Transfer Learning for Lyrics-to-Audio Alignment Using Wav2Vec and WavLM

Wentao Fan[1]*, **Kangyi Zhang**[1]*

[1]Tandon School of Engineering, New York Univeristy
wf2205@nyu.edu, kz2643@nyu.edu

Abstract

Lyrics-to-audio alignment is a critical task in music information retrieval, enabling applications such as karaoke, automated subtitle generation, and enhanced music streaming services. In this work, we built upon the Wav2Vec 2.0 and WavLM models, leveraging transfer learning and advanced processing techniques to achieve accurate lyrics alignment. By combining vocal separation, segmentation, forced alignment, and fine-tuning, we developed a pipeline that achieves strong results on the Jamendo V1 dataset. Our streamlined approach highlights the potential for effective alignment using simplified workflows and shorter training durations. The repository and structured dataset are shared openly to encourage further development in the field.

Introduction

Lyrics-to-audio alignment is a critical task in the field of music information retrieval, enabling numerous applications such as karaoke systems, automated subtitle generation, and enhanced music streaming services. This task involves synchronizing lyrics text with the corresponding singing audio, which is especially challenging due to variations in singing styles, tempo fluctuations, and the presence of instrumental accompaniments. These complexities require robust and adaptable models to address the intricacies of polyphonic music.

Recent advancements in self-supervised learning have introduced powerful speech models such as Wav2Vec 2.0 and WavLM. Wav2Vec 2.0, developed by Facebook (now Meta), revolutionized the field of speech processing by introducing a framework capable of learning meaningful representations from raw audio with minimal supervision. Its robustness to noise and ability to perform well with limited labeled data make it a strong candidate for adaptation to music-related tasks (Baevski et al. 2020). WavLM builds upon these advancements by incorporating masked speech prediction and learning universal representations, demonstrating exceptional performance across diverse speech tasks(Chen et al. 2022). These models excel at capturing nuanced acoustic patterns, providing a solid foundation for adapting to singing audio.

*These authors contributed equally.

Our interest in this domain aligns with the goals of the MIREX Lyrics-to-Audio Alignment competition (Music Information Retrieval Evaluation eXchange (MIREX) 2024), which sets benchmarks for evaluating alignment systems and fosters innovation in the field. During the process of learning relevant knowledge and gaining deeper understanding, we initiated practical experiments, aiming to showcase the potential of transfer learning techniques by leveraging Wav2Vec 2.0 and WavLM to overcome challenges in lyrics synchronization. These challenges include handling polyphonic complexity, diverse vocal styles, and irregular tempos, which are key obstacles to achieving fine-grained alignment.

Our research significantly builds upon the foundation of the lyrics-to-audio alignment system on Github (Mikezzb 2023). Using this as a base, we developed our system to include vocal separation, segmentation, recognition and forced alignment techniques, and system evaluation. We fine-tuned Wav2Vec 2.0 and WavLM through transfer learning on datasets meticulously pre-processed for this task. The acquisition, processing, and use of the data sets are completely different, and we have updated the model to help the refinement, in addition to optimizing the implementation details and expanding on more language models.

Regarding datasets, we initially planned to use the DALI dataset in Zenodo (https://zenodo.org/records/2577915), given its rich features and high-quality alignment data, making it a solid foundation for lyrics alignment studies. However, after obtaining the data through communication with its creator, we encountered challenges due to its reliance on outdated tools such as `youtube_dl`. To address these issues, we updated the code to use `yt_dlp`, assisting the original author in modernizing the toolchain. Despite this, we found the DALI dataset's acquisition process to be complex and prone to dependency conflicts, making it less accessible for beginners.

To overcome these limitations, we adopted a more straightforward method to obtain datasets by registering accounts on karaoke platforms and and other digital music platforms, including mainstream music streaming applications.Through these platforms, we downloaded and captured audio-lyric pairs directly, streamlining data collection while maintaining high-quality annotations.

Our complete codebase is available at

- **https://github.com/WF2205/Lyrics-audio-Alignment.**

In the following sections, we present a detailed literature review, outline our methodology, and discuss the results and insights gained from this study. This research demonstrates the transformative potential of self-supervised learning models in addressing long-standing challenges in lyrics-to-audio alignment, paving the way for further advancements in music information retrieval.

