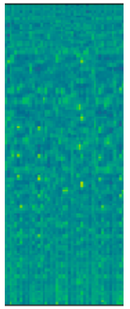


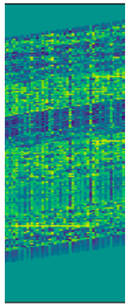
Approaches

[Ullrich+ 2014]
[Grill+ 2015]

(Repetition) (Homogeneity) Novelty Duration



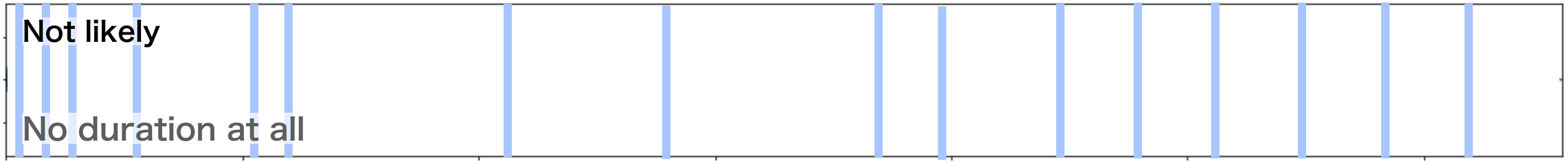


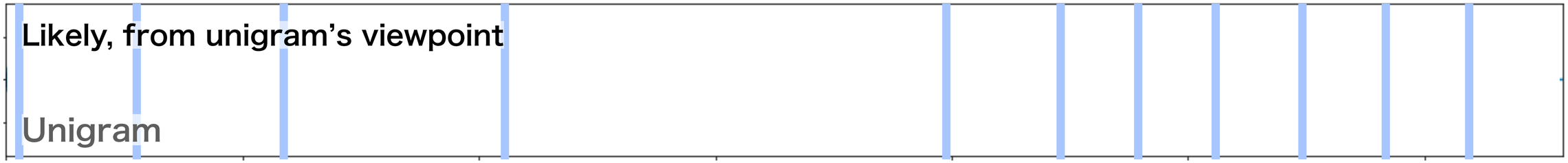


Approaches

[Levy+ 2006]
[Smith+2016]

Repetition Homogeneity Novelty Duration

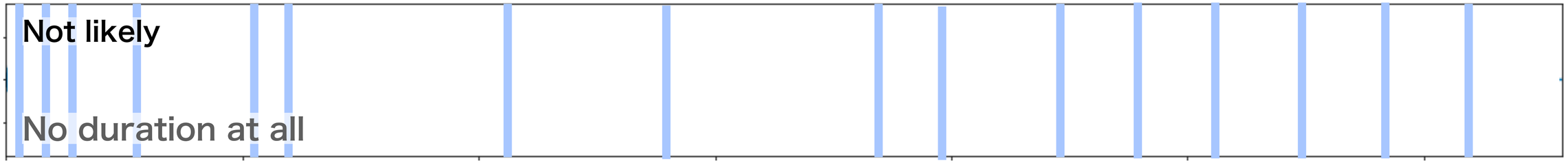


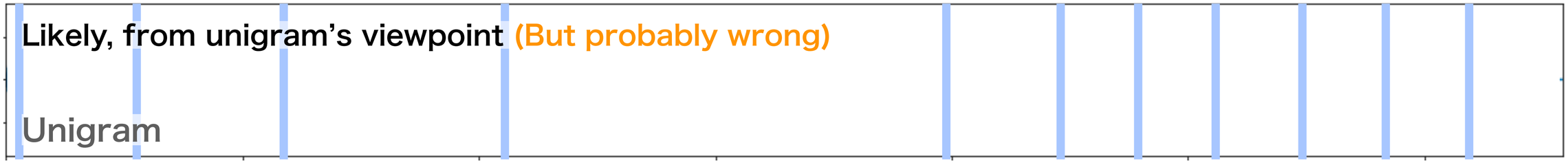


Approaches

[Levy+ 2006]
[Smith+2016]

