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

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

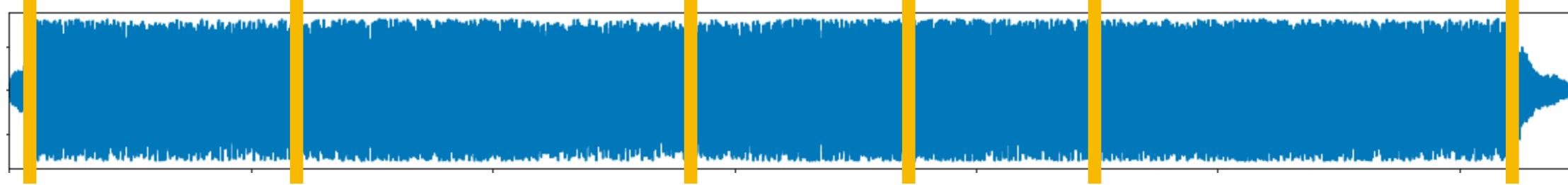
Akira Maezawa Yamaha Corporation

Introduction Problem setting

Input - Music audio signal

تحتافه مخاطرها فطريبا الأرابيا الأطراب والمتعاد مستريع مصابية فتحمد وتنجل والمقرف المتحاجين المالية ألكم والمترافي والم

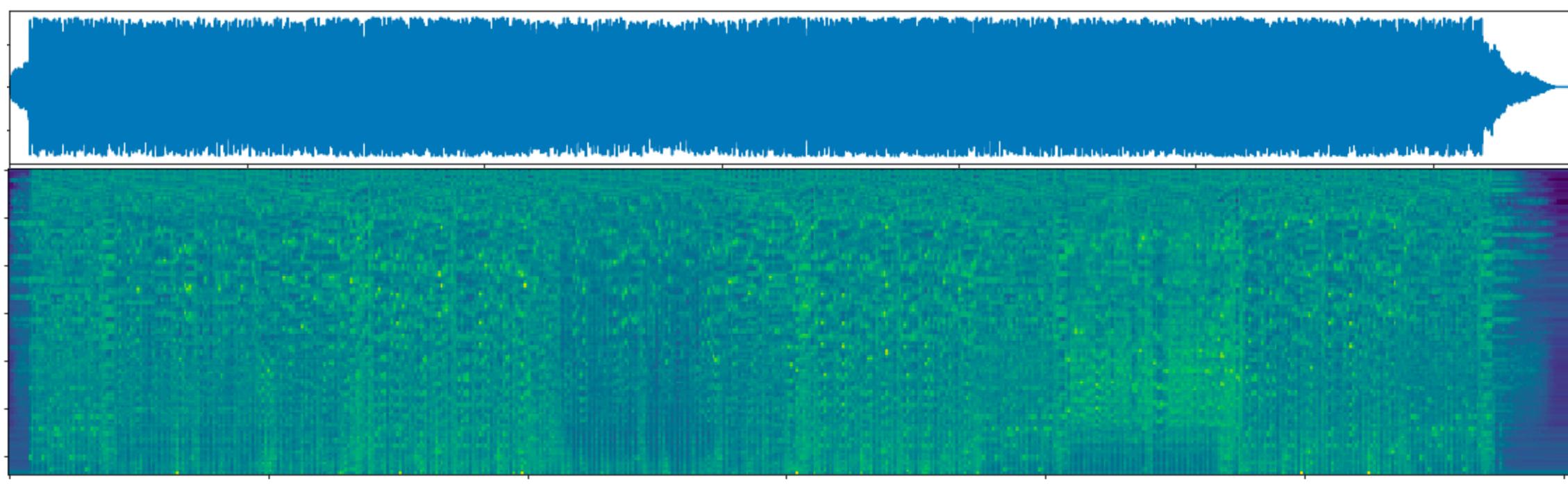
Output - Locations of structural boundaries (e.g., Verse, Chorus)





Repetition

Homogeneity



MSLS (Mel-scale log spectrogram)

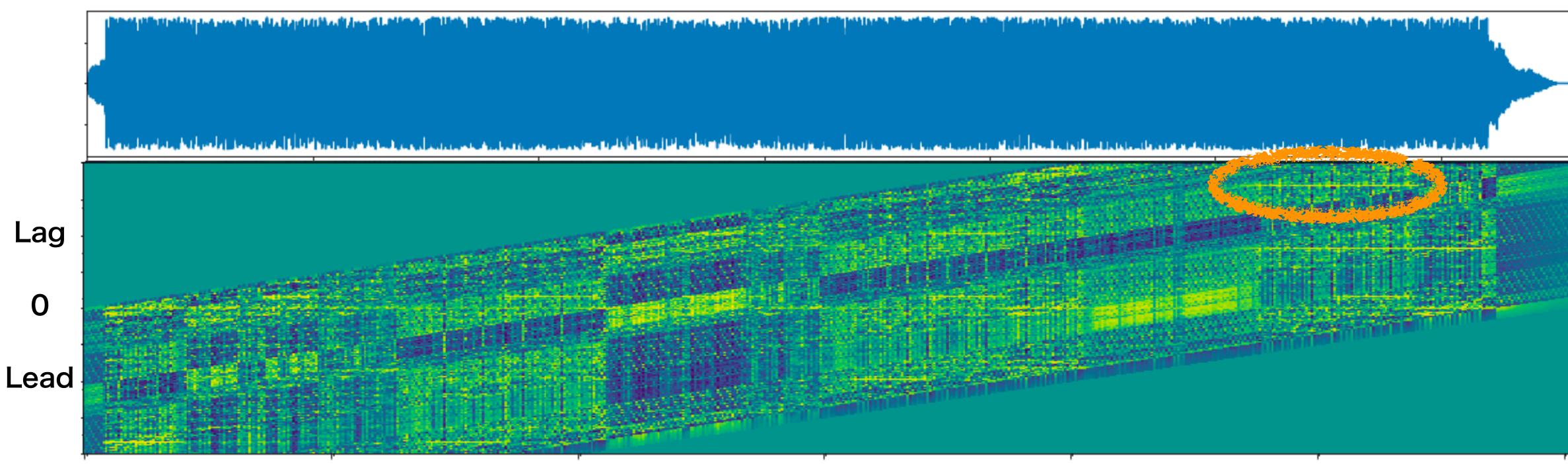
[Paulus+2010]

Novelty

Duration



Homogeneity Repetition



SSM (Self-similarity matrix)

[Paulus+2010] [Goto2003]

