Emir Demirel

Topic: Automatic Lyrics Transcription and Alignment

Supervisors : Simon Dixon, Sven Ahlback

Queen Mary University of London

MIP-Frontiers

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068

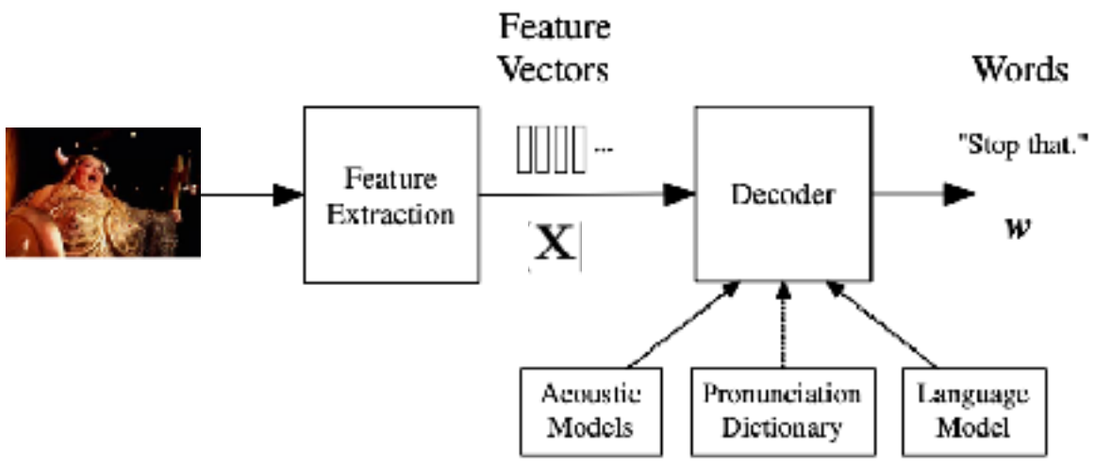
PART 2 :

Low Resource Audio-to-Lyrics Alignment

Contents

- Automatic Speech Recognition

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{w}) \max_{\mathbf{Q} \in Q_w} P(\mathbf{X}|\mathbf{Q}) P(\mathbf{Q}|\mathbf{w}),$$



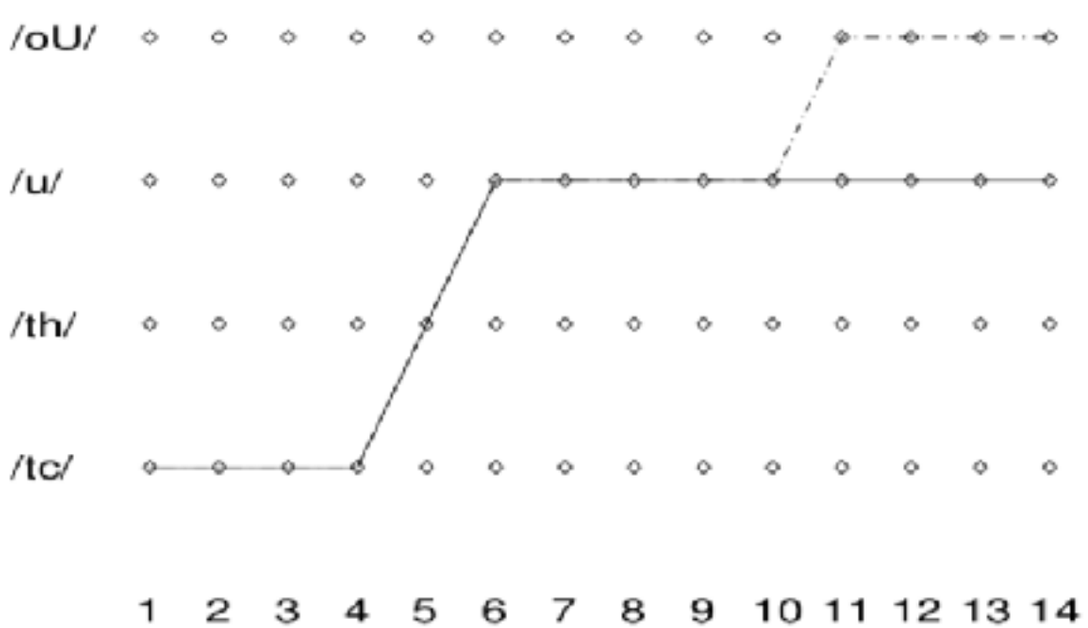
1 Architecture of a HMM-based Recogniser.

- Automatic Speech Recognition

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{w}) \max_{\mathbf{Q} \in Q_w} P(\mathbf{X}|\mathbf{Q}) P(\mathbf{Q}|\mathbf{w}),$$

w is given.

Audio-to-lyrics alignment



Forced Alignment. The correct utterance is two"; the recognized utterance is two oh".

- State-of-the-art in lyrics alignment applies forced alignment at a single pass with a beam size of 3000 *.

- This could be memory exhaustive for a recording of few minutes long.