Literature Review

Synchronizing lyrics with audio is a critical task in music information retrieval (MIR), underpinning applications such as karaoke, transcription tools, and music streaming platforms. Despite its importance, the acoustic and rhythmic variations in songs present significant challenges, demanding innovative solutions. This review examines key methodologies, datasets, and models, highlighting recent advancements and identifying persistent gaps.

Foundational Approaches

Early research into lyrics-to-audio alignment relied on traditional methods such as Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs). These techniques provided foundational insights by aligning audio and textual data using phonetic models. However, their performance in polyphonic music was hindered by the inability to effectively separate vocal signals from instrumental accompaniment (Dzhambazov 2017).

To overcome these limitations, researchers incorporated domain-specific knowledge. For instance, the integration of rhythmic and melodic structures into probabilistic models has been shown to improve alignment accuracy. Dzhambazov (Dzhambazov 2017) utilized Dynamic Bayesian Networks to represent dependencies between phonemes and musical events, demonstrating the potential of probabilistic frameworks.

Datasets and Their Influence

As early as 2014, MedleyDB has supported the evaluation of vocal separation techniques (Bittner et al. 2014). Then the availability of annotated datasets has significantly propelled this domain forward. The DALI dataset, comprising over 5,000 annotated tracks, serves as a cornerstone for training and evaluation (Meseguer-Brocal, Cohen-Hadria, and Peeters 2018). Complementing this, the MulJam dataset enables multilingual lyrics transcription, though imbalances in language representation present challenges (Huang and Benetos 2024). Together, these datasets provide diverse contexts for testing and refining alignment methodologies.

Advancements in Self-Supervised Learning

Self-supervised learning has redefined the landscape of lyrics-to-audio alignment. Models like Wav2Vec 2.0 (Baevski et al. 2020) and WavLM (Chen et al. 2022) leverage large-scale unlabeled data to learn robust audio representations. Wav2Vec 2.0 employs contrastive learning on masked audio representations, while WavLM incorporates speech denoising and prediction tasks, enhancing performance in noisy and polyphonic environments. These advancements underscore the utility of self-supervised frameworks in tackling alignment challenges.

Hierarchical and Multimodal Methods

Hierarchical approaches have emerged as a promising avenue for precise alignment. Lee et al. (Lee and Scott 2017) proposed a two-step process involving coarse alignment of lyric paragraphs to audio segments, followed by fine-grained word-level alignment. This approach utilizes synthesized speech and integrates multimodal features such as pitch and rhythm, demonstrating its adaptability across musical styles. Real-time systems further build on this foundation, employing phonetic posteriorgrams (PPGs) and chroma features to achieve low-latency alignment (Park et al. 2024).

Challenges and Future Directions

Despite these advancements, significant challenges persist. Multilingual alignment remains underexplored due to dataset imbalances and the complexity of language-specific phonetic models. Real-time processing systems, while effective, often rely on specialized acoustic models that limit scalability. Addressing these issues will require continued innovation in dataset development and model adaptability.

Furthermore, evaluation metrics such as Average Absolute Error (AAE), Median Absolute Error (MAE), and Word Error Rate (WER) provide essential benchmarks for assessing model performance (Meseguer-Brocal, Cohen-Hadria, and Peeters 2018; Lee and Scott 2017). However, a standardized framework for comparing models across datasets and tasks remains an unmet need.

The field of lyrics-to-audio alignment has achieved remarkable progress through self-supervised learning, probabilistic modeling, and hierarchical methods. Yet, the road ahead requires addressing multilingual capabilities, real-time scalability, and dataset diversity. Our proposed model synthesizes the most effective strategies identified in existing research while addressing critical limitations in the field. By integrating self-supervised learning from frameworks like Wav2Vec 2.0 (Baevski et al. 2020) and WavLM (Chen et al. 2022), and adopting hierarchical approaches as demonstrated by Lee et al. (Lee and Scott 2017), our system achieves both robust and fine-grained alignment across diverse musical contexts. Inspired by recent works, including the open-source project Lyrics-Sync, our design builds on practical implementations, providing a solid foundation for future improvements and real-world applications.

Method

Transfer Learning

Transfer learning is a machine learning technique where a model developed for a specific task is reused as the starting point for a model on a related task (Pan and Yang 2009). It has gained prominence in domains with limited labeled data, such as music information retrieval and audio analysis.

