Audio-Lyrics Alignment Dataset for Italian Arias

In memoriam of Rodolfo Delmonte 03.05.1946 – 12.03.2026

Pushkar Jajoria[1], Arianna Graciotti[2], Giovanna Casali[3], Jesujoba O. Alabi[1], Rodolfo Delmonte[5], Angelo Pompilio[3], Rocco Tripodi[5], James McDermott[4], Dietrich Klakow[1]

[1]Saarland University, Germany, [2]University of Groningen, The Netherlands, [3]University of Bologna, Italy, [4]University of Galway, Ireland, [5]Ca' Foscari University of Venice, Italy

pjajoria@lsv.uni-saarland.de, rocco.tripodi@unive.it, james.mcdermott@universityofgalway.ie

Abstract

Aligning song lyrics with sung audio is challenging, especially for languages and music styles where annotated datasets are scarce. We address this gap by presenting the first dataset of Italian opera arias annotated with lyrics and time-stamps per word. The dataset comprises of 24 arias drawn from well-known operas of the 18th to 20th centuries with a total audio duration of nearly two hours. We benchmark both music alignment models and speech forced alignment models and show that existing methods face significant challenges on this dataset, with performance dropping by 45% compared to other datasets. Multilingual and speech-based models exhibit relatively better performance on this dataset. We also evaluate few-shot fine-tuning on the new dataset and find that, while it yields only marginal overall improvement, it produces localized gains on specific arias, suggesting that limited exposure helps the model adapt to some patterns but cannot fully overcome differences in language or musical style.

Keywords: Italian Aria, Low-Resource Dataset, Audio-Lyrics Alignment

1. Introduction

The automatic alignment of song lyrics with their corresponding audio is an important task in the field of music information retrieval (Mesaros and Virtanen, 2008; Mauch et al., 2010; Fujihara and Goto, 2012). Audio-lyrics alignment means mapping textual lyrics to precise temporal locations in a song's audio signal, enabling applications such as music subtitling and lyrics-based audio search (Müller et al., 2007; Fujihara and Goto, 2012; He et al., 2023). In recent years, substantial progress has been achieved for audio-lyrics alignment task in high-resource languages such as English and for popular genres like pop (Stoller et al., 2019; Gupta et al., 2020), however other languages and styles remain a relatively underexplored challenge.

One genre that remains underexplored is the Italian opera, despite its central role in the development of Western vocal music (McCartney and Rorem, 2010). Aria is a fundamental piece of the Italian opera tradition—the most memorable pieces of the opera repertoire on which the tradition of Italian opera singing is based. The function of the aria is to express the emotions and feelings of a character throughout the opera, through the use of well-defined and attractive phrases with a strong emotional impact, designed to highlight the singer's technical abilities and capture the audience's attention (Gervás and Torrente, 2022; Fernicola et al., 2020; Zhang et al., 2022). The main difference between opera singing and pop singing lies in the vocal technique and the sound produced. Opera singing relies on a cultivated vocal technique—trained rather than natural—which shapes the voice to achieve specific artistic effects, sometimes at the expense of text intelligibility.

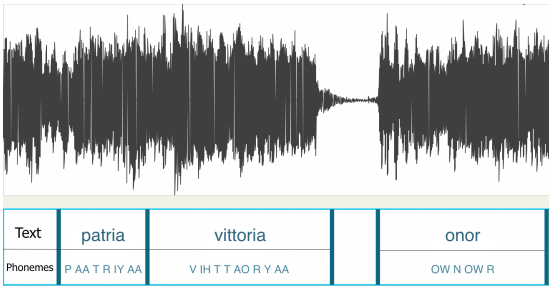


Figure 1: Aligning the text in the aria to the audio waveform to the precise timings.

Recent advances in language and speech modeling, as well as deep learning–based cross-modal approaches, have opened new possibilities for improving audio-lyrics alignment across languages (Bain et al., 2023; **?**; Pratap et al., 2024). However, existing methods (Gupta et al., 2020; Huang et al., 2022; Bain et al., 2023) often rely on language-specific resources, such as phonetic dictionaries or large annotated datasets, which are readily available for high-resource languages and genres but scarce for Italian. Additionally, most publicly available lyrics–audio datasets focus on popular genres such as pop, raising uncertainty about whether models trained on these datasets will transfer effectively to less-studied genres like arias.

