Reprogramming Pre-trained Model for Lyrics Transcription Task

Jiarui Xu
College of Design
Georgia Institute of Technology
Atlanta, USA
jxu605@gatech.edu

Alexander G Lerch
College of Design
Georgia Institute of Technology
Atlanta, USA
alexander.lerch@gatech.edu

***Abstract*—This article explores the challenges and solutions in automatic lyrics transcription, highlighting the potential of Reprogramming as a remedy for data scarcity in this field. Leveraging the Whisper model and three reprogramming techniques—Noise, CNN, and U-Net—the study reveals CNN's superiority in transcription accuracy. Key insights uncover the impact of pitch modulation, fundamental frequency differences in vocals, and background music's frequency distribution on transcription precision. The research paves the way for enhanced automatic lyrics transcription methodologies, emphasizing CNN's efficacy and the influence of nuanced audio elements.**

***Index Terms*—Automatic Lyrics Transcription, Reprogramming, Whisper Model, CNN, U-Net, Transcription, Frequency**

I. Introduction

Speech recognition has always been a challenging task, and according to previous research and experiments, recognizing and transcribing lyrics is even more challenging than speech recognition [1]. Even for humans, recognizing lyrics is sometimes not an easy task. Therefore, if computers could assist in transcribing lyrics, it would greatly aid in annotating, editing, aligning lyrics in music or video, among other aspects! However, the current level of lyrics transcription has not yet reached a sufficiently high level to be applicable in real-life scenarios.

Overall, the development of lyrics transcription shows a very promising trend, and as long as this trend continues, it will increasingly play a pivotal role in the music industry. There has been some progress in the transcription of pure vocals [2] [3] [4]. The release of some useful datasets has also to some extent stimulated research in this field [16] [13]. Meanwhile, for music that includes both vocals and instrumental elements, in other words, real music, there has also been some research progress and breakthroughs in lyrics transcription [5] [6].

One of the most significant challenges facing Automatic Lyrics Transcription (ALT) research is the lack of useful evaluation and training data, resulting in results being singular and non-reproducible [7]. While there have been efforts in this direction in the DAMP dataset [13], and attempts through adding annotations [10] [11] [12] to improve this data scarcity dilemma, the effect remains marginal.

Reprogramming, as an emerging technology, seems well-suited to address this data scarcity predicament. Model reprogramming takes a pre-trained model as the "source domain" without making any modifications to it. Instead, it preprocesses only parts of the input and output to expand the methods of transfer learning. Reprogramming was initially proposed by Elsayed in 2018 [15]. These scientists demonstrated a highly accurate way to solve the MNIST/CIFAR-10 image classification problem by introducing an input transformation function to reprogram a pretrained ImageNet model. Reprogramming requires modifying the model's input, which can be achieved through a trainable input reprogramming model. Consequently, reprogramming can adapt to many complex scenarios, enabling pretrained models to adapt to different types of input data and new tasks. Due to operating solely on the model's input and output, the entire reprogramming model has a significant advantage in reducing training complexity and data requirements. In recent years, reprogramming has found significant applications in various domains, including medical image classification [9], language processing [8], among others. Some results even suggest that reprogramming can outperform existing methods. Thus, reprogramming appears feasible and relatively advantageous compared to other methods in situations with limited high-quality training data.

II. Motivation

The motivation is the exploration of methods to improve automatic lyrics transcription, a challenging task within the field of audio processing. I aim to address the limitations and challenges faced in accurately transcribing lyrics from music or audio recordings.

The core drive is the proposal and exploration of a technique called Reprogramming to improve automatic lyrics transcription. Reprogramming leverages pre-trained models and modifies their input/output mechanisms, allowing these models to adapt to new tasks or different types of input data. This method could be particularly promising in mitigating the issue of data scarcity, which has been a significant bottleneck

in advancing accurate lyrics transcription.

The limitations of existing evaluation datasets showcases the developments in machine learning methodologies such as transfer learning and reprogramming, and emphasizes the need for innovative approaches to tackle the challenges faced in accurate lyrics transcription.

