Multilingual lyrics-to-audio alignment

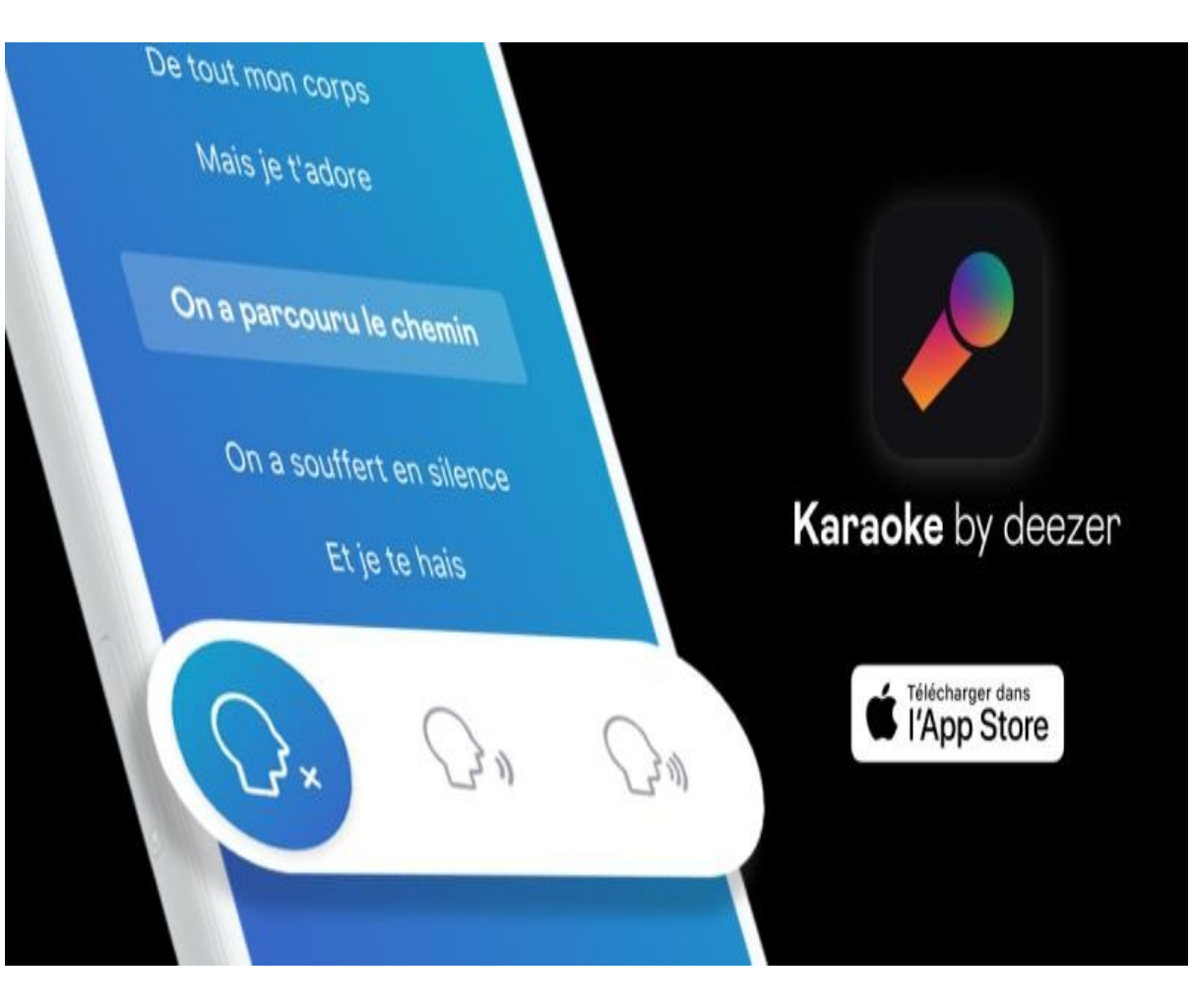
Andrea Vaglio[†*], Romain Hennequin[*], Manuel Moussallam[*], Gaël Richard[†], Florence d'Alché-Buc[†]