To address these limitations, this paper introduces an audio-lyrics alignment dataset for Italian arias, providing a resource that enables research on an underrepresented linguistic and musical con-

text and facilitates evaluation of alignment models in a novel genre. Figure 1 shows an example from our dataset, consisting of an audio clip and its corresponding Italian text and phoneme sequences, which are already mapped to the segments they cover. The objective of an alignment model is to automatically establish such mappings. Using this dataset, we evaluate the performance of existing alignment models, including approaches based on both speech and audio representations. Our results highlight the limitations of current methods when applied to underrepresented languages and genres, and demonstrate the need for adaptation to specific linguistic and musical contexts. Our contributions are as follows:

1. **Novel audio-lyrics alignment dataset:** We introduce a new lyrics–audio dataset, the first focused on Italian Arias, addressing the lack of resources beyond popular genres.
2. **Benchmarking existing alignment models:** We evaluate state-of-the-art speech- and audio-based lyrics/text–alignment models on our dataset, providing the first benchmark for Italian Arias.
3. **Evaluation of Few-Shot Learning on Aria Dataset:** We use part of the dataset to comment on the few-shot learning capabilities on Aria Dataset.

Our code for finetuning and experiments is available[1]. We will also release and share the dataset, more information can be found on our Github repo.

2. Related Work

2.1. Audio-Lyrics Datasets

Several datasets exist for audio–lyrics alignment, created using either human annotation or weak labeling techniques Rafii et al. (2017); Lee and Scott (2017); Mauch et al. (2012). Many of the human-annotated datasets provide word-level or even finer-grained time-stamp labels, enabling precise alignment and evaluation. Most existing work, however, has focused on English language and popular music genres such as pop and rock. For example, the MUSDB18 dataset offers word-level annotations for 45 English pop songs (Rafii et al., 2017). Other notable contributions include just 8 and 20 songs with annotations, respectively (Lee and Scott, 2017; Mauch et al., 2012). A notable exception is the DALI dataset (Meseguer-Brocal et al., 2018), which contains 5358 songs annotated with weak labels using a teacher-student machine learning paradigm, rather than purely human annotations. It is common for researchers to use this dataset for training and use smaller human annotated datasets for evaluation. Another popular dataset is Jamendo Lyrics dataset made up of 20 English songs with word-level timing annotations. It has recent been extended (JamendoLyrics++) to about 100 songs across diverse genres and languages with careful manual corrections to timestamps (Kick et al., 2025; Durand et al., 2023).

Despite these efforts, there is a shortage of multilingual and genre-diverse datasets, especially for non-English and non-popular styles including classical, and traditional styles. This limitation motivates our work, which focuses on creating a new dataset that addresses these gaps for Italian arias.

2.2. Audio-Lyrics Alignment Models

Audio-lyrics alignment is often framed as a forced-alignment task, and solved with a dynamic programming algorithm such as Dynamic Time Warping (Lee and Scott, 2017). Systems may take advantage of multiple specialised audio analysis components (Fujihara and Goto, 2012). Some systems use a hierarchical approach, first aligning at a coarse level (section or line) and then at a finer level (word or syllable) (Lee and Scott, 2017).

In the absence of a large dataset of audio-lyrics alignment datasets, some researchers have tried to adapt speech models to the singing voice, but with limited success (Mesaros and Virtanen, 2009; Kruspe, 2016). Most of the recent approaches in audio-lyrics alignment use a machine learning model to predict timestamps by training them on singing voice datasets (Stoller et al., 2019; Gupta et al., 2020). Another approach for learning the phoneme alignment was proposed by Schulze-Forster et al. (2021) which, instead of adapting the data intensive models (Stoller et al., 2019; Gupta et al., 2020), learns an acoustic model without direct supervision as a side effect. This produces state of the art results in both source separation and the audio-lyrics alignment task.

In this paper, we benchmark existing pre-trained audio-lyrics alignment models including speech and audio models to evaluate their generalization ability to our new dataset.

3. Aria Dataset

In this section, we present information about our new *Aria Dataset* that consists of 24 arias with a total audio time of 1 hour and 53 minutes. The mean duration of the audio in the dataset is 284 seconds while the minimum and maximum durations are 125 and 570 seconds respectively.

[1] https://github.com/pushkarjajoria/lyrics-aligner

Composer	Opera	No. of Arias	Period	Genre
Handel	*Giulio Cesare*	6	early 18th c.	opera seria
Mozart	*Le nozze di Figaro*, *Don Giovanni*	4	late 18th c.	opera buffa
Rossini	*Il barbiere di Siviglia*	3	early 19th c.	opera buffa
Bellini	*I Puritani*, *Norma*	2	early 19th c.	melodramma
Verdi	*Rigoletto*, *La Traviata*, *Il Trovatore*	8	mid-19th c.	melodramma
Puccini	*Turandot*	1	early 20th c.	melodramma

Table 1: Summary of the 24 arias included in the dataset, spanning operas in *Italian*. The selection covers major composers and different genres from the early 18th century to the early 20th century.