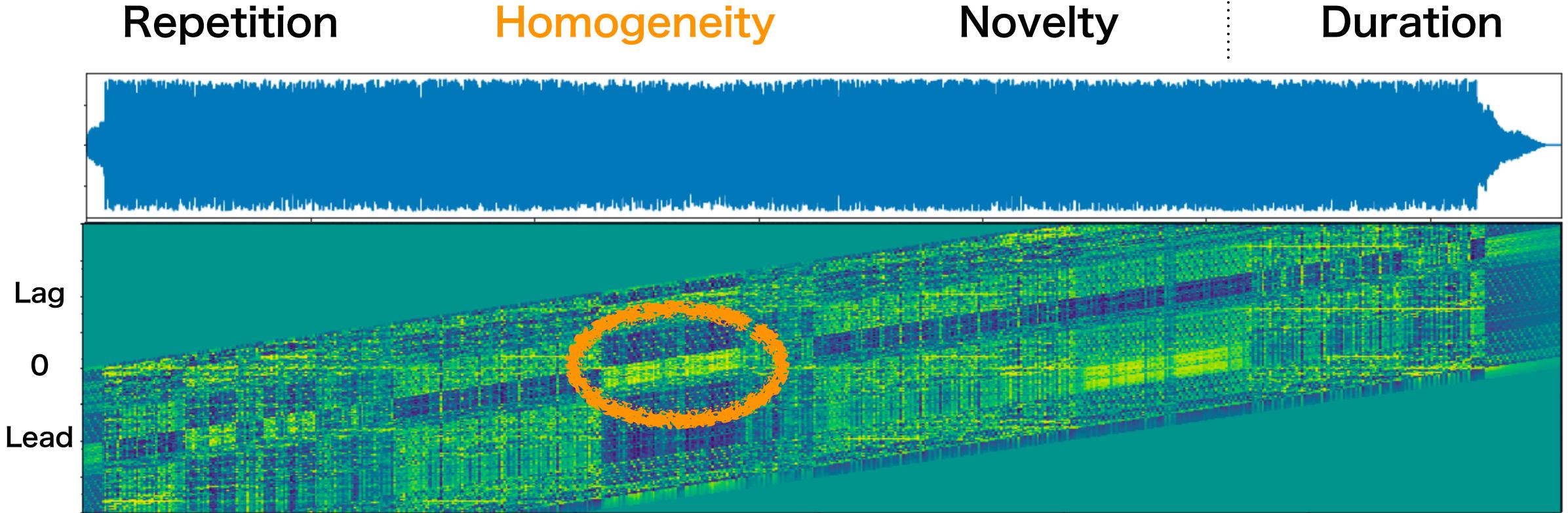


Duration





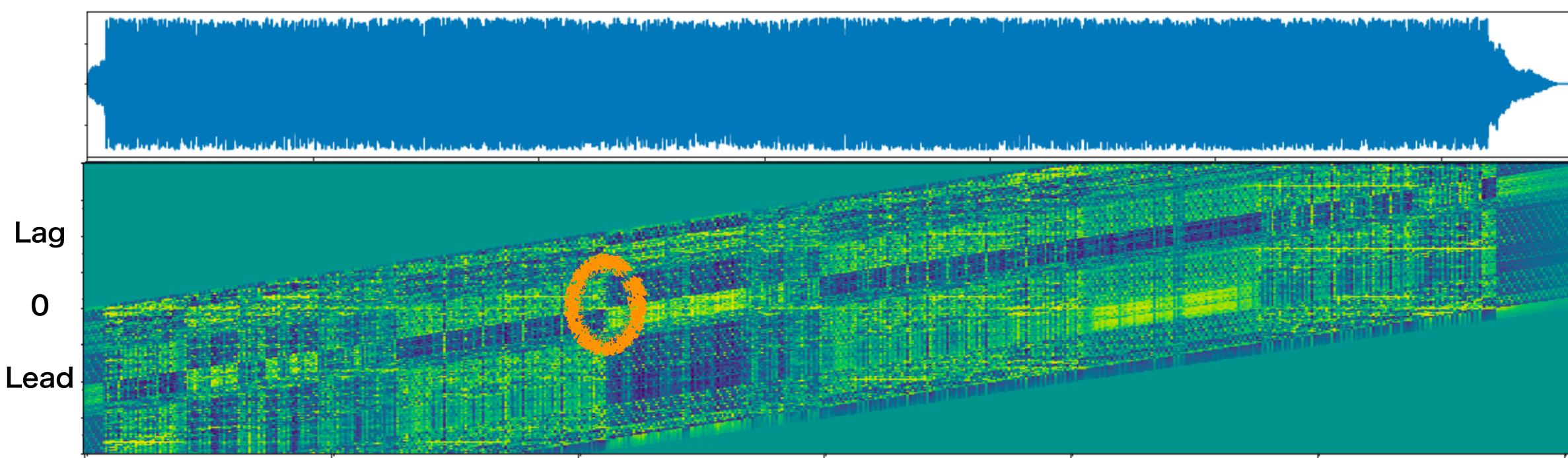
Repetition Homogeneity



[Paulus+2010] [Cooper+2003]



Homogeneity Repetition



[Paulus+2010] [Foote2000]



