End-to-End Lyrics Alignment Using An Audio-to-Character Recognition Model

Session AASP-L7: Music Information Retrieval

ICASSP 2019

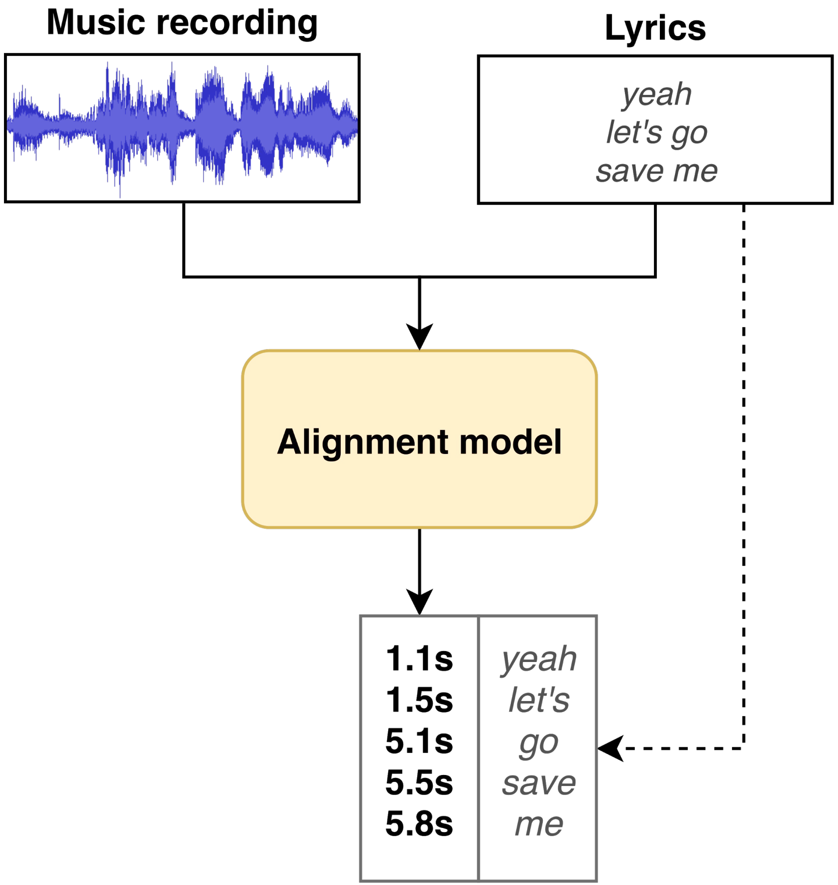
Daniel Stoller[1], Simon Durand[2], Sebastian Ewert[2]

[1]Queen Mary University of London [2]Spotify

Motivation

Lyrics alignment

Given a music recording and lyrics text, predict the time at which each word is sung in the recording



Current lyrics alignment methods

1

Poor performance,
lacking robustness

Multiple-second alignment errors are common [1]

Systems break down when accompaniment is introduced

2

Conceptually
complex

Many interdependent processing stages [2,3]

Manual specification of prior knowledge

3

Require **highly precise annotations**

Real-world datasets cannot be used

→ Lack of training data limits performance

1. MIREX Lyrics alignment results, https://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results, 2017
2. Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi Gokuno, “LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics”, IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 6, pp. 1252–1261, 2011
3. Hiromasa Fujihara and Masataka Goto, “Lyrics-to-audio alignment and its application”, in Dagstuhl Follow-Ups. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2012, vol. 3.

