MULTILINGUAL GRAPHEME-TO-PHONEME AND LYRICS-TO-AUDIO ALIGNMENT

Wanpeng Fan **Jiaye Zhu** **Peng Zhong**

Guangzhou Huancheng Culture Media Co., Ltd.
{fanwanpeng, zhujiaye, zhongpeng}@52tt.com

ABSTRACT

We present our system for MIREX 2024 Lyrics-to-Audio Alignment task. Our system utilizes separated vocal tracks as input and a joint training objective with pitch prediction to train an acoustic module. We introduce pitch extraction and voice activity detection (VAD) module in the alignment pipeline to further augment the result of the trained model, and improve the overall performance of lyrics-to-phonemes transcription to retain sufficient alignment in multilingual application scenarios. The experimental results show that our system can perform well in multilingual lyric alignment scenarios.

1. INTRODUCTION

Lyrics-to-Audio Alignment (LAA) is one of the fundamental tasks in music information processing and understanding. With the increasing attention in music generative models [1,2], LAA has become essential in automated data pre-processing for large-scale datasets.

A typical LAA system comprises a pre-processing module, an acoustic processing module, an alignment model, and finally a post-processing module. It requires the system to effectively synchronize the timestamps between a singing audio segment and its corresponding lyrics in various granularities such as sentence, phrase, word, or phoneme level, depending on the use case. The difficulty of designing and implementing LAA arises with the scale of alignment, i.e., a phoneme-level alignment finds the smallest alignment in time and is the hardest to achieve. Furthermore, the languages supported by LAA should possibly cover the common languages in music as opposed to only English or Chinese. We target our design at *multilingual* lyrics alignment at the *phoneme level* to support most use cases.

Previous methods align lyrics at the word level [3], in which only the starting and ending timestamps are provided for each word in the lyrics. We find this setup insufficient for downstream tasks that require finer-level alignment at phonemes. For example, in music diffusion models that adopt a text-to-speech approach [4, 5], lacking phoneme-level alignment may cause blurry pronunciation artifacts. Phoneme-to-audio timestamps may also benefit downstream tasks utilizing coarse alignments, such as word or sentence level. In this work, we aim to align phonemes in our system. Phoneme alignment also enables our system to perform multilingual LAA. We use a Grapheme-to-Phoneme (G2P) conversion to transcribe multiple languages, including English, Chinese, German, French, etc.

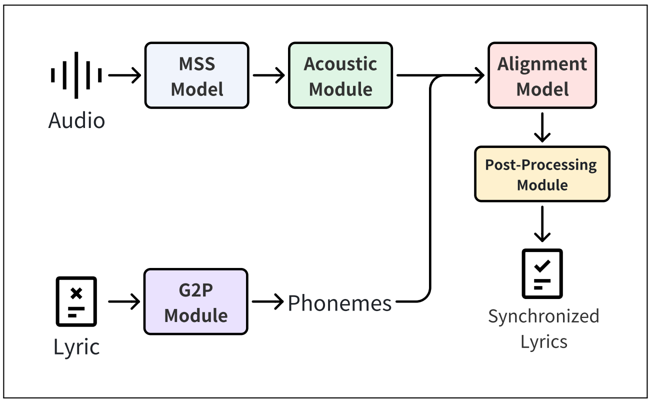

Figure 1. Our proposing system **FZZ** aligns phonemes with pre-trained audio embeddings to achieve multilingual and phoneme-level lyrics-to-audio alignment.

Most LAA systems are prone to errors when predicting the boundaries of lyrics. The long-tail voice in singing is often truncated by an early-predicted ending timestamp, resulting in misalignment. To address this issue, we introduce a joint training objective with pitch prediction in the acoustic module, and Conformer [6] modules to simultaneously predict alignment and boundary. We also redesigned the post-processing pipeline, incorporating common singing features like the fundamental frequency (F0) to improve the overall alignment performance.

In the following sections, we discuss the detailed design choices of our LAA pipeline and provide quantitative comparisons against previous methods to show the effectiveness of our method.