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

ISMIR 2020
Virtual Conference
October 11-15 2020

Lyrics-to-audio alignment

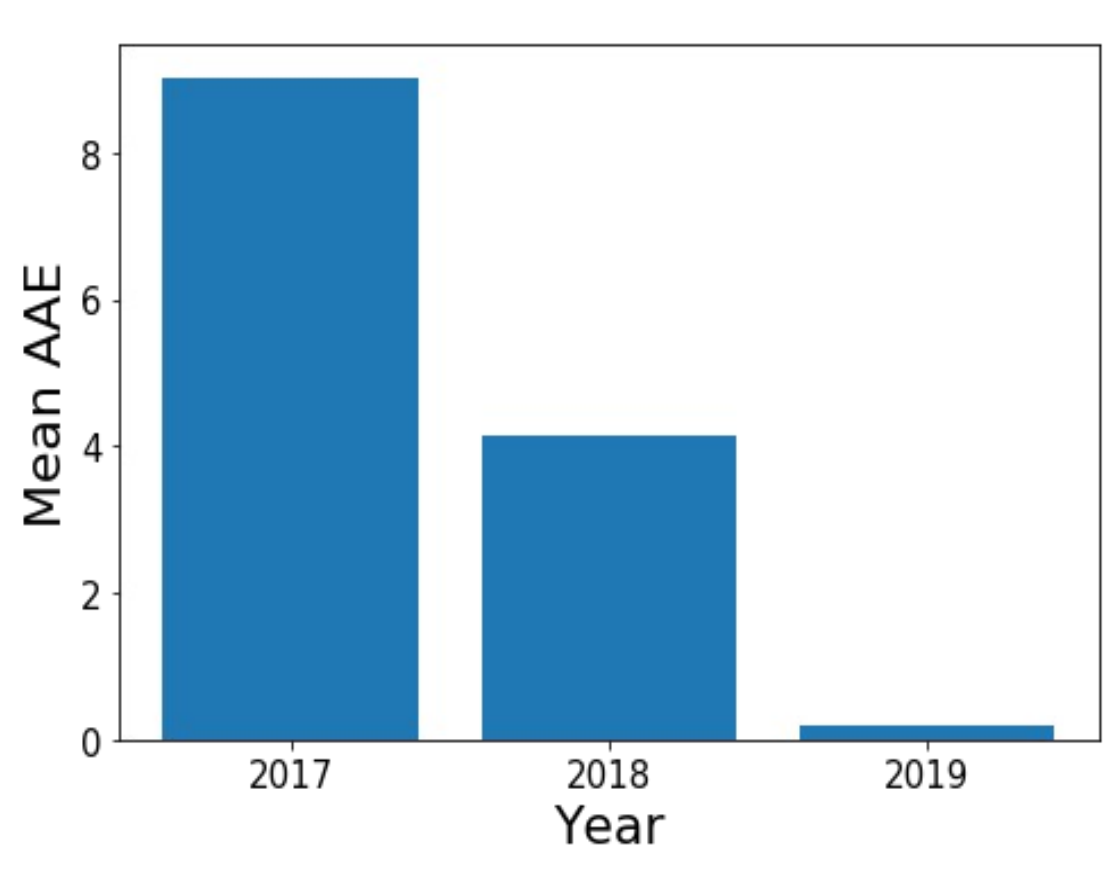
Synchronizing lyrics text units such as paragraph, line or word to the **timed position** of their appearance in the audio signal



Efficient alignment methods proposed recently [Sto18, Gu20]

