EXPANDING THE VOCADITO DATASET BY EXPERIMENTING WITH AUTOMATIC LYRICS & PHONEMES ALIGNMENT ALGORITHMS

Thomas Le Roux
Student at the Sound & Music Computing Master's degree
Universitat Pompeu Fabra
Barcelona
thomas.leroux01@estudiant.upf.edu

March 28, 2023

ABSTRACT

In this project, we apply 2 state of the art algorithms to a selected MIR task : lyrics and phoneme alignments for a monophonic singing task. Those algorithms are applied on a dataset with no ground truth for comparison. We therefore chose to recreate a false ground-truth by using an onset detection algorithm. The results obtained are visually satisfying, and when compared to the "false ground truth", we nearly retrieve the papers results applied on a new dataset.

Keywords Lyrics alignment · Phonemes alignment

1 Introduction

With the advent of machine learning use in scientific study of music, there is a strong need for well annotated, important and diverse dataset. The more informations available, the more accurate the models can be.

Voice is presumed to be the oldest musical instrument, and singing has been universally present in our human culture. In order to analyze and understand singing, scientists have gathered and published various datasets, each one having specificities. While some of them are more annotated toward on lyrics transcription (DAMP[1]), some of them are singing skill evaluation [Wilkins et al., 2018], or the comparison between speech and singing (NHSS [2]) ...

Vocadito [Bittner et al., 2021] is a dataset gathering short singing extracts, containing several annotations (cf section 2.1).

When presenting an example of their data collected on Vocadito, the authors mention that : "Lyric time-alignments were labeled here for demonstration purposes, but are not part of vocadito", cf Fig. 1

However, the authors did not provided the lyric alignment in the dataset. Those annotations would "add another analysis dimension possibility" to Vocadito. It could help understanding where a vocal note onset or offset should occur and why. This is a problem that have not been solved computationally yet, and that could be interesting to investigate.

In this project, we are going to apply several state of the art algorithm related to lyrics alignment on our dataset :

- Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation by [Schulze-Forster et al., 2020][3]
- Improving lyrics alignment through joint pitch detection by [Huang et al., 2022][4]

[1] https://ccrma.stanford.edu/damp/
[2] https://hltnus.github.io/NHSSDatabase/
[3] https://github.com/schufo/lyrics-aligner#readme
[4] https://github.com/jhuang448/LyricsAlignment-MTL

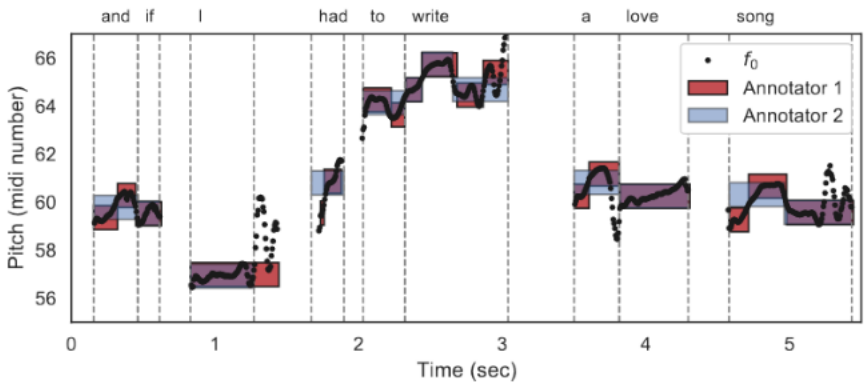

Figure 4. Annotations for the first 5.5 seconds of track 6 of vocadito. The plot shows f_0 annotations (black), note annotations by annotators 1 and 2 in red/blue respectively (overlaps shown in purple), and lyrics above the plot. Lyric time-alignments were labeled here for demonstration purposes, but are not part of vocadito.

Figure 1: Extract from the Vocadito paper showcasing what we want to reproduce

The main goal of this project is to attempt time-alignment of lyrics on the whole dataset.

The side goal is to create a Colab example for K. Schulze-Forster's algorithms, that doesn't exist yet. It will be shared with the author, and could potentially be used as a resource to give an example.

2 Methods / Algorithms

2.1 Vocadito

Vocadito is a small dataset released in 2021. , aiming at providing ground truth *f0*, note annotated by several professional annotators and lyrics for monophonic singing, sung in various languages with varying levels of training .

The dataset is agenced like this :

- Annotations
  - F0 : Folder containing the *f0* values computed for the whole song, with associated time stamps. The files have the name *vocadito_x_f0.csv*, with x being the file corresponding to the audio.
  - Lyrics : Folder containing the lyrics, under the name *vocadito_x_lyrics.txt*, with x being the file corresponding to the audio. The files only contain the lyrics, no time stamp or onsets.
  - Notes : Folder containing the notes annotated by 2 experts, under the name *vocadito_x_notesAy.csv*, with x being the file corresponding to the audio, and y being the annotators identifying number. The notes annotated have a time stamp, a frequency and a duration associated.
- Audio : Folder containing all the audio files. They are all under the name *vocadito_x.wav*, with x being the way to link with the annotations file.