The proposed methods—Noise Reprogramming, CNN Reprogramming, and U-Net Reprogramming—demonstrate attempts to enhance the transcription process by modifying the input data for the Whisper model, a speech processing system trained extensively on vast amounts of internet audio data. These methods seek to optimize the input spectrograms to improve the accuracy of lyrics transcription.

Furthermore, I aim to evaluate the effectiveness of these reprogramming techniques using the Word Error Rate (WER) metric, commonly used to measure the accuracy of automatic speech recognition systems. The goal is to reduce WER by employing reprogramming techniques, thereby enhancing the accuracy of automatic lyrics transcription.

III. Related work

The pursuit of accurate lyrics transcription faces its own set of challenges. A scarcity of publicly available evaluation data has hindered progress in this arena [7]. Efforts like Dabike and Barker's annotations for DAMP subset and the Jamendo (lyrics) dataset [10] aimed to alleviate this issue. However, commonly used evaluation sets, such as Hansen and Mauch's, despite their limitations, persist as benchmarks [11] [12].

Concurrently, in the realm of machine learning, recent breakthroughs have unfolded new vistas in transfer learning methodologies, diverging from traditional approaches to unlock innovative paradigms. The focus has shifted to exploring groundbreaking techniques such as reprogramming. This evolution extends beyond the domain of lyrics transcription, transcending into diverse applications where understanding complex audio data becomes an art.

Earlier research emphasized audio speech transcription (AST) from audio tracks, paving the way for contemporary advancements, notably in model reprogramming. Researchers like Emir et al. [14] proposed Convolutional Neural Networks (CNNs) with datasets highlighting the potential of modern neural networks in instrument recognition.

Despite commendable efforts in data curation with datasets like IRMAS, MedleyDB, challenges persist in dataset completeness and diversity. This propelled alternative methodologies, particularly reprogramming, inspired by adversarial machine learning. Model reprogramming, exemplified by Elsayed et al. and Tsai et al., aims to adapt pre-trained models for novel tasks by manipulating input and output data [15]. Its success extends across various fields, including biomedical image classification [9] and natural language processing [18], showcasing its adaptability and potential in addressing data scarcity.

IV. Proposed Method

A. Pre-trained Model

The Whisper model is a speech processing system trained extensively to predict transcripts from vast amounts of internet audio data. Scaled to 680,000 hours of multilingual and multitask supervision, these models demonstrate robust generalization on standard benchmarks, often matching or surpassing prior fully supervised results. Notably, they perform remarkably well in a zero-shot transfer setting without requiring fine-tuning. The models' performance approaches human accuracy and robustness, releasing models and code to advance further research in robust speech processing [17].

It attributes the difference to the distinct capabilities measured during testing due to variations in training methodologies. Human performance reflects out-of-distribution generalization, while machine learning models' evaluation relies on in-distribution generalization. The Whisper models, trained on diverse audio distributions and evaluated in a zero-shot setting, have the potential to align better with human behavior compared to existing systems [17].

Whisper models exhibit distinct robustness properties compared to supervised LibriSpeech models, outperforming them significantly on diverse datasets. Even the smallest Whisper model competes well with supervised models on different datasets, matching human accuracy and robustness [17].

The specific Whisper model employed in this study is whisper-small in English. The reason for choosing Whisper to do the reprogramming task is that this model, unlike DeepSpeech model, is comparatively new and easy-handling model with detailed implementation.