3.1. Data Collection

The Italian opera repertoire was selected by two musicologists (co-authors of this work). The arias includes works of different genres, voices of different registers in terms of range and timbre, and different vocal traditions. The arias provides a substantial sample of the different styles of opera singing. They were selected balancing the aforementioned criteria and selecting from the opera repertoire that exemplify the current operatic repertoire that enjoys both institutional prominence and audience appeal.

The audio for the dataset was collected by downloading the files from the web, primarily from YouTube. The selection process involved identifying audios/videos that matched the intended version. It was ensured that the recordings were sufficiently high quality, complete, uncut and, with a clear vocal presence. We resample these files to 16kHz. A Python script provided with the dataset downloads the audio files from their sources.

Word	Start Time (s)	End Time (s)
suoni	2.24	3.17
la	3.27	3.37
tromba	3.47	4.04
e	4.14	4.24
intrepido	4.34	5.94
io	6.13	6.64
pugnerò	6.74	7.53
da	7.63	8.01

Table 2: Bellini Puritani: Suoni La Tromba sample annotations

The annotation for an aria consists of each word in the aria, along with its start and end times. A sample of these annotations can be seen in Table 2. We used Praat (Boersma and Weenink, 2025) to create files in Text-Grid format to standardize the annotation process and make it easier to map the frames in the audio files to words and their corresponding start and end times.

Each aria was manually annotated by a single annotator and each annotation was validated and corrected by a musicologist. The annotation process took on average 1 hour and 20 minutes to annotate one minute of music, bringing the total annotation time to 19 work days[2]. And for each of the 24 arias, we provide the lyrics in a text file, together with a tsv file containing the word-level timestamps denoted by start time and end time.

3.2. Phoneme transcription

The phonetic transcription was performed using ARPAbet[3] codes. ARPAbet is a version of IPA symbols that allows discretization and easy comparisons. ARPAbet was originally introduced for American English and encompasses 10 vowels, 5 vowel diphthongs, and 24 consonant symbols. Each ARPAbet symbol consists of one or two ASCII uppercase characters.

We used the version for Italian introduced by Arango et al. (2021) which includes the 30 basic phonemes of Italian together with the additional 20 geminates (longer sounds often represented by doubled letters) which are characteristic of the language. Translation of a written word to ARPAbet was done using the Italian version of the SPARSAR system (System for Poetry Analysis and Reading) (Delmonte, 2019), which produces an internal phonetic representation and translates it both to IPA and to ARPAbet.

However, the Italian language has changed since the historic period represented by our dataset, i.e. 18th-early 20th century. The lexicon has changed, and the richness of syntactic structures has gradually reduced. These changes are notable in the 24 arias: they not only introduce many archaic words which no longer appear in today's writing; but also, being sung poetry, they are characterized by a frequent use of *apocope* (omission of a final sound in a word) which makes phonetic transcription more difficult. Furthermore, correct transcription of a word to its ARPAbet representation requires not only turning graphemes into phonemes, but correct syllable division and word

[2]Considering an 8 hour work day.

[3]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

stress. Similarly to English, Italian possesses ambiguous trisyllabic words which place word stress on different syllables, depending on grammatical category. Thus, different phonemes may ensue from the same grapheme. To correctly process such kind of text required building a specialized module to accompany the Italian version of the SPARSAR system, as listed below:

1. **archaic words**, A special lexicon for archaic words has been created to account for them, e.g.: conturba, sparafucile, desire, core, dimonio, maledivami, securo, stromento, usasi, soglio, cittade, solinga, sovente, pingere, effigiò;
2. **apocope words**, A vowel was tentatively added and the result searched in the lexicon, e.g.: cor, error, null, quell, altr, poss, uom, pensier, va, oprar, muor, giovin, cagion, ch, or, far, perir, gioir, piacer, volar, amor;
3. **words omitting "v", "gl", "i"**, Some of these contracted words had to be specifically encoded as exceptions, e.g.: godea, vedea, pascea, sentìa, dee, ei, ne, de,
4. **special cases** This cases required a specialized module, e.g.: vo, deggio, degg'io, pei.

We will only comment more extensively on one special case, "deggio" or "degg'io", where the double palatal voiced affricate /JH/ in the first case is followed by the false "i" vowel used to turn velar stop /G/ into a palatal affricate in front of three full vowels |O, A, U|. But in the second case the apostrophe prevents the false vowel to appear and the word cannot be turned from grapheme to phoneme in the same manner: it requires the second word "io" to be taken into account. Thus a specific set of rules had to be inserted in the module to account for this exception. The reason is due to the fact that in the first case word stress falls on first syllable /DEY1/ nd the normal Italian orthography would be "devo" meaning "must"; while in the second case word stress falls on the second word |io| /IY1_OW/ and the two words would be "devo io", "must I".