(*) Gupta, Chitralekha, Emre Yılmaz, and Haizhou Li. "Automatic Lyrics Alignment and Transcription in Polyphonic Music: Does Background Music Help?." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020

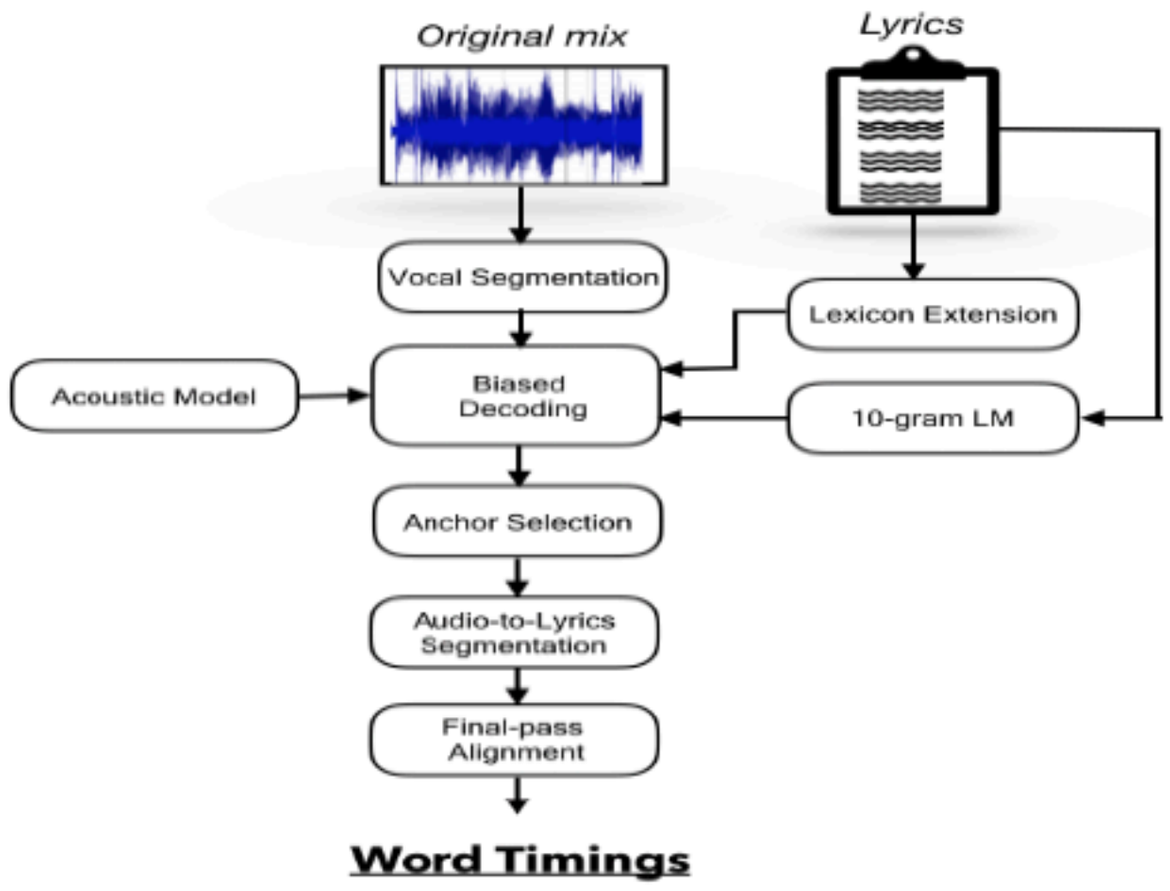


Fig. 1: Pipeline of our lyrics transcription and alignment system

Vocal Segmentation :

1) Vocal Source Separation

2) Log-energy based Vocal Activity Detection (VAD)

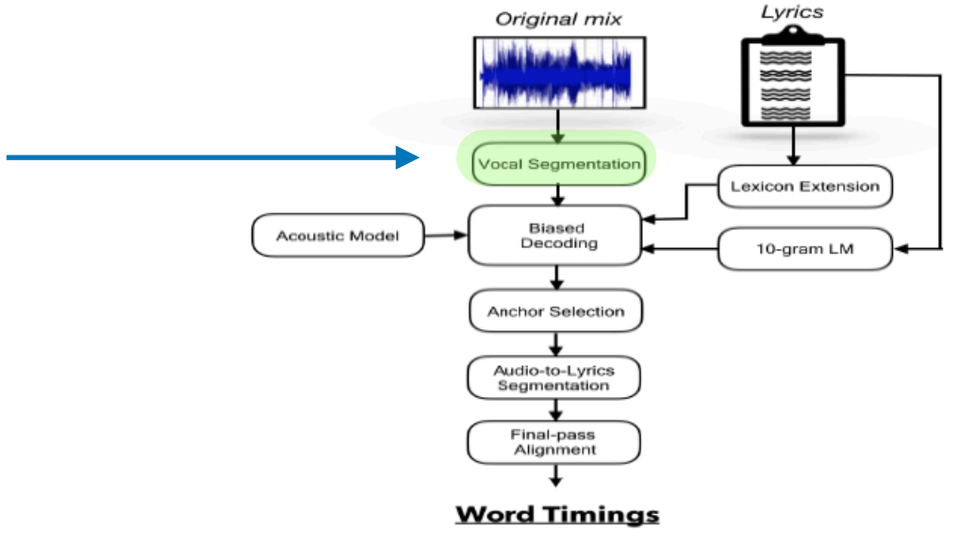
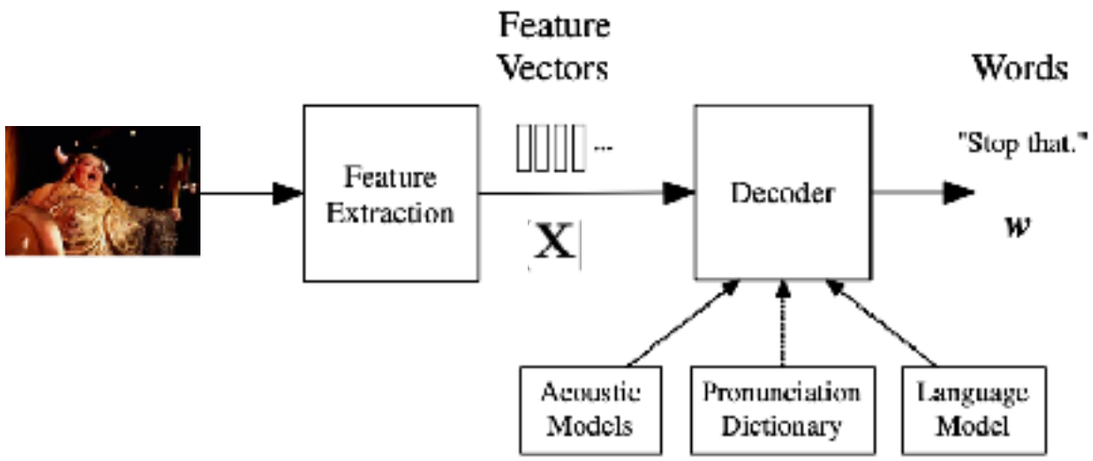