2.2 Paper #1 - Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation

This paper has been published in 2021, and brings to the table :

- Novel approach to lyrics alignment at phoneme level and lyrics-informed singing voice separation, based on a new mecanism : DTW-attention.
- Extension of the MUSDB[5] dataset with lyrics transcripts.
- Experimental evaluation for the approach introduced.

The approach is based on a newly introduced DTW-attention algorithm (attention-based Dynamic Time Warping algorithm), and matches the observed phoneme sequences (inputted data) to the observed audio frames. For the

[5] `https://sigsep.github.io/datasets/musdb.html`

phonemes alignment, he reaches a Percentage of Correctly Aligned Segments (PCAS) 85.94% for monophonic singing. For the word alignment, the method achieves up to 97% accuracy (0.3 seconds of tolerance).

2.3 Paper #2 - Improving lyrics alignment through joint pitch detection

This paper has been published in 2022, and is the current state of the art for lyrics alignment. The particularity of the method used is that a multi-task method is used, where the lyrics alignment is the primary goal and the pitch detection is a secondary goal. This idea is inspired by the works of [Paredes et al., 2012, Hung et al., 2019] that showed the multi-task learning can lead to improved accuracy, and that pitch could be a good candidate. This method achieves up to 95% accuracy for the Percentage of Correct Onsets (PCO) (0.3 seconds of tolerance).

2.4 Evaluation metric

Because the Vocadito dataset doesn't have a ground truth value, we choose to come with a "mock-up" ground-truth : the onsets.

Because our dataset is exclusively composed of monophonic singing, the only onsets are when singing occurs. However, onsets can also happen at other times (when changing syllable, when there's a sudden vibrato ... We need to find a way to not considerate those, and to only identify the correct onsets to compare them with the boundaries obtained from our algorithms. This is simply done by considering the closest onset to the boundary that we study. This can lead to some errors, but it still works pretty well for our small task.

We use Librosa's onsets detection function [6], and we will compare the 2 values in time.

We will have 2 main metrics :

- Mean difference between calculated between onsets and beginning of the word (in ms)
- F0-Score for correctly identified boundaries (algorithm vs ground truth (onset from librosa). For the F0 score, we tolerate up to 0.3s of delay, as it seems to be a standard measure for this task [Mauch et al., 2011]

3 Results

We can the results obtained in Table 1.

	Before eliminating outlier(s)				**After eliminating outlier(s)**					**Paper results**
	Mean Diff	**Variance**	**Mean F0-score**	**F0-variance**	**Mean Diff**	**Variance**	**Mean F0-score**	**F0-variance**		
Algo 1	101.9 ms	17934	96.49%	156.5	81.46 ms	1635	98.40%	24.11		Up to 93%
Algo 2	105.8 ms	126614	95.5%	232.7	48.9 ms	620	98.1	14		Up to 97%

Table 1: Results obtained for our 2 algorithms

3.1 Algo #1 :

For this algorithm, the process requires more steps :

- Setting up the environment is longer (requires creating a Conda environment within Python)
- The steps described in the GitHub asks the user to manually input the word list to an online *word-to-phoneme* database hosted by Carnegie Mellon University [7], and then to move the results in the coding environment. Using BeautifulSoup, a Python library that aims at parsing html data and by sending the right requests, we manage to automatize this process.
- The CMU database has issues with some words present in the lyrics (chinese characters but also *à* (french character) and *trempez-la* (issue with the linking '-' character).

After extracting the onsets from librosa, and applying the algorithm to our dataset, we found the following results that are presented in Fig. 2.