The output of the SPARSAR system is an ARPAbet representation of the lyrics of each aria, which will be used in our proposed model. A sample of the output of the model is shown in Table 3.

Word	ARPAbet
suoni	S W OW N IY
la	L AA
tromba	T R AO M B AA

Table 3: Example phonetic transcription as per Italian pronunciation

4. Evaluation and Benchmarking

Given the Italian arias dataset, we conduct benchmarking experiments on existing alignment models, including both music alignment models and speech forced alignment model. The following subsections describe the five benchmarked models, evaluation metrics we used, and the results of benchmarking.

4.1. Models

We selected five models based on recency, code availability, adaptability to working with a new dataset, and multilingual capability. They fit into two major categories of music alignment models and speech forced alignment models which we describe below.

Music alignment models We include two models pretrained on English language music, ALT (Gupta et al., 2020) and PLLA (Schulze-Forster et al., 2021). We select these models to evaluate if such models trained on English language and on music can generalize to other genres and languages. In addition, we include LA-JPD (Huang et al., 2022), a multilingual music alignment model to evaluate the ability of an existing multilingual model on our dataset. LA-JPD was trained on multilingual corpora of DALI v2.0 which consists of five languages including Italian.

Speech alignment models We also evaluate two speech forced alignment models, MMS-FA (Ashraf, 2024) and WhisperX (Bain et al., 2023). Both of these models are based on massively multilingual transformer models pre-trained on a large amount of speech datasets including Italian. MMS-FA is based on MMS (300M parameter) (Pratap et al., 2024), a speech encoder, while WhisperX is based on Whisper (**?**), a massively multilingual speech recognition model. To convert Whisper to WhisperX, the authors transcribe the input audio into segments of words. And they input these transcriptions along with the audio to a forced aligner to generate the timestamps. For our evaluation, we bypass the transcription step and directly provide a single segment with the full lyrics to have a fair comparison with other alignment models. Since music alignment models either requires separated vocals at inference time or do so internally (LA-JPD, PLLA); to ensure fairness between the speech and music alignment models and to remove the impact of background music; we also evaluated the two speech models, WhisperX and MMS-FA, on the separated vocals using Spleeter (Hennequin et al., 2020) denoted using 'SV'.

Model Name	Multilingual	RMSE (↓) [s]	MAE (↓) [s]	MedAE (↓) [s]	$PCO_{0.3}$ (↑) [%]
Music alignment models					
PLLA (Schulze-Forster et al., 2021)	X	$25.24_{\pm 7.7}$	$20.94_{\pm 6.7}$	$18.36_{\pm 6.5}$	$30.45_{\pm 5.8}$
LA-JPD (Huang et al., 2022)	✓	$\mathbf{9.44_{\pm 3.1}}$	$\mathbf{6.14_{\pm 2.3}}$	$\mathbf{3.46_{\pm 1.4}}$	$52.98_{\pm 5.6}$
ALT (Gupta et al., 2019)	X	$18.49_{\pm 4.3}$	$13.51_{\pm 3.7}$	$9.58_{\pm 3.3}$	$20.69_{\pm 3.5}$
Speech Models					
MMS-FA (Ashraf, 2024)	✓	$11.47_{\pm 3.2}$	$7.45_{\pm 2.8}$	$3.99_{\pm 2.1}$	$\mathbf{55.43_{\pm 5.9}}$
MMS-FA-SV (Ashraf, 2024)	✓	$11.26_{\pm 3.3}$	$7.30_{\pm 2.8}$	$3.98_{\pm 2.1}$	$\mathbf{56.02_{\pm 6.0}}$
WhisperX (Bain et al., 2023)	✓	$39.43_{\pm 12.7}$	$34.61_{\pm 12.1}$	$33.37_{\pm 12.7}$	$23.34_{\pm 4.4}$
WhisperX-SV (Bain et al., 2023)	✓	$41.35_{\pm 13.5}$	$36.02_{\pm 12.6}$	$34.64_{\pm 12.9}$	$28.60_{\pm 5.7}$

Table 4: Alignment performance of different models on the Aria dataset. Values are reported as mean ± standard error across all arias. 'SV' denotes separated vocals were provided as input. The 'Multilingual' column indicates whether the model's training was multilingual.