Fig. 1: Pipeline of our lyrics transcription and alignment system



l Architecture of a HMM-based Recogniser.

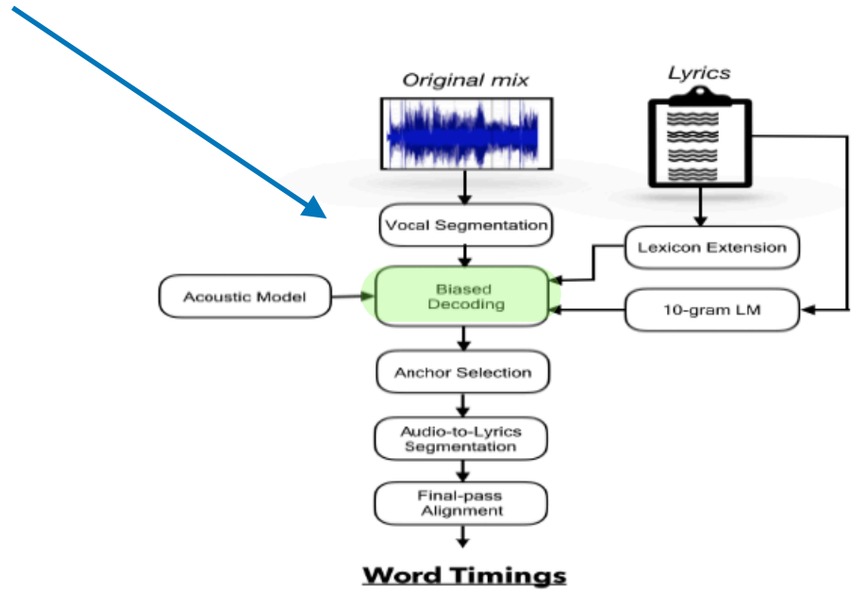
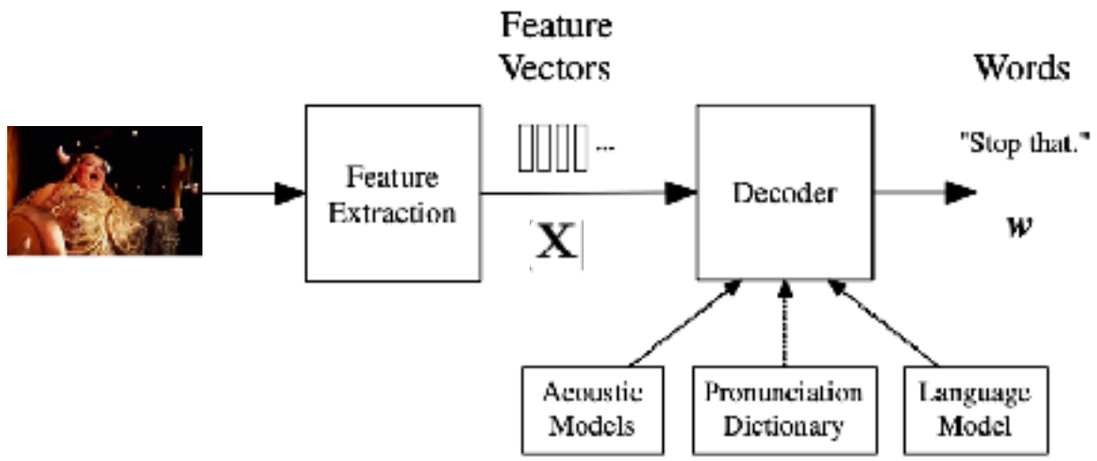


Fig. 1: Pipeline of our lyrics transcription and alignment system



Architecture of a HMM-based Recogniser.