In our project, transfer learning was applied to fine-tune Wav2Vec 2.0 and WavLM, pre-trained on vast amounts

of speech data, for lyrics-to-audio alignment. These models were retrained using datasets containing singing vocals to ensure their acoustic representations captured the nuances of polyphonic music. This fine-tuning process involved training the models on connectionist temporal classification (CTC) loss for mapping audio features to lyric tokens, enabling precise word-level alignment.

Vocals Extraction Using HTDemucs Model

Vocal separation is a critical preprocessing step in lyrics alignment. HTDemucs, a high-performance deep learning model for music source separation, was employed to extract vocal tracks from polyphonic music. HTDemucs extends Demucs by incorporating Hybrid Transformers for enhanced accuracy in separating vocals and instruments (Rouard and Peeters 2021).

We processed each song using HTDemucs, generating high-fidelity vocal tracks free from instrumental interference. These vocal tracks were then resampled to 16 kHz for uniformity and prepared for subsequent segmentation and alignment tasks.

Segmentation

Lyrics segmentation ensures that audio segments correspond to meaningful lyric phrases, facilitating efficient alignment. The segmentation process involved parsing lyric files (LRC format) to extract timestamps and text using regular expressions. Each lyric line was mapped to a time interval using the following workflow:

1. Parse timestamp-text pairs from LRC files using regex patterns.
2. Convert timestamps to seconds for precise alignment with audio.
3. Extract audio segments corresponding to each lyric line from the vocal tracks generated by HTDemucs.

Speech Recognition Using Wav2Vec 2.0 and WavLM

These models learn representations from raw waveforms, capturing fine-grained acoustic details. For this project:

1. Wav2Vec 2.0: Adapted to the singing domain, leveraging its robust speech-to-text capabilities.
2. WavLM: Enhanced with masked speech prediction for improved performance on noisy datasets.

Alignment Using Forced Alignment

The segmented audio and corresponding lyric lines were exported as WAV files, forming the input for alignment tasks.

Forced alignment is a technique used to align audio signals with corresponding textual transcriptions at a fine-grained, frame-level resolution. According to Rotem Rousso (Rotem Rousso , Eyal Cohen , Joseph Keshet , Eleanor Chodroff 2024), there are three primary methods for achieving forced alignment:

1. Montreal Forced Aligner (MFA)
2. Massively Multilingual Speech (MMS) Model
3. ASR-based Phoneme Recognition Model

Montreal Forced Aligner (MFA)

- MFA wraps the Kaldi ASR toolkit and utilizes traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) acoustic models.
- The training pipeline refines phone models progressively: monophone $\rightarrow$ triphone $\rightarrow$ LDA+MLLT $\rightarrow$ SAT with fMLLR.
- Uses 39 MFCC features extracted every 10 ms.
- Alignment applies a pronunciation lexicon to map words to phonemes, using the Viterbi algorithm to find the most likely alignment path with a 10 ms resolution.

Massively Multilingual Speech (MMS) Model

- Based on a multilingual wav2vec 2.0 architecture, covering over 1,000 languages.
- Trained through self-supervised pre-training on large unlabeled audio corpora and fine-tuned with a CTC objective for ASR tasks.
- Relies on CTC-based alignment, which lacks explicit frame-level boundaries, making forced alignment challenging without additional strategies.

Phoneme Recognition Model (ASR-based Approach)

- Employs a wav2vec 2.0-based ASR model to generate phoneme-level transcriptions with timestamps.
- Aligns recognized phonemes to the reference lyrics' phoneme sequence using dynamic programming, treating alignment as an optimization problem.

Chosen Method: We will use the Phoneme Recognition Model, leveraging the strengths of wav2vec 2.0 and WavLM for high-quality phoneme-level alignments.

Evaluation

Evaluation of the alignment system was conducted using multiple metrics:

Connectionist Temporal Classification (CTC) Loss: Evaluates the model's ability to map sequential data (audio features to lyric tokens) without frame-level alignment.

Alignment Precision Metrics:

- Average Absolute Error (AAE): Measures the average time difference between predicted and actual timestamps, reflecting overall alignment accuracy.
- Median Absolute Error (MAE): Reports the median of absolute time errors, reducing the impact of extreme outliers.
- Percentage of Correct Segments (PCS): Calculates the proportion of correctly aligned segments, where alignment is considered correct if timestamps fall within a specified tolerance window.
- Percentage of Correct Onsets (PCO): Evaluates the accuracy of predicted onset times, with correctness determined by tolerance thresholds.