Processing Output The output from PLLA and MMS-FA was processed without any challenges. For ALT the output sometimes swapped words from lyrics with *BREATH**. To match the number of word onset in labels with the number of predicted onsets we considered this also part of the annotation while computing the four metrics. For a single aria *'Mozart Nozze Non so più'* ALT did not produce the same number of words as in the labels. We exclude this aria from all evaluation for ALT model. LA-JPD required us to provide IPA phoneme and vocals separated from the music and background noise. We converted our ARPAbet phonemes into IPA using the *phonecodes* python package and we used Spleeter for voice separation.

4.2. Evaluation Metrics

We evaluate word onset alignment for each of the five models against ground truth. We use similar evaluation setup as Cheng et al. (2025). The evaluation metrics used include: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MedAE) and Percentage of Correct Onsets ($PCO_{0.3}$) with a tolerance window τ = 0.3 (see equation 1). All results are averaged over all 24 arias in the dataset to provide the mean ± standard error.

RMSE penalizes larger deviations more strongly, providing a sense of the scale of alignment errors in seconds. While MAE and MedAE measures the average and median absolute difference between predicted and true onsets. Unlike RMSE, they weights all errors equally, thus giving a more robust summary of overall predictive accuracy without being overly influenced by outliers. PCO_τ measures the proportion of words whose predicted onset times fall within a pre-defined tolerance threshold τ of the ground truth onsets. For example, $PCO_{0.3}$ with $\tau = 0.3$ represents the percentage of words aligned within 300ms of their true onset times. This metric offers a more interpretable view of accuracy. Formally, PCO is defined as,

$$PCO_\tau = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}_{|u^i_{\text{pred}} - u^i_{\text{ref}}| < \tau} \quad (1)$$

where, N is total number of words and u^i denotes the onset time for word 'i'.

4.3. Results and Discussion

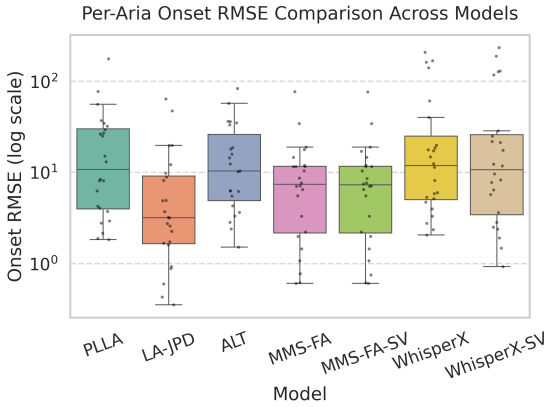


Figure 2: RMSE values across models for the 24 arias. The y-axis uses a log scale due to large performance variations across arias.

We report the RMSE, MAE, MedAE, and $PCO_{0.3}$ for all models in Table 4. Among them, LA-JPD performs best in RMSE, MAE, and MedAE, while MMS-FA-SV achieves the highest PCO score. However, lower RMSE, MAE, and MedAE values do not necessarily correspond to higher PCO values, as

seen with LA-JPD and MMS-FA. As shown in Figure 2, the RMSE values vary widely across arias, exceeding a difference of 100 in some cases. WhisperX, in particular, exhibits a few strong outliers, resulting in a higher variance and lower overall mean performance.

Furthermore, among the evaluated model, LA-JPD, a multilingual music alignment model, demonstrates superior performance across RMSE, MAE, and MedAE, emphasizing the importance of multilingual training. MMS-FA, a model not trained exclusively on musical data, achieves the highest PCO of 55.4%, which increases slightly to 56% when using separated vocals (MMS-FA-SV). In contrast, both music alignment models without multilingual exposure, PLLA and ALT, perform suboptimally with a PCO of 30% and 20%. ALT showing a particularly sharp drop in PCO from 94% to 20% (Cheng et al., 2025) when compared against the Jamendo dataset.

Vocal separation provides a clear gain for WhisperX, improving PCO from 23.3% to 28.6%, whereas the effect on MMS-FA (55.4% to 56%) is marginal and not statistically significant. This suggests that while input preprocessing can offer some improvements, it cannot fully address the model's limitations in handling complex musical structures.

One of the central questions we posed was whether alignment models trained on high-resource, popular-genre datasets can generalize effectively to low-resource languages and musical styles such as Italian arias. The results from Table 4 indicate only partial success. None of the evaluated models reached the performance levels reported on the Jamendo dataset by Cheng et al. (2025), where the PCO of 94% drops to 53% on the aria dataset, a 44% decline that when viewed through the lens of gross errors (predictions deviating by more than 0.3s), represents an approximately 8-fold increase from just 6% gross errors on Jamendo to 47% on the aria dataset.