Focus only on **the English language**, for which annotated data is abundant

Ability to **generalize to other languages?**



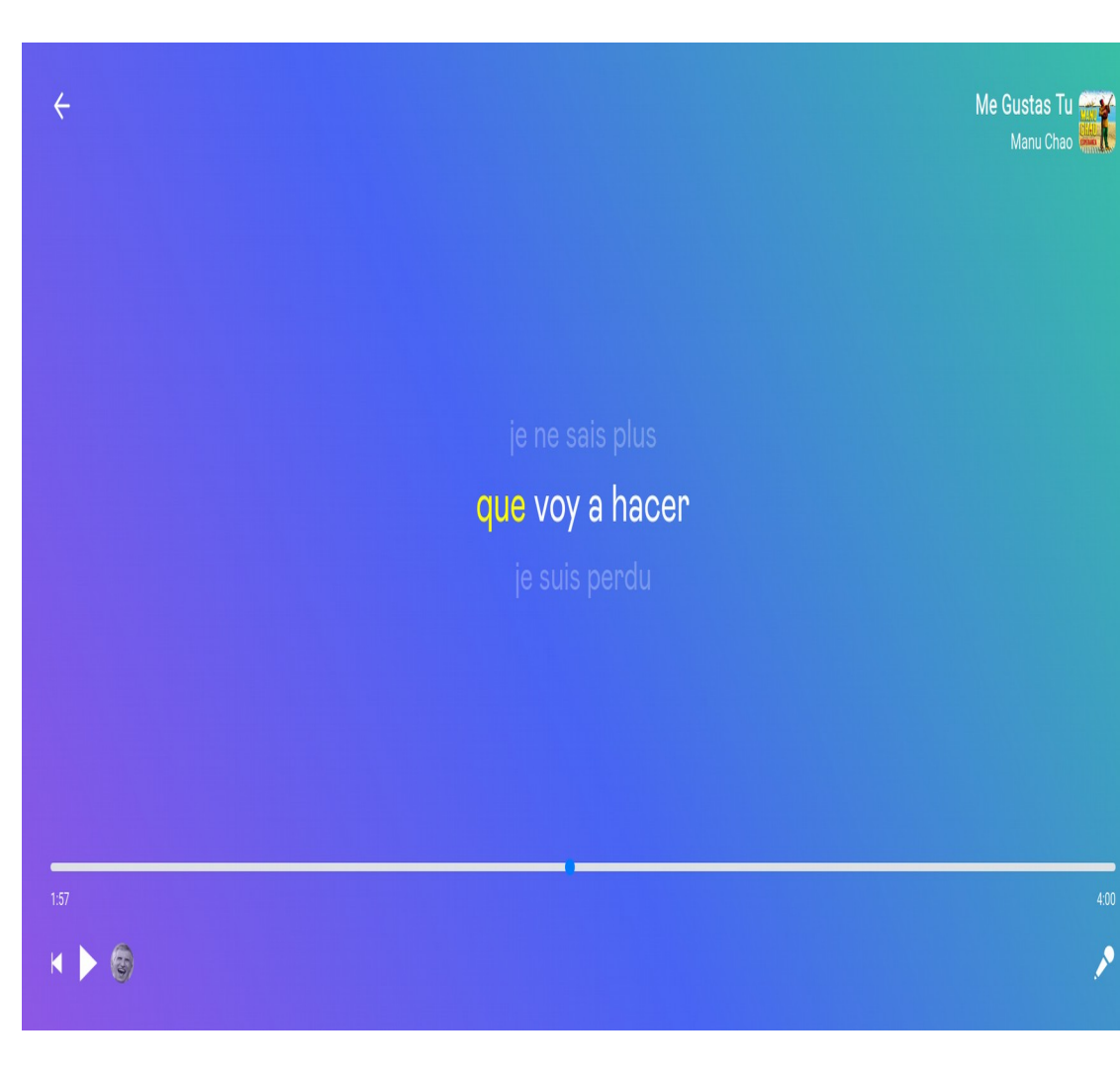
Best MIREX scores on Mauch dataset*

* Taken from https://www.music-ir.org/mirex/wili/MIREX_HOME

Objectives

Propose an alignment model than can handle multiple languages

First attempt at creating a language independent lyrics-to-audio alignment system