The evaluation script, implemented in `Evaluation.ipynb`, validated the model's performance on word-level alignment tasks across diverse datasets, demonstrating the system's robustness and accuracy.

Experiments

Datasets

Training Dataset: We have constructed our own dataset consisting of approximately 300 English songs, each paired with word-synchronized lyrics. These songs were acquired from Tencent Music and represent some of the most popular English-language tracks from the past two decades. The data preprocessing steps include vocal extraction to isolate the singing voice from instrumental components and segmenting the audio according to the provided time-aligned lyrics. The result is a refined training dataset suitable for phoneme-level alignment model training.

Testing Dataset: For evaluation, we will use 20 English songs from the Jamendo dataset (f90 2024). The Jamendo dataset is often employed as a standard benchmark in music information retrieval tasks, including lyrics-to-audio alignment, as recognized by the Music Information Retrieval Evaluation eXchange (Music Information Retrieval Evaluation eXchange (MIREX) 2024). This ensures that our evaluation is consistent with established community standards.

Model Configuration

We will use two state-of-the-art models:

1. **Wav2Vec 2.0:** A self-supervised speech representation model pre-trained on large unlabeled corpora and fine-tuned for ASR tasks using a CTC objective.
2. **WavLM:** An enhanced variant of wav2vec models, leveraging additional training objectives for improved robustness and accuracy in speech-related tasks.

Training and Inference

Training

1. **Preprocessing:** We begin by extracting vocals from the mixed audio tracks and slicing the audio based on the lyrics. This involves ensuring that each audio clip corresponds closely to the associated lyric lines.
2. **Metadata Construction:** Using the processed audio segments and their associated transcriptions, we create a metadata file (CSV) that maps each audio file to its corresponding text. This serves as input for the training pipeline.
3. **Training Arguments:** We specify hyperparameters such as weight decay, learning rate, and num train epochs. We also use gradient accumulation and mixed-precision training (fp16) to optimize GPU resource usage and training speed.
4. **Fine-Tuning:** Using Hugging Face's Transformers library and Trainer API, we will fine-tune Wav2Vec 2.0 and WavLM models on our processed music dataset. The models will learn to output phoneme sequences with timestamps that align with the sung lyrics.

Inference (Forced Alignment)

1. **Audio Preprocessing:** For a given test audio clip, the vocals are extracted to isolate the singing voice from instrumental components. The extracted vocals are then downsampled to 16,000 Hz to match the input requirements of the ASR model. Segmentation is performed to split the audio into manageable chunks. This ensures efficient processing and alignment.
2. **Phoneme Recognition:** The preprocessed audio chunks are fed into the fine-tuned ASR model (Wav2Vec 2.0 or WavLM). The model outputs a phoneme-level transcript, where each phoneme is associated with a probability and a timestamp (time frame). The emission probabilities from the model represent the likelihood of each phoneme or blank token at each frame.
3. **Phoneme Matching:** The recognized phoneme sequence is aligned to the reference lyrics' phoneme sequence using a dynamic programming approach. A **trellis matrix** is constructed, where each cell represents the best cumulative alignment score for matching a token (phoneme) up to a specific time frame. The alignment is performed in two steps:

   Trellis Construction: The trellis matrix is filled using emission probabilities from the ASR model. Each cell considers whether to stay on the current phoneme (adding the blank token's score) or advance to the next phoneme (adding the token's score).

   Backtracking: Starting from the last token, the optimal alignment path is reconstructed by tracing back through the trellis. This path determines the most likely time frame for each phoneme.
4. **Output:** Generate phoneme-level forced alignments, providing start and end times for each phoneme (and word).

Models for Comparison

We compare our fine-tuned models against:

- **wav2vec2-960h (Meta):** A baseline Wav2Vec 2.0 model trained on 960 hours of English speech.
- **WavLM (Microsoft):** A robust self-supervised model known for excellent performance across speech tasks.
- **MMS Model:** A multilingual wav2vec2-based model, demonstrating broad language coverage but evaluated here on English singing voices.