B. Reprogramming

a) Baseline: Employed the Whisper model directly without altering any parameters for lyrics transcription.

b) Noise Reprogramming: Noise reprogramming involves adding a learnable noise component to the input spectrogram of music audio. This technique aims to translate target data into a representation compatible with Whisper model. By adding noise to the spectrogram rather than the time domain signal, it seeks to match the complexity of music audio. The operation introduces a universal noise component that remains independent of the input, altering the input spectrogram to generate a new spectrogram, which serves as the input for the whisper model.

c) CNN Reprogramming: In CNN reprogramming, a trainable CNN structure is employed as an input transformation for the input spectrogram. This method replaces the superposition of noise in noise reprogramming. The CNN comprises two 2D convolutional layers without max-pooling, allowing the input transformation directly compatible with the pre-trained model. The key distinction from noise reprogramming is that CNN reprogramming applies learnable transformations to the input itself, using CNN layers rather than adding independent noise components.

d) U-Net Reprogramming: U-Net reprogramming is an extension of CNN reprogramming, introducing a U-Net architecture by transforming audio spectrograms. U-Net, initially developed for biomedical image segmentation, consists of a contraction path capturing context and an expansion path reconstructing features back to input resolution. It leverages convolutional layers, skip connections, and upsampling blocks to represent multi-scale features, allowing high-level and detailed representations. This is the first proposal to use a U-Net structure for reprogramming, aiming to do lyrics transcription effectively by processing the spectrogram. The structure includes three convolutional layers in both paths, utilizing batch normalization, ReLU activation, max-pooling, and upsampling.

C. Evaluation

The evaluation method I used is Word Error Rate (WER). The Word Error Rate (WER) is a metric used to evaluate the accuracy of automatic speech recognition (ASR) systems. It measures the difference between the recognized words generated by an ASR system and the reference transcript (the ground truth).

WER calculates the number of substitutions, insertions, and deletions required to transform the recognized text into the reference text, normalized by the total number of words in the reference text. It's often expressed as a percentage.

The formula for WER is:

$$WER = \frac{S + D + I}{N}$$

Where:
- S represents the number of substitutions (incorrect words).
- D represents the number of deletions (missing words in the recognized text).
- I represents the number of insertions (extra words in the recognized text).
- N is the total number of words in the reference text.

The lower the Word Error Rate, the better the accuracy of the ASR system, as it indicates fewer errors in the recognized text compared to the reference.

V. Experiment and Result

A. Dataset

The experiment is based on DAMP-Sing! 300x30x2 [13]. It is a collection of 34,620 interpretations (i.e. performances) covering 302 different songs. All the songs are sang by female and male separately. To do the preprocessing of the dataset, I cut the data in the dataset into more than 80,000 fragments. Then I filtered the specific symbols like "&", "$", etc in the ground truth of the dataset manually. I also convert the sampling rate of the dataset to 16000 Hz to fit the Pre-trained Whisper Model.

B. Experiment Details

The loss function I used is CrossEntropyLoss() because Whisper used this function to do the Loss calculation. The Optimizer I used is Adam. Researchers used Momentum Optimizer and Adam Optimizer to do these task most of the time. After trying, Adam optimizer did a much better job than Momentum. The learning rate is 0.01. All of the three reprogramming: Noise, CNN, U-Net are processed on spectrogram of the audio not the audio itself because the input of Whisper is spectrogram, using spectrogram will have a direct effect of the output.

For CNN reprogramming, I tried four types of CNN with the ReLU activation:

- two 2D convolutional layers with a receptive field of 3 × 3, a stride of 1 × 1, and a padding size of 1 × 1.
- two 2D convolutional layers with a receptive field of 4 × 4, a stride of 1 × 1, and a padding size of 1 × 1.
- two 2D convolutional layers with a receptive field of 5 × 5, a stride of 1 × 1, and a padding size of 2 × 2.
- two 2D convolutional layers with a receptive field of 2 × 2, a stride of 1 × 1, and a padding size of 1 × 1.

For U-Net reprogramming, the structure of it is that it consist of 3 convolutional layers with contraction path and expansion path. For each convolutional layer, there are a batch normalization layer and ReLU activation. A 2 × 2 max-pooling is applied in the CNN layer of the contraction path, upsampling in the expansion path.

Utilized 10% data for testing, 10% data for validation, 80% data for training.

C. Result

As Shown in the Fig. 1 and Fig. 2. Though the WER still has some distance from state of arts, which is 15.38% [14], it is worth discussing because I used less data and parameters in training and I used a different approach.

VI. Discussion

From the result, I noticed that two interesting point. 1. U-Net does not perform as good as CNN, 2. Famale vocal's transcription's WER is significantly lower than Male vocal's

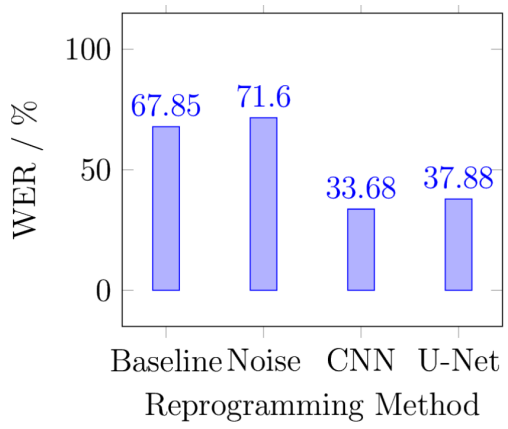

Fig. 1. Overall Result

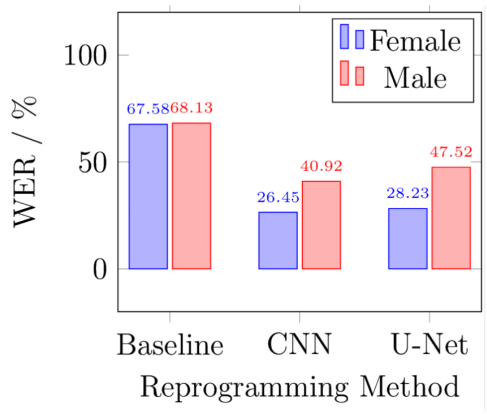

Fig. 2. Female & Male Reult

transcription's WER.

There are several reasons to explain why U-Net not performs as good as CNN:

- CNN Simplicity:
  Compared to Complex U-Net, CNN is more suitable. The model does not benefit significantly from the additional complexity provided by U-Net.

- Feature Relevance:
  U-Net is designed to capture both high-level and detailed features in images. If the lyrics transcription task doesn't require such a detailed representation, CNN may perform better by focusing on relevant features.

- Spatial vs. Sequential Data:
  U-Net is particularly effective in tasks involving spatial relationships in images. If lyrics transcription is more sequential and doesn't benefit significantly from spatial relationships, CNN might be more suitable.

- Data Availability:
  U-Net may require more training data than CNN.

For the next question: Famale vocal's transcription's WER is significantly lower than Male vocal's transcription's WER, I thought I need to do some follow up experiments and analysis to find the reason.

VII. Follow Up Experiments and Analysis

A. *Frequency Scaling*

One of the difference between female voice and male voice is their fundamental frequency are different. According to Opera Pitch Range, male voice fundamental frequency range from 123Hz to 440Hz, female voice fundamental frequency range from 247Hz to 880Hz. Therefore, if we scale the frequency of the input audio, how the WER will change?

Based on this question, I did follow up experiments for the relationship between WER and the frequency scaling. The result is:

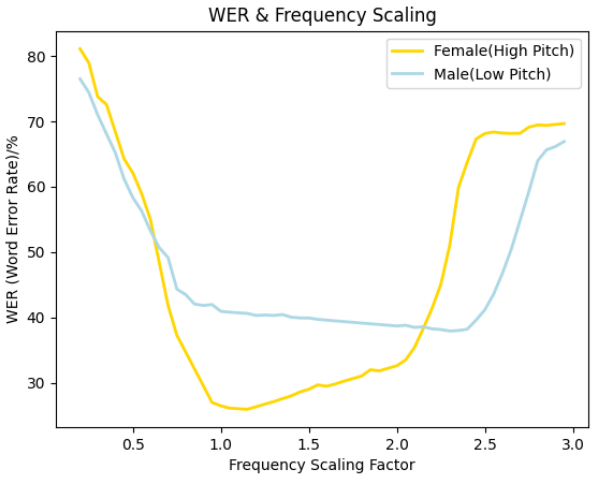

Fig. 3. Frequency Scaling of Input vs WER

I did the experiments on CNN model because this model performed the best among the other models. The horizontal axis is the frequency scaling factor, so 1.0 on the horizontal axis is the original frequency of an audio chip and 2.0 on the horizontal axis is raising every second's frequency to twice the original frequency in the audio clip.

It is interesting to notice that the WER is related to the frequency of the audio input. In addition, female has the best result around 1.2*original frequency; male has the best result around 2.4*original frequency.

Comparing the original audio and the reprogrammed audio's spectrum shown in Fig. 4 and Fig. 5. I found that the reprogramming audio highlighted some parts and diluted some parts. Also, it let the audio's frequency come up a little. Those things proved that frequency changing affect lyrics

transcription's result.

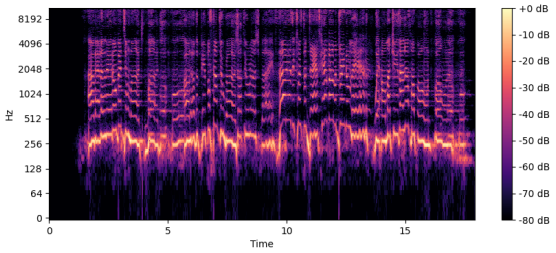

Fig. 4. Original Audio's spectrum

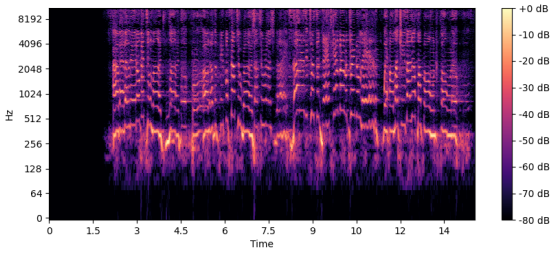

Fig. 5. Reprogrammed Audio's spectrum

Therefore, I can have these follow up conclusion:

- Appropriate frequency scaling helps the audio fit in Reprogramming.
- Female and Male's difference in WER partially caused by frequency distribution's difference.

B. Adding Background Music

Another follow up question for lyrics transcription task is about background music. Background music can result in worse lyrics transcription but the frequency distribution on the spectrum of each background music are different, so I do some following up experiments to see if the frequency distribution of background music matters in lyrics transcription.

According to [14], music accompaniment can be regarded as noise. Therefore, I used white noise with the peak intensity fluctuation equivalent to highest sound intensity of the audio to represent background music in my dataset.

According to Opera Pitch Range, male voice fundamental frequency range from 123Hz to 440Hz, female voice fundamental frequency range from 247Hz to 880Hz.

In the experiment, I first blocked 440Hz and the frequency above as white noise. Next step, I blocked 247Hz and the frequency below as white noice. Finally, I blocked the frequency from 247Hz to 440 Hz as white noise.

The result shown in Fig. 6.

Therefore, I got the conclusion:

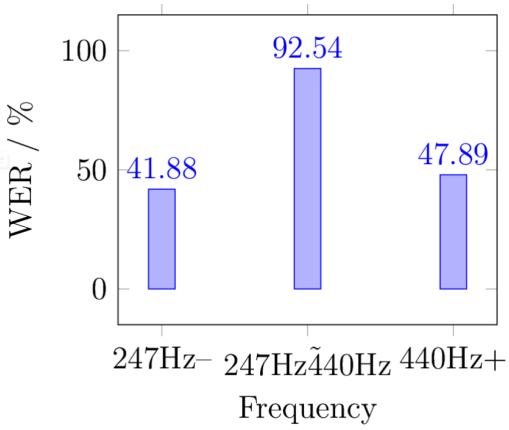

Fig. 6. WER When BGM Covered Different Frequency

- Background music's frequency distribution matters when we do Lyrics Transcription.
- The frequency range where vocal's fundamental frequency in, contains most information.

VIII. Novelty

The novelty of this paper:

- Utilize Reprogramming on Whisper model to do lyrics transcription and successfully improved the performance of Whisper model.
- Discover that CNN does better job in reprogramming on whisper model to do lyrics transcription than U-Net.
- Find out appropriate frequency scaling helps the audio fit in Reprogramming.
- Figure out female and Male's difference in WER partially caused by frequency distribution's difference.
- Find that background music's frequency distribution matters in lyrics transcription.
- Proved that the frequency range where vocal's fundamental frequency in, contains most information.

These discoveries offer clear guidance for future researchers aiming to preprocess audio inputs for lyrics transcription. They also pave the way for researchers interested in transcribing lyrics within a background music context, introducing a new avenue for exploration: the distribution of background music across the spectrum. Furthermore, I present data showcasing the percentage of information between 247Hz and 440Hz.

IX. Conclusion

In conclusion, for reprogramming on whisper model for lyrics transcription task, CNN did the best job. Appropriate frequency scaling helps the audio fit in Reprogramming. Female and Male's difference in WER partially caused by frequency distribution's difference. Background music's frequency distribution matters when we do lyrics transcription. The frequency range where vocal's fundamental frequency in, contains most information.

REFERENCES

[1] P. A. Fine and J. Ginsborg, "Making myself understood: Perceived factors affecting the intelligibility of sung text," Frontiers in Psychology, vol. 5, p. 809, 2014.

[2] G. R. Dabike and J. Barker, "Automatic lyrics transcription from karaoke vocal tracks: Resources and a baseline system," in Interspeech, 2019.

[3] E. Demirel, S. Ahlback, and S. Dixon, "Automatic ¨ lyrics transcription using dilated convolutional neural networks with self-attention," in International Joint Conference on Neural Networks (IJCNN), 2020.

[4] E. Demirel, S. Ahlback, and S. Dixon, "Computational ¨ pronunciation analysis in sung utterances," in European Conference on Signal Processing (EUSIPCO), 2021.

[5] ——, "Low resource audio-to-lyrics alignment from polyphonic music recordings," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[6] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics transcription in polyphonic music: Does background music help?" in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[7] A. M. Kruspe, "Training phoneme models for singing with "songified" speech data." in International Society for Music Information Retrieval Conference (ISMIR), 2015.

[8] Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., "Adversarial reprogramming of text classification neural networks." In: EMNLP-IJCNLP, 2019. with "songified" speech data." in International Society for Music Information Retrieval Conference (ISMIR), 2015.

[9] Tsai, Y.Y., Chen, P.Y., Ho, T.Y., "Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources." In: ICML. PMLR, 2020.

[10] D. Stoller, S. Durand, and S. Ewert, "End-toend lyrics alignment for polyphonic music using an audio-to-character recognition model," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[11] J. K. Hansen, "Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients," in Ninth Sound and Music Computing Conference (SMC), 2012.

[12] M. Mauch, H. Fujihara, and M. Goto, "Integrating additional chord information into HMM-based lyricsto-audio alignment," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 200–210, 2011.

[13] "Smule sing! 300x30x2 dataset," accessed April, 2021, https://ccrma.stanford.edu/damp/.

[14] Demirel E, Ahlbäck S, Dixon S. MSTRE-Net: Multistreaming acoustic modeling for automatic lyrics transcription[J]. arXiv preprint arXiv:2108.02625, 2021.

[15] Elsayed, G.F., Goodfellow, I., Sohl-Dickstein, J. "Adversarial reprogramming of neural networks." In: ICLR, 2019.

[16] 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," in International Society for Music Information Retrieval Conference (ISMIR), 2019.

[17] Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//International Conference on Machine Learning. PMLR, 2023: 28492-28518.

[18] Hambardzumyan, K., Khachatrian, H., May, J. "WARP: Word-level Adversarial ReProgramming." In: IJCNLP, 2021.