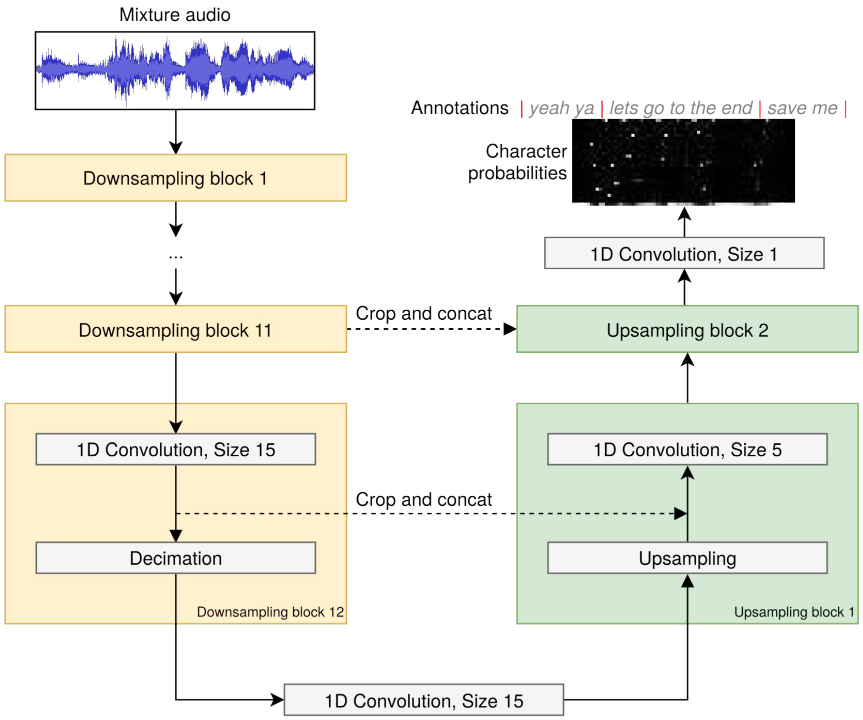
End-to-End Learning

Acoustic model

Adapt the Wave-U-Net [4] for source separation by removing some upsampling blocks

- Directly from waveform to character probabilities

- Acquires features at multiple time-scales without prior knowledge



Overview of our acoustic model.

4. Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end source separation", in Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 2018, vol. 19, pp. 334–340.

Weak label training

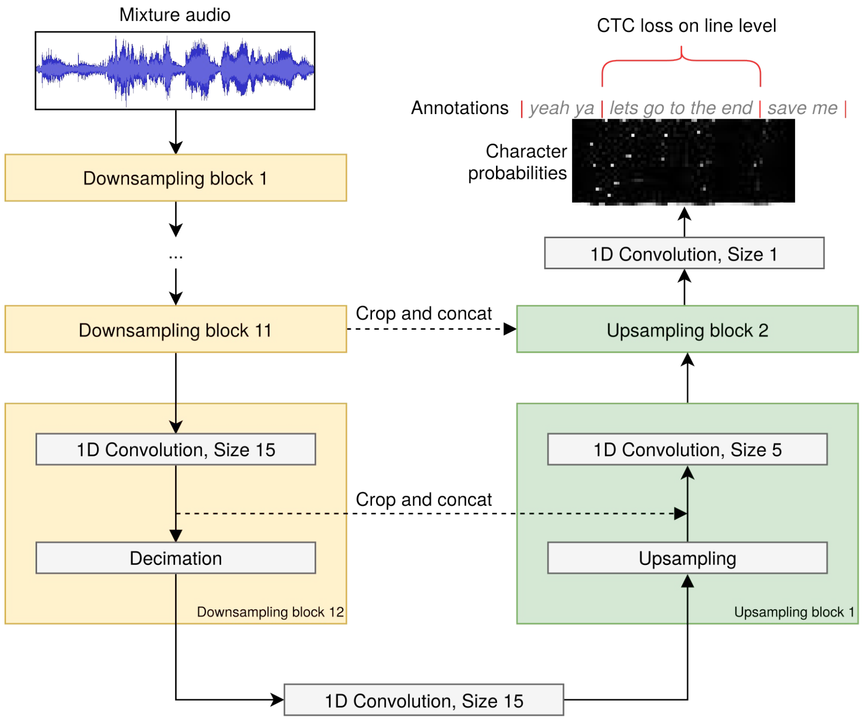
Goal: Train the model **using only line-level alignments**

Solution: Maximise likelihood of each lyrical line using a **CTC loss**:

$$p(y|x) = \sum_{\hat{y} \in \hat{C}^T, B(\hat{y})=y} \prod_{t=1}^{T} P_{t,\hat{y}_t}$$

Benefits:

- No frame-by-frame annotations needed before or during training
- Soft instead of hard alignment