Overall, these findings suggest that current alignment models exhibit some generalization ability, but they still face substantial challenges in transferring learned representations to new languages and musical styles such as Italian arias. The relatively strong performance of MMS-FA models also raises the possibility that pretrained speech architectures may offer better cross-domain generalization for music alignment tasks.

5. Few Shot Learning

The results in Section 4 showed us that models trained on high-resource genre and languages struggle to generalize to the Italian aria dataset. Having established this gap, in this section we try to answer the questions if training an alignment on few examples of in-domain data (Italian arias) helps the model generalize to this new dataset. Hence, we explore the few-shot learning ability of PLLA.

The PLLA model introduced by Schulze-Forster et al. (2021) is a phoneme level lyrics alignment model. The model was trained via multi-task learning where the main task is singing voice separation and audio-lyrics alignment is learned as an intermediary step. For our experiments, we focus on the alignment component of this model and ignore voice separation component. Therefore, We fine-tune this model with a few examples on Italian Arias to study the few shot learning capabilities of the PLLA model and learn if the new data can improve performance of this model.

5.1. Data Processing and Augmentation

For us to fine-tune the PLLA model, the raw audio and textual data were preprocessed to produce phoneme-level alignment labels compatible with the pre-trained PLLA alignment model. All the audio files were downsampled to 16 kHz and converted to single-channel format.

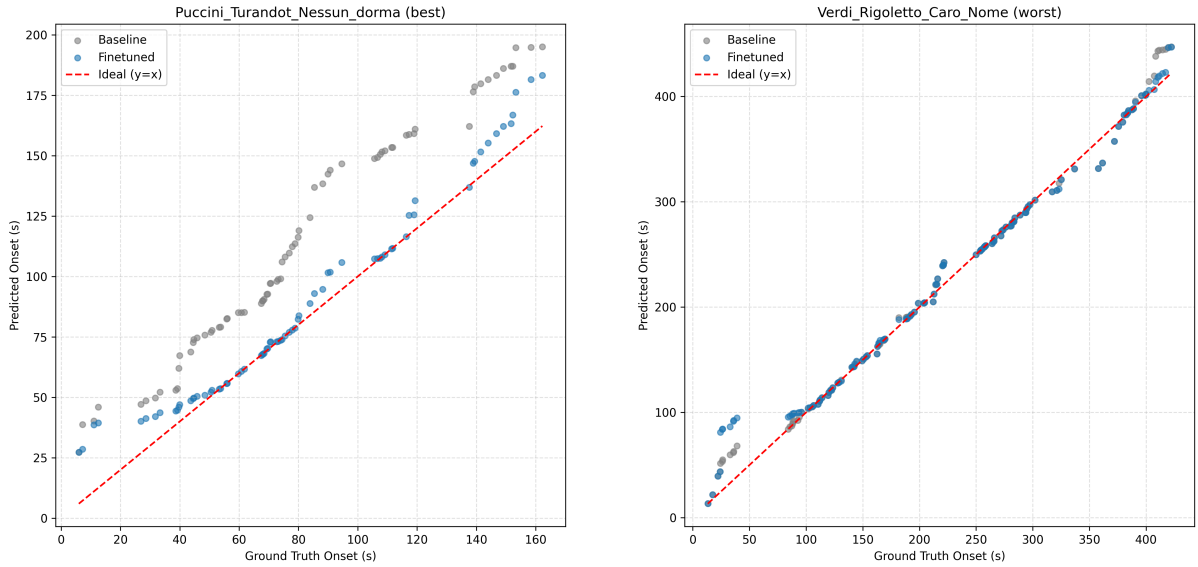
The lyrics corresponding to each recording were tokenized into words and subsequently converted into phoneme sequences (see Section 3.2). Since the alignment model works at phoneme level timestamps, we estimated phoneme-level timestamps from word level timestamps (see Section 3.1) through uniform interpolation between the start and end times of each word. We use the end times to compute the start of silence token '<' that is inserted between each word in the lyrics. Each phoneme was thus assigned an interpolated onset and offset time that preserves the duration and order of the phonemes within that word. For each Short-Time Fourier Transform (STFT) frame of the audio, the active phonemes were determined based on their interpolated timestamps, producing an '*alpha tensor*' that indicates which phoneme(s) are active in every frame producing the ground-truth alignment label, while the model predicts a corresponding alpha matrix during fine-tuning; the alignment loss is computed between these two representations.