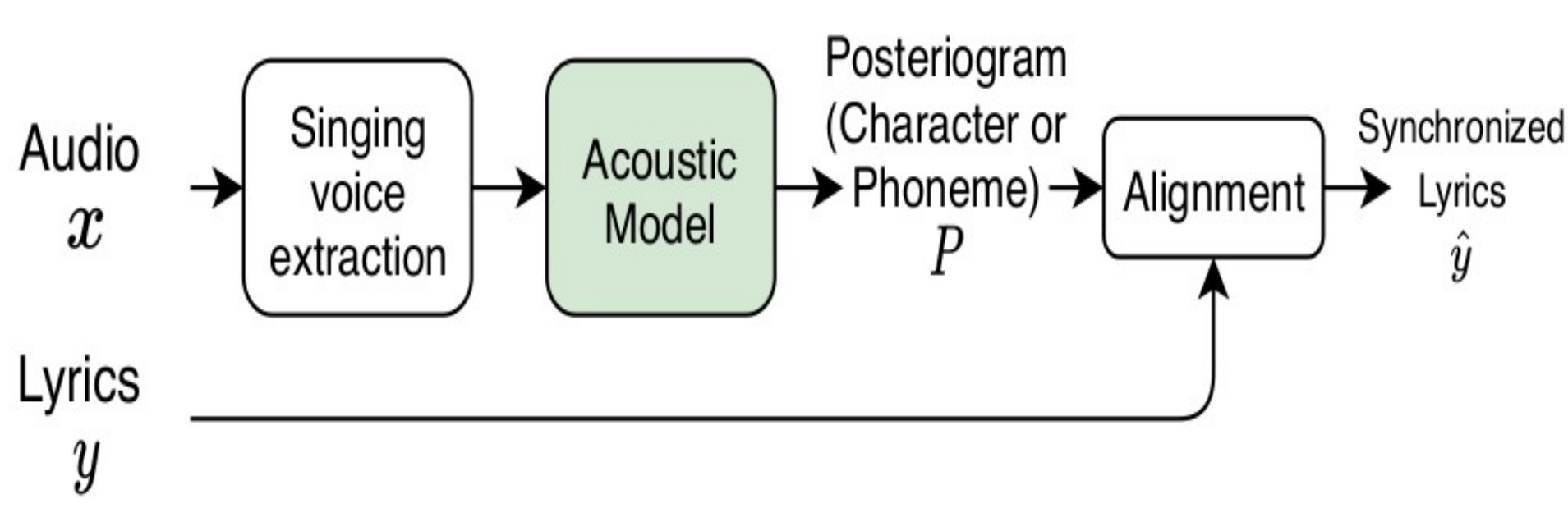


Study's workplan

Reviewing the **fitness of state-of-the-art systems to the multilingual framework**

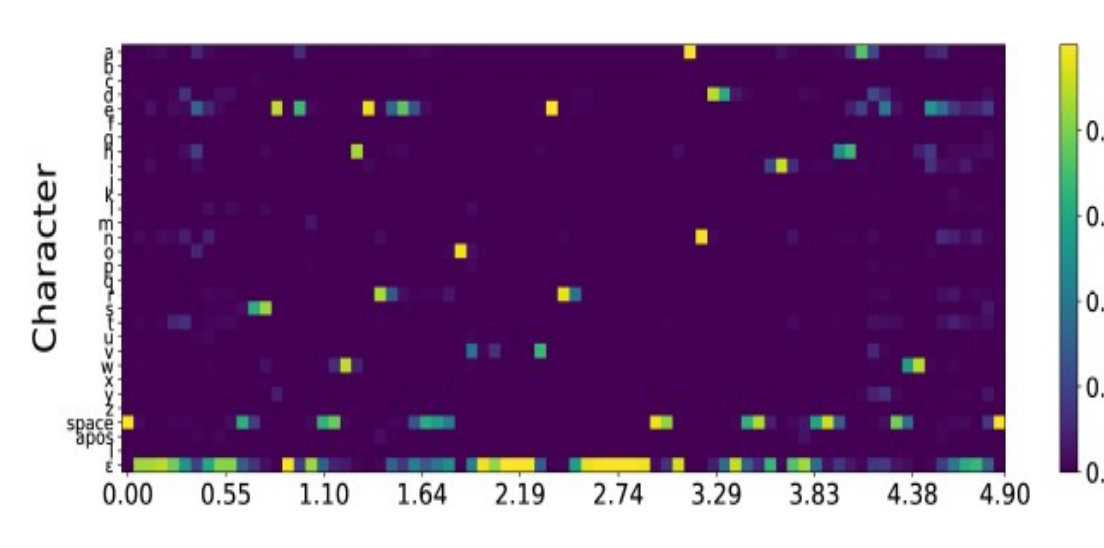
Study **two key features** likely to allow multilingual generalization
- **Intermediate representation**
- **Design of the training dataset**