Results

Testing Methods

To evalutate the goodness of our results, we use the evaluation standards provided by the MIREX as we mentioned. The details are as follows:

- **Average absolute error/mean (ABE/AAM)**: It measures the time displacement between the actual timestamp and its estimate at the beginning and the end of each lyrical unit. The error is then averaged over all individual errors. An error in absolute terms has the drawback that the perception of an error with the same duration can be different depending on the tempo of the song.
- **Percentage of correct segments (PCS)**: The perceptual dependence on tempo is mitigated by measuring the percentage of the total length of the segments, labeled correctly to the total duration of the song.

Table 1: Results on the Jamendo V1 Dataset

Group	Average Absolute Error	Median Absolute Error	Percentage of Correct Segments	Percentage of Correct Onsets with Tolerance
FZZ1	0.547	0.047	0.686	0.912
NUS (Baseline)	0.217	0.046	0.751	0.945
ourModel based on wav2vec2-960h	0.626	0.084	0.673	0.897

- **Percentage of correct estimates according to a tolerance window (PCE)**: A metric that takes into consideration that the onset displacements from ground truth below a certain threshold could be tolerated by human listeners. We use 0.3 seconds as the tolerance window.

Evaluation Dataset: Jamendo V1

The Jamendo V1 dataset consists of 20 English songs with annotated lyrics, specifically curated for evaluating lyrics-to-audio alignment systems. It has been extensively used in MIREX evaluations to ensure consistent comparisons across submissions (Music Information Retrieval Evaluation eXchange (MIREX) 2024). This dataset includes audio tracks paired with corresponding lyric annotations, enabling a comprehensive evaluation using metrics such as Average Absolute Error (AAE), Median Absolute Error (MAE), Percentage of Correct Segments (PCS), and Percentage of Correct Onsets (PCO). Table 1 shows the results achieved by the submissions on this dataset.=

For submissions, the FZZ1 group used WavLM with a Conformer architecture, while the baseline submission from NUS employed a genre-informed silence model combined with a phoneme-based model. Details about the models and results can be accessed on the MIREX official website: https://www.music-ir.org/mirex/wiki/2024:Lyrics-to-Audio_Alignment_Results#Submissions.

In our experiments, we built upon the well-adapted Wav2Vec 2.0 960h model and WavLM model, fine-tuning it for a duration of **two hours**. The results achieved by our system are comparable to those of FZZ1 and even approach the NUS baseline. While our model does not outperform the leading submissions, it offers simplicity and efficiency. The entire process is streamlined, requiring minimal manual intervention, and our approach demands significantly less training time.

Moreover, the pipeline we developed allows for the model to be saved and reused, providing flexibility for future improvements. With access to larger and more diverse datasets, along with extended training durations, we believe that our model has significant potential for further enhancement. This efficient and scalable approach demonstrates the feasibility of achieving strong performance in lyrics-to-audio alignment without excessive computational resources.

Fine-Tuning Results

In this project, the baseline model used is the Wav2Vec 2.0 960h model, a pre-trained and fine-tuned version of Wav2Vec 2.0 developed by Meta. This model has proven to be highly effective, achieving over 87% correct segments, making it a strong benchmark for lyrics-to-audio alignment tasks. For base models without fine-tuning, the results are significantly less reliable, demonstrating the importance of task-specific adaptation through fine-tuning.

The MMS model, while slightly less effective than Wav2Vec 2.0 960h, still achieves robust performance, showcasing its capabilities in handling polyphonic and complex audio.

Model	ABE	AAM	PCS	PCE
wav2vec2-960h-fineturn-checkpoint-12000	**0.626**	**0.084**	**0.673**	**0.897**
wav2vec2-960h (Meta)	0.684	0.109	0.658	0.871
wav2vec2-base	91.724	89.841	0.0	0.0
wavlm-base-plus (Microsoft)	91.727	89.845	0.0	0.0
MMS Model (pytorch)	1.459	0.297	0.367	0.767
NUS (MIREX 2024)	0.217	0.046	0.751	0.945

Table 2: Performance metrics of various models.

Testing Results

After finetuning on the base wav2vec2 model, blow is the result we get, we have around 90 percent of correct segments. Even though it's not as good as the wav2vec2-960h model from Meta nor the winning team from NUS, it is still a massive improvement from the base model we originally have.