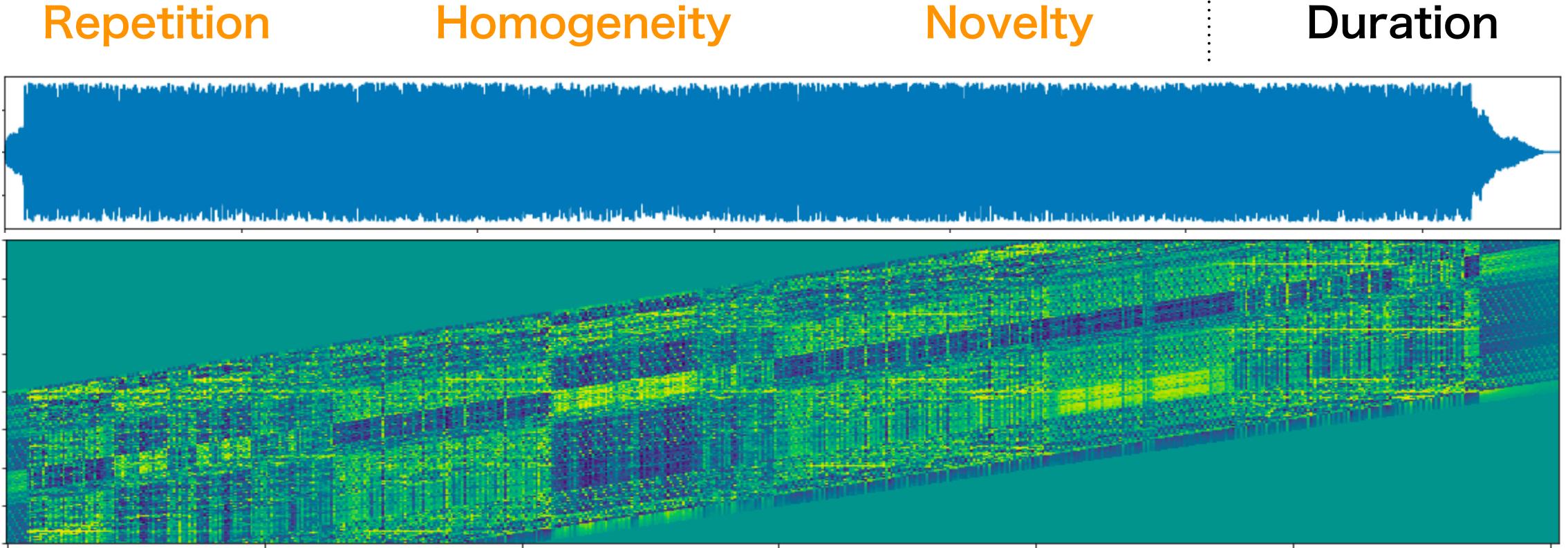
Duration



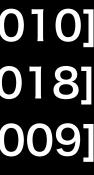


Repetition Homogeneity

أر الأرب بأذاريا أورار بلارد الأرجعيا أنترجموه فسار وبالبلا



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



(Homogeneity) (Repetition)

والمكافأ الملك ومراد الأعماقة واعتادها والمتاج الأرجعة أتتحدد فمار وغادا وملطة وقرب والتم تبعيد الأباليا والشطة

المحيد والكرافية الأرار والكرافي والمتعاد مستحد والمتكر فيتحدث والمحاج والمتحاج والمتحاج المحاج والمحاج والكرام والماع والمحرج

[Ullrich+ 2014] [Grill+ 2015]



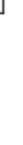
Duration

متأبيته ومعرفا ليتبار ومتعاقبه		la ta la milla da tata da facilia da na sur la	و النامل المنام المرب الأور من الأوار من الأورا به الحق	Just 1			































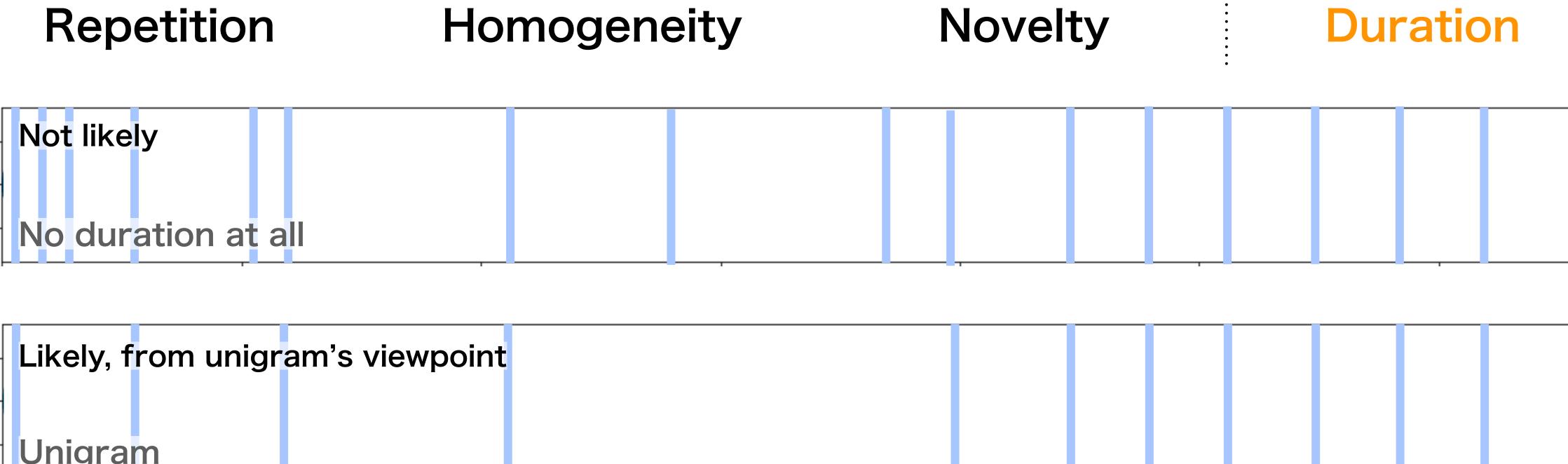








Repetition



[Likely, f	rom unigr	am's viewpoint	
	Unigra	n		

[Levy+ 2006] [Smith+2016]