Build a biased 20-gram LM

- Using the input song lyrics only

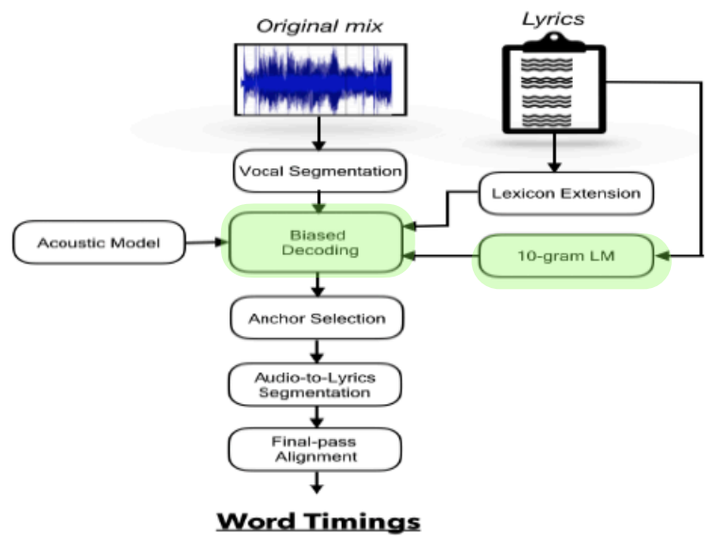
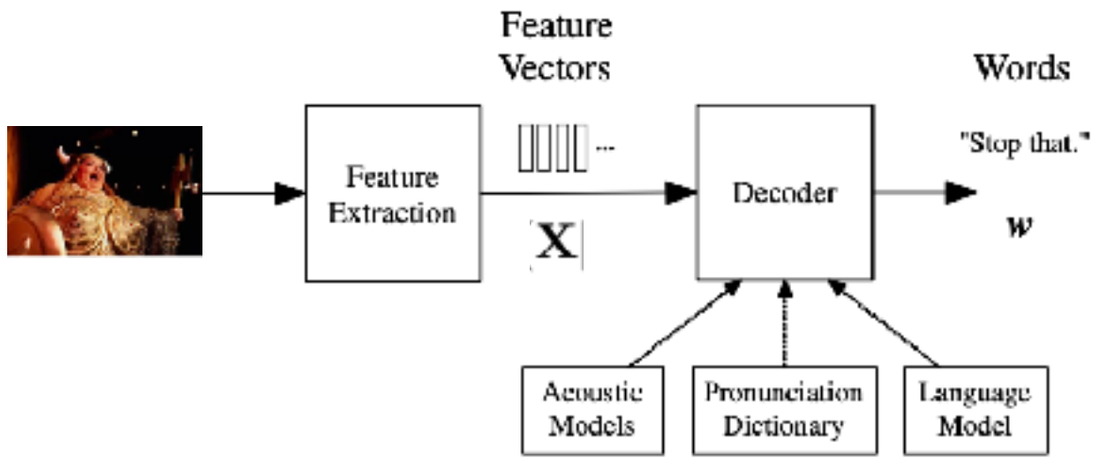


Fig. 1: Pipeline of our lyrics transcription and alignment system



Architecture of a HMM-based Recogniser.

Lexicon Extension

- Use a G2P model for out-of-vocabulary words

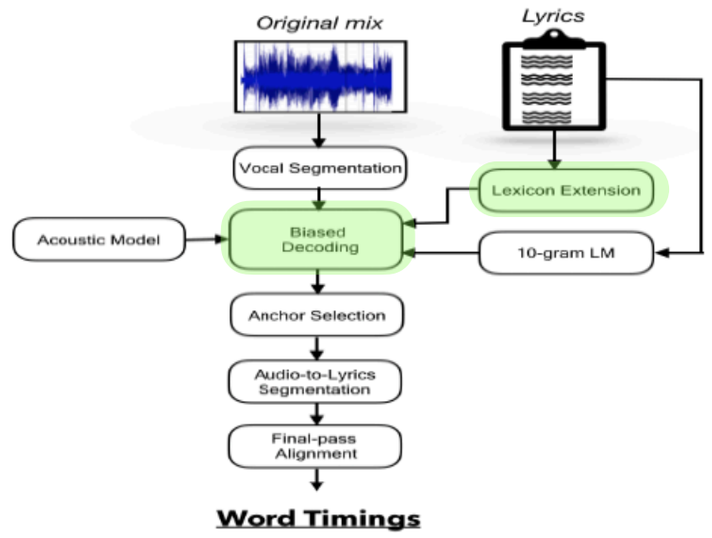
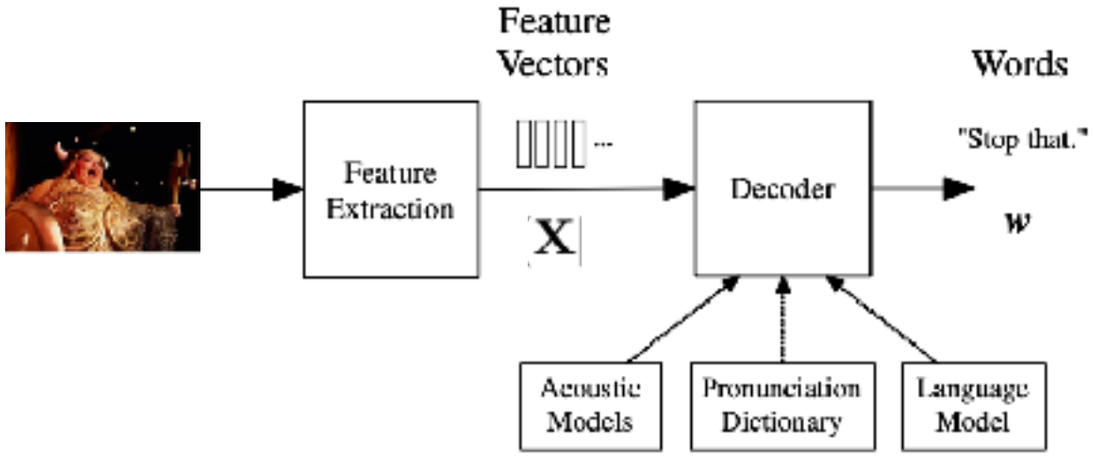


Fig. 1: Pipeline of our lyrics transcription and alignment system



Architecture of a HMM-based Recogniser.

Acoustic Model

ALTA - (A)utomatic (L)yrics (T)ranscription & (A)lignment

A kaldi recipe for automatic lyrics transcription and audio-to-lyrics alignment tasks.