Attribution: Wanpeng Fan, Jiaye Zhu, and Peng Zhong, “Multilingual Grapheme-to-Phoneme and Lyrics-to-Audio Alignment”, in *Extended Abstracts for the Late-Breaking Demo Session of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

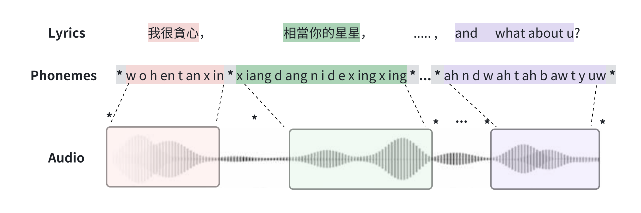
Lyrics
我很貪心，
相當你的星星，
.....，
and what about u?
Phonemes
* w o h en t an x in * x iang d ang n i d e x ing x ing * ... * ah n d w ah t ah b aw t y uw *
Audio
*
*
*
...
*
*

Figure 2. Overview of lyric-to-audio alignment with mixed languages. We use CMU-dict style phonemes as unified pronunciation representation. <*> denotes the sentence separator.

2. PROPOSED METHOD

We provide a general overview of our system **FZZ** in Figure 1. For multilingual inputs, we first transcribe the lyrics into unified phoneme vocabulary with a G2P module. The audio input is first processed by music source separation (MSS) to obtain the vocal track, followed by an acoustic understanding module to extract vocal features. The vocal features are aligned with phonemes, and finally we rectify the predicted boundaries via a carefully designed post-processing schedule.

2.1 Transcribing Lyrics

To first align the phonemes from different families of language, one can choose the International Phonetic Alphabet (IPA) or Carnegie-Mellon University pronouncing dictionary (CMU-Dict) as the vocabulary set. Previous attempts [7, 8] with IPA directly use raw lyrics as input without transcription and predict IPA sequences by a deep neural network. Although the training is similar to machine translation and no explicit linguistic expert knowledge is required, the complexity of IPA sequences makes it difficult to learn consonant-vowel-audio correspondence in the alignment model. A large vocabulary set of 163 characters and more numbers of tokens require a larger parameterized model to learn their semantics, which is computationally heavy and data-expensive. By contrast, CMU-dict has a vocabulary of 39 characters and cannot represent the exact pronunciation of every language. We choose CMU-dict as our phoneme vocabulary as we do not target the exact alignment to the smallest pronunciation units. We refer to CMU-dict characters as phonemes unless otherwise noted.

An overview of multilingual alignment with phonemes is shown in Figure 2. Lyrics lines from a single piece of audio composed of multiple languages are converted into a unified CMU-dict representation. We add an extra placeholder token < * > at the start and end of sequence tokens of each lyric line to represent any out-of-vocabulary voice or sound, increasing our method's robustness.

2.2 Joint Training

Inspired by K.W.Cheuk et al. [9], introducing pitch prediction in the training process of LAA benefits the final performance of the system. Figure 3 shows the joint training in our system, where pitch prediction can form an implicit alignment to help identify melodic vocal components.

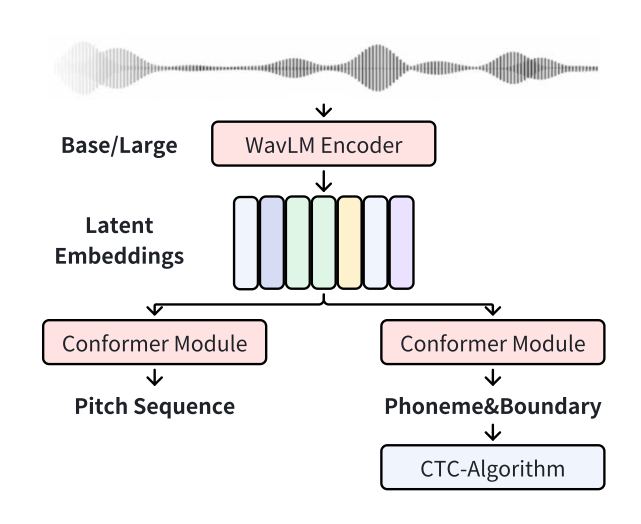

Figure 3. Overview of joint training of pitch and phoneme boundary prediction. An auxiliary pitch prediction target can augment the final performance of the alignment.