Overview of the proposed system



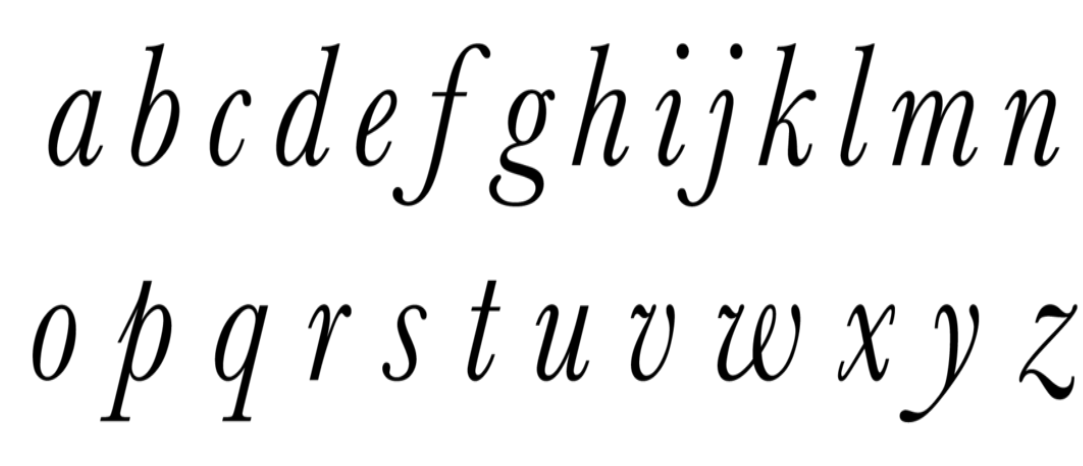
Acoustic model: RNN composed of **BILSTM layers**

Trained with a **Connectionist Temporal Classification (CTC) loss** [Gr06]