We also augment the training dataset with four different data augmentation techniques: pitch shift, additive noise, reverb, and frequency masking. The data augmentation code[4] is adapted from Abdullah et al. (2025). For pitch shift we randomly sample a number of semitones from {-8, -6, -4, -2, 2, 4, 6, 8}. We add Gaussian noise, sampled with a noising factor of 0.01. For reverb, we use a random sample from the room impulse response and noise database from openSLR (Ko et al., 2017) to apply it to the aria. For frequency masking, we ran-

[4] https://github.com/badrex/arabic-dialect-identification

Model Name	RMSE (↓) [s]	MAE (↓) [s]	MedAE (↓) [s]	$PCO_{0.3}$ (↑) [%]
Baseline PLLA (Schulze-Forster et al., 2021)	$23.77_{\pm 3.0}$	$19.95_{\pm 2.7}$	$17.23_{\pm 2.3}$	$26.38_{\pm 1.5}$
Finetuned PLLA	$21.32_{\pm 3.3}$	$17.44_{\pm 2.9}$	$14.91_{\pm 2.3}$	$27.37_{\pm 2.1}$

Table 5: Alignment performance of baseline and finetuned PLLA model. Values are reported as mean ± standard error across all five folds on the test set.



(a) Best improvement: Puccini Turandor Nessun Dorma

(b) Worst decline: Verdi Rigoletto Caro Nome

Figure 3: Comparison of predicted word onset times for the finetuned and baseline models on two arias. The left plot shows the case with the largest improvement from finetuning, and the right shows the most decline in performance, across all 5 folds. Both arias were part of the test set. The line y=x corresponds to a perfect match between ground truth and predictions.

domly mask a portion of the frequency spectrum to simulate the loss of certain frequency bands.

5.2. Training

To evaluate the few shot learning capabilities of the PLLA model on the new dataset we divide the 24 arias into 17:7 arias as per the 70:30 split for 5 random folds. Note that we only do this while fine-tuning the model and while benchmarking existing methods in Section 4 we use the full dataset for evaluation. We trained the model for 15 epochs with a learning rate of 1e-5. We found the training to to unstable while optimizing every step and decided to accumulate the gradients of 4 steps before backpropagating the model. The comparison between the fine-tuned model and a the baseline model is provided in the Table 5.

5.3. Discussion

Even though we see a slight improvements across all metrics based on the results from Table 5, it is inconclusive to say if the model truly learned to align this new language and genre of music. We provide a two examples of arias from the test data to visualize the effect of fine tuning in Figure 3. We see evidence of learning on some arias where the model aligns the middle part of the aria while drifting at the start and the end. While the drop is not significant, we did see a performance drop on a few arias after finetuning. While the improvements across all metrics in Table 5 are modest, the consistent direction of change indicates partial adaptation to the new domain. Figure 3 shows that fine-tuning notably improves local alignment for some arias (e.g., *Puccini Turandot Nessun dorma*), while others (e.g., *Verdi Rigoletto Caro Nome*) exhibit drift in the initial predictions. This suggests that while the model benefits from training on the new language and genre, it still struggles to learn general features to improve the performance across the dataset.

6. Conclusion

In this work, we present the *Aria Dataset*, a novel resource for audio and lyrics alignment that addresses the scarcity of datasets for non-contemporary music. Unlike existing resources that focus on popular music, this is the first dataset dedicated entirely to Italian arias. The dataset is curated by two musicologists (co-authors of this paper), and word-level time-stamps underwent manual annotation with expert validation. Through benchmarking, we demonstrate that existing alignment and speech models face significant challenges when applied to this dataset, with LA-JPD's performance dropping by 44% compared to previous benchmarks in audio-lyrics alignment. Furthermore, our results indicate that in the context of this new dataset, forced alignment speech model outperform specialized music alignment models. Additionally, we find that few-shot learning adaptation on one music alignment model yields only minor overall improvements, yet shows promising signs of better alignment on certain arias, indicating a step in the right direction. This highlights the need for further research to develop more versatile alignment models capable of handling diverse musical genres.

7. Future Work

While the dataset with 24 arias provides insight into the performance of lyrics alignment systems, further effort would be needed to allow models to use these samples during training. The current dataset size limits its direct use for training deep learning models, which typically require large-scale annotation. Future effort should focus on expanding the dataset both in scale and variety, incorporating a broader set of genres and, languages.

Although we observed gains within the range of variance and thus not statistically conclusive, the consistent reduction in RMSE, MAE, and MedAE suggests that the model is learning in the right direction. This behavior aligns with expectations for few-shot adaptation, where improvements are gradual and may vary across samples depending on acoustic and linguistic similarity to the fine-tuning data. Future work could address these limitations by using parameter-efficient tuning methods such as adapters or LoRA (Hu et al., 2021).

8. Acknowledgments

Pushkar Jajoria was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. This work was also supported by the Polifonia Project of the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004746.

We are grateful to Badr M. Abdullah for his feedback and guidance on this project. We also thank Aravind Krishnan, Janaki Viswanathan, and other members of the LSV Lab at Saarland University for their feedback on the manuscript.