If you use this repository, please cite it as follows:

```
@inproceedings{demirel2020,
  title={Automatic lyrics transcription using dilated convolutional neural networks with self-attent
  author={Demirel, Emir and Ahlback, Sven and Dixon, Simon},
  booktitle={International Joint Conference on Neural Networks},
  publisher={IEEE},
  year={2020}
}
```

Link to paper : https://arxiv.org/abs/2007.06486

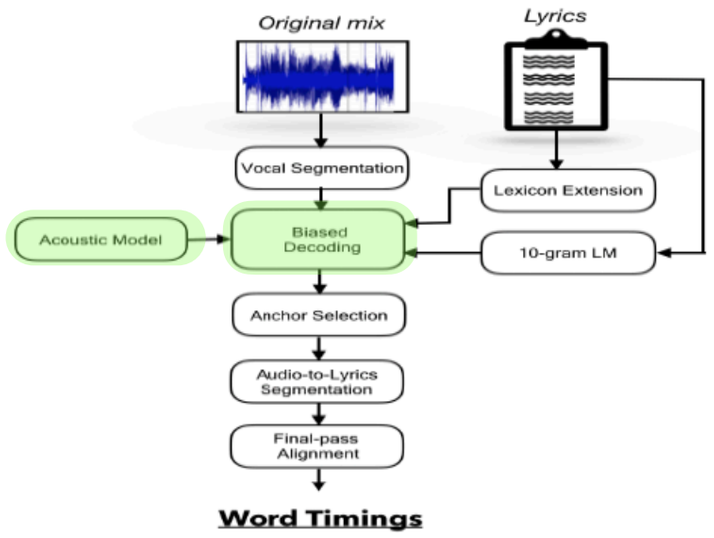


Fig. 1: Pipeline of our lyrics transcription and alignment system

Anchor Selection

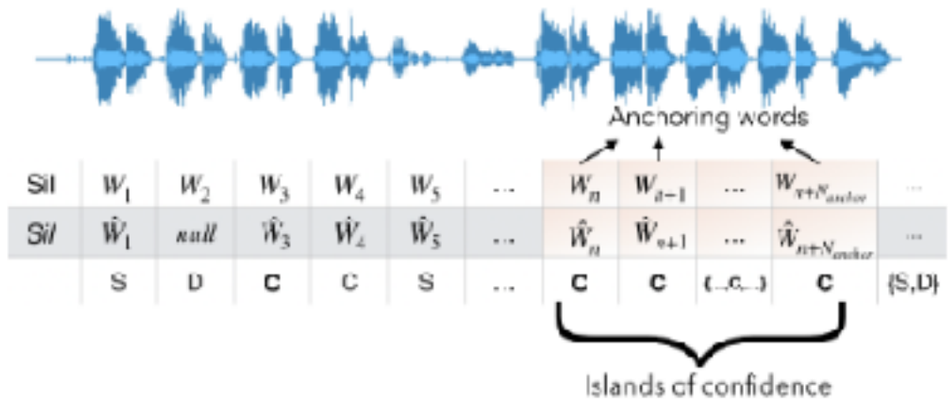


Fig. 2: Anchor selection. W_n and $\hat{W}_n$ are the reference and predicted words respectively. D and S stand for word deletions and substitutions after text alignment. C are the labels for correctly recognized (matching) words.

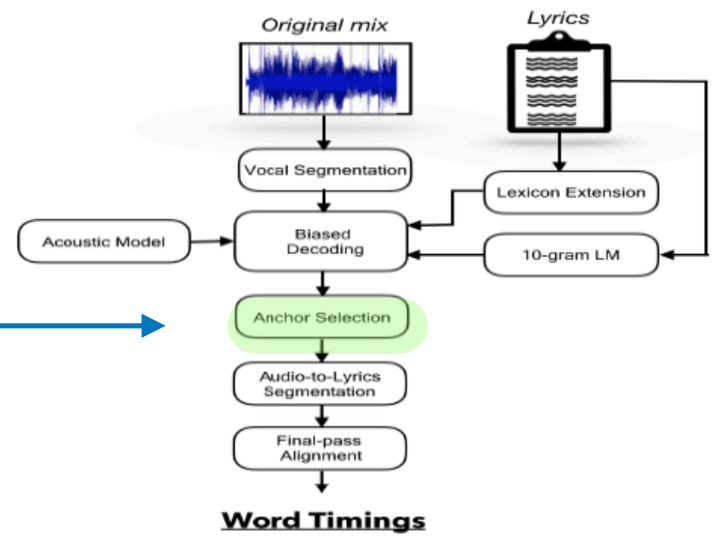


Fig. 1: Pipeline of our lyrics transcription and alignment system

Audio-to-Lyrics Segmentation

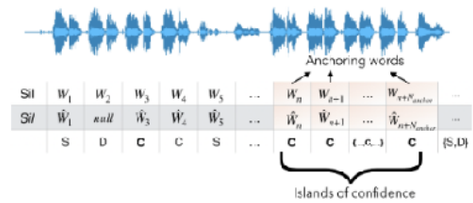
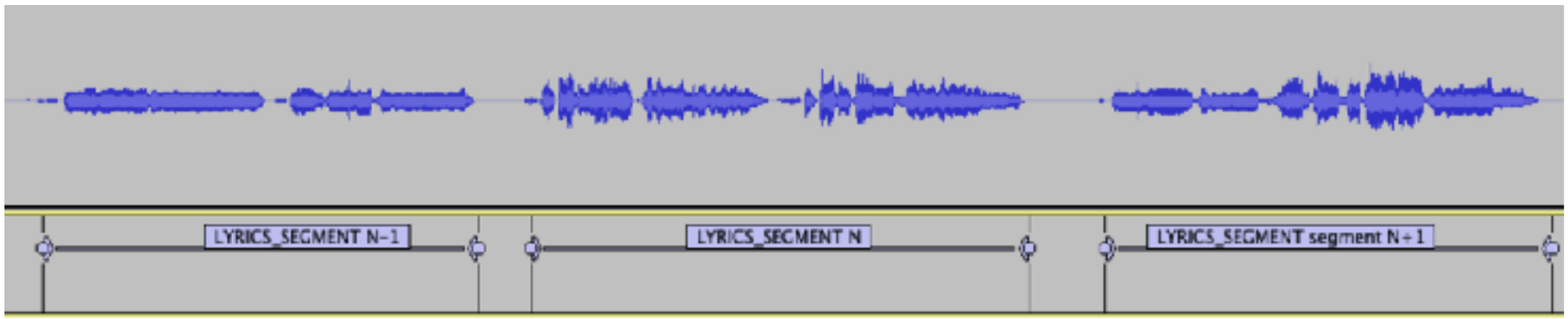