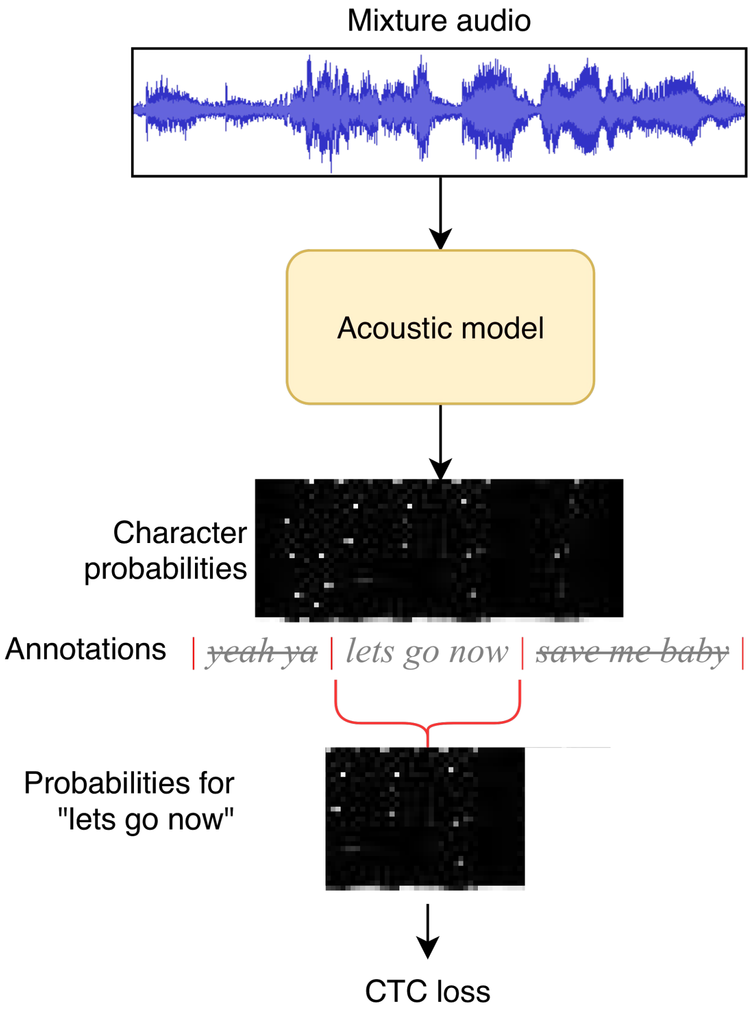


Overview of our acoustic model.

Dataset

Creating training samples

1. Evenly sample sections from dataset
2. Create example for each lyrical line within output window
3. Apply loss only to outputs made between start and end of lyrical line

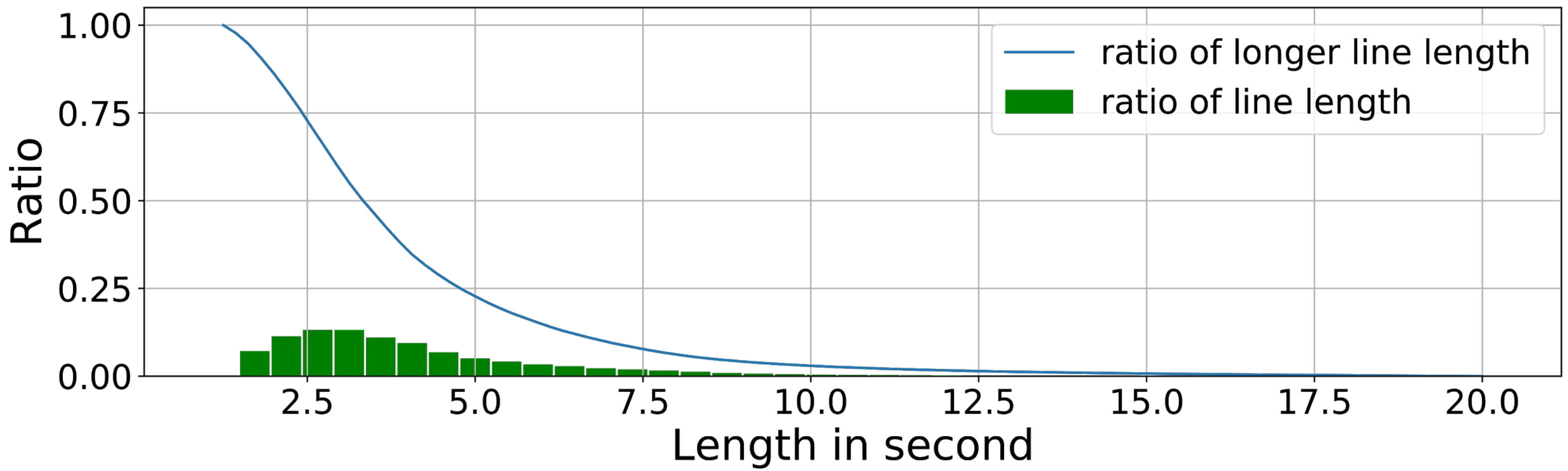


Dataset

44k songs, various genres, English lyrics

Most lyrical lines are quite short

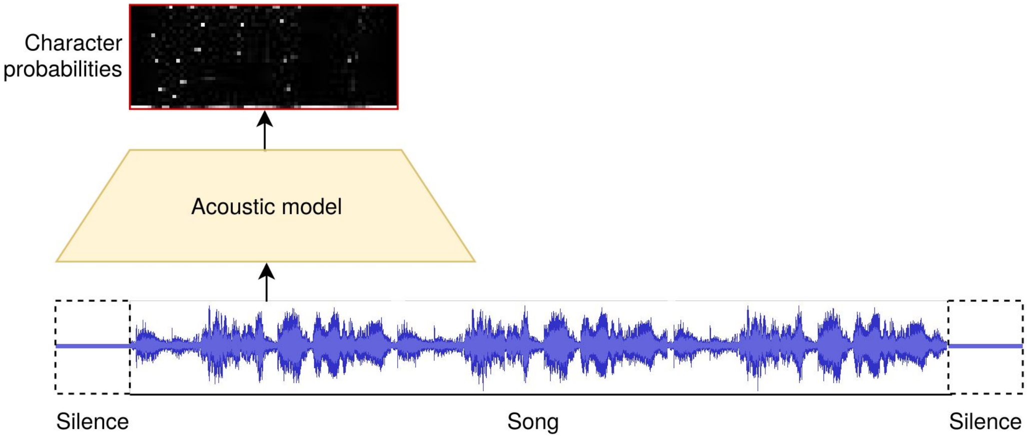
→ Model with 15s music input, 10s lyrics output



Prediction

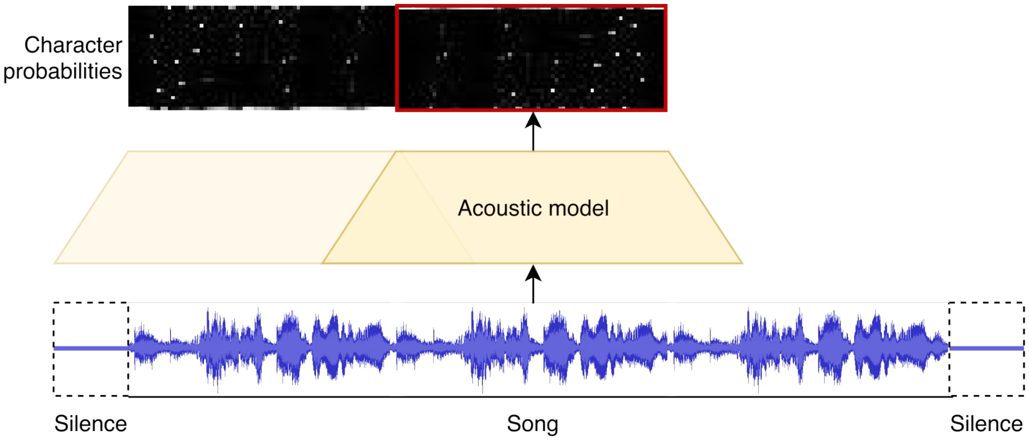
Predicting characters

1. Insert silence at start and end of song
2. “Slide” acoustic model across song
3. Collect character probabilities



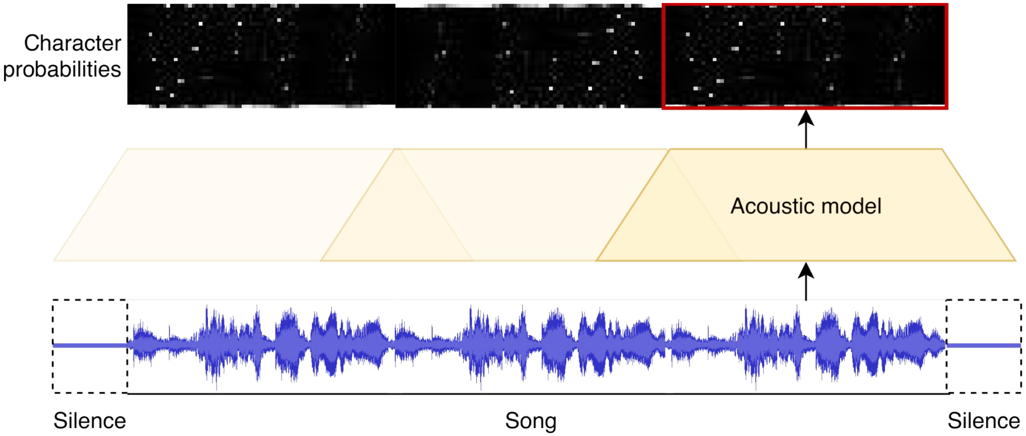
Predicting characters

1. Insert silence at start and end of song
2. “Slide” acoustic model across song
3. Collect character probabilities



Predicting characters

1. Insert silence at start and end of song
2. “Slide” acoustic model across song
3. Collect character probabilities



Predicting alignments

Find alignment with **maximum probability** under acoustic model:

$$\tilde{y} := \underset{\hat{y} \in \hat{C}^T, B(\hat{y}) = y}{\arg\max} \prod_{t=1}^{T} P_{t,\hat{y}_t}$$

Dynamic programming for exact solution in O(TL) time

- T = No. of time frames in the song
- L = Length of lyrics sequence

Predicting alignments

Find alignment with **maximum probability** under acoustic model:

$$\tilde{y} := \underset{\hat{y} \in \hat{\mathcal{C}}^T, B(\hat{y}) = y}{\arg\max} \prod_{t=1}^{T} P_{t, \hat{y}_t}$$

Dynamic programming for exact solution in O(TL) time

- T = No. of time frames in the song
- L = Length of lyrics sequence

Predicting lyrics

Find lyrics with

1. Maximum probability (beam search) or
2. Additional language model

Evaluation

Evaluation datasets

Annotations of word onset times from:

Mauch [5]

- 20 Songs, restrictive copyright
- Pop
- Polyphonic
- English

Jamendo (new dataset)

- 20 Songs, Creative Commons license
- 10 Genres (Western)
- Polyphonic
- English

Released for public usage at
https://github.com/f90/jamendolyrics

5. Matthias Mauch, Hiromasa Fujihara, and Masataka Goto, "Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations", in Proceedings of the Sound Music Computing Conference (SMC), 2010, pp.9–16

Alignment results

	Mauch						Jamendo
Metric	AK1	AK2	AK3	DMS1	DMS2	Ours	Ours
AE							
Perc							

Metrics:

- Average absolute error (**AE**)
- Average percentage of time in a song that the predicted position in the lyrics is correct (**Perc**)

Alignment results

	Mauch						Jamendo
Metric	AK1	AK2	AK3	DMS1	DMS2	Ours	Ours
AE	17.70	22.23	9.03	14.91	11.64	**0.35**	**0.82**
Perc	8.5	2.4	15.4	3.8	13.8	**77.2**	**70.4**

In comparison to MIREX 2017 methods:

- **alignment errors (AE) reduced more than ten-fold.**
- correct word predicted over 70% of the time (Perc)

Jamendo more difficult than Mauch, also due to Hip-Hop

Further improvement when using voice separation first

Transcription results

		Mauch		Jamendo	
Model	Decoder	WER	CER	WER	CER
Ours	Beam	80.4	**48.9**	84.4	**49.2**
Ours	LM	**70.9***	49.4*	**77.8**	50.2

* after optimising the language model on Mauch dataset

Half of characters in the lyrics correctly transcribed
High WER, but lower than previous works (94.5 [6], 77.1 [7])

LM improves WER

Jamendo again more difficult than Mauch dataset

6. Annamaria Mesaros and Tuomas Virtanen, "Automatic recognition of lyrics in singing", EURASIP Journal on Audio, Speech,and Music Processing, vol. 2010, no. 1, pp. 1, 2010.
7. Che-Ping Tsai, Yi-Lin Tuan, and Lin-shan Lee, "Transcribing lyrics from commercial song audio: the first step towards singing content processing," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5749–5753

Demo


baby youve heard me
when i get to warwick avenue

Conclusion

Conclusion

- Simple & end-to-end
- Strong alignment accuracy
- Supports weak labels
- Transcription needs more work