First intermediate representation considered is a **character set**

Limited for multilingual framework

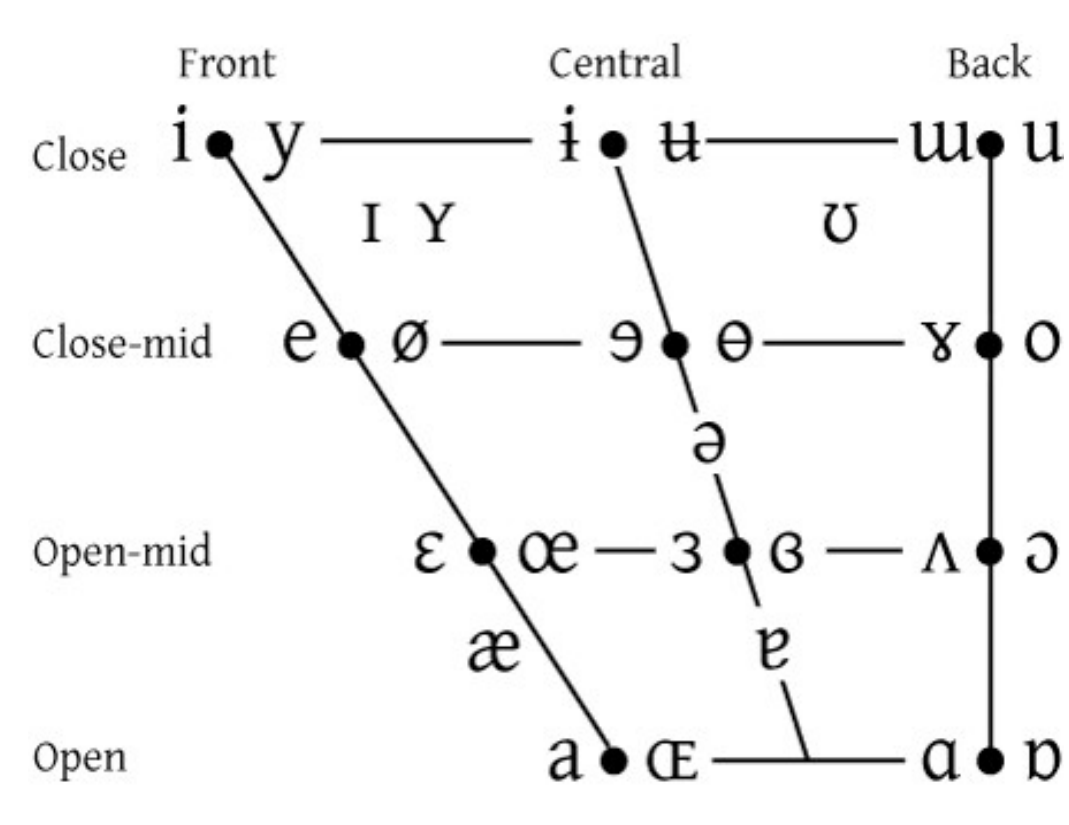


Latin alphabet⁺

Second intermediate representation is an **universal phoneme set**

Exploit **similarities between sounds** across langages [Sch01]

Based on the **International Phonetic Alphabet (IPA)**



IPA vowels chart*

⁺ Taken from https://en.wikipedia.org/wiki/Latin_alphabet, work in the public domain

* Created by Nohat Grendelkhan, released under the GNU Free Documentation License, taken from https://commons.wikimedia.org/wiki/File:Ipa-chart-vowels.png

Datasets

Various **language-subsets** of DALI dataset* [MbCh18]

English dataset is the largest one

Zero-resource languages, only used for evaluation

Dataset 5lang created for multilingual training

Language	# Phonemes	Train (h)	Test (h)
English	44 (5)	192.7	31.5
German	44 (1)	17.4	2.3
French	42 (0)	8.9	0.9
Spanish	35 (3)	8.4	1.1
Italian	33 (0)	8.5	1.2
Portuguese	37 (0)	X	1.8
Polish	31 (2)	X	4.2
Finnish	25 (0)	X	3.1
Dutch	41 (2)	X	3.1

Description of DALI language subset datasets and corresponding phoneme dictionary sizes