Fig. 2: Anchor selection. W_n and $\hat{W}_n$ are the reference and predicted words respectively. D and S stand for word deletions and substitutions after text alignment. C are the labels for correctly recognized (matching) words.



Start from the 1st word, progress linearly and segment once **k** number of anchoring words are met.

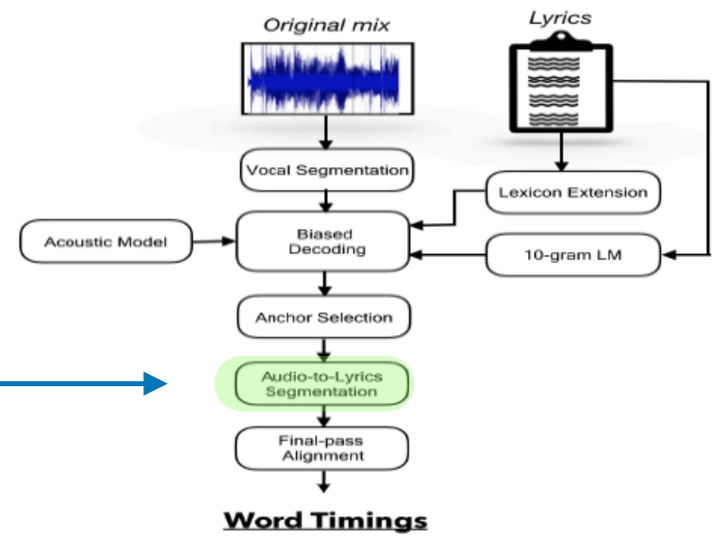
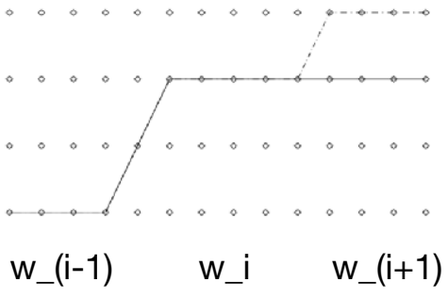


Fig. 1: Pipeline of our lyrics transcription and alignment system

Audio-to-Lyrics Segmentation

Final-pass Alignment



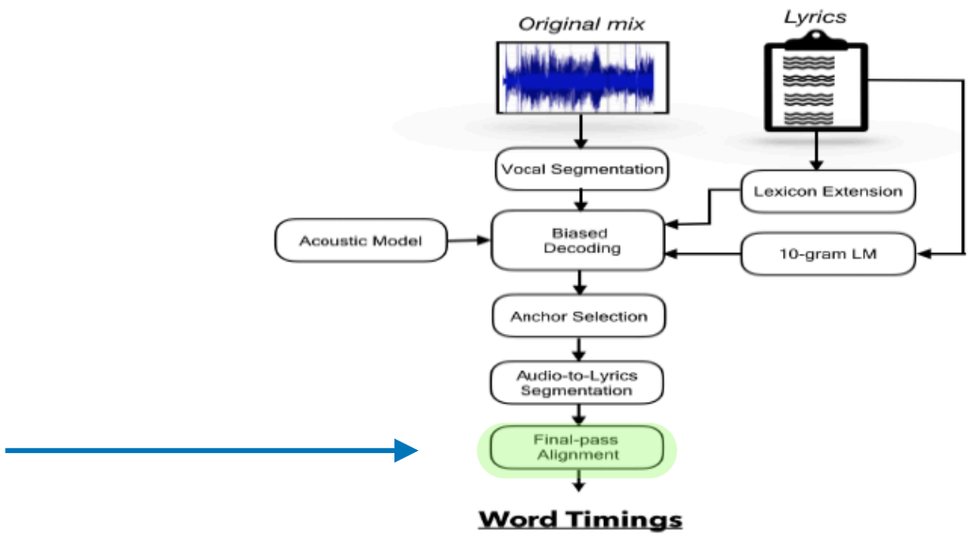


Fig. 1: Pipeline of our lyrics transcription and alignment system

Evaluation Set:

- ***JamendoLyrics*** dataset

(*https://github.com/f90/jamendolyrics*)

- 20 songs, not mainstream, open-source
- Polyphonic music - various genre including pop, hiphop, reggae, rock, metal
- English

Models

Trained on …

- DE1 - vocals separated with ‘demucs'
- DE2 - vocals separated with ‘spleeter’

DAMP-Sing! 300x30x2

- SD1 - E-2-E alignment polyphonic *
- SD2 - E-2-E alignment polyphonic separated with ‘U-net' *

Private Dataset

- GC1 - TDNN-f polyphonic **
- GC2 - TDNN-f vocals separated with ‘U-net’ **
- VA - E-2-E alignment separated with ‘spleeter’ ***

DALI_v1.0

(*) Stoller, Daniel, Simon Durand, and Sebastian Ewert. "End-to-End Lyrics Alignment Using An Audio-to-Character Recognition Model." (2019).

