LYRIC TRANSCRIPTION IN NOISY ENVIRONMENT

Kwang Bin Lee
KAIST
klee166@kaist.ac.kr

Zsombor Banfi
KAIST
bzsoma@kaist.ac.kr

Hye Bin Ahn
KAIST
ahb9900@kaist.ac.kr

ABSTRACT

We advanced lyric transcription in noisy environments by fine-tuning a pre-trained Whisper-medium model and analyzing its layer activations. Using the noise-augmented JamendoLyrics dataset and noise samples from the MUSAN dataset with varying signal-to-noise ratios (SNRs), we evaluated performance based on Word Error Rate (WER) on a test subset and additional data. Results indicate that fine-tuning at SNRs above -2 dB yields stable performance, while SNRs of -3 dB cause instability, and SNRs below -3 dB lead to overfitting. The fine-tuned model achieved a WER of 18.05%, significantly outperforming the baseline model with a Speech Enhancement (SE) module, which had a WER of 63.45%. Additionally, our analysis of the Whisper-medium model's activations revealed a general decrease in Mean Squared Difference (MSD) in both the last convolutional and transformer layers, indicating improved alignment of the model's response to clean and noisy data. These findings demonstrate the effectiveness of fine-tuning with noise-augmented data and underscore the potential for further optimization in low SNR conditions.

1. INTRODUCTION

Automatic lyric transcription in noisy environments is a challenging task with significant applications in music information retrieval, assistive technologies, and digital archiving. Traditional speech recognition systems often struggle with background noise, leading to high error rates. Recent advancements in deep learning have opened new avenues for improving transcription accuracy under such adverse conditions.

The field of automatic speech recognition (ASR) has seen significant progress over the years, particularly with the advent of deep learning techniques. Despite these advancements, the robustness of ASR systems in noisy environments remains a critical issue. To deal with the noise robustness of asr model, research on noise-robust ASR often involves combining clean and noisy data to improve performance. One notable approach is multi-condition training (MCT) [2], which mixes clean and noisy speech data during training to enhance the model's robustness. This technique has shown significant improvements in reducing WER under various noise conditions.

Due to the effectiveness of ASR techniques in handling diverse audio conditions and improving transcription accuracy, we leverage the Whisper-medium model [3], a state-of-the-art end-to-end ASR model known for robust speech recognition. However, the Whisper model is not inherently robust to noise, necessitating fine-tuning for improved performance in noisy environments.

Our approach addresses these challenges through noise-augmented training. Inspired by noise-robust training methods and noise-augmented data generation techniques, we use the JamendoLyrics dataset combined with noise samples from the MUSAN dataset, incorporating varying signal-to-noise ratios (SNRs) to enhance robustness and generalization.

A significant innovation is the fine-tuning of the Whisper model with songs in noisy environments, such as cafes. Our methodology focuses on fine-tuning in real-world noisy conditions, demonstrating superior performance compared to the baseline larger model. We rigorously assess performance using Word Error Rate (WER) on both a designated subset and additional noisy data.

In summary, our contributions are listed as follows:

- We introduced noise augmentation using the MUSAN dataset to simulate various noisy environments, enhancing the robustness and generalization capabilities of the model for lyric transcription.
- We fine-tuned the Whisper-medium model on the noise-augmented JamendoLyrics dataset, using advanced techniques to improve computational efficiency and effectiveness in sequence-to-sequence tasks.
- We evaluated the impact of a Speech Enhancement module by comparing model performance with and without this preprocessing step, demonstrating the significant benefits of fine-tuning on noise-augmented data.
- We conducted an activation analysis of the Whisper model, comparing its behavior in clean versus noisy environments to provide insights into the model's response under different conditions, revealing improved alignment in the fine-tuned model.

This research demonstrates the potential of fine-tuned models in handling noisy environments and underscores

the importance of tailored training strategies. Our work paves the way for more reliable and accurate lyric transcription systems capable of operating effectively in real-world settings, offering substantial advancements in music information retrieval.

2. DATASETS AND METRICS

2.1 Datasets

We finetuned the Whisper model using the JamendoLyrics dataset, augmented with noise samples from the MUSAN dataset [4] to simulate various noisy environments. The JamendoLyrics dataset [1] provides a substantial collection of song-lyrics pairs essential for lyric transcription. This dataset contains 80 songs spanning different genres and languages, with lyrics time-aligned on a word-by-word level, including start and end times aligned with the music.