Structure To facilitate reproducibility, the dataset and alignment files were organized into a structured repository. The lyrics and audio files were stored as follows:

- **Lyrics Files**: All processed lyrics in LRC format are stored in the `songs_en/lrc` and `songs_en/lrc_processed` folder.
- **Audio Files**: Original WAV/OGG files are stored in the `songs_en/songs` folder.
- **Generated Audio Files**: The extracted vocal files and segemented audio files are stored in the `output_en/vocal` and `output_en/splits` folder.
- **Repository Structure**:

The sample dataset and preprocessing pipeline have been uploaded to an open-source repository to enhance accessibility and reproducibility. Due to file size limitations—the original dataset exceeds **30 gigabytes**—only a subset of sample files has been included. The repository is publicly available at: https://github.com/WF2205/Lyrics-audio-Alignment.

This structure enables a streamlined process for managing and analyzing the lyrics-to-audio alignment dataset, providing a strong foundation for further fine-tuning and evaluations.

Listing 1: Folder Structure of the Dataset and Generated Files

```
dataset/
|
|---- output-en/              # Directory for generated audio files
|      |---- splits/          # Segmented audio files
|      |---- vocal/           # Extracted vocal files
|      |---- metadata.csv     # Metadata file for tracking
|      |---- metadata_new.csv# Updated reformatted metadata file
|
|---- songs_en/               # Directory for original dataset files
       |---- lrc/             # Original lyrics files in LRC format
       |---- lrc_processed/   # Processed lyrics files for alignment
       |---- songs/           # Original WAV/OGG audio files
       |---- songs_track/     # Additional track-related data
```

Fine-Tuning Results and Analysis The fine-tuning process was mainly conducted on the Wav2Vec 2.0 960h model. Figure 1 and Figure 2 show the training and validation loss curves over the training steps.

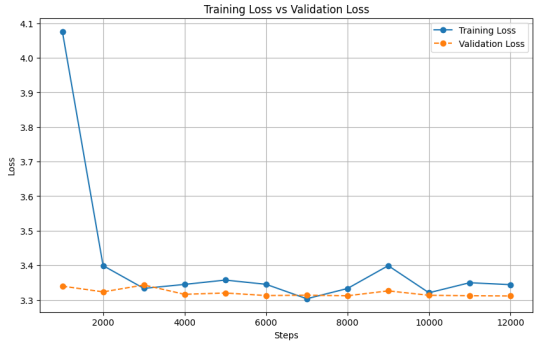

Figure 1: Training and Validation Loss when fine-turning base model

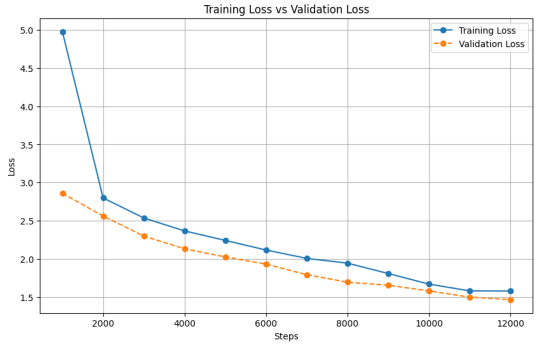

Figure 2: Extended Training and Validation Loss when fine-turning Wav2Vec 2.0 960h model

The loss curves in Figure 1 indicate that both training and validation loss stabilize quickly after the initial steps, reflecting the model's ability to adapt efficiently during fine-tuning. Figure 2 shows further fine-tuning over a longer duration, with a continuous decrease in both training and validation loss, highlighting the potential for improved alignment performance with additional training.

The results suggest that:

- The Wav2Vec 2.0 model base fine-tuned on the dataset achieves rapid convergence, indicating efficient adaptation to singing voice tasks.
- Extended training shows potential for further improving alignment accuracy, as evident from the declining loss values.
- **With more training time and access to additional annotated data, the model could achieve results comparable to state-of-the-art systems**.

Conclusion

In this study, we addressed the challenges of lyrics-to-audio alignment by leveraging state-of-the-art self-supervised models and an efficient processing pipeline. Building on the Wav2Vec 2.0 960h model, we performed two hours of fine-tuning, achieving results comparable to leading systems such as FZZ1 and the NUS baseline on the Jamendo V1 dataset. Despite achieving slightly lower alignment accuracy than the best-performing submissions, our approach is significantly simpler, requiring less training time and minimal manual intervention.

The pipeline includes robust vocal extraction using HT-Demucs, precise segmentation of lyrics and audio, and forced alignment with fine-tuned models. We also demonstrated the potential for further improvement by extending training and incorporating additional annotated datasets. The entire pipeline is shared in an open-source repository, enabling reproducibility and fostering future innovation.

Through this work, we emphasize the importance of streamlined workflows for advancing music information retrieval. By focusing on accessibility and scalability, our approach lays the groundwork for broader adoption and refinement of lyrics-to-audio alignment technologies.

References

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv preprint*, arXiv:2006.11477.

Bittner, R. M.; Raffel, C.; Salamon, J.; and Bello, J. P. 2014. MedleyDB: A Multitrack Dataset for Annotation-Intensive

MIR Research. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 155–160. Taipei, Taiwan.

Chen, S.; Wang, C.; Wu, Y.; and Liu, S. e. a. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *arXiv preprint*, arXiv:2110.13900.

Dzhambazov, G. 2017. *Knowledge-Based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals*. Doctoral dissertation, Universitat Pompeu Fabra.

f90. 2024. JamendoLyrics: A Dataset of Song Lyrics with Aligned Audio Features. https://github.com/f90/jamendolyrics. Accessed: 2024-12-18.

Huang, J.; and Benetos, E. 2024. Towards Building an End-to-End Multilingual Automatic Lyrics Transcription Model. *arXiv preprint*, arXiv:2406.17618.

Lee, S. W.; and Scott, J. 2017. Word Level Lyrics-Audio Synchronization Using Separated Vocals. In *ICASSP Proceedings*. New Orleans, LA, USA: IEEE.

Meseguer-Brocal, G.; Cohen-Hadria, A.; and Peeters, G. 2018. DALI: A Large Dataset of Synchronized Audio, Lyrics, and Notes. In *ISMIR Proceedings*. Paris, France: ISMIR.

Mikezzb. 2023. Lyrics-Sync Repository. https://github.com/mikezzb/lyrics-sync.

Music Information Retrieval Evaluation eXchange (MIREX). 2024. Lyrics-to-Audio Alignment Task. https://www.music-ir.org/mirex/wiki/2024:Lyrics-to-Audio_Alignment. Accessed: 2024-01-10.

Pan, S. J.; and Yang, Q. 2009. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359.

Park, J.; Yong, S.; Kwon, T.; and Nam, J. 2024. A Real-Time Lyrics Alignment System Using Chroma and Phonetic Features for Classical Vocal Performance. *arXiv preprint*, arXiv:2401.09200.

Rotem Rousso , Eyal Cohen , Joseph Keshet , Eleanor Chodroff. 2024. Paper Title. *arXiv preprint arXiv:2406.19363*.

Rouard, S.; and Peeters, G. 2021. Hybrid Transformer Based Source Separation for Better Separability of Sources. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 202–209. Online: ISMIR.

Appendix: Example Visualization

To provide users with an intuitive understanding of the alignment results, an example visualization is included. Users can generate similar results by dragging the `song_name.lrc` file and the corresponding `song_name.ogg/wav` file into the web-based tool available at https://mikezzb.github.io/lrc-player/. This tool displays the synchronized lyrics in real time, enabling a clear perception of the alignment quality.

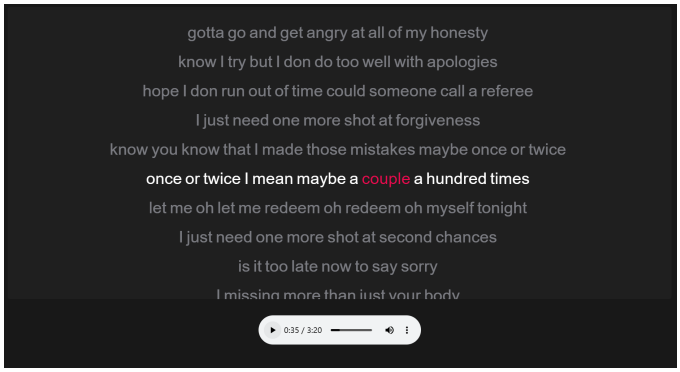

Figure 3