We follow the structural design of Unconstrained Forced Aligner [6] to build the acoustic module for the joint training objective. As shown in Figure 3, a vocal segment is passed into pre-trained WavLM [10] encoder for latent representations. Two separate Conformer [11] modules predict the corresponding pitch sequence and phoneme boundaries respectively. For pitch prediction, we use a floating number ranged [0, 127] and normalized to represent the pitch targets, and then optimize by a regression objective.

2.3 Pre and Post-Processing

Musical structure can be complicated. For example, voice may not be present in the Intro and Outro, and sometimes pure instrumental segments can last tens of seconds long. Following the previous practices, we apply source separation to operate only on the vocal track and decouple the model from learning dependencies from accompaniments. For trimming non-vocal segments, we use voice activity detection (VAD) as pre-processing to make a coarse alignment of an approximate lyric location before feeding to the model to avoid the lack of generalization on long silent audio segments.

We observe the alignment model often suffers from the voice boundary, especially at the end, where long-tail voice and breathing sounds are misclassified. We correct the ending timestamp by an F0 post-processing process. Energy-based VAD may mistake breathing or other noise for voice. We use the auxiliary pitch sequence output from our alignment model to further check the inclusion of non-pitched sound.

Methods	Abs. Err	CS%	CE%
GGL1 [3]	0.33	N/A	0.94
GGL2 [3]	0.22	N/A	0.94
ZWZL2 [12]	0.61	N/A	0.88
FZZ (Ours, English)	0.34	**0.77**	**0.94**
FZZ (Ours, Multi-Lang)	0.36	0.72	0.88

Table 1. Quntitative comparison to baseline methods on the Jamendo dataset. Our method outperforms all other methods and is able to perform multilingual.

3. RESULTS

We evaluate our method on the Jamendo dataset [13] with 80 songs in German, French, and English. We compare to previous methods GGL1, GGL2 [3], and ZWZL2 [12] reporting average absolute error (Abs. Error), percentage of correct segments (CS%), and percentage of correct estimates with tolerance (CE%) metrics.

In Table 1, our method outperforms or is on par with all previous methods on word-level CE despite being a phoneme-level LAA, showing our design effectively mitigates the boundary errors commonly seen before. For other languages, our method reaches a high-performance level with CS and CE 72%, 88%, respectively. Due to our system's ability to align lyrics in multiple languages and phoneme levels, it will be able to be applied in more complex downstream scenarios related to music generation.

4. REFERENCES

[1] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Long-form music generation with latent diffusion," *arXiv preprint arXiv:2404.10301*, 2024.

[3] X. Gao, C. Gupta, and H. Li, "Lyrics transcription and lyrics-to-audio alignment with music-informed acoustic models," in *MIREX*. MIREX, 2021.

[4] Y. Wu, J. Shi, T. Qian, D. Gao, and Q. Jin, "Phoneix: Acoustic feature processing strategy for enhanced singing pronunciation with phoneme distribution predictor," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] B. Maman, J. Zeitler, M. Müller, and A. H. Bermano, "Performance conditioning for diffusion-based multi-instrument music synthesis," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5045–5049.

[6] J. Lian, C. Feng, N. Farooqi, S. Li, A. Kashyap, C. J. Cho, P. Wu, R. Netzorg, T. Li, and G. K. Anumanchipalli, "Unconstrained dysfluency modeling for dysfluent speech transcription and detection," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[7] M. N. Sundararaman, A. Kumar, and J. Vepa, "Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript," *arXiv preprint arXiv:2102.00804*, 2021.

[8] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché Buc, "Multilingual lyrics-to-audio alignment," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[9] K. W. Cheuk, K. Choi, Q. Kong, B. Li, M. Won, J.-C. Wang, Y.-N. Hung, and D. Herremans, "Jointist: Simultaneous improvement of multi-instrument transcription and music source separation via joint training," *arXiv preprint arXiv:2302.00286*, 2023.

[10] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[11] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[12] B. Zhang, W. Wang, E. Zhao, and S. Lui, "Lyrics-to-audio alignment for dynamic lyric generation," in *Music Inf. Retrieval Eval. eXchange Audio-Lyrics Alignment Challenge*, 2022.

[13] S. Durand, D. Stoller, and S. Ewert, "Contrastive learning-based audio to lyrics alignment for multiple languages," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.