Repetition Homogeneity Novelty Duration

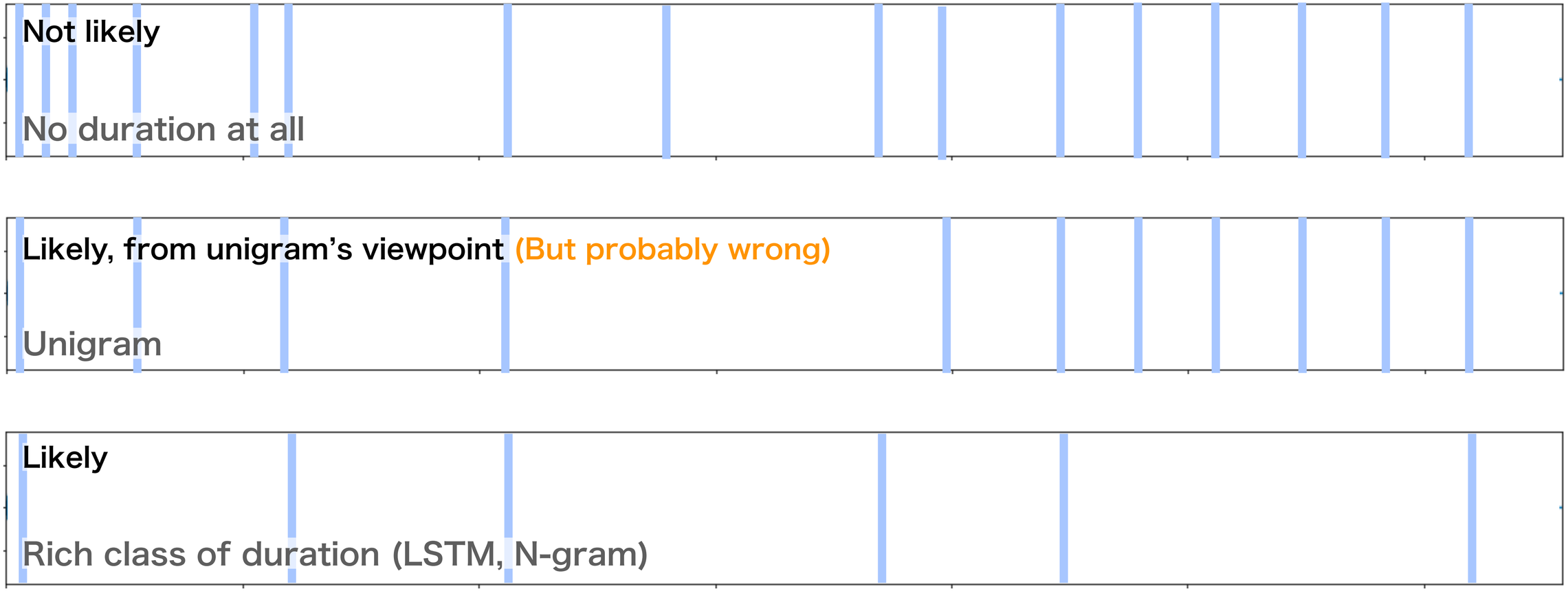




Approaches

[Levy+ 2006]
[Smith+2016]

Repetition Homogeneity Novelty **Duration**

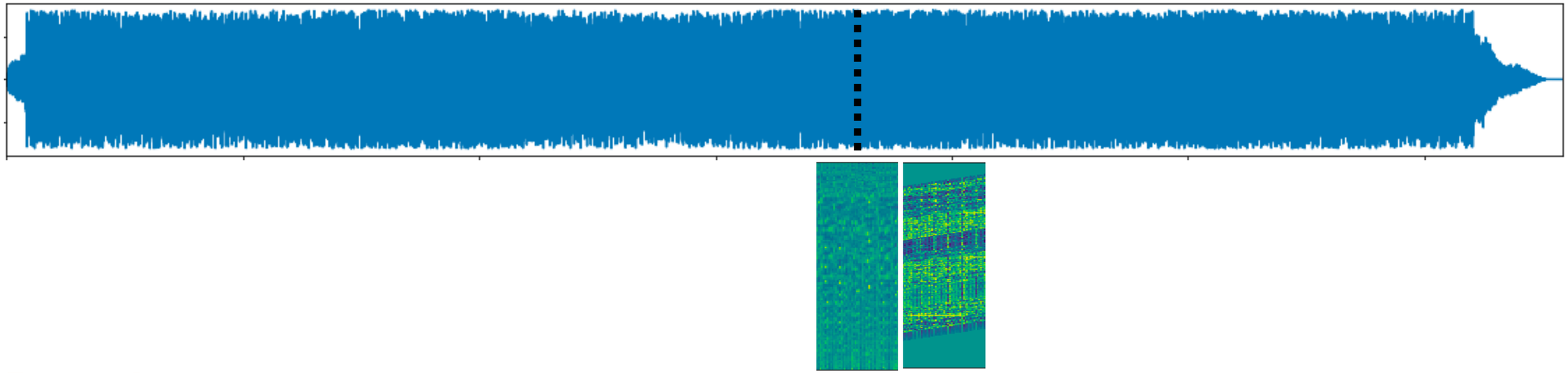


Main Contribution: Enabling incorporation of elaborate duration models

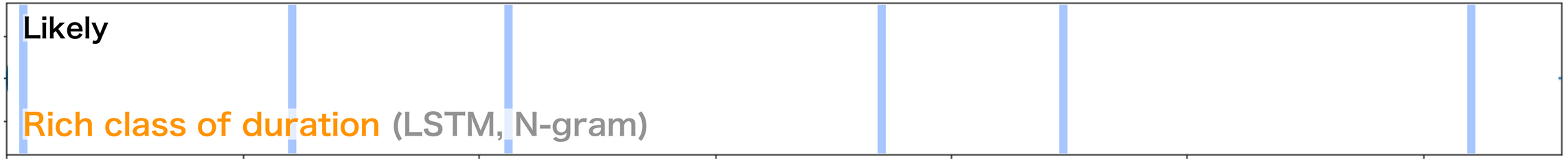
Our Approach

[Ullrich+ 2014]
[Grill+ 2015]