[6] `https://librosa.org/doc/main/generated/librosa.onset.onset_detect.html`

[7] `http://www.speech.cs.cmu.edu/tools/lextool.html`

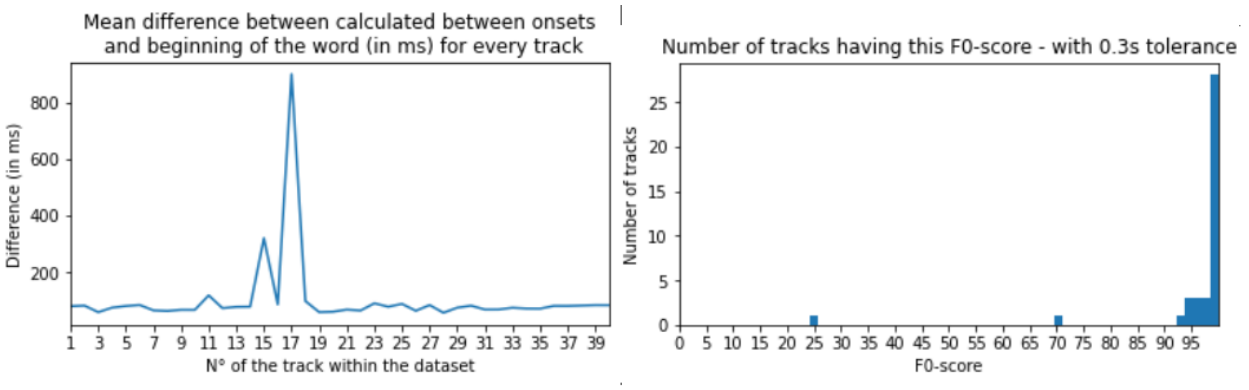

Figure 2: Results for the Algorithm #1

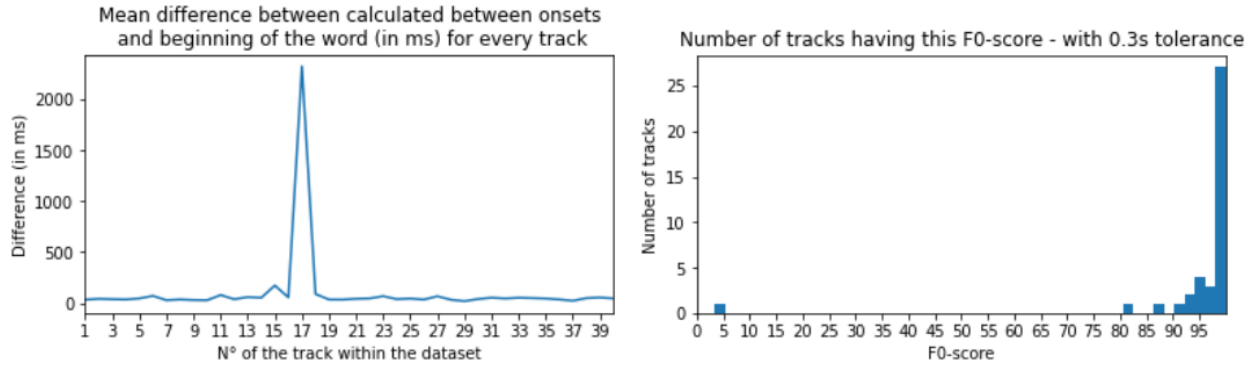

Figure 3: Results for the Algorithm #1

3.2 Algo #2 :

The process for this algorithm is rather straight forward, as the tutorial given with the paper is fairly clear, and doesn't require any automation.

After extracting the onsets from librosa, and applying the algorithm to our dataset, we found the following results that are presented in Fig. 3

4 Conclusion

4.1 Overall results

We can see in our results that we have outliers. We decide to remove them, and we can see that our results fairly improve.

Those outliers are probably due to the onset detection in librosa and by the way we link this onset with the beginning of the word detected by our algorithm.

The result we obtain when we try to reproduce those example given in Vocadito's paper are visible Fig. 4 and Fig. 5

We can see that it seems fairly accurate with what we have. The onsets seems to be rather close, and it looks like we are pretty accurate with the pitch information that could indicate the beginning of a word.

4.2 To conclude

We have implemented 2 state of the art algorithms, and have applied them on a dataset, where ground truth data was not present. The boundaries found at word level seem to be pretty accurate, and to be within the 0.3s tolerance margin.

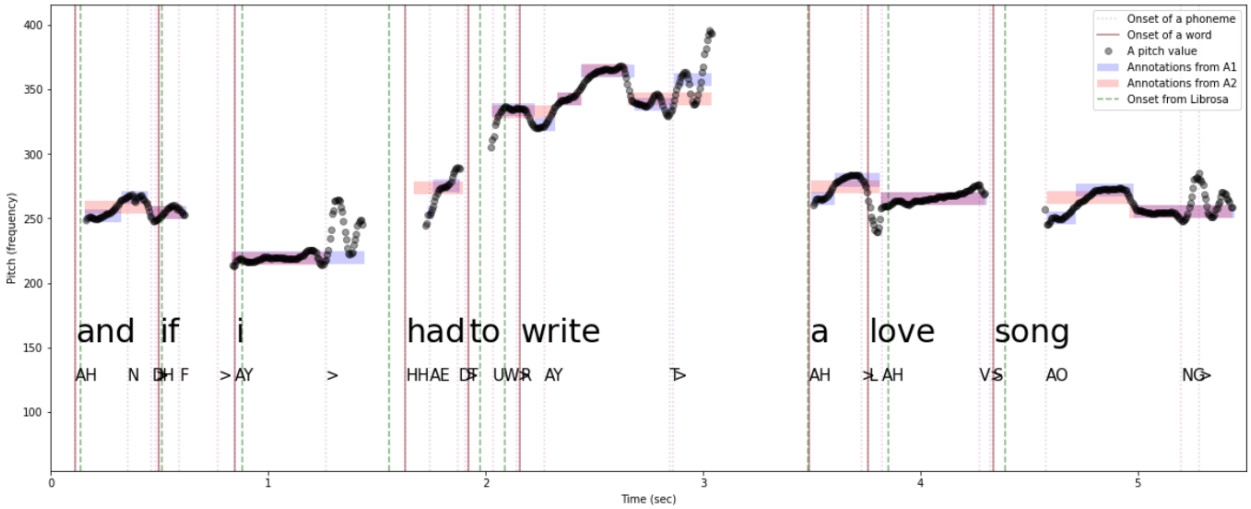

Figure 4: Extract from the Vocadito paper s with algorithm #1

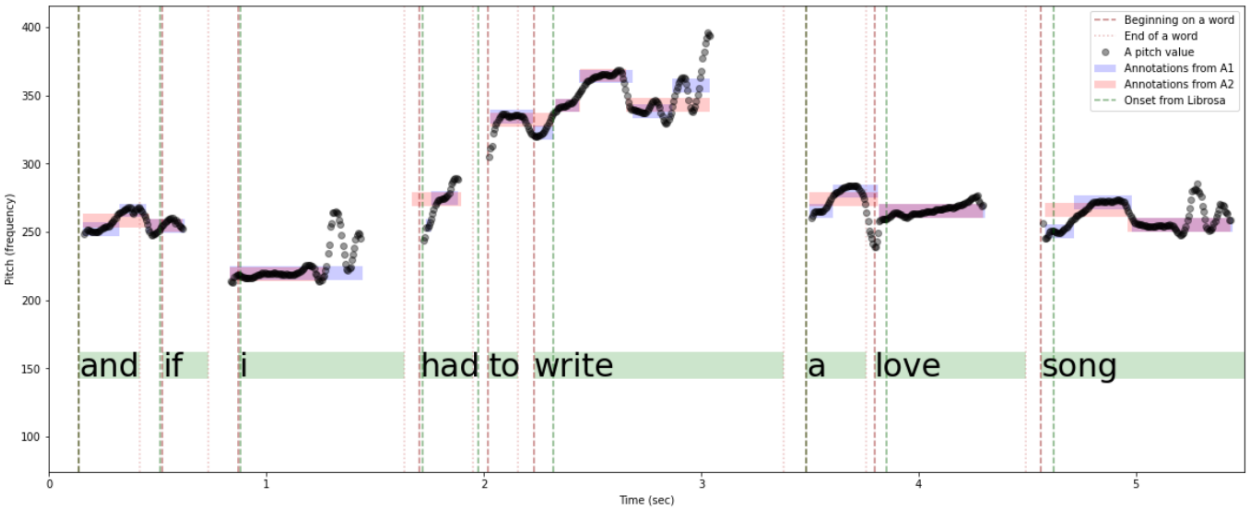

Figure 5: Extract from the Vocadito paper s with algorithm #2

For the phoneme level, it's more complex as the data is a bit more messy. We didn't developed a fake ground truth in order to have evaluation metrics to calculate the phonemes alignment, we therefore can only conclude by a visual evaluation on some samples. From what we've seen, it seems to be pretty accurate, and to fall within the 0.3s tolerance margin.

It would've interesting to manually annotate this data in order to compare it to ground truth, and have reliable results. Since we faked the ground truth in this study, we can't really draw proper conclusions.

References

Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. Vocalset: A singing voice dataset. In *ISMIR*, pages 468–474, 2018.

Rachel M Bittner, Katherine Pasalo, Juan José Bosch, Gabriel Meseguer-Brocal, and David Rubinstein. vocadito: A dataset of solo vocals with f_0, note, and lyric annotations. *arXiv preprint arXiv:2110.05580*, 2021.

Kilian Schulze-Forster, Clement SJ Doire, Gaël Richard, and Roland Badeau. Joint phoneme alignment and text-informed speech separation on highly corrupted speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7274–7278. IEEE, 2020.

Jiawen Huang, Emmanouil Benetos, and Sebastian Ewert. Improving lyrics alignment through joint pitch detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 451–455. IEEE, 2022.

Bernardino Romera Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. Exploiting unrelated tasks in multi-task learning. In *Artificial intelligence and statistics*, pages 951–959. PMLR, 2012.

Yun-Ning Hung, Yi-An Chen, and Yi-Hsuan Yang. Multitask learning for frame-level instrument recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–385. IEEE, 2019.

Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Integrating additional chord information into hmm-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):200–210, 2011.