9. Bibliographical References

Badr M. Abdullah, Matthew Baas, Bernd Möbius, and Dietrich Klakow. 2025. Voice conversion improves cross-domain robustness for spoken Arabic dialect identification. In *Interspeech*.

Javi Arango, Alec DeCaprio, Sunwoo Baik, Luca De Nardis, Stefanie Shattuck-Hufnagel, and Maria Gabriella Di Benedetto. 2021. Estimation of the frequency of occurrence of Italian phonemes in text. *arXiv preprint arXiv:2101.06147*.

Mahmoud Ashraf. 2024. mms-300m-1130-forced-aligner. Accessed: 2025-10-12.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech*.

Tian Cheng, Tomayasu Nakano, and Masataka Goto. 2025. Improving lyrics-to-audio alignment using frame-wise phoneme labels with masked cross entropy loss. In *DAFx*.

Rodolfo Delmonte. 2019. Poetry and speech synthesis: Sparsar recites. *RiCOGNIZIONI: Rivista di Lingue e Letterature Straniere e Culture Moderne*, 6(11):75–95.

Francesco Fernicola, Shibingfeng Zhang, Federico Garcea, Paolo Bonora, Alberto Barrón-Cedeno, et al. 2020. Ariemozione: Identifying emotions in opera verses. In *CEUR WORKSHOP PROCEEDINGS*, volume 2769, pages 58–63. CEUR Workshop Proceedings.

Hiromasa Fujihara and Masataka Goto. 2012. Lyrics-to-audio alignment and its application. In *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Pablo Gervás and Alvaro Torrente. 2022. Emotional interpretation of opera seria: Impact of specifics of drama structure (position paper). In *KDIR*, pages 330–336.

Chitralekha Gupta, Emre Yilmaz, and Haizhou Li. 2019. Automatic lyrics transcription in polyphonic music: Does background music help? *ArXiv*, abs/1909.10200.

Chitralekha Gupta, Emre Yılmaz, and Haizhou Li. 2020. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500. IEEE.

Zihao He, Weituo Hao, Wei-Tsung Lu, Changyou Chen, Kristina Lerman, and Xuchen Song. 2023. ALCAP: Alignment-augmented music captioner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16501–16512, Singapore. Association for Computational Linguistics.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154. Deezer Research.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models.

Jiawen Huang, Emmanouil Benetos, and Sebastian Ewert. 2022. Improving lyrics alignment through joint pitch detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 451–455. IEEE.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224.

Anna Kruspe. 2016. Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing. In *International Society for Music Information Retrieval Conference*.

Sang Won Lee and Jeffrey Scott. 2017. Word level lyrics-audio synchronization using separated vocals. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650.

Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. 2010. Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations. In *Proceedings of the 7th Sound and Music Computing Conference (SMC 2010)*, pages 9–16.

Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. 2012. Integrating additional chord information into hmm-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:200–210.

Paul McCartney and Ned Rorem. 2010. Oxford history of western music: Richard Taruskin. In *Oxford History of Western Music*.

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

Annamaria Mesaros and Tuomas Virtanen. 2009. Adaptation of a speech recognizer for singing voice. In *2009 17th European Signal Processing Conference*, pages 1779–1783.

Meinard Müller, Frank Kurth, David Damm, Christian Fremerey, and Michael Clausen. 2007. Lyrics-based audio retrieval and multimodal navigation in music collections. In *International conference on theory and practice of digital libraries*, pages 112–123. Springer.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2017. The MUSDB18 corpus for music separation.

Kilian Schulze-Forster, Clement S. J. Doire, Gaël Richard, and Roland Badeau. 2021. Phoneme level lyrics alignment and text-informed singing voice separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2382–2395.

Daniel Stoller, Simon Durand, and Sebastian Ewert. 2019. End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185. IEEE.

Shibingfeng Zhang, Francesco Fernicola, Federico Garcea, Paolo Bonora, and Alberto Barrón-Cedeño. 2022. Ariemozione 2.0: Identifying emotions in opera verses and arias. *IJCoL: Italian Journal of Computational Linguistics*, 8(8-2).

10. Language Resource References

Paul Boersma and David Weenink. 2025. Praat: doing phonetics by computer.

Simon Durand, Daniel Stoller, and Sebastian Ewert. 2023. Contrastive learning-based audio to lyrics alignment for multiple languages. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece.

Timon Kick, Florian Grötschla, Luca A. Lanzendörfer, and Roger Wattenhofer. 2025. Contrastive lyrics alignment with a timestamp-informed loss. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. 2018. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm.