2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

VietLyrics: A Large-Scale Dataset and Models for Vietnamese Automatic Lyrics Transcription

Nguyen Quoc Anh* and Bernard Cheng† and Kelvin Soh‡
* National University of Singapore, Singapore
E-mail: e1124714@u.nus.edu Tel/Fax: +65-80706568
† National University of Singapore, Singapore
E-mail: e1124449@u.nus.edu Tel/Fax: +65-98557620
‡ National University of Singapore, Singapore
E-mail: e1124460@u.nus.edu Tel/Fax: +65-91186401

Abstract—**Automatic Lyrics Transcription (ALT) for Vietnamese music presents unique challenges due to its tonal complexity and dialectal variations, but remains largely unexplored due to the lack of a dedicated dataset. Therefore, we curated the first large-scale Vietnamese ALT dataset (VietLyrics), comprising 647 hours of songs with line-level aligned lyrics and metadata to address these issues. Our evaluation of current ASR-based approaches reveal significant limitations, including frequent transcription errors and hallucinations in non-vocal segments. To improve performance, we fine-tuned Whisper models on the VietLyrics dataset, achieving superior results compared to existing multilingual ALT systems, including LyricWhiz. We publicly release VietLyrics and our models, aiming to advance Vietnamese music computing research while demonstrating the potential of this approach for ALT in low-resource language and music.**

I. Introduction

In the modern music industry, Automatic Lyrics Transcription (ALT) has become increasingly relevant and important. It describes the process of leveraging deep learning models to convert vocal recordings into written text, making song lyrics accessible and searchable. This enhances the user experience on streaming platforms, aids in music analysis, and supports subtitle creation for music videos. It also promotes inclusivity for the hearing impaired and non-native speakers. As the industry evolves, accurate lyrics transcription is essential for greater engagement and accessibility.

Despite advances in Western music computing research, Vietnamese music research lags behind, especially lyrics transcription, which remains mostly unexplored due to the lack of a large-scale, high-quality Vietnamese ALT dataset. Without technology like ALT or auto-tagging, Vietnamese media companies and streaming platforms lack the necessary tools to analyze, classify, and describe their content. This makes it difficult to build standardized datasets, hindering further research and innovation in the field, ultimately leaving Vietnamese music computing research stagnant.

This paper aims at tackling the task of automatic lyric translation for the Vietnamese language. Accurate transcription of Vietnamese is arguably more difficult than English due to additional language characteristics, such as changes in tonality leading to different meanings. Vietnamese also contains a multitude of regional dialects, each with unique pronunciation patterns and differences in vocabulary. Furthermore, changes in the speaker's pitch, accent, and emotionality can affect tonality, altering the diacritic mark, and thus the meaning of the transcribed word. For example, Tằm may sound similar to Tắm during singing, but they have completely different meanings. These complexities pose significant challenges in accurately transcribing Vietnamese vocals, especially when it is intertwined with melodies on the same track.

In this work, these challenges are addressed by making the following key contributions:

- The first large-scale Vietnamese ALT dataset, VietLyrics was constructed, which comprises 647 hours of songs with line-level aligned lyrics, supplemented with AI-predicted metadata such as gender and genre. The dataset is publicly released for research purposes, in compliance with Vietnamese copyright law.
- The Whisper [1] architecture was further fine-tuned using the VietLyrics dataset, significantly improving performance over existing multilingual ALT/ASR systems. Three model variants of different sizes are released to support further research and development in Vietnamese lyrics transcription.

II. Related Works

The following sections outline the current challenges in Vietnamese ALT and discuss the state-of-the-art solution (SOTA), detailed in Section II-B.

A. Dataset Scarcity

Table I presents a non-exhaustive list of open-source song-lyric datasets. There is a lack of large-scale non-English song datasets in general, due to copyright law. These datasets circumvent this issue by collecting music exclusively under Creative Commons licenses or in the Public Domain [2], [4], [6], leveraging AI-generated lyrics [3], and incorporating karaoke vocals voluntarily shared by app users [5]. None of these

TABLE I
OPEN-SOURCE DATASETS

Dataset	Languages	Songs (Hours)
DALI-full [2]	30 (mostly EN)	5,358
DALI-train [2]	1 (EN)	3,913
MulJam [3]	6 (EN, FR, DE, ES, IT, RU)	6,031 (381.9h)
JamendoLyrics [4]	4 (EN, ES, DE, FR)	80
DSing30 [5]	1 (EN)	4,324 (149.1h)
MUSDB18 [6]	1 (EN)	150 (10h)
VietLyrics-full (Ours)	1 (VN)	8,440 (647.1h)

include Vietnamese, based on our findings, which leads to the lack of Vietnamese ALT models.

B. Limitations in Existing Solutions

Although open source and commercial solutions are available for ALT or Automatic Speech Recognition (ASR) tasks, the general performance on Vietnamese music is error-prone. Various solutions were tested, including Whisper, LyricWhiz, PhoWhisper and one closed-source commercial API. While all solutions resulted in some degree of substitution, insertion, deletion errors, along with long stretches of hallucination during periods with only melody and no vocals, ASR application Whisper stands out because it wrongly recognizes the audio as Thai or other languages regularly.

Among existing ALT frameworks, LyricWhiz [3] is the current SOTA for multilingual ALT, leveraging both Whisper for transcription and GPT-4 for refining outputs. Whisper serves as the "ear" by converting audio into text, while GPT-4 functions as the "brain", selecting the most accurate lyrics. While LyricWhiz has demonstrated strong performance across various prominent languages, its effectiveness on Vietnamese lyrics is limited. The replicated baseline model struggles with Vietnamese audio, as shown in the qualitative results in Table II. Note that other solutions also suffer from similar errors. Most errors stem from the inherent complexity of Vietnamese phonetics and the fact that Whisper, being primarily an ASR model, is not optimized for ALT tasks.

III. Proposed Method

A. Dataset

Around 647.1 hours of Vietnamese song audio with accompanying lyrics were scraped from zingmp3.vn, a Vietnamese music streaming site, to form the VietLyrics dataset. This dataset will be released publicly in the form of metadata, scraped links and scraping code, with author attributions to respect Vietnamese Intellectual Property Law. Quantitative and qualitative analysis shows that VietLyrics is diverse in genres, instruments, regional dialects, and gender.

1) Data Collection & Processing: Out of 120,000 raw URLs collected by querying Vietnamese song titles using the zingmp3.vn search function, the dataset was filtered down to 17,000 entries. Heuristics were applied to exclude instrumental tracks, songs with English titles, titles containing numbers, and other irrelevant entries, ensuring the retained songs were predominantly in Vietnamese. For instance, searching for songs with numbers in their titles often resulted in tracks from other languages. To maximize dataset diversity, duplicate song names and potential cover versions were removed. A total of 12,758 songs and their metadata were successfully scraped without encountering significant web security challenges. Given that zingmp3.vn is region-restricted, a VPN should be used to access the platform from abroad.

8,440 of 12,758 songs were found to have line-level lyric transcriptions that require further processing to ensure usability. This involved removing line breaks, punctuation, redundant timestamps, and introductory content from lyric transcriptions. Some songs were also identified as having incomplete transcriptions, in which the lyrics stopped before the final chorus repeats. To address this, the corresponding audio was trimmed to match the transcription length, plus an additional 10 seconds, ensuring no vocal segments were left without accompanying lyrics. Lastly, the sample rate for all audio tracks was standardized to 16 kHz.

Under Article 25, Clause 1(a) of the Vietnamese Intellectual Property Law (amended June 16, 2022), copyrighted musical works may be reproduced for academic and research purposes without permission or fees, provided the original creators are credited [7]. To respect copyright, we also scraped 2 other lyrics websites [8], [9] for songwriters and artists for proper attributions, to the best of our abilities. We publicly release all metadata (including song names, links, and author attributions) and scraping code for academic research, allowing future studies to recreate the dataset, train, and benchmark for automatic transcription or other tasks.

2) Data Analysis: The dataset includes songs from over 4,000 unique artists, lasting on average 4.6 minutes/song. The average singing speed is found to be 90.1 Words Per Minute (WPM), less than half the Vietnamese spoken rate of 190 WPM [10]. The dataset includes various genres, as shown in Fig. 1. However, metadata extraction from zingmp3 alone resulted in many songs lacking genre information. Since Vietnamese music research lacks standardized genre classification tools or an agreed-upon genre ontology, Contrastive Learning of Musical Representations (CLMR) [11] was used to infer the top 50 MagnaTagATune tags [12]. The result was of weak alignment, with the 95th percentile confidence score reaching only 0.22. This suggests that this collection of Vietnamese music does not fit well within existing Western music's tag ontology, highlighting the need for further research in Vietnamese music genre classification.

Vietnamese dialects (which include both accents and variations in word choices, idioms, sentence structure, etc.) differ significantly across regions, a difference even more pronounced in singing, making accent classification valuable for music analysis. A pre-trained Wav2Vec2-base model [13], [14] fine-tuned on Vietnamese Speech corpora [15], [16] was used to predict gender [17] and dialect [18].

We extracted and ran inference on three 10-second vocal-only

TABLE II
QUALITATIVE ANALYSIS ON BASELINE LYRICSWHIZ'S PREDICTION

Ground Truth	**Prediction**	**Source Song**	**Analysis**
(No Vocals)	Thành thành chuyện Thành thành; Hãy subscribe cho kênh Ghiền Mì Gõ Để không bỏ lỡ những video hấp dẫn;...	Ừ Thì Anh Sai - Thiên Tú	Long stretches of complete hallucination during no-vocals, due to the ASR model is not addapted into ALT
(No Vocals)	Lời Tất Này Lời Tất Này; What the hell;...	365 Ngày Yêu (Remix) - Lương Gia Huy	Insertion of hallucinative random phrases
Nước mắt em rơi cho cõi lòng em nắm tan	Ngược mắt em rơi cho cõi lòng em đã tan	Anh Biết Không Anh - Lưu Ánh Loan	Substitutions of single words that can reasonably be misheard
Chẳng lẽ ta mất nhau để em đây khổ sầu; Thì ta đã lìa xa;...	Chẳng lẽ ta mất nhau để em đây khô sâu; Thì ta đã lịa xa;...	Anh Biết Không Anh - Lưu Ánh Loan	Substitutions or deletion of a single word's diacritic that can reasonably be mis-transcribed due to regional accents.

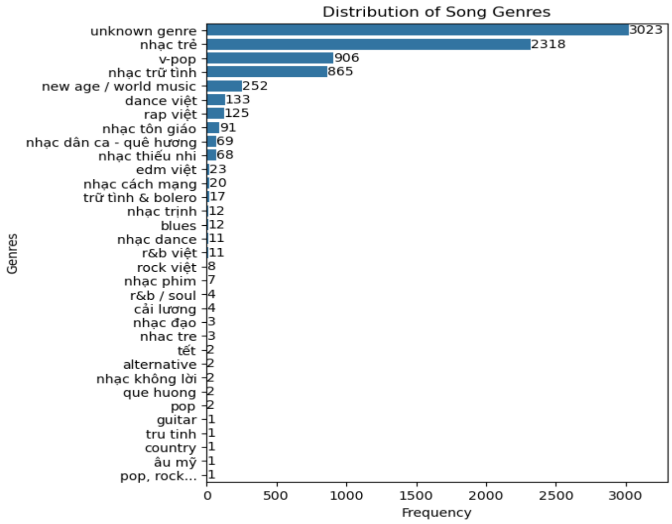

Fig. 1. Distribution of songs by genres. Extracted songs covers a wide range of genres, however a lot is missing.

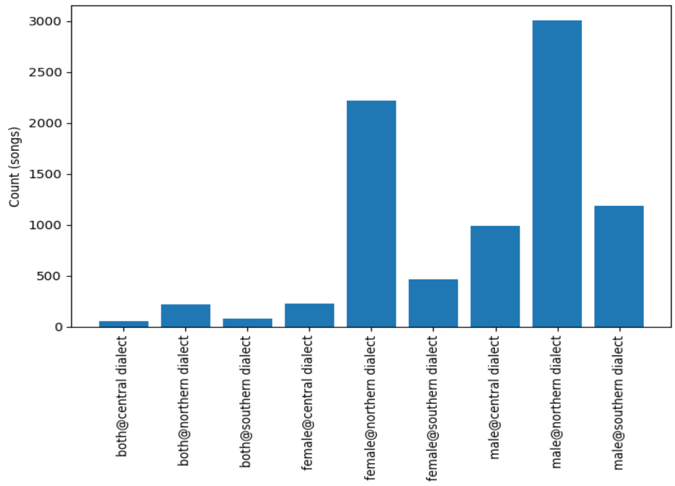

Fig. 2. Distribution of the dataset based on predicted genders and regional dialects.

segments (isolated using Demucs v3 [19]) and determined the final classification through majority voting. For gender classification specifically, a "both" category was introduced for cases where at least one segment had a confidence score between 0.4 and 0.6, indicating ambiguity due to a unisex voice, absence of vocals, or a mix of male and female vocals. As shown in Fig. 2, the results indicate a distribution skewed towards Northern dialects, while Central dialects are underrepresented. Male-sung songs were more prevalent than those classified as "female" or "both". This data imbalance is hard to address, as gender and accent metadata were unavailable during data collection. (Note that this observed skew can also stem from Wav2Vec2 model bias.)

A qualitative examination revealed unique challenges, including lyrics in multiple languages (e.g., English, Mandarin in Latin characters), nonsensical vocalizations like "là la la ..." (*Kết Thúc Vậy Sao* by Ngô Trác Lâm), and inconsistencies in whether background vocals are transcribed (*Phút Giây Muộn Màng* by Bắc San Ho, *Người Không Cần Yêu* by Khánh Đơn). These issues may require further processing to address.

After using both manual qualitative sampling and quantitative exploration, the VietLyrics dataset is found to be sufficiently diverse, covering a large range of Vietnamese genres, traditional and modern instruments, regional music, and various dialects from across the country.

B. Modeling

Since the current SOTA LyricWhiz model does not effectively handle Vietnamese songs, our approach is to fine-tune an ASR model end-to-end to serve as an ALT system. The Whisper model [1] was chosen as the starting point, as it is renowned for its SOTA performance in multilingual tasks and exceptional fine-tuning capabilities.
Three Whisper model variants of different sizes (small, medium and large-v2) were fine-tuned and made available as checkpoints on Hugging Face [20].

1) Data Pre-processing: Audio segmentation of songs is performed using the provided ground truth timestamps from the song lyrics to chunk both the audio and lyrics into 30-second segments. This is required as Whisper processes audio in 30-second chunks [1].

To enhance training stability, audio and lyrics were sliced to keep only segments with corresponding ground truth transcription. This removes silences and parts without vocals, reducing dataset size and accelerating training while preserving meaningful content.

2) Fine-Tuning: The dataset of 8,440 Vietnamese songs was divided into 7,440 training and 1,000 evaluation. All Whisper variants were fine-tuned using consistent hyper-parameters: batch_size = 8, gradient_accumulation_steps = 2, learning_rate = 1e-5, warmup_steps = 500, and max_steps = 5000. Training was conducted in FP16 precision to optimize computational efficiency. Our approach involved iterative fine-tuning, starting with smaller models (small and medium variants) to experiment with and refine our pre-processing techniques. After validating these improvements on the smaller models, we applied them to fine-tune the larger model, ensuring that the methodologies and pre-processing strategies generalized effectively across different model scales such as scaling from Whisper-small, to Whisper-medium and to Whisper-large-v2.

IV. Experiments and Results

A. Finetuned Models

LyricWhiz was evaluated as a baseline for lyrics transcription, given its state-of-the-art performance in prior studies [3], though it had not been tested on Vietnamese songs. PhoWhisper [21], an ASR model fine-tuned on Vietnamese audio, was also assessed with the expectation of better performance on Vietnamese datasets. On VietLyrics test set, LyricWhiz produced a Word Error Rate (WER) of 49.22%, while PhoWhisper-large achieved 38.3%, shown in Table III. However, both models fell short of the accuracy needed for effective Automatic Lyrics Transcription.

All Whisper models were fine-tuned on a dataset consisting of 7,440 training samples and 1,000 validation samples. The reported metrics were evaluated on the 1,000 validation samples. For context, Whisper was trained on 691 hours of Vietnamese ASR data, PhoWhisper was fine-tuned on 844 hours, and ours was fine-tuned on about 570 hours of songs. Our fine-tuned Whisper models were evaluated using two key metrics: Word Error Rate (WER) and Character Error Rate (CER). CER is particularly useful in accounting for cases where the model's predictions are close but incorrect due to minor diacritic errors, allowing such predictions to be considered as partially correct.

The results are summarized in Table III. As shown, scaling the model size significantly improves performance, with the best-performing Whisper-large-v2 model achieving a WER of 24.61% and a CER of 17.14%. The reported metrics demonstrate that Whisper-large-v2 is a robust ASR model suitable to be used for ALT for a low-resource language/music like Vietnamese. The relatively lower CER indicates that the model can accurately predict partial words, with errors arising from character or diacritic replacements.

Through qualitative analysis, Whisper-large-v2 was found to perform well on various instruments, music genres, and styles. As for gender and dialects, Whisper-large-v2 was found to have performed equally well in most cases, see Fig. 3. Except for "both" gendered voices with Southern dialect, which we consider to be insufficient data to have any conclusive assertion, the model performs worse on female Southern dialect, a problem that needs to be addressed in future research.

We also experimented with training and evaluating only on vocals, separated using Demucs v3 [19]. However, the improvement observed after source separation was minimal for both models, with 1.3% WER improvement for whisper-medium and 0.5% WER improvement for whisper-large-v2. Given the negligible benefits and the added complexity of these extra pre-processing steps, the decision was made to exclude these steps from the final approach.

Lower-case evaluation was also found to produce better performance metrics compared to case-sensitive evaluation, with WER decreasing by up to 5% on large, as shown in Table IV. This intuitive result suggests that relying on the LLM-based transcriber for capitalization may not be the optimal approach. Instead, more effective methods for inferring line breaks could involve post-processing techniques, such as detecting silence between singing segments or leveraging the approaches introduced in Cífka et al. [22].

TABLE III
Model Performance
WER (%) and CER (%) Results.

Model	WER	CER
Baseline: LyricWhiz	49.22	38.55
Baseline: PhoWhisper-large	38.3	27.5
Whisper-small (Fine-tuned)	34.91	24.82
Whisper-medium (Fine-tuned)	26.42	17.03
Whisper-large-v2 (Fine-tuned)	24.61	17.14
Whisper-medium (Demucs, Fine-tuned)	25.12	16.34
Whisper-large-v2 (Demucs, Fine-tuned)	24.11	16.21

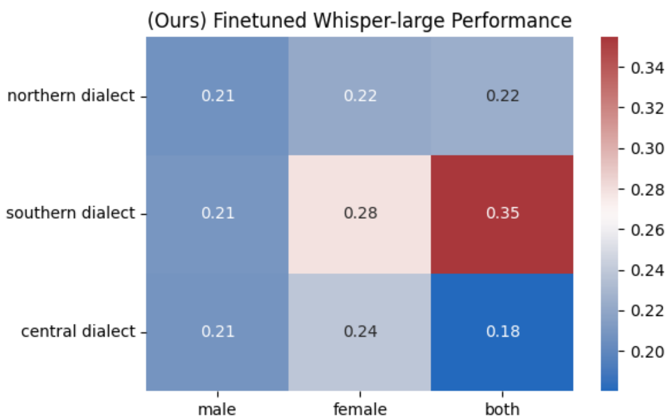

Fig. 3. Our Fine-tuned Whisper-large performance WER (%), evaluated based on predicted genders and regional dialects.

B. <nospeech> suppression

Since hallucination from non-vocal audio is a major challenge in ASR models, experiments were conducted to suppress <nospeech> tokens. Following the original Whisper paper [1], tokens were filtered when the <nospeech> probability

TABLE IV
MODEL PERFORMANCE
WER (%) RESULTS FOR CASE-SENSITIVE VS LOWERCASE

Model	WER (Case-sensitive)	WER (Lowercase)
Whisper-medium	26.42	23.15
Whisper-large-v2	24.61	20.52

exceeded 0.6 and the log confidence probability was below -1. These experiments were evaluated on 100 randomly sampled songs from the 1,000-song validation set. PhoWhisper was tested under various configurations, with vocals-only tracks extracted using Demucs. As shown in Table V, for PhoWhisper-small, combining source-separated audio with `<nospeech>` suppression provided the best performance. However, for Whisper large, testing on raw audio without suppression performed best, with our fine-tuned Whisper significantly outperforming others.

This outcome is consistent with the findings in Cífka et al. [22], where the transcriptions from only vocals are less accurate than those from raw data for large models. This is preferable, as it enables deploying an end-to-end model without dependence on source separation performance. Furthermore, qualitative reviews of our finetuned Whisper's inferences found no instances of long hallucinations, resulting in the discontinuation of further silence suppression experiments.

TABLE V
SILENCE SUPPRESSION PERFORMANCE
WER (%) AND CER (%) RESULTS FOR LOWERCASE.
NS MEANS <NOSPEECH> TOKEN IS SUPPRESSED.

Model	Input	NS	WER	CER
Whisper-large-v2 (Ours)	Raw Data	No	20.1	16.0
PhoWhisper-large	Raw Data	No	38.3	27.5
PhoWhisper-large	Only Vocals	Yes	41.8	31.5
PhoWhisper-small	Raw Data	No	54.2	38.4
PhoWhisper-small	Raw Data	Yes	54.4	39.1
PhoWhisper-small	Only Vocals	No	50.2	35.3
PhoWhisper-small	Only Vocals	Yes	48.7	34.5

V. CONCLUSION

We collected VietLyrics, a large-scale diverse dataset of Vietnamese music with lyrics, and fine-tuned the Whisper models [1] to develop an Automatic Lyrics Transcription system. Our best performing model, Whisper-large-v2, achieved a WER of 24.61% for case-sensitive lyrics and 20.52% for lowercase lyrics, demonstrating its effectiveness in transcribing Vietnamese songs of various dialects, genres and genders. We publicly release the VietLyrics dataset [23] and models [24]–[26], hoping that this work inspires further research and improvements in Automatic Lyrics Transcription and music computing in general, for the Vietnamese language and music.

VI. ACKNOWLEDGMENT

We would like to thank Zhao Junchuan and Nguyen Thi Minh Thu for their invaluable help and support throughout this research.

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[2] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm.," 2018. DOI: 10.5281/ZENODO.1492443. [Online]. Available: https://zenodo.org/record/1492443.

[3] L. Zhuo, R. Yuan, J. Pan, *et al.*, "Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to chatgpt," *arXiv preprint arXiv:2306.17103*, 2023.

[4] 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, Jun. 2023, pp. 1–5. DOI: 10.1109/icassp49357.2023.10096725. [Online]. Available: http://dx.doi.org/10.1109/ICASSP49357.2023.10096725.

[5] G. Roa Dabike and J. Barker, "Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019)*, 2019.

[6] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, *The MUSDB18 corpus for music separation*, Dec. 2017. DOI: 10.5281/zenodo.1117372. [Online]. Available: https://doi.org/10.5281/zenodo.1117372.

[7] *Law no. 07/2022/qh15 of june 16, 2022, amending and supplementing a number of articles of the law on intellectual property, viet nam, wipo lex.* [Online]. Available: https://www.wipo.int/wipolex/en/legislation/details/21740.

[8] [Online]. Available: https://lyricvn.com/.

[9] [Online]. Available: https://loibaihat.biz/.

[10] B. Dağ, "10 fastest spoken languages in the world," 2024. [Online]. Available: https://lingopie.com/blog/fastest-spoken-languages-in-the-world/.

[11] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," *CoRR*, vol. abs/2103.09410, 2021. arXiv: 2103.09410. [Online]. Available: https://arxiv.org/abs/2103.09410.

[12] E. Law, K. West, M. Mandel, M. Bay, and J. Downie, "Evaluation of algorithms using games: The case of music tagging.," Jan. 2009, pp. 387–392.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. arXiv: 2006.11477. [Online]. Available: https://arxiv.org/abs/2006.11477.

[14] V. L. B. Nguyen, *Nguyenvulebinh/wav2vec2-base-vi·hugging face*. [Online]. Available: https://huggingface.co/nguyenvulebinh/wav2vec2-base-vi#vlsp-2020-asr-dataset.

[15] P. T. Anh, *Vlsp2020 asr dataset*. [Online]. Available: https://huggingface.co/datasets/doof-ferb/vlsp2020_vinai_100h.

[16] L. T. T. Tran, H.-G. Kim, H. M. La, and S. Van Pham, "Automatic speech recognition of vietnamese for a new large-scale corpus," *Electronics*, vol. 13, no. 5, 2024, ISSN: 2079-9292. DOI: 10.3390/electronics13050977. [Online]. Available: https://www.mdpi.com/2079-9292/13/5/977.

[17] T. T. Nguyen, *Thangtrungnguyen/vietnamese-female-male-voice-classification-model, hugging face*. [Online]. Available: https://huggingface.co/thangtrungnguyen/vietnamese-female-male-voice-classification-model.

[18] T. T. Nguyen, *Thangtrungnguyen/vietnamese-regional-voice-classification-model, hugging face*. [Online]. Available: https://huggingface.co/thangtrungnguyen/vietnamese-regional-voice-classification-model.

[19] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 23*, 2023.

[20] OpenAI, HuggingFace, *Openai's whisper model card*, https://huggingface.co/openai/whisper-large-v2, 2024.

[21] T.-T. Le, L. T. Nguyen, and D. Q. Nguyen, "PhoWhisper: Automatic Speech Recognition for Vietnamese," in *Proceedings of the ICLR 2024 Tiny Papers track*, 2024.

[22] O. Cífka, H. Schreiber, L. Miner, and F.-R. Stöter, *Lyrics transcription for humans: A readability-aware benchmark*, 2024. arXiv: 2408.06370 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2408.06370.

[23] BatmanofZuhandArrgh, *Batmanofzuhandarrgh/vietlyrics: Vietlyrics: A large-scale dataset and models for vietnamese automatic lyrics transcription*. [Online]. Available: https://github.com/BatmanofZuhandArrgh/VietLyrics.

[24] Kelvin Soh, Bernard Cheng Zheng Zhuan, Nguyen Quoc Anh, *Vietnamese automatic lyrics transcription - whisper-large-v2*, https://huggingface.co/xyzDivergence/whisper-large-v2-vietnamese-lyrics-transcription, 2024.

[25] Kelvin Soh, Bernard Cheng Zheng Zhuan, Nguyen Quoc Anh, *Vietnamese automatic lyrics transcription - whisper-medium*, https://huggingface.co/xyzDivergence/whisper-medium-vietnamese-lyrics-transcription, 2024.

[26] Kelvin Soh, Bernard Cheng Zheng Zhuan, Nguyen Quoc Anh, *Vietnamese automatic lyrics transcription - whisper-large-v2*, https://huggingface.co/xyzDivergence/whisper-small-vietnamese-lyrics-transcription, 2024.