Audio-based Detection of Explicit Content in Music

Andrea Vaglio[†*], Romain Hennequin[*],
Manuel Moussallam[*], Gaël Richard[†],
Florence d'Alché-Buc[†]

[*]Deezer Research & Development
[†]LTCI, Télecom Paris, IP Paris
research@deezer.com

ICASSP 2020
Virtual Conference
May 4-8 2020

Explicit content detection:

Given a piece of music, detect if music contains explicit content. **Binary classification task**

For example: strong language or depictions of violence, sex or substance abuse

Particularly **sensitive** for streaming services

Explicit content detection:

Still a **manual** task (following general guidelines such as parental advisory label)

- Slow and hard to scale to industrial-size catalog

Few automatic approaches and only based on **preexisting lyrics** [MMC+05]

[MMC + 05] Jose PG Mahedero, Álvaro MartÍnez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon. Natural language processing of lyrics. In ACM, 2005.

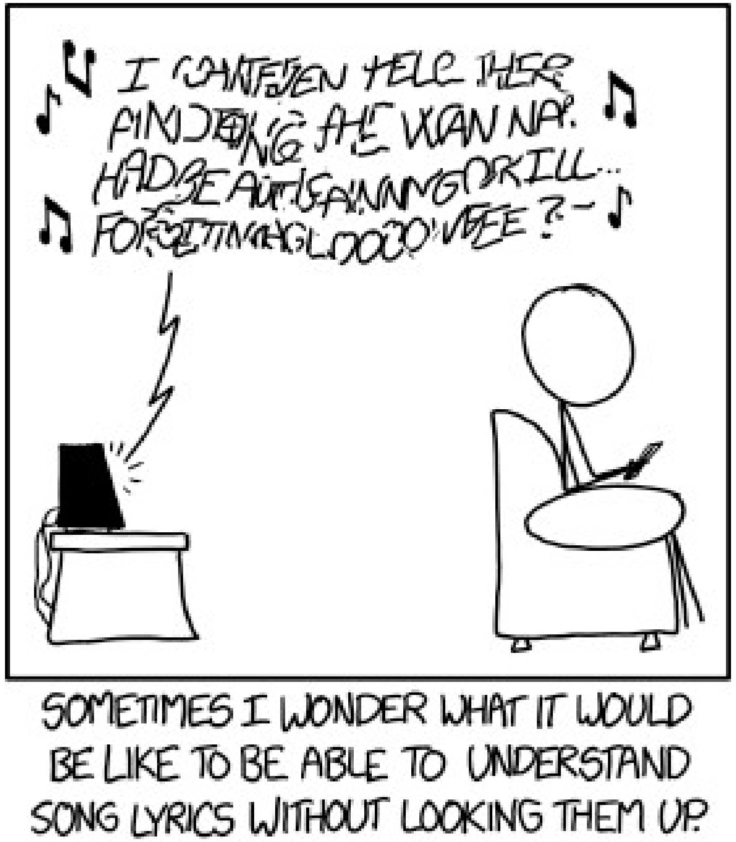
Lyrics transcription:

Singing voice recognition Algorithms inspired from **ASR**

ASR good results [Amo16] , singing voice **not so well** [Sto18] …

Lyrics transcription complicated problem with **specific limitations**

- Singing voice properties differ greatly than those of speech [Mes12]
- Music is (mainly) polyphonic



[Amo16] Dario Amodei and al. Deep speech 2 : End-to-end speech recognition in english and mandarin. In ICML, 2016.

[Sto18] Daniel Stoller, and al.. End-to- end Lyrics Alignment for Polyphonic Music Using an Audio- to-Character Recognition Model. In ICASSP, 2018.

[Mes12] Anna Mesaros. Singing Voice Recognition for Music Information Retrieval. PhD thesis, Tampere university of technology, 2012.

A Keyword spotting approach:

When lyrics available, dictionary-based methods with **suitable keywords** perform well [Fe19]

KeyWord Spotting (KWS) well researched in speech [MKM14]

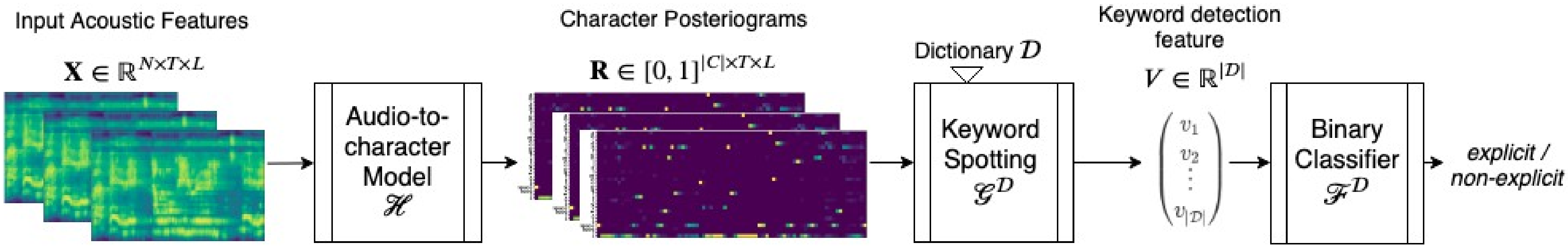
In singing case, research sparse and still **highly challenging**

First only audio explicit content detection system in the music domain!

[Fe19] Michael Fell, Elena Cabrio, Michele Corazza, and Fabien GanDon. Comparing Automated Methods to Detect Explicit Content in Song Lyrics. In RANLP 2019.

[MKM14] Anupam Mandal, KR Prasanna Kumar, and Pabitra Mitra. Recent developments in spoken term detection: a survey. In International Journal of Speech Technology, 2014.

Our modular method:



Given a song, vocal are extracted using spleeter [Hen19], downsampled to 16 kHz and converted to mono

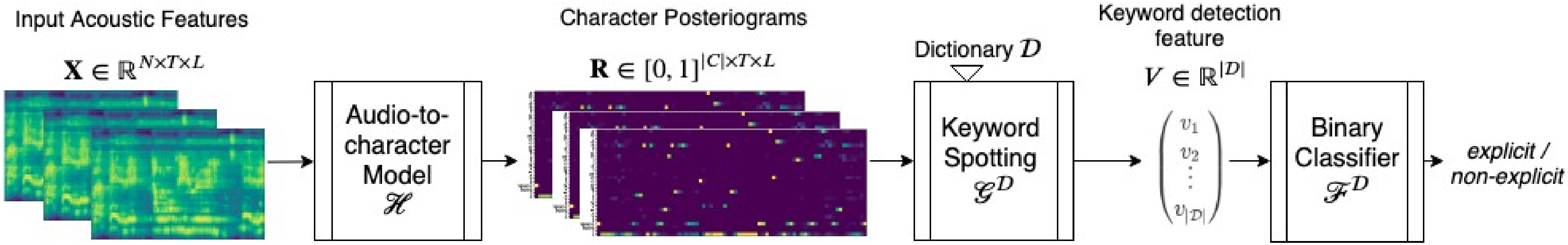
$$\boxed{\mathscr{L}^{\mathcal{D}}(\mathbf{X}) = \mathscr{F} \circ \mathscr{G}^{\mathcal{D}} \circ \mathscr{H}(\mathbf{X})}$$

Vocal track sliced in *L* segment of same size *T*

For each segment, mel spectrogram are computed

[Hen19] Spleeter : A Fast And State-of-the Art Music Source Separation Tool With Pre-trained Models. Romain Hennequin and Anis Khlif and Felix Voituret and Manuel Moussalam. In Late-Breaking/Demo ISMIR 2019.

Training of our system:



Only $\mathscr{H}$ and $\mathscr{F}$ need to be trained

Learning $\mathscr{H}$ can be done using training dataset $\{(X^i, u^i)_{i=1}^{n_{seg}}\}$

Learning $\mathscr{F}$ requires to apply the preprocess $\mathscr{G}^{\mathcal{D}} \circ \mathscr{H}$ to the training dataset $\{(\mathbf{X}^i, y^i)_{i=1}^{n_{songs}}\}$

$$\boxed{\mathscr{L}^{\mathcal{D}}(\mathbf{X}) = \mathscr{F} \circ \mathscr{G}^{\mathcal{D}} \circ \mathscr{H}(\mathbf{X})}$$

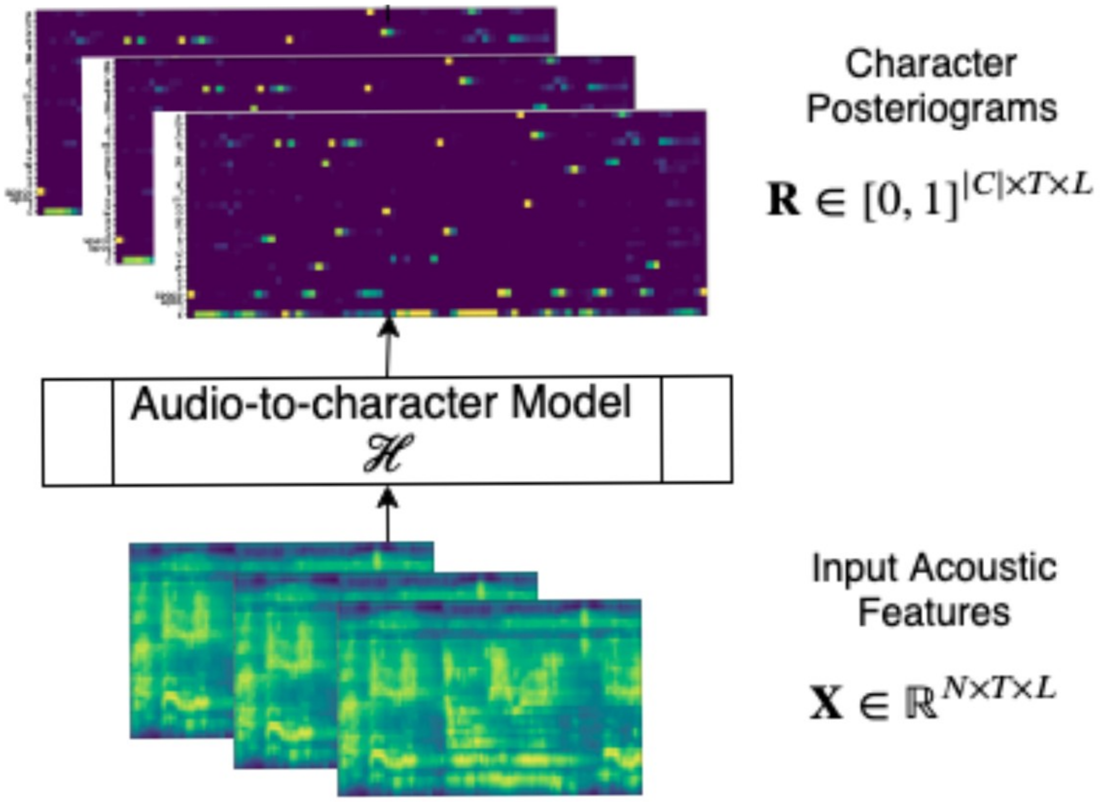
Training datasets **don't have songs in common**

Acoustic model $\mathscr{H}$:

Audio-to-character end-to-end model, great results for lyrics alignment [SDE18]

No need of expert knowledge (e.g pronunciation dictionary)

Trained with **DALI** dataset: +4000 songs with line-level annotations



[SDE18] Daniel Stoller, Simon Durand, and Sebastian Ewert. End-to-end Lyrics Alignment for Polyphonic Music Using an Audio-to-Character Recognition Model. In ICASSP, 2018.

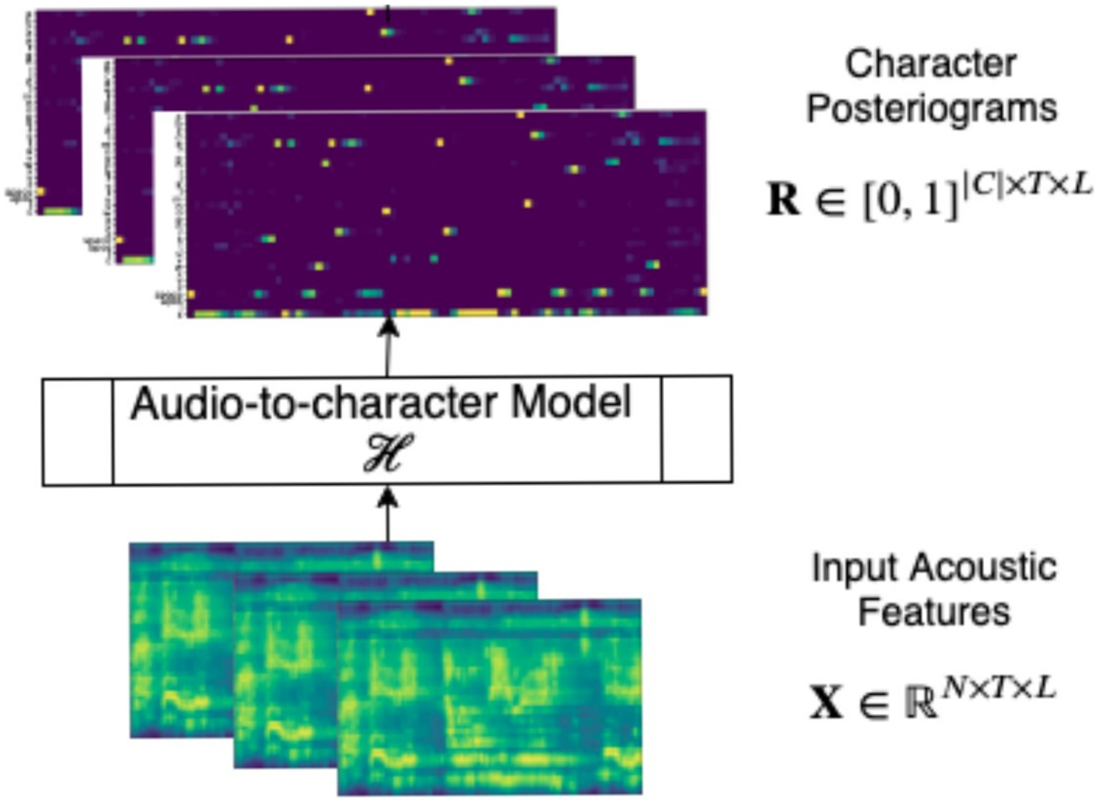
deezer

TELECOM Paris

Acoustic model $\mathscr{H}$:

Architecture **CRNN** trained with a **Connectionist Temporal Classification (CTC) loss**

- Works with unsynchronized annotations
- Avoid first step of forced alignment using intermediate models (suboptimal model performance)



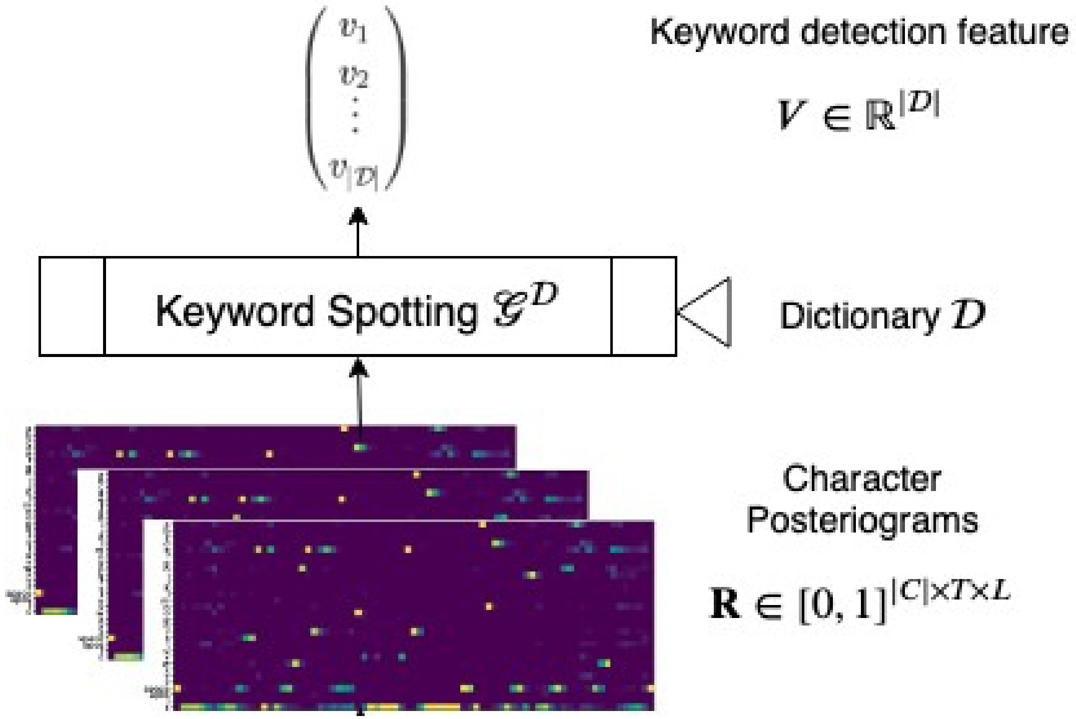
Keyword spotting $\mathscr{G}^{\mathcal{D}}$:

Dictionary dataset: 24250 non-explicit tracks and 24250 explicit tracks, **genre balanced**

$\mathcal{D}$ automatically generated [Kim19], restricted to 128 words

KWS algorithm based on **CTC-based decoding function** [Hwa15]

Keywords can be easily added to $\mathcal{D}$ without retraining the model



[Kim19] Jayong Kim and Y Yi Mun, A hybrid modeling approach for an automated lyrics-rating system for adolescents. In ECIR, 2019.

[Hwa15] Kyuyeon Hwang et al. Online Keyword Spotting with a Character-Level Recurrent Neural Network. In Arxiv, 2015.

Explicit content detection $\mathscr{F}$:

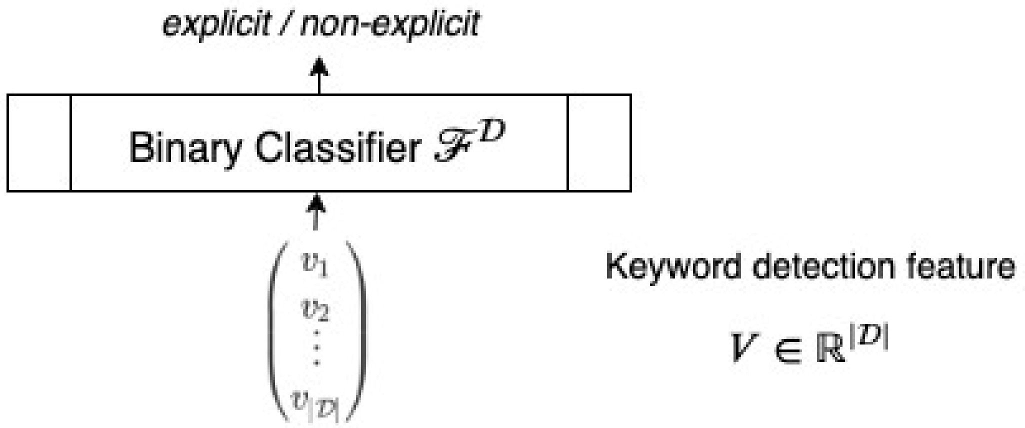
Explicit Dataset: 2600 non-explicit and 2600 explicit tracks, genre balanced

Architecture: **Random Forest**

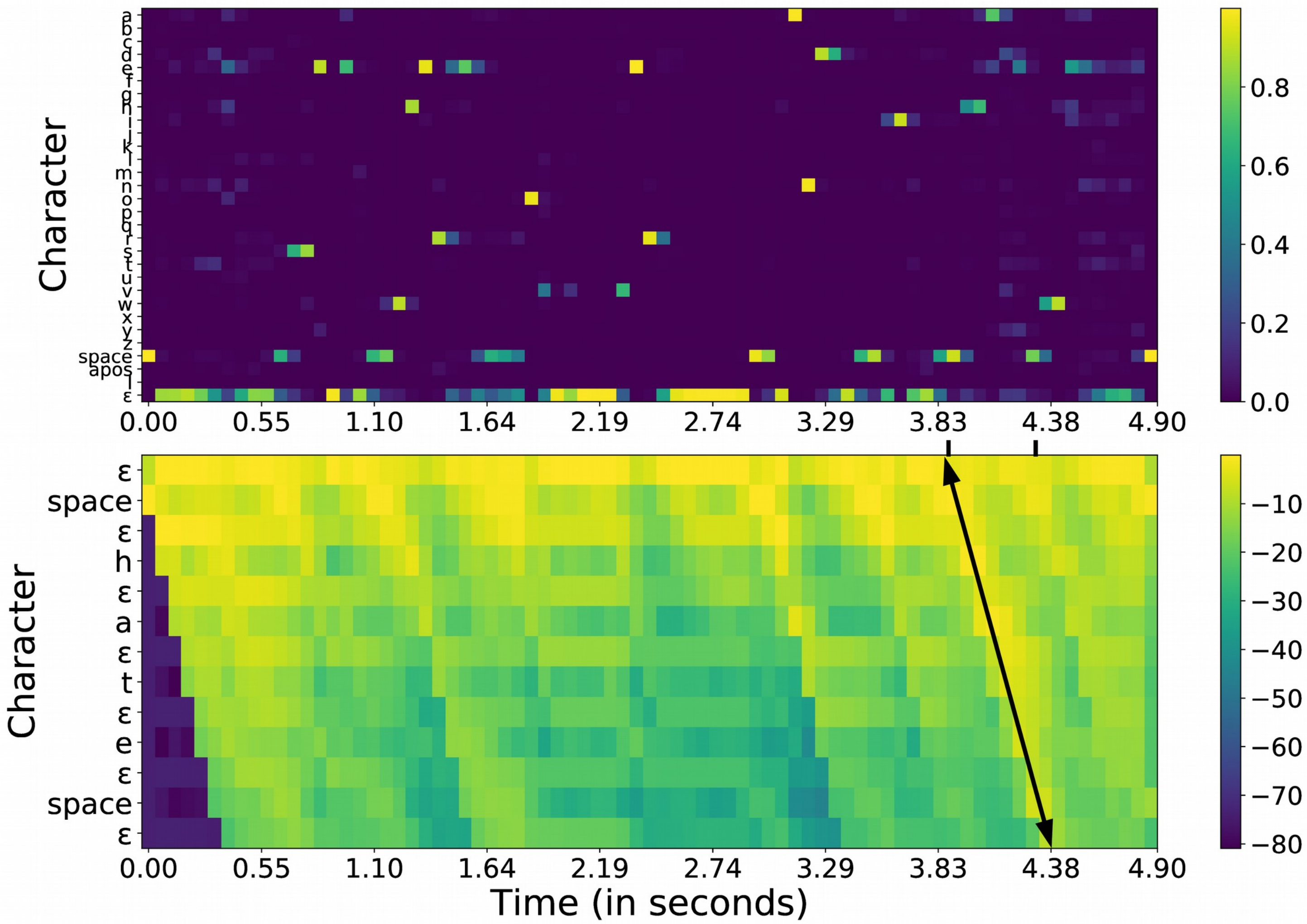
- Hyperparameters tuned using Random search and Grid seach

Number of dictionary words tuned on validation set

- 32 best parameters



Transcription / KWS results:



A positive sample for keyword ”hate”.

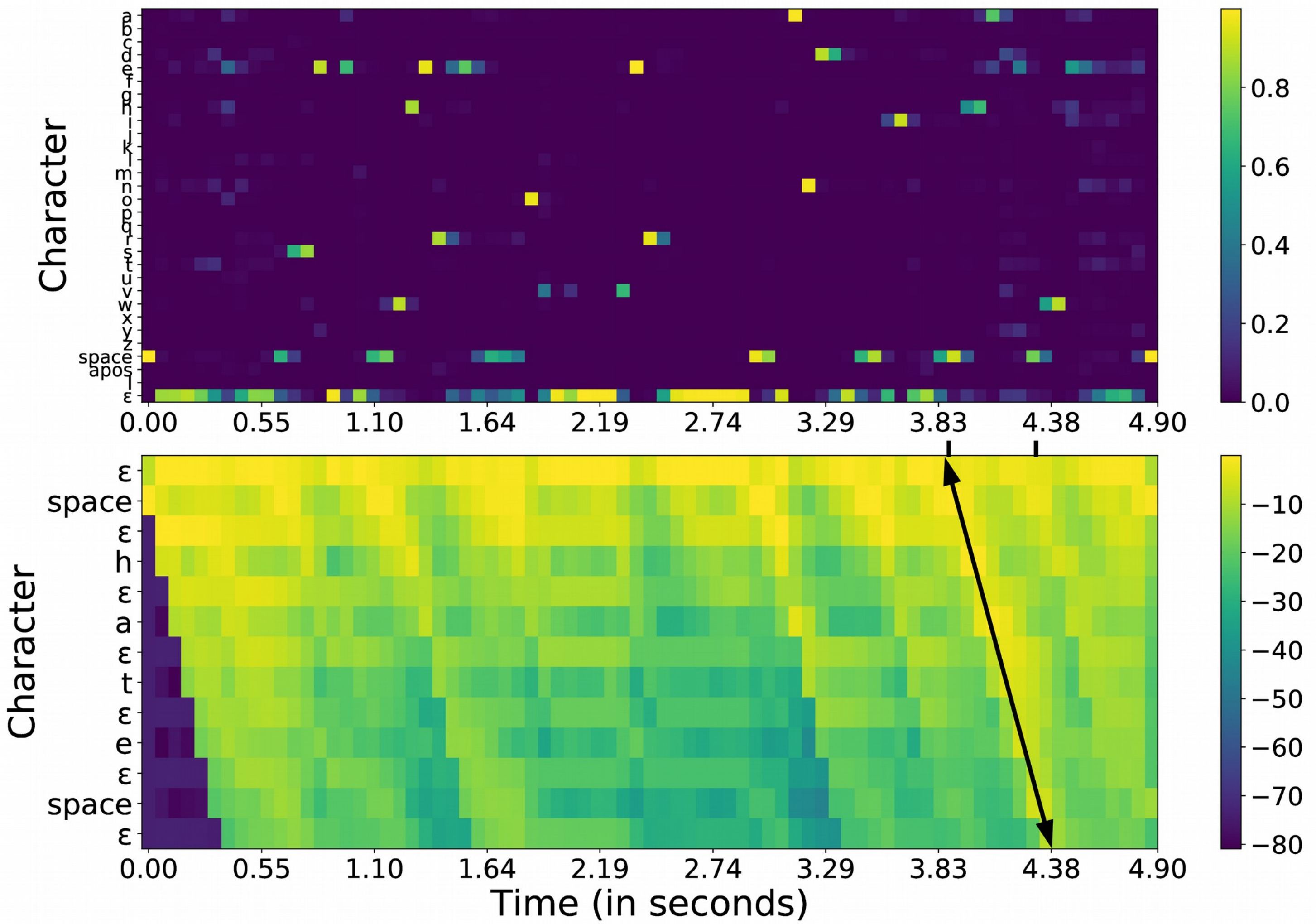
Top: Posteriogram R_ℓ inferred by acoustic model $\mathscr{H}$

Bottom: Decoding matrix

Ground truth: ”to see we’re over and i hate when”

Transcription with beam search: ”e se where over and i hae we”

Transcription / KWS results:



Transcription results:

Character error rate = 0.47

On par with State-Of-The-Art (SOTA)

KWS results:

75 % keywords ROC-AUC > 0.81

V carries information about presence of keywords

Explicit content detection results:

Baseline:
- Lyrics informed oracles (Dictionary lookup). Song explicit if contains at least one keyword of $\mathcal{D}$
- End-to-end naive architecture (CRNN)

Precision, recall, F1 on explicit class

Metrics	Audio baseline	Our system	Lyrics baseline
Precision	.61 (.02)	.63 (.02)	.65 (.02)
Recall	.59 (.02)	.71 (.02)	.84 (.02)
F1-score	.60 (.02)	.67 (.02)	.73 (.02)

Table 1. Results for explicit detection task on the test set (standard deviation in parenthesis)

Explicit content detection results:

Our model **significantly outperformed** naive architecture

Yet not equivalent to the lyrics-informed scenario, the results show **validity of the method**

Metrics	Audio baseline	Our system	Lyrics baseline
Precision	.61 (.02)	.63 (.02)	.65 (.02)
Recall	.59 (.02)	.71 (.02)	.84 (.02)
F1-score	.60 (.02)	.67 (.02)	.73 (.02)

Table 1. Results for explicit detection task on the test set (standard deviation in parenthesis)

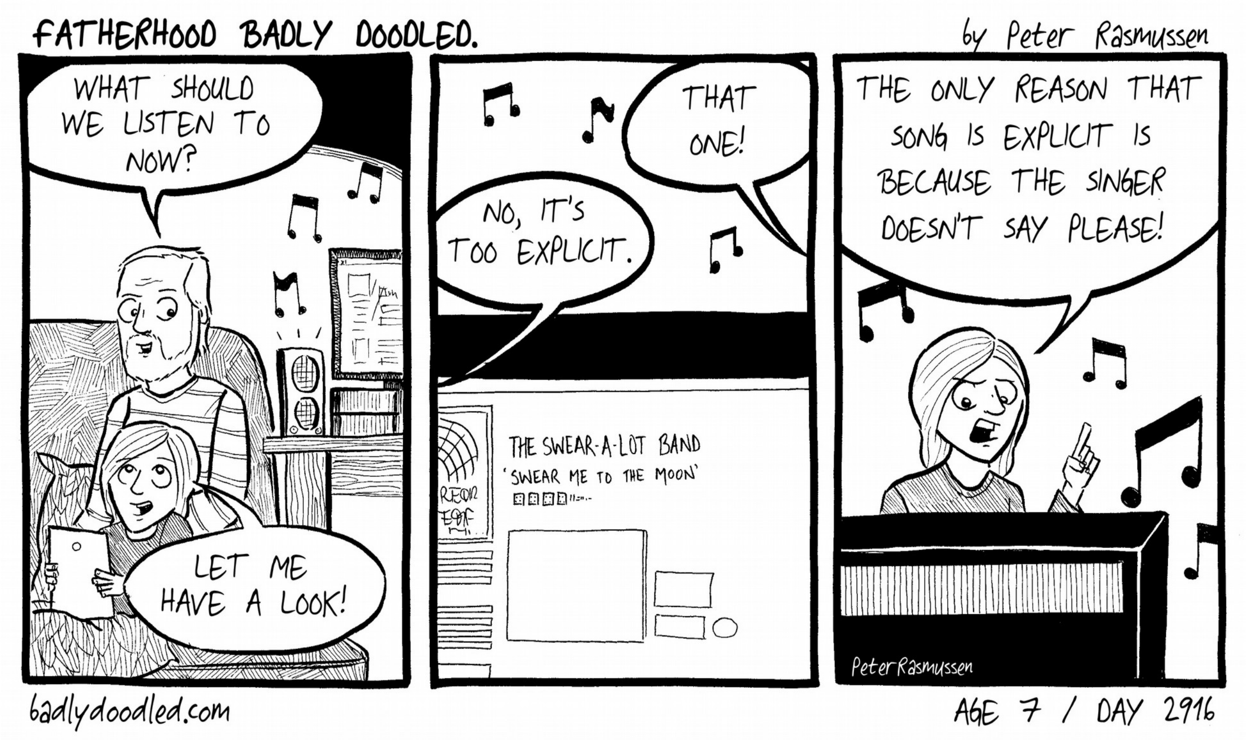
Conclusion:

Novel task of explicit musical content detection from audio only

Despite the task being challenging, our proposed modular approach yield **promising results**.

System's decision can be easily **explained**

- Nice property given the sensitivity of the task




That's all Folks!