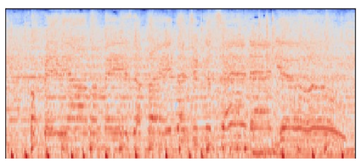
End-to-End Lyrics Transcription Informed by Pitch and Onset Estimation

Tengyu Deng Eita Nakamura Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Japan

Background

Lyrics Transcription



spectrogram

→

"but baby i'm amazed of the hate that you can send and you"

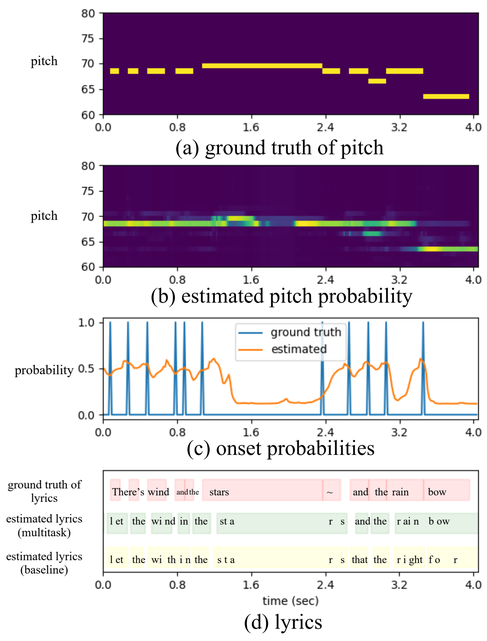
Lyrics

Use speech recognition methods?
Musical features are ignored

↓

Multitask with pitch recognition

Proposed Method



Transfer Learning with ASR Corpus

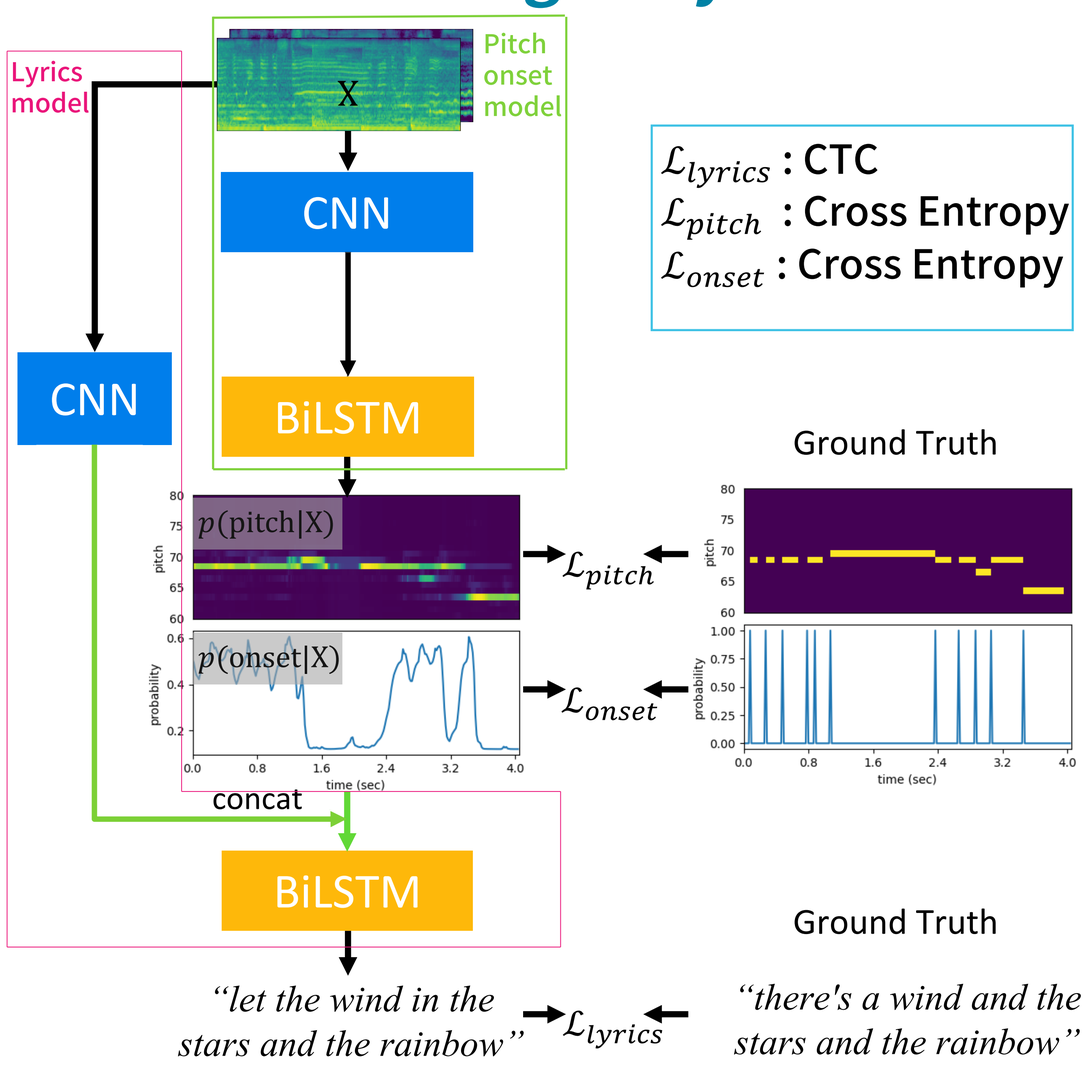
- Not enough lyrics data
- Drastic pitch changes
- Time-stretching nature of singing voice

↓

Transfer learning with ASR corpus

Learn lyrics transcription model with ASR corpus, and fine-tune with singing data

Multitask Learning of Lyrics Transcription and Pitch and Onset Recognition



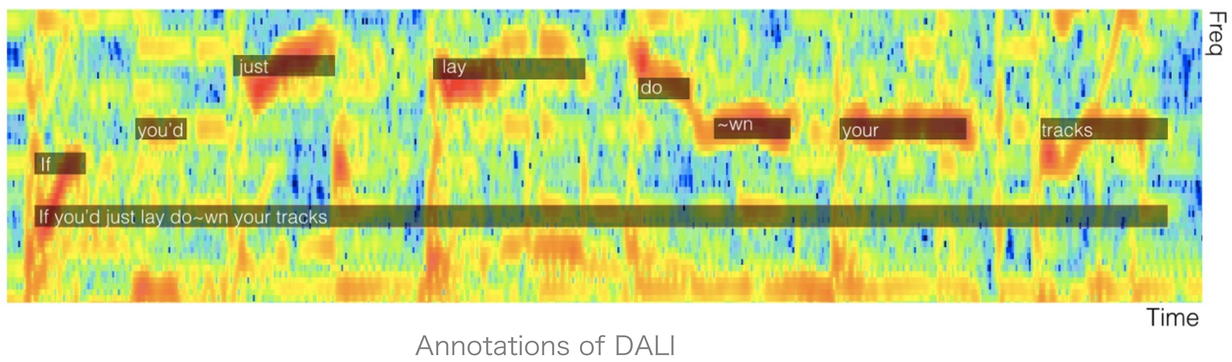
$\mathcal{L}_{lyrics}$: CTC
$\mathcal{L}_{pitch}$: Cross Entropy
$\mathcal{L}_{onset}$: Cross Entropy

Model

- Based on CRNN
- Multitask Learning
- Learn lyrics model with results of pitch recognition
- Transfer learning from ASR corpus

Data

- DALI[1]: lyrics transcription dataset
- With fine-grained lyrics and pitch annotations



Annotations of DALI

Data Pre-processing

- Remove tracks with bad annotations
- Using 2485 tracks from 5358 tracks of DALI
- 2237 train: 122 validation: 126 test
- Jamendo dataset for test
- For transfer learning, use LibriSpeech Data

Experiments

- With zero dummy pitch information
- With ground truth pitch information (Oracle)
- With estimated pitch information (multitask)
- Pitch recognition only

Results

Method	DALI-test	Jamendo
Ours (baseline)	69.22	77.3
Ours (oracle)	**64.41**	/
Ours (multi-task)	68.29	**76.2**
Wave-U-Net[2]	/	77.8

WER for each model (%)

	COn (%)			COnP (%)		
	precision	recall	F value	precision	recall	F value
Ours(baseline)	53.21	**30.99**	**38.77**	36.92	**21.49**	**26.90**
Ours(multi-task)	**59.84**	28.69	38.41	**40.49**	19.57	26.14
VOCANO[3]	18.78	20.45	19.07	7.46	7.71	7.40

COn and COnP F-values for each model

- Lyrics transcription accuracy is improved with pitch information
- Pitch detection accuracy remains the same

Future Works

- Evaluate lyrics alignment accuracy with proposed method
- Take more interaction between lyrics and pitch into consideration
- Data augmentation with pitch shift and time stretch

References
[1] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. In *Proc. of the 19th International Society for Music Information Retrieval Conference*, pages 431–437, Paris, France, 2018.
[2] Daniel Stoller, Simon Durand, and Sebastian Ewert. End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–185, 2019.
[3] Jui-Yang Hsu and Li Su. Vocano: A note transcription framework for singing voice in polyphonic music. In *Proc. of the 22nd International Society for Music Information Retrieval Conference*, pages 293–300, Online, 2021.