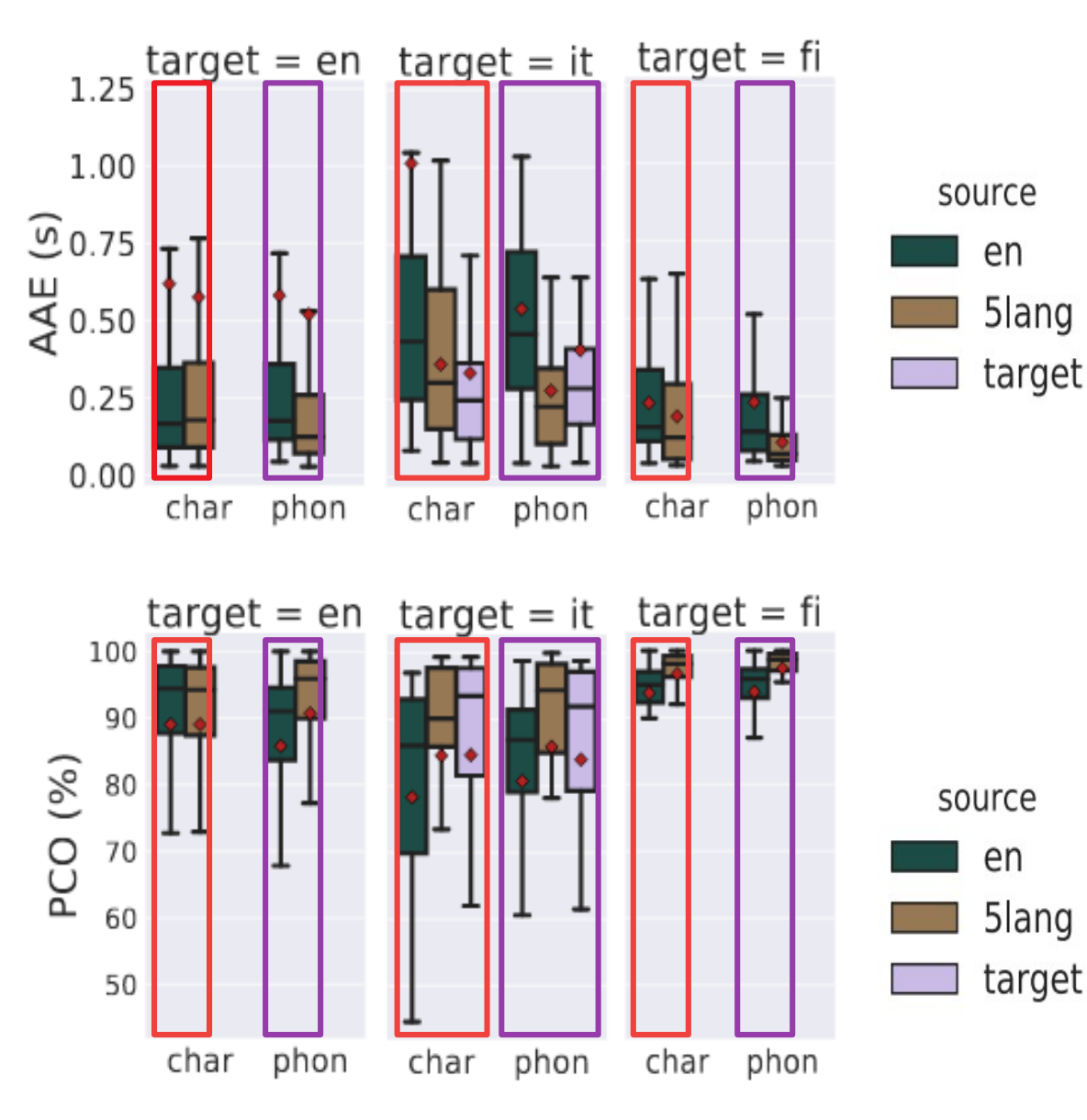
* Dali ids belonging to each dataset are made publicly avalaible at https://github.com/deezer/MultilingualLyricsToAudioAlignment

Results

Lyrics-to-audio evaluation on DALI language subset datasets for phoneme and character based architectures. AAE is better if smaller, PCO is better if larger

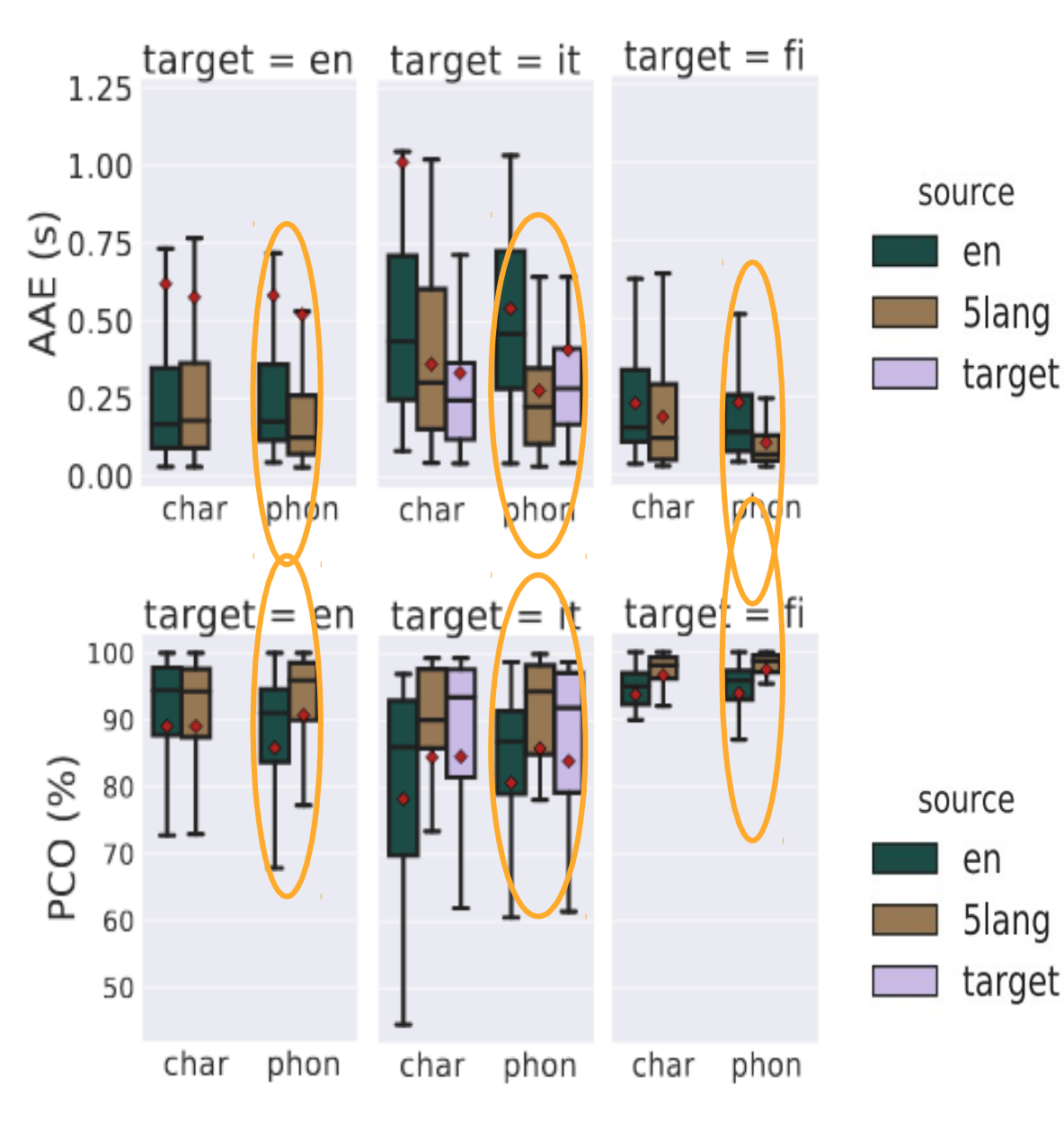
Multilingual training set helps for both character and phoneme architectures

Diverse training set **improves performances in ALL cases**



Phonemes over characters as an intermediate representation has **better performances**

Phoneme representation helps transfer knowledge between languages



Conclusion

Extend state-of-the-art methods to **multilingual context**

Learning from diverse data and using **an universal phoneme set** yield the best generalization performances

* Image free of licence, taken from https://pixabay.com/fr/illustrations/jeune-fille-chant-musique-femmes-1292241/

Limitations

Small set of languages considered
- Only in latin script

Additional experiments on a wider, **more diverse set of songs of various scripts** remain to be conducted

Paper available
https://research.deezer.com/

References

- [Sto18] 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.
- [MbCh18] 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 ISMIR, 2018.
- [Gu20] Chitralekha Gupta, Emre Yılmazand Haizhou Li. Automatic Lyrics Transcription in Polyphonic Music : Does Background Music Help? In ICASSP, 2020.
- [Gr06] Alexander Graves and al. Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Network. In ACM, 2006.
- [Sch01] Tanja Schulz and Alex Waibel. Language-Independent and Language-Adaptive Acoustic Modeling for Speech Recognition. In speech communication, 2001.

deezer TELECOM Paris