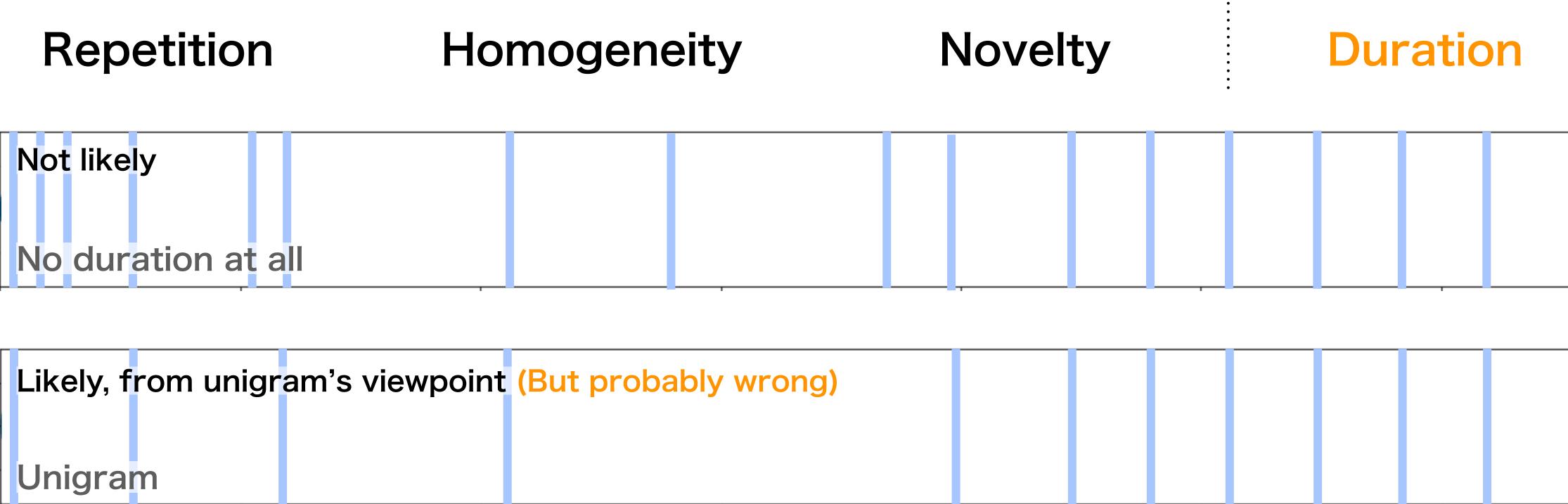








Repetition



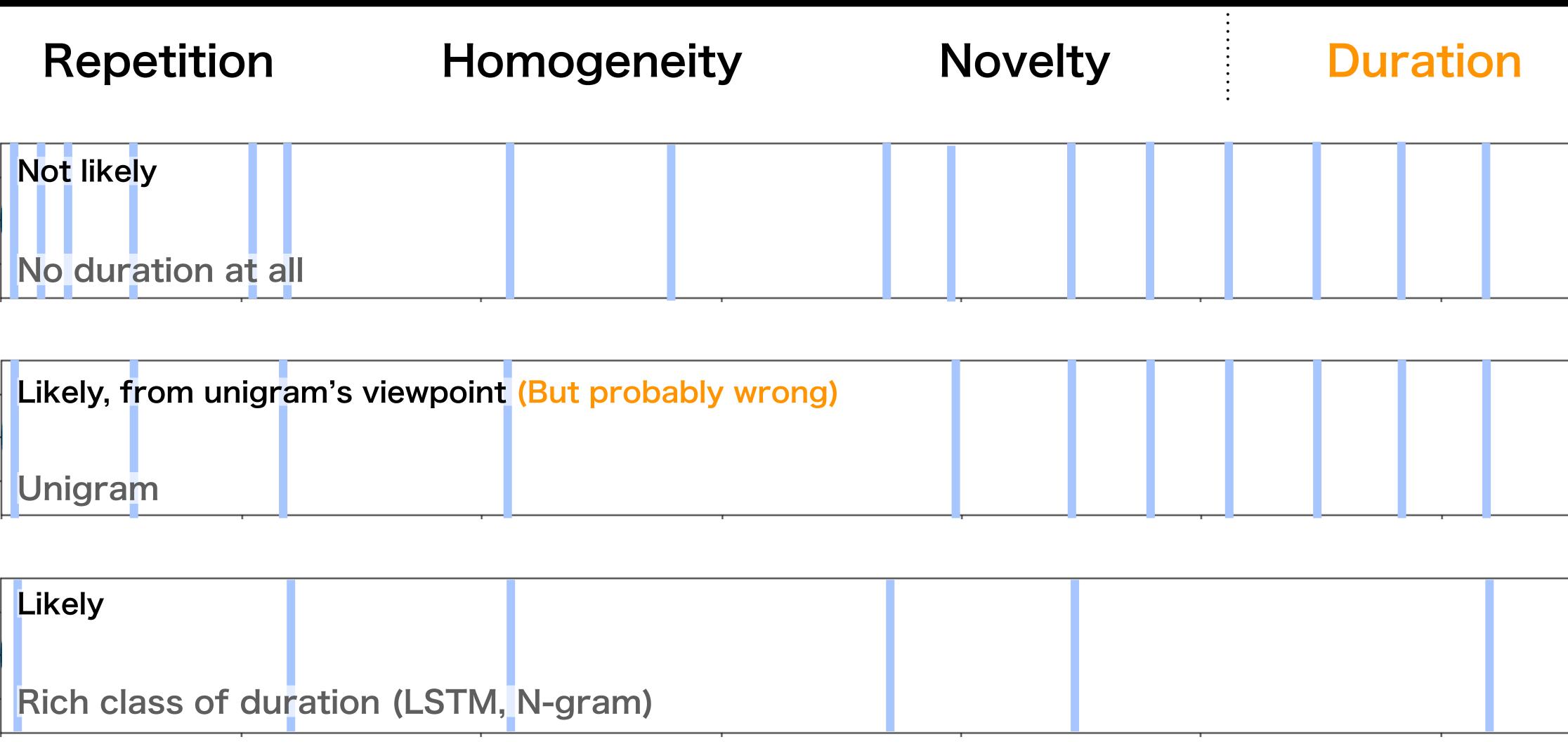
[Levy+ 2006] [Smith+2016]











[Levy+ 2006] [Smith+2016]

Main Contribution: Enabling incorporation of elaborate duration models









Our Approach

Homogeneity (Repetition)

فيطر بالأبل بالمتك الكاتب بالمكافلات الكريميين الأردي بأطلال وجماعه الأرجعة أتحدث ومعتر المراجع والمأقية الكرا تربقا مريحية وأربية بسيدياتهم تبعجم وأبارتها والمأسطان وتعرير والطائل أأكارك ويؤار محافظت فأطر وبالطراب الطراب والطرفان والمتعار ومستعيرهم بيرك فسط متحد والمحران ويتعرف فالمراجع فالانتها والمتكار المائية

Duration

Likely

Rich class of duration (LSTM, N-gram)

[Ullrich+ 2014] [Grill+ 2015]

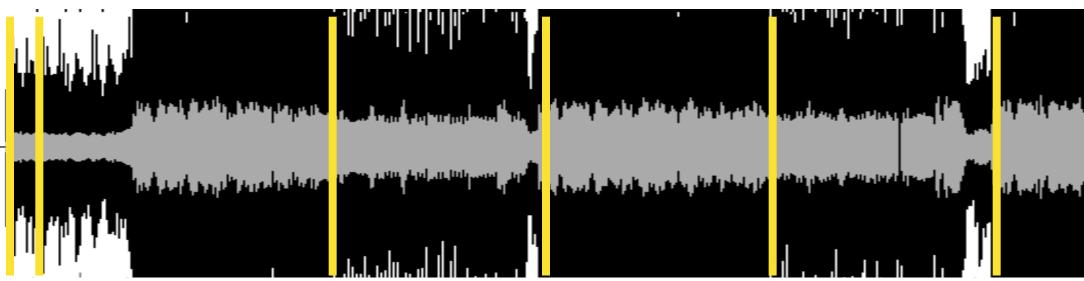
Novelty

	(th) (n) at hit is an institute of		inians (halideaned (the divisi)	
	in i line i state e the internet	sinessis adeitatel	<u></u>	

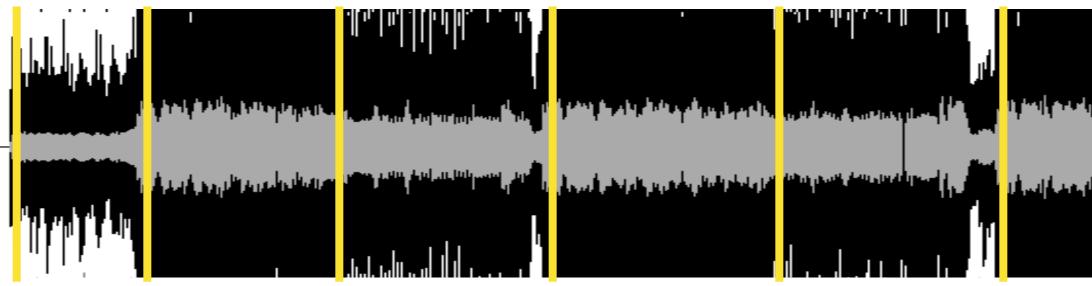


An example

DNN boundary



DNN boundary + LSTM duration + homogeneity



	and total at a sec	land manifer the state of the	ka kalena atra atra da d	ju La sulta sulta su		
ا ارالا ارولاي					ľ	