(**) Gupta, Chitralekha, Emre Yılmaz, and Haizhou Li. "Automatic Lyrics Alignment and Transcription in Polyphonic Music: Does Background Music Help?." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020

(***) Vaglio, Andrea, et al. "Multilingual lyrics-to-audio alignment."

Audio-to-Lyrics Alignment

Evaluation Metrics

- Average absolute error/deviation (mean, median) *
- Percentage of correct estimates (according to a tolerance window go 0.3 sec)**

(*) Mesaros, Annamaria, and Tuomas Virtanen. "Automatic alignment of music audio and lyrics." *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*. 2008.

(**) Mauch, Matthias, Hiromasa Fujihara, and Masataka Goto. "Integrating additional chord information into HMM-based lyrics-to-audio alignment." *IEEE Transactions on Audio, Speech, and Language Processing* (2011):

Audio-to-Lyrics Alignment

Results

	Mean AE	Median AE	PCS
SD1 [7]	0.82	0.10	0.85
SD2 [7]	0.39	0.10	0.87
VA [5]	0.37	N/A	0.92
GC1 [8]	**0.22**	**0.05**	**0.94**
DE1	0.31	**0.05**	0.93
DE2	0.38	**0.05**	0.90

Table 1: Lyrics alignment results on the Jamendo dataset

Audio-to-Lyrics Alignment

Memory Consumption

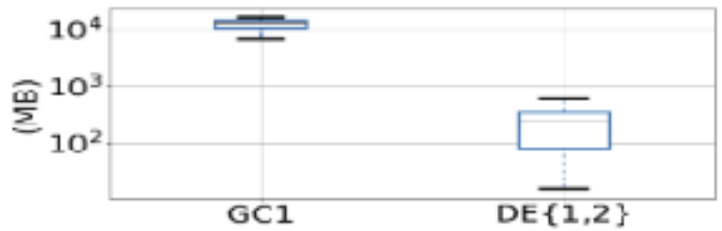


Fig. 3: Memory usage on RAM in megabytes (MB)

	GC1	DE{1,2}
Mean (Std.%)	13,740 (8.8%)	**343** (31%)
Max	16,745	**748**

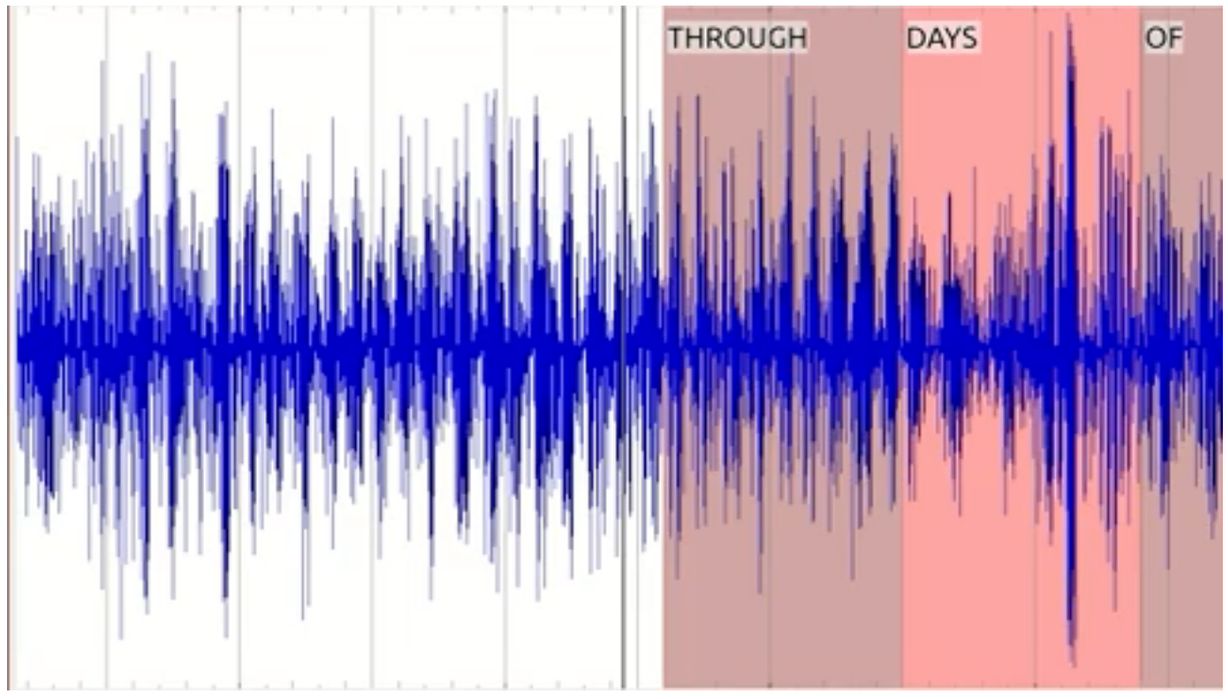
Table 2: Statistics on memory usage in MB