(Repetition) Homogeneity Novelty



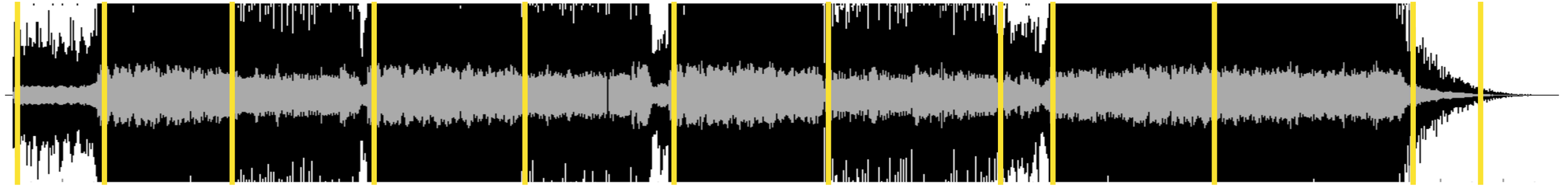
Duration



An example

DNN boundary

DNN boundary + LSTM duration + homogeneity

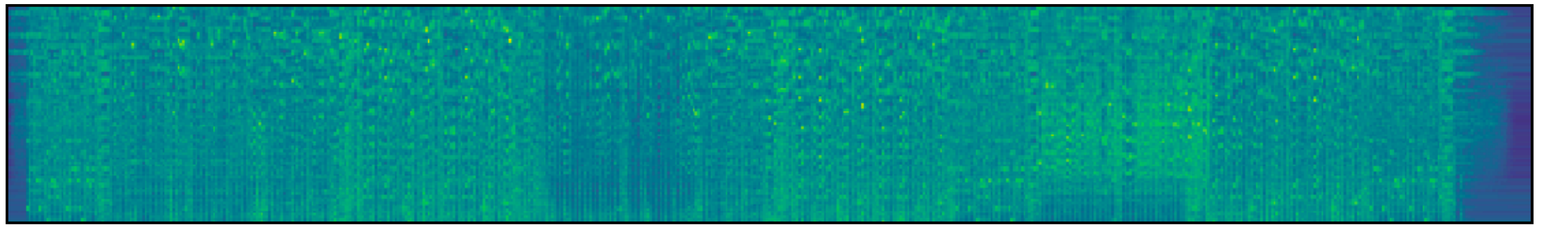


Our method

Overview

MSLS

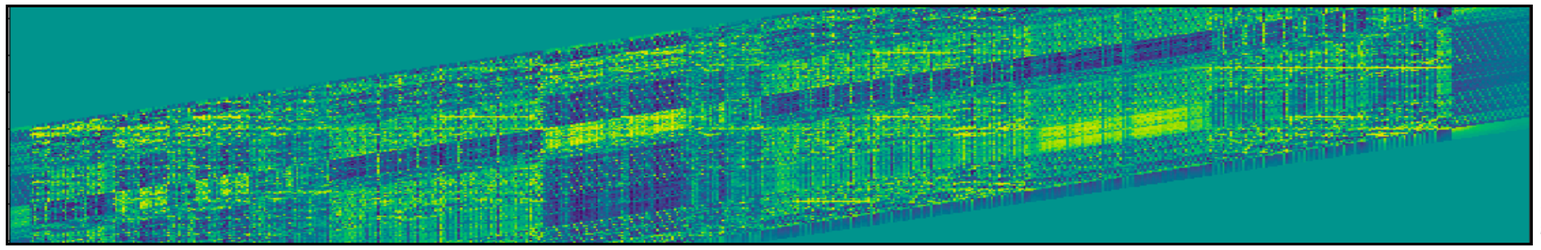
sliced at
8th-note level



128 dims
0 to 11kHz

SSM (MSLS)

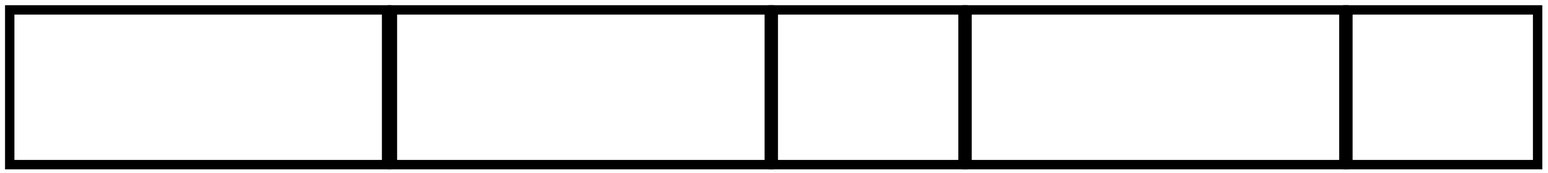
sliced at
8th-note level



lead/lag of
200 beats

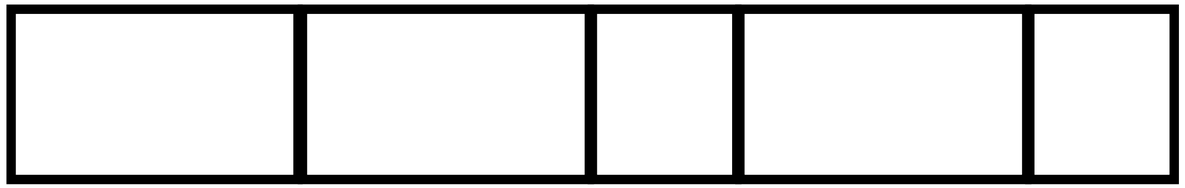
B

Boundary positions
in beats



Our method
Overview

B
Boundary
positions in beats



$$B = \arg\max_{\hat{B}} f_B(\hat{B}) + \alpha f_D(\hat{B}) + \beta f_H(\hat{B})$$

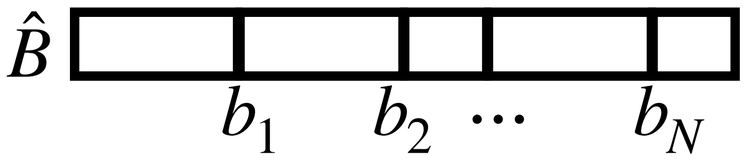
$f_B(\hat{B})$ **Boundary fitness**

$f_D(\hat{B})$ **Segment duration fitness**

$f_H(\hat{B})$ **Timbre homogeneity fitness**

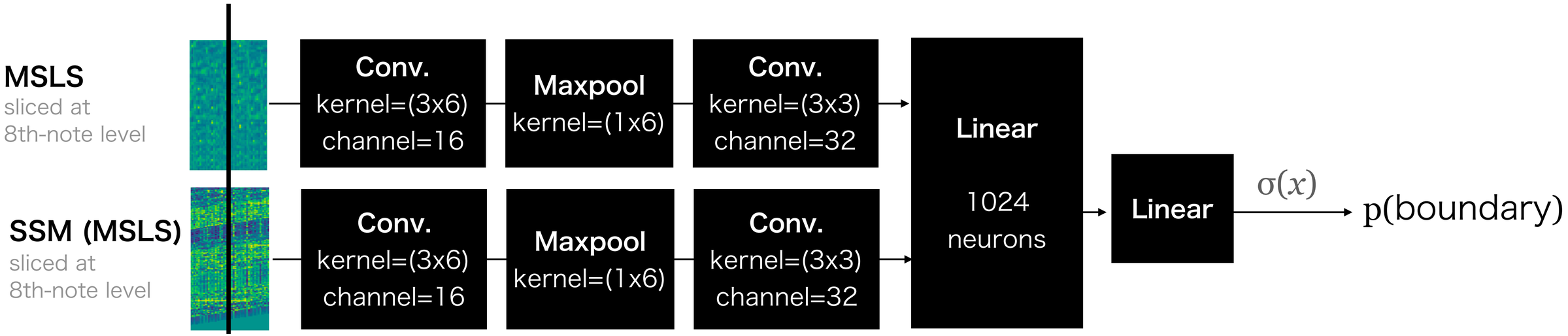
Find B using beam-search

Our method
Boundary fitness

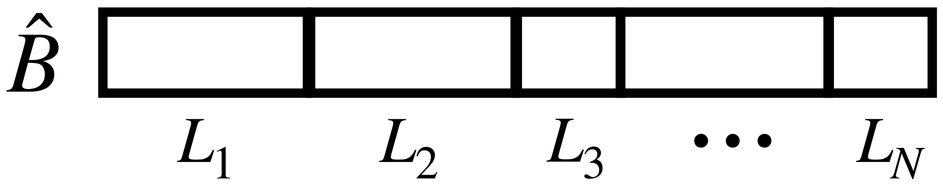


$f_B(\hat{B})$ = log-probability of boundary at $\{b_1, \cdots, b_N\}$
- log-probability of **non**-boundary at $\{b_1, \cdots, b_N\}$

Beat b_n+/-16 beats



Our method
Duration fitness



$$f_D(\hat{B}) = \log\left(\prod_i p(L_i \mid L_1 \cdots L_{i-1})\right)$$

n-gram

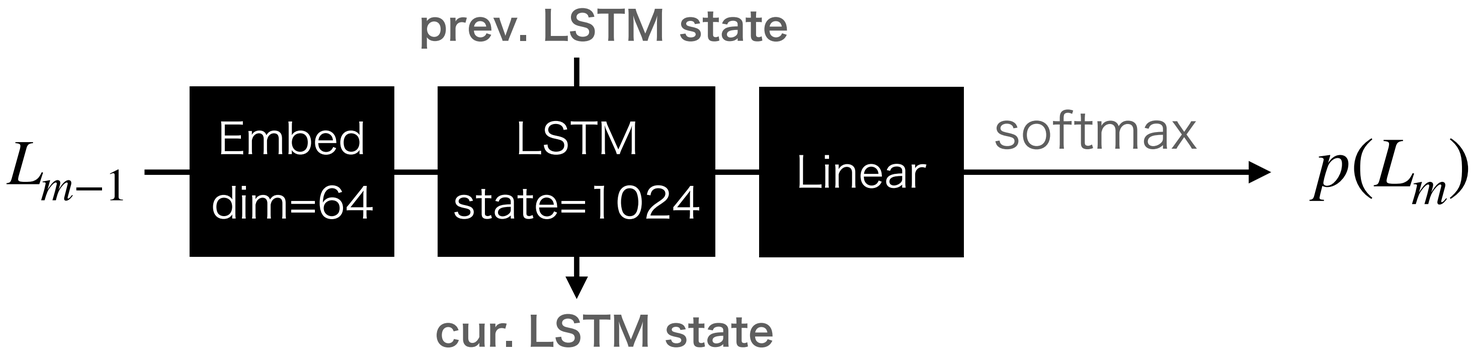
$$p(L_m \mid L_1 \ldots L_{m-1})$$

$$= p(L_m \mid L_{m-n+1} \ldots L_{m-1})$$

(n+1)-th order Markov assumption

LSTM

$$p(L_m \mid L_1 \ldots L_{m-1}) = \mathrm{LSTM}(L_1 \cdots L_{m-1})$$



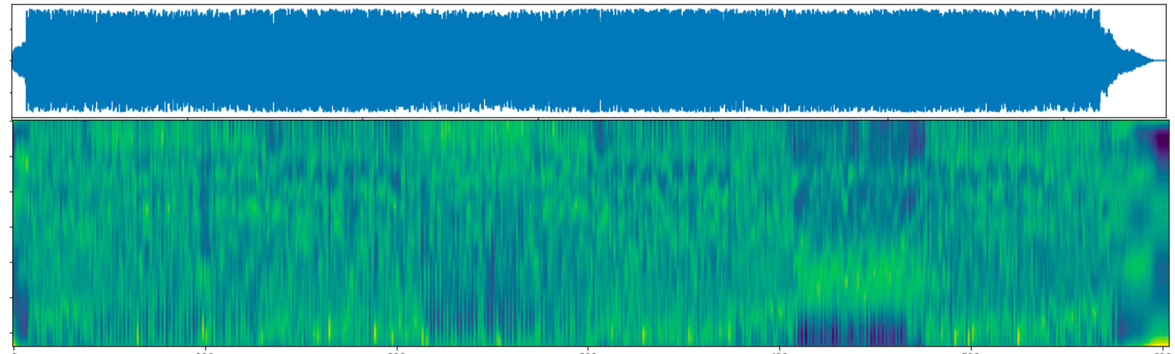
(Edge cases L_1 and L_N are treated differently)

Our method

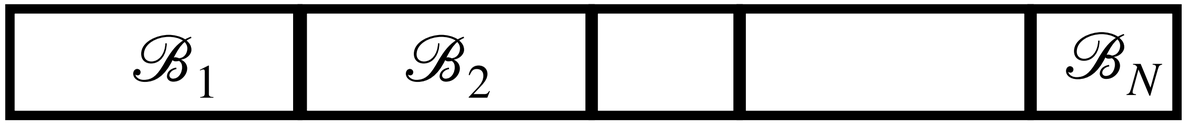
Timbre homogeneity fitness

$X(t)$

MSLS **smoothed**
by a Hanning window



$\hat{B}$



$$f_H(\hat{B}) = -\sum_n \mathrm{Tr}(\mathrm{Cov}(X(\mathscr{B}_n))$$

n Variance inside segment n,
summed over dimensions

Evaluation
Experimental Conditions

- **Training data**
  - 410 songs from JP + US hit-charts, with in-house labels
  - In-house 7700 MIDI data with structural annotations

- **Validation data**
  - First album of the Beatles w/ Isophonics label [Mauch+2009]

- **Test data**
  - RWC Popular [Goto+2002]
  - SALAMI [Smith+2011]
    - Use only Internet Archives (more degraded compared to commercial audio)
  - Beatles w/ Isophonics label (all BUT the first album) [Mauch+2009]

Evaluation
Training details

1. **Train each component individually**
   - Boundary fitness trained on real audio + synthesized MIDI
   - Duration fitness trained on MIDI
2. **Optimize α and β using Bayesian Optimization**

Evaluation
Experiments

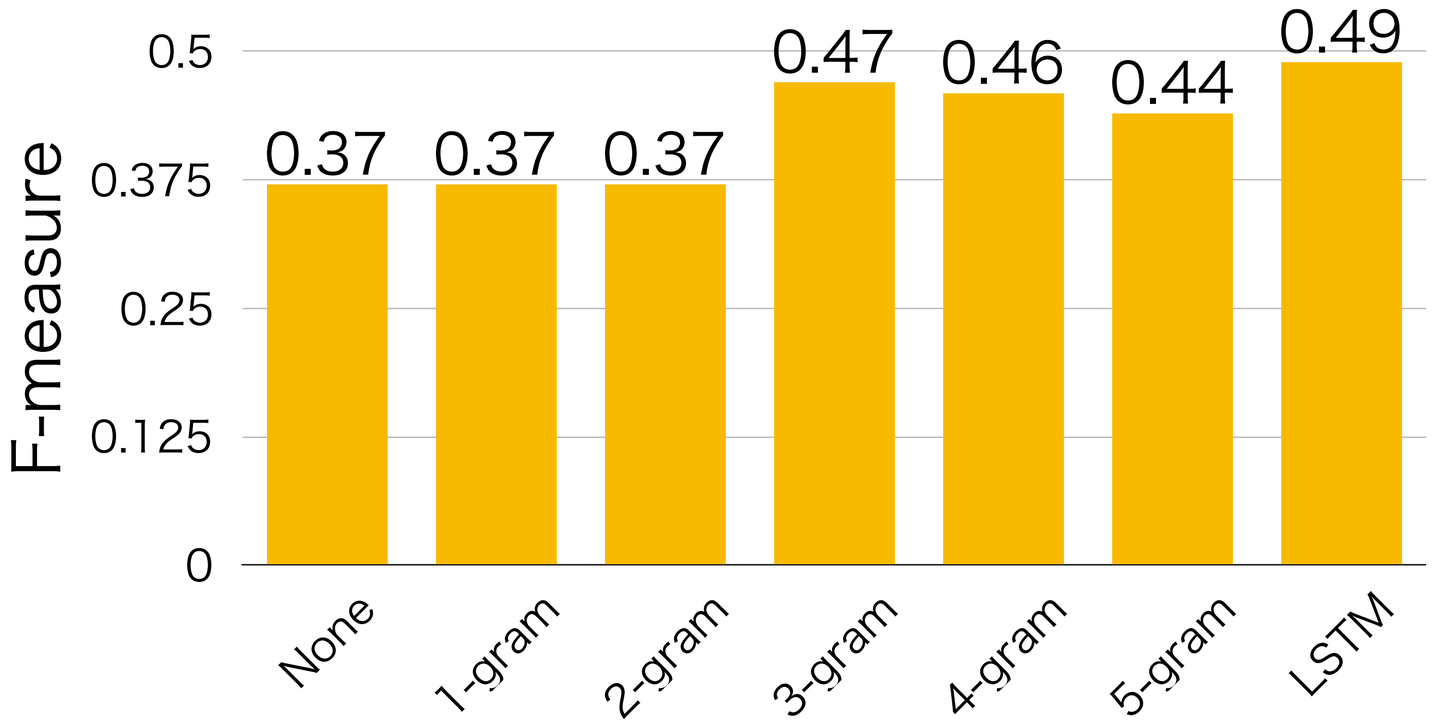
1. **Does duration model help?**
   1. Train boundary fitness model, and various duration models
      - 1-gram, 2-gram … 5-gram
        - Katz backoff (k=1); Unseen unigram duration assigned log-likelihood of -100
      - LSTM
   2. Evaluate the F-measure with 0.5s threshold using mir_eval [Raffel+14]

2. **How does each fitness contribute as the duration model becomes more expressive?**
   - Compare best weights α, β

$$B = \arg\max_{\hat{B}} f_B(\hat{B}) + \alpha f_D(\hat{B}) + \beta f_H(\hat{B})$$

3. **How does our method compare with other methods?**

Evaluation
Results



- duration model contributes significantly
- ...but only for higher-order models that can take into account more than three past durations

Duration model is useful for boundary detection, for expressive models >2-gram (incl. LSTM)

Evaluation
Experiments

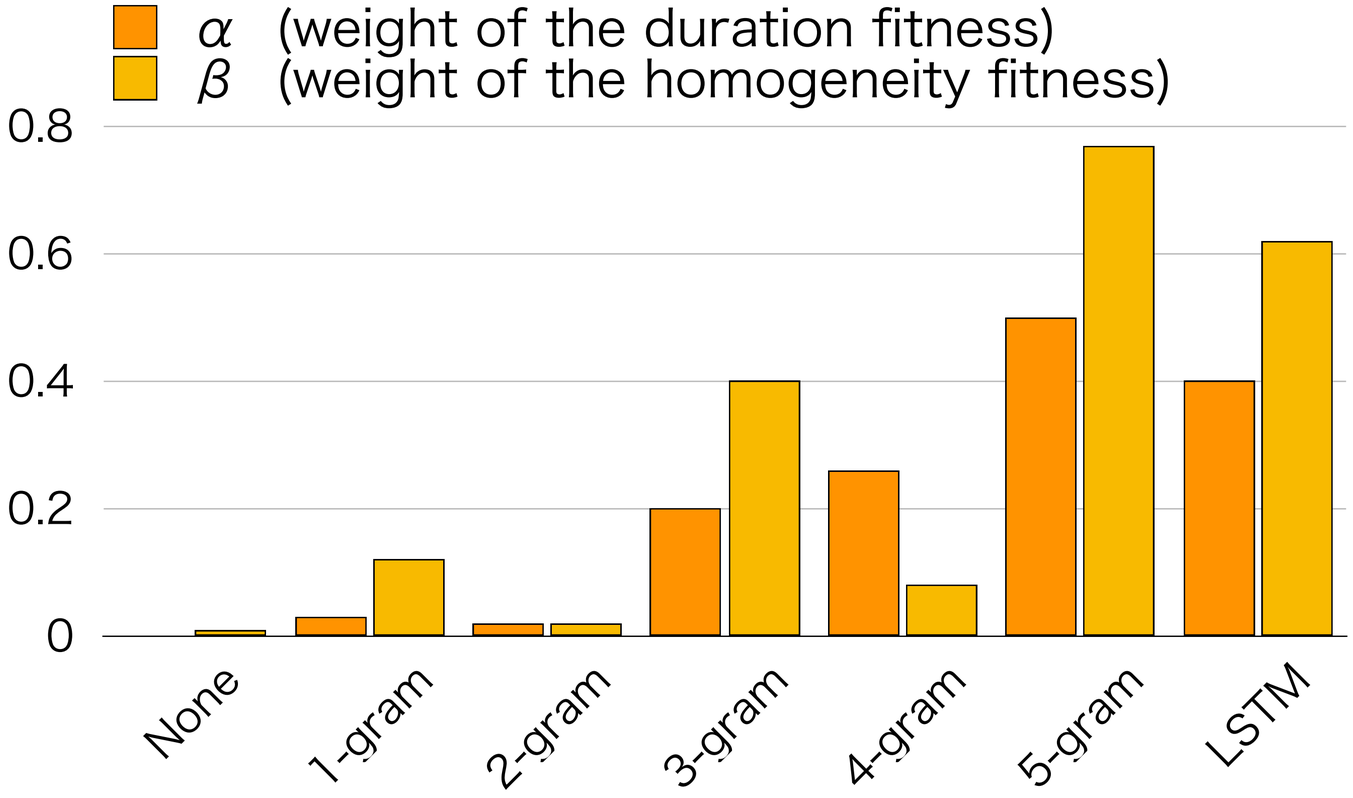
1. **Does duration model help?**
   1. Train boundary fitness model, and various duration models
      - 1-gram, 2-gram … 5-gram
        - Katz backoff (k=1); Unseen unigram duration assigned log-likelihood of -100
      - LSTM
   2. Evaluate the F-measure with 0.5s threshold using mir_eval [Raffel+14]

2. **How does each fitness contribute as the duration model becomes more expressive?**
   - Compare best weights α, β

$$B = \arg\max_{\hat{B}} f_B(\hat{B}) + \alpha f_D(\hat{B}) + \beta f_H(\hat{B})$$

3. **How does our method compare with other methods?**

Evaluation
Results



- Less expressive duration models does little good
- Strong contribution for N>2-gram
- Homogeneity inside a segment counteracts excessive reliance on duration

Duration and homogeneity contribute more as the duration model becomes more expressive

Evaluation
Experiments

1. **Does duration model help?**
   1. Train boundary fitness model, and various duration models
      - 1-gram, 2-gram … 5-gram
        - Katz backoff (k=1); Unseen unigram duration assigned log-likelihood of -100
      - LSTM
   2. Evaluate the F-measure with 0.5s threshold using mir_eval [Raffel+14]

2. **How does each fitness contribute as the duration model becomes more expressive?**
   - Compare best weights α, β

$$B = \arg\max_{\hat{B}} f_B(\hat{B}) + \alpha f_D(\hat{B}) + \beta f_H(\hat{B})$$

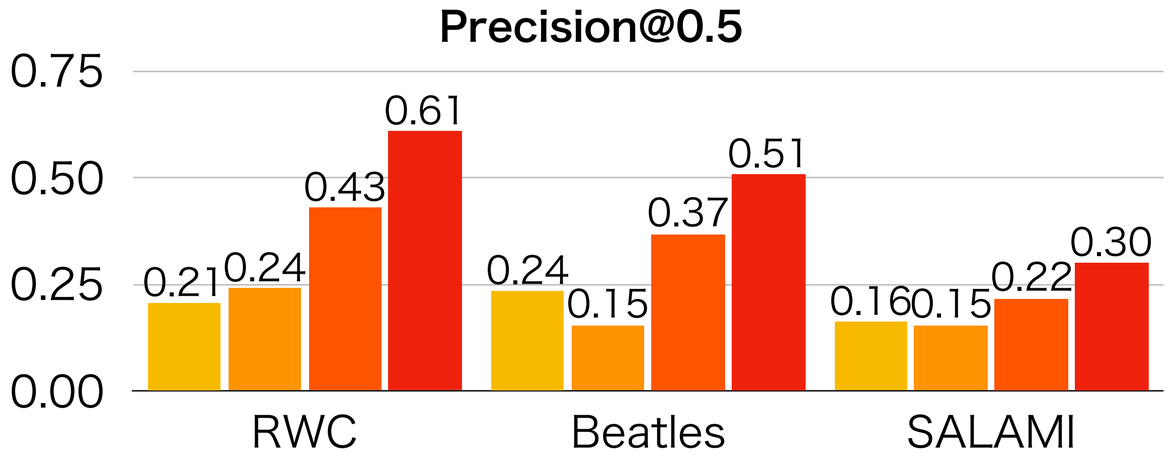
3. **How does our method compare with other methods?**

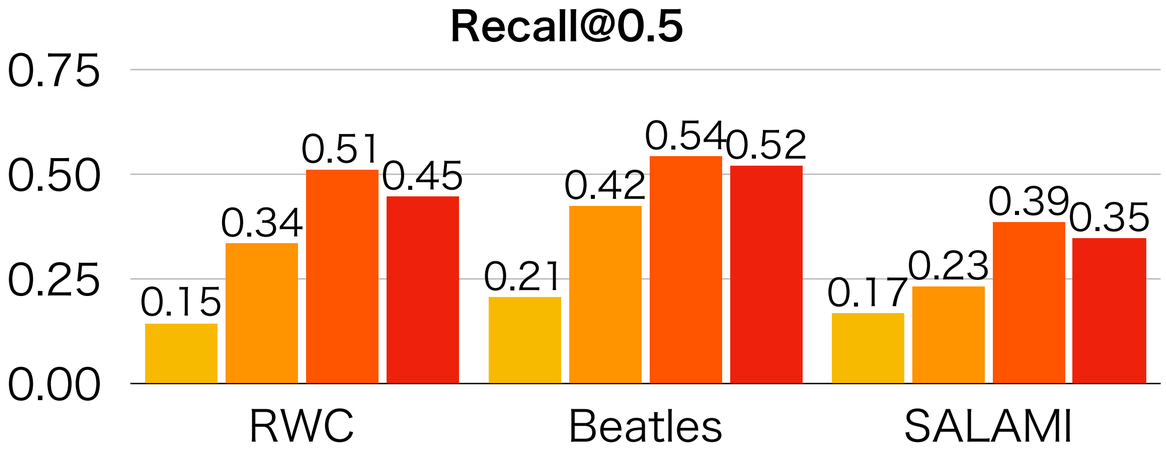
Evaluation
Conditions

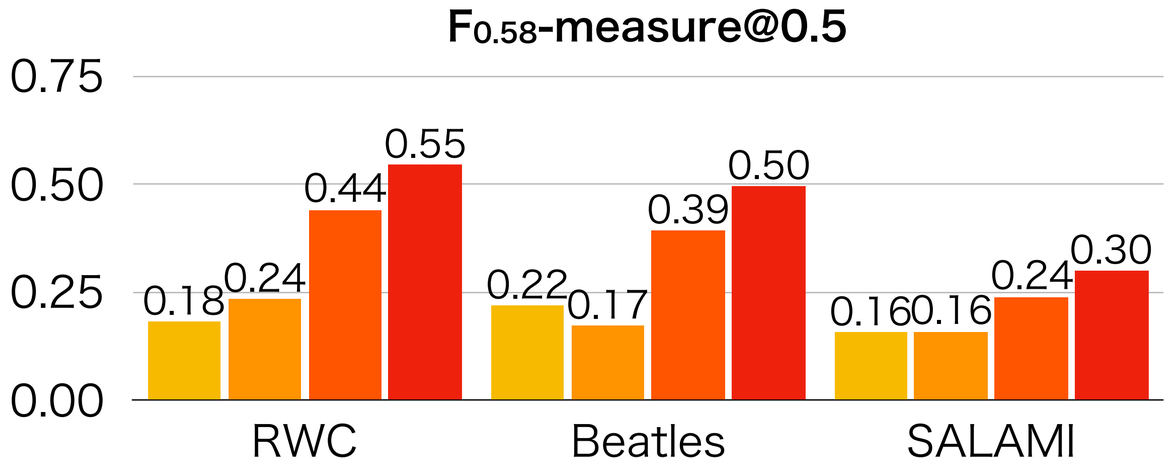
- **Conditions**
  - MSAF (PCP/SF) [Nieto+16][Fujishima99][Serra+2014]
  - Spectral Clustering [McFee+14]
  - Peak pick [Grill+15], using boundary fitness model
    - Threshold optimized on validation data
  - Proposed method

- **Metrics**
  - Precision, Recall, $F_{0.58}$ measure [Nieto+14], F-measure
  - All 0.5s threshold

Evaluation
Results







(Similar results for F-measure)

- DNN is good at finding candidates
- Duration+homogeneity serves to **increase P significantly** while **slightly decreasing R**

Conclusion

- **CNN-based boundary detection**, with **elaborate segment duration** models (various ngrams and LSTM), and a simple **homogeneity** model
- Beam-search to combine multiple hypotheses sources
- Evaluation showed **homogeneity and duration models helps**, **provided that duration model is expressive enough (>2gram)**

- Future work - combine more expressive model of homogeneity and an explicit expressive model of repetition