<mark> ℓ,</mark> , , , , , , , , , , , , , , , , , ,			
ինը հեղել Գրնիստող Գեծև-Միսի Գերերությո	i ng sanga Lasi na dika		1
		հենցինացենն հենցիները ներկոնըներին էրիներին Ա	

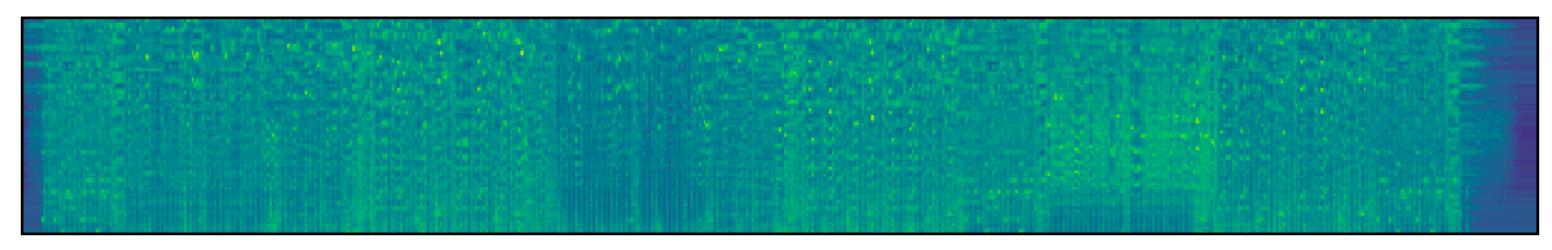
Our method Overview

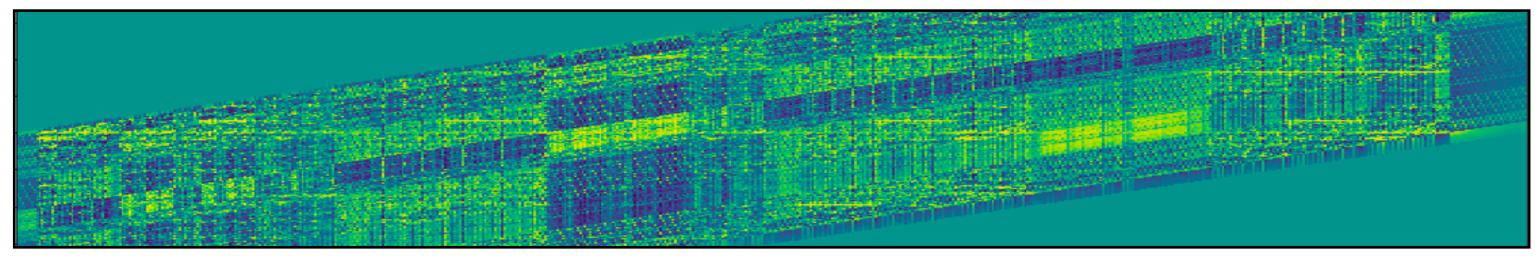
MSLS

sliced at 8th-note level

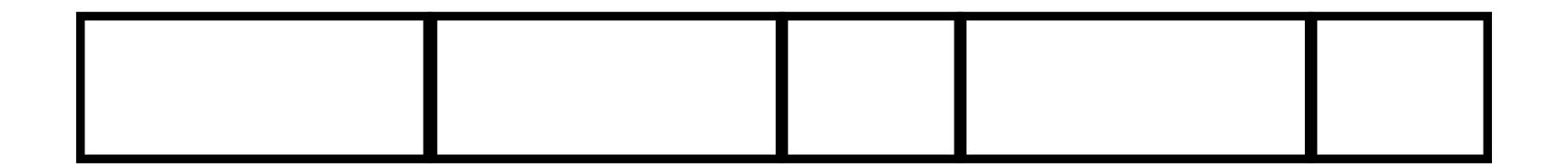
SSM (MSLS)

sliced at 8th-note level





B Boundary positions in beats

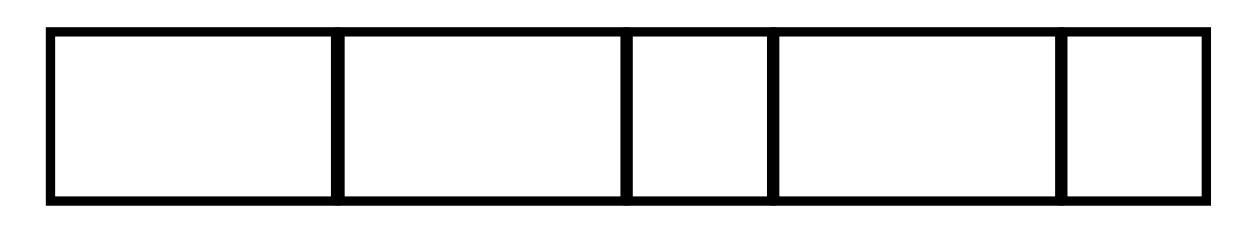


128 dims 0 to 11kHz

lead/lag of 200 beats

Our method Overview

В Boundary positions in beats

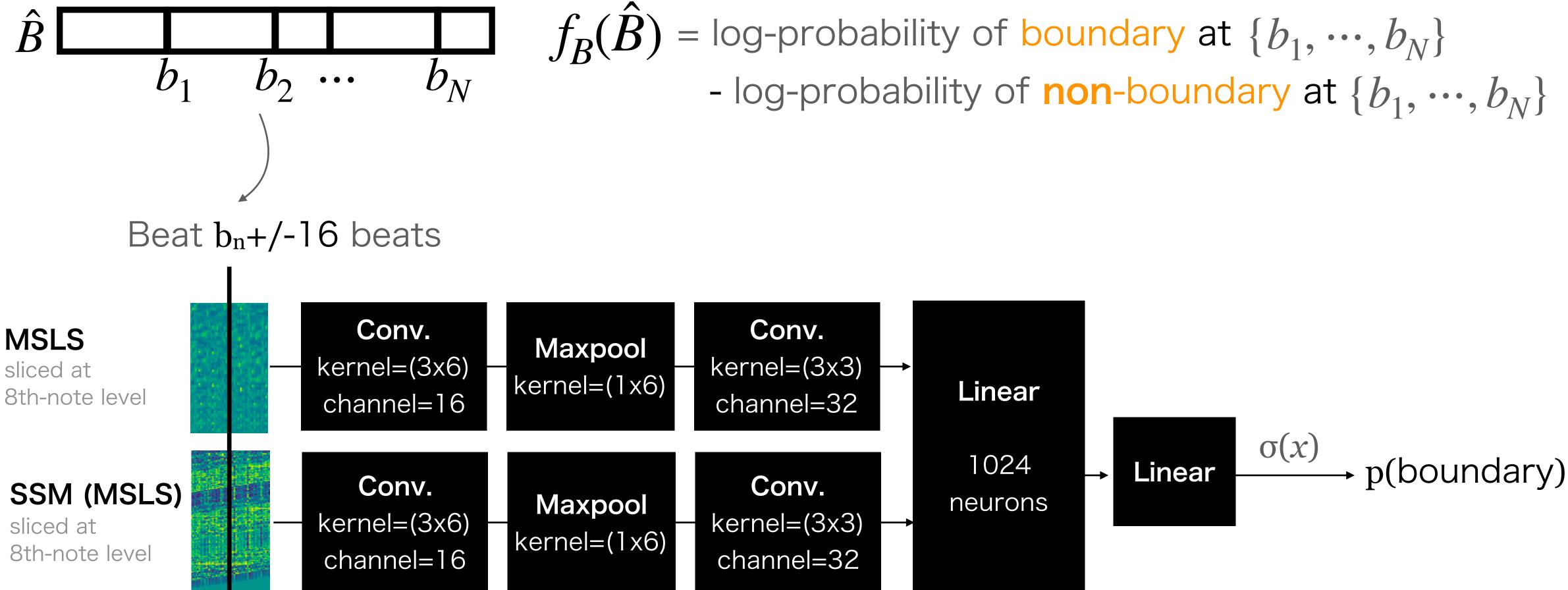


 $B = \arg \max_{\hat{B}} f_B(\hat{B}) + \alpha f_D(\hat{B}) + \beta f_H(\hat{B})$

Boundary fitness

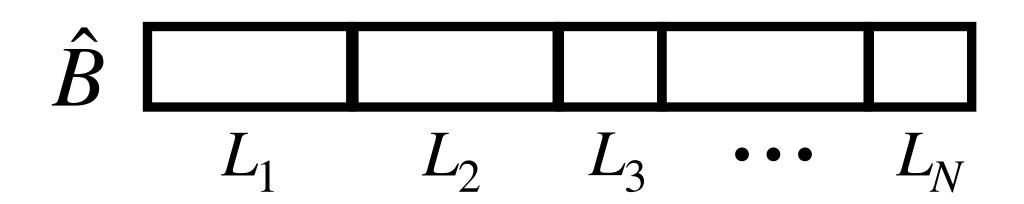
- **Segment duration fitness**
- **Timbre homogeneity fitness**
- Find *B* using beam-search

Our method Boundary fitness





Our method Duration fitness





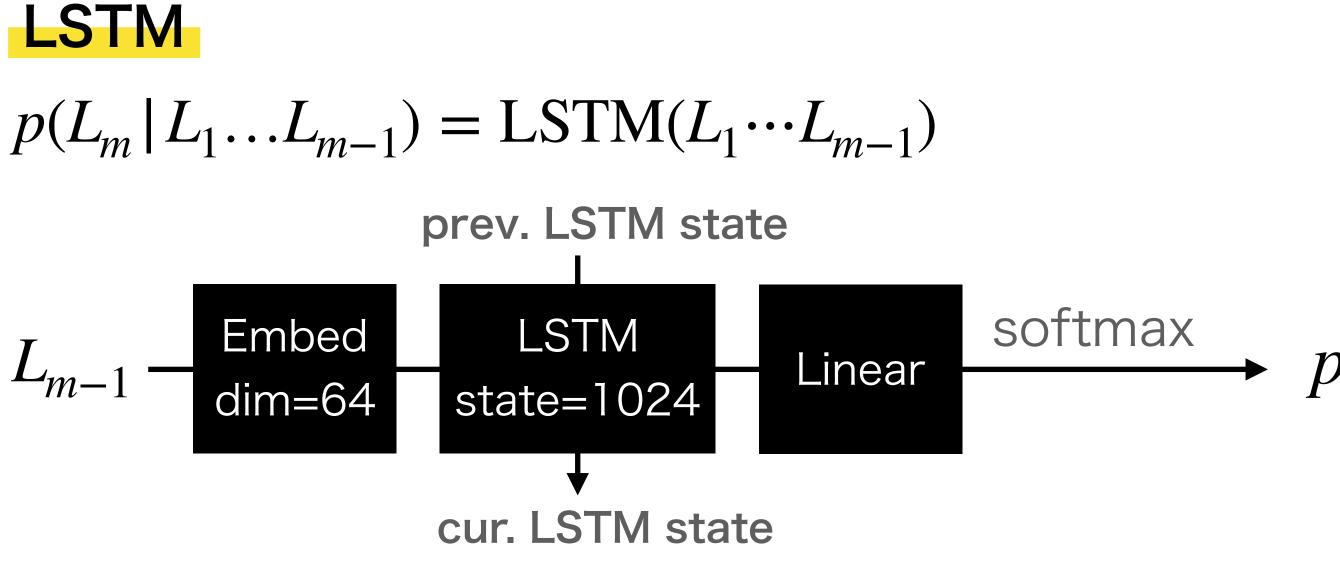
 $p(L_m | L_1 ... L_{m-1})$

 $= p(L_m | L_{m-n+1} \dots L_{m-1})$

(n+1)-th order Markov assumption

(Edge cases L_1 and L_N are treated differently)