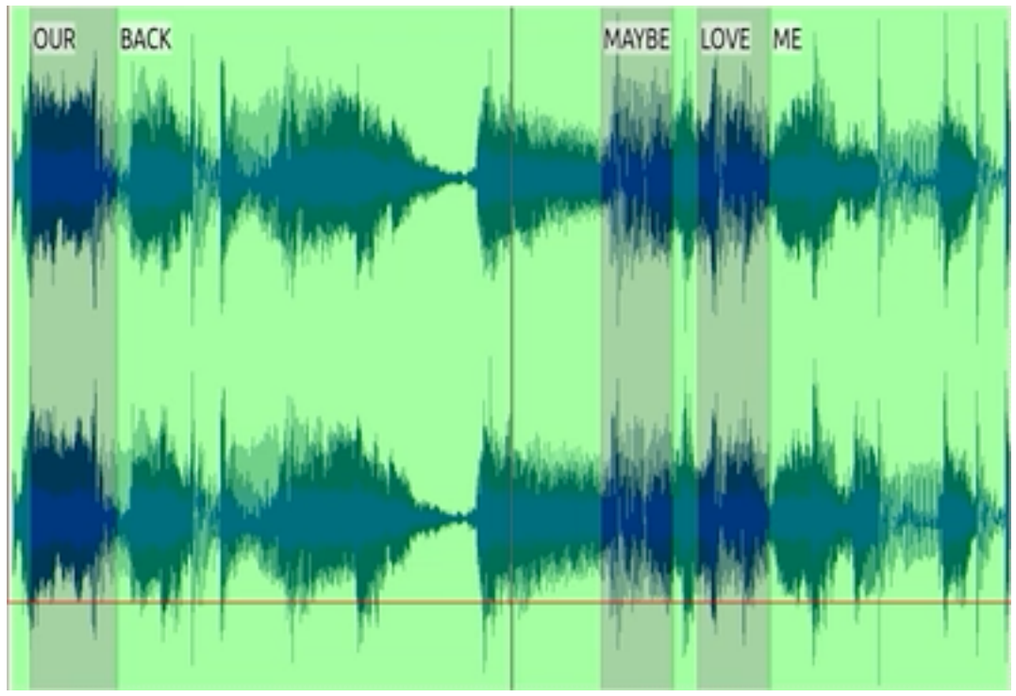
Automatic Lyrics Transcription

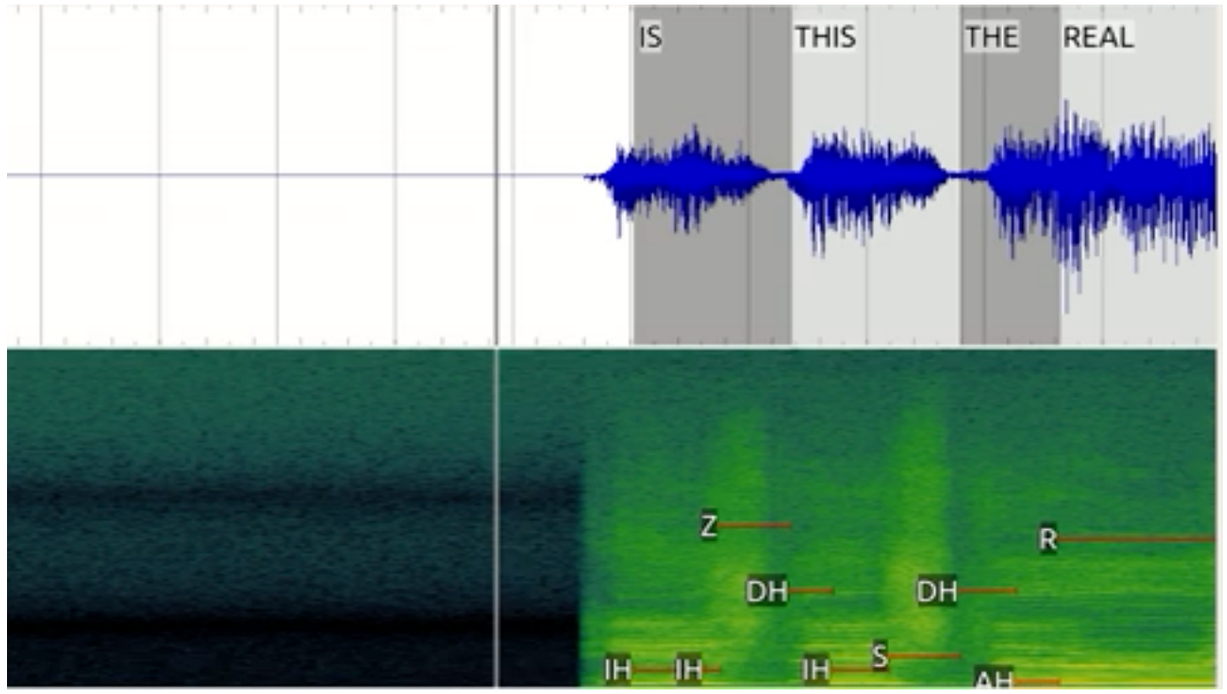
Results

	WER		*CER*	
	Mauch	Jamendo	Mauch	Jamendo
SD1 [7]	70.09	77.80	48.90	49.20
GC1 [8]	***44.02***	*59.57*	N/A	N/A
GC2 [8]	78.85	71.83	N/A	N/A
DE1 - VAR	60.92	62.55	44.15	47.02
DE1 - segmented	50.44	55.47	38.65	42.11
DE2 - VAR	57.36	**51.76**	**41.52**	**37.26**
DE2 - segmented	49.92	**44.52**	**38.**41	**32.90**

Table 3: Lyrics transcription results


THROUGH
DAYS
OF


OUR
BACK
MAYBE
LOVE
ME


IS
THIS
THE
REAL
Z
R
DH
DH
IH
IH
IH
S
AH

Conclusion

- Competitive results with the s.o.t.a
- A tool for automatically generating sentence-level annotations for lyrics transcription.
- Possibility for lyrics alignment in zero-resource languages.
- Importance of vocal source separation.
- Reported best transcription results on a public benchmark evaluation dataset.
- Red AI vs. Green AI ?

Emir Demirel

Topic: Automatic Lyrics Transcription and Alignment

Supervisor: Simon Dixon

Queen Mary University of London

e.demirel@qmul.ac.uk

Q&A?