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

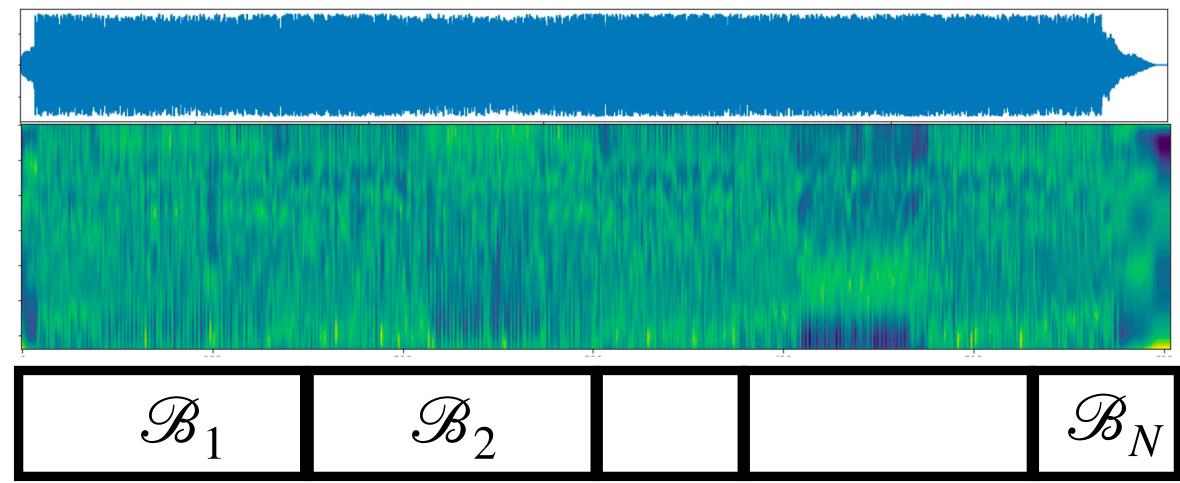




Our method Timbre homogeneity fitness

X(t)MSLS smoothed by a Hanning window





 $f_H(\hat{B}) = -\sum_{n} \operatorname{Tr}(\operatorname{Cov}(X(\mathscr{B}_n)))$

Variance inside segment n, 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]
 - Beatles w/ Isophonics label (all BUT the first album) [Mauch+2009]

• Use only Internet Archives (more degraded compared to commercial audio)

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?

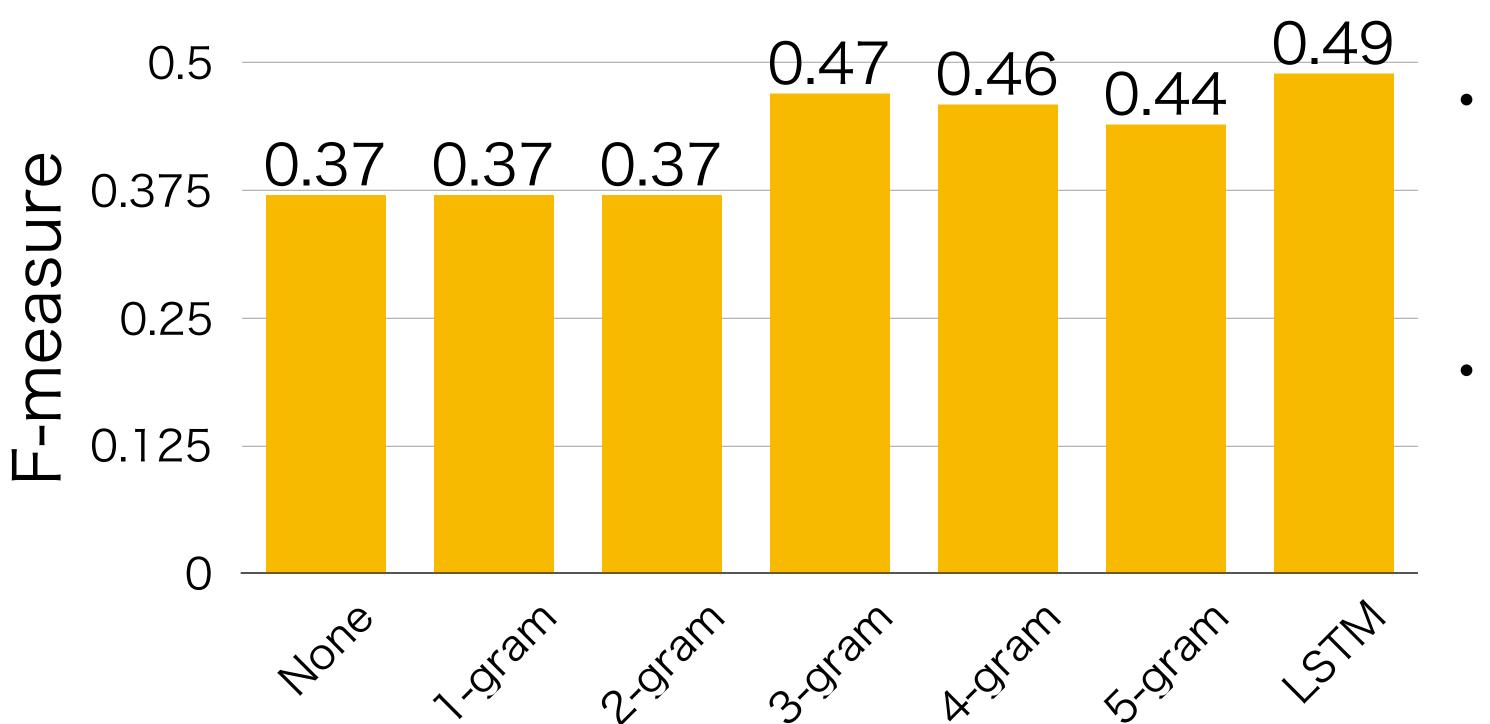
- 1. Train boundary fitness model, and various duration models
 - 1-gram, 2-gram ... 5-gram
 - Katz backoff (k=1); Unseen unigram duration assigned log-likelihood of -100
 - LSTM

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

3. How does our method compare with other methods?

Evaluation Results



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

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



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]

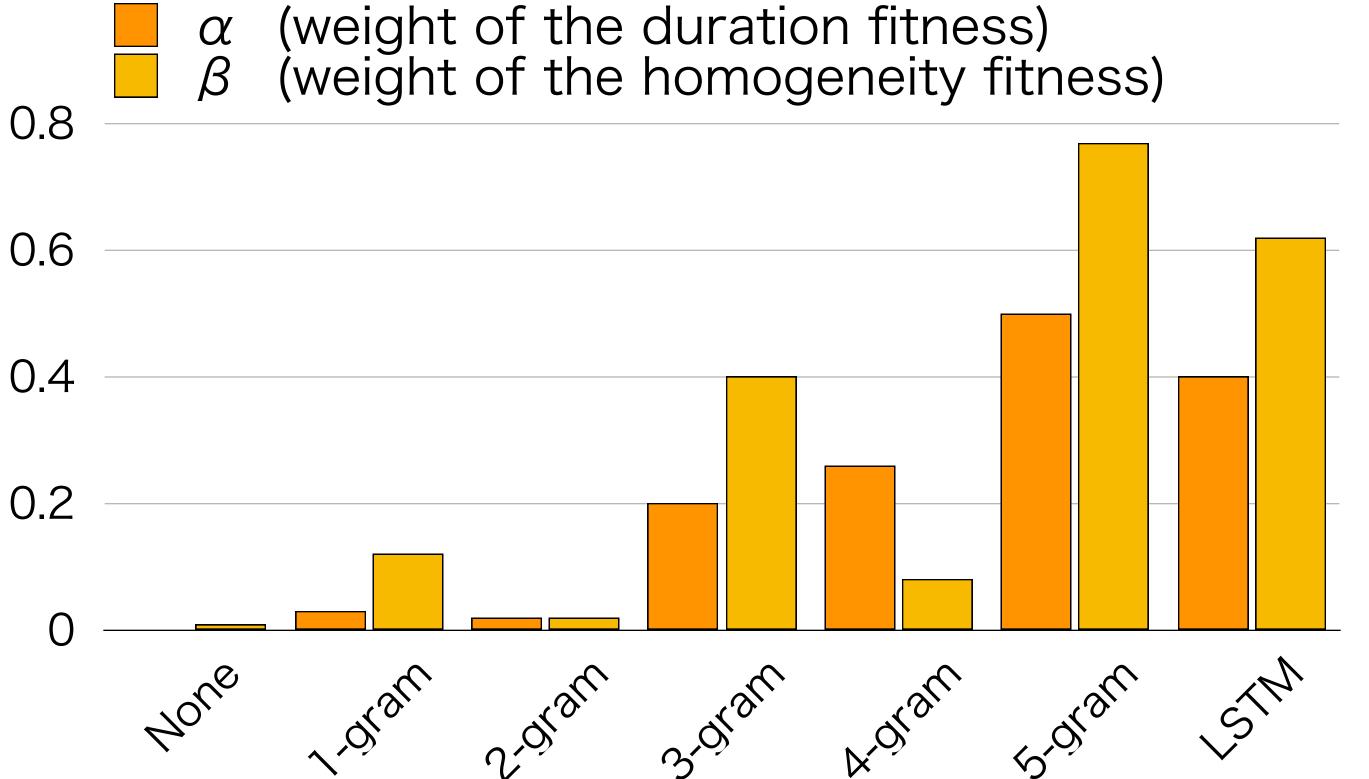
Compare best weights α, β

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

3. How does our method compare with other methods?

- 2. How does each fitness contribute as the duration model becomes more expressive?
 - $\beta f_H(\hat{B})$

Evaluation Results



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

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



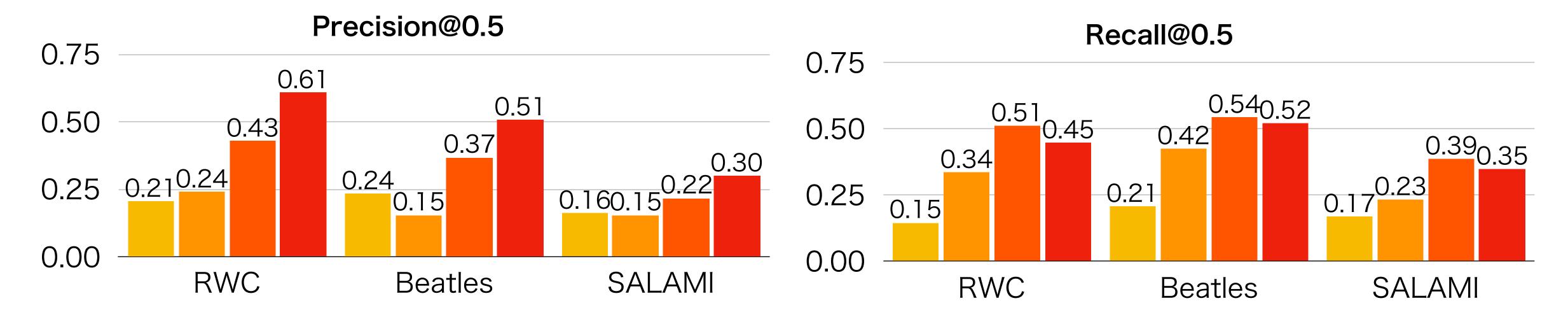
Experiments

- 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 $B = \arg \max_{\hat{p}} f_B(\hat{B}) + \alpha f_D(\hat{B}) + \beta f_H(\hat{B})$
- 3. How does our method compare with other methods?

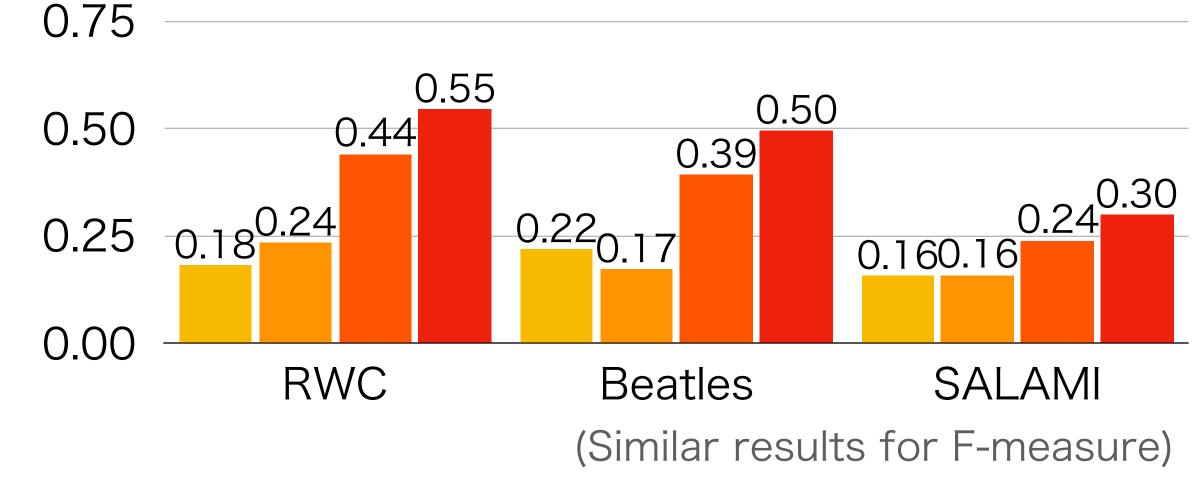
Evaluation Conditions

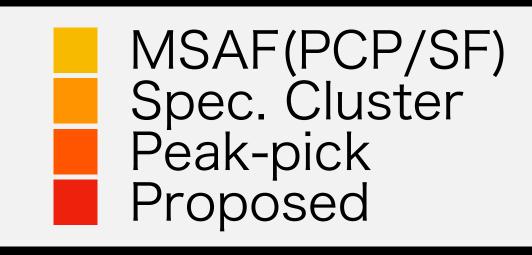
- Conditions
 - MSAF (PCP/SF) [Nieto+16][Fujishima99][Serra+2014]
 - Spectral Clustering [McFee+14]
 - Peak pick [Grill+15], using boundary fitness model
 - Threshold optimized on validation data
 - Proposed method
- Metrics
 - Precision, Recall, F_{0.58} measure [Nieto+14], F-measure
 - All 0.5s threshold

Evaluation Results



F_{0.58}-measure@0.5





- 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