To enhance robustness against background noise, we incorporated noise samples from the MUSAN dataset, which includes diverse noise types such as music, speech, and environmental sounds. This portion of the MUSAN corpus contains 929 files of assorted noises, with a total duration of about 6 hours.

To create varied noisy conditions for training, we applied a range of signal-to-noise ratios (SNRs) from -5 to 5 dB. The original training and test splits from the JamendoLyrics dataset were maintained for consistency. After fine-tuning the model with this augmented dataset, we tested our model's performance on our own data: a less noisy song recorded with a phone in a cafe and poor-quality concert recordings from YouTube.

This approach allowed us to develop a resilient lyric transcription system capable of handling real-world noise conditions effectively, thereby enhancing the robustness and generalization capabilities of our model.

2.2 Metrics: Word Error Rate

For evaluating the performance of our lyric transcription model, we employed the Word Error Rate (WER), a widely used metric in speech recognition tasks. WER measures the difference between the transcribed text produced by the model and the reference text, considering insertions, deletions, and substitutions. It is calculated as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference. A lower WER indicates better transcription accuracy. By using WER, we were able to quantitatively assess the model's performance across different noise conditions and configurations, providing a clear measure of its effectiveness in handling noisy environments.

3. APPROACH

3.1 Noise Augmentation

To simulate noisy environments, we augmented the JamendoLyrics dataset with noise samples from the MUSAN dataset. Noise was added to the clean lyrics data with SNRs ranging from -5 dB to 5 dB, generating a diverse set of noisy conditions. This augmentation aimed to enhance the model's robustness and generalization capabilities.

3.2 Fine-tuning Detail

The fine-tuning process involved training the Whisper-medium model using the noise-augmented dataset. We hypothesized that leveraging this robust foundational model with our noisy data samples would yield satisfactory results. Unsuccessful experiments with publicly available Whisper models fine-tuned on noise-free samples, indicated that noise addition to the samples is crucial for achieving better task accuracy.

The training involved fine-tuning the Feature Extractor and Conditional Generator modules of the Whisper model while keeping the Tokenizer module static. Given that our dataset contained songs in multiple languages, it was important to set the tokenizer prefix token accurately for each label during training. This adjustment proved to have a significant impact on performance. The model was optimized by minimizing the cross-entropy objective function.

3.3 Speech Enhancement

Speech enhancement (SE) is an algorithm designed to improve the quality and intelligibility of speech signals by reducing background noise and other distortions. The SE module, specifically the pre-trained WaveUNet [5], was used to preprocess the noisy audio data before feeding it into the baseline Whisper model and compared it with the noise-robust fine-tuned model. We investigated how the SE module affects the robustness of the transcription system in various noise environments, providing insights into its practical applications in 5.2.

3.4 Whisper: Clean vs. Noisy Activation Analysis

We analyzed the behavior of the Whisper Model by comparing its response to music in a clean environment versus music in a noisy environment. This comparison was conducted by computing the Mean Squared Difference (MSD) in activations of different layers within the Whisper Model between the clean and noisy inputs.

For the convolutional layers, MSD_{conv} was calculated using the following formula:

$$\text{MSD}_{\text{conv}} = \frac{1}{N} \sum_{i=1}^{N} \left(A_{\text{clean},i} - A_{\text{noisy},i}\right)^2$$

where N is the total number of activations in the convolutional layer, $A_{\text{clean},i}$ represents the activation for the i-th neuron in the clean input, and $A_{\text{noisy},i}$ represents the activation for the i-th neuron in the noisy input.

For the transformer layers, we first computed the average activation across all attention heads for each transformer layer. $\text{MSD}_{\text{trans}}$ was then calculated using the following formula:

$$\text{MSD}_{\text{trans}} = \frac{1}{N} \sum_{i=1}^{N} \left(\bar{A}_{\text{clean},i} - \bar{A}_{\text{noisy},i} \right)^2$$

where N is the total number of activations in the transformer layer, $\bar{A}_{\text{clean},i}$ represents the average activation for the i-th neuron across all attention heads in the clean input, and $\bar{A}_{\text{noisy},i}$ represents the average activation for the i-th neuron across all attention heads in the noisy input.

These calculations allowed us to quantify the differences in the model's internal representations when processing clean versus noisy audio inputs, providing insights into the model's robustness to noise.

4. EXPERIMENTS

4.1 Model Configuration

We utilized HuggingFace's sequence2sequence trainer, which offers an easy-to-use interface for fine-tuning transformer models. The training and evaluation batch sizes were set to 32 and 5 respectively. We observed that using a training batch size smaller than 32 often resulted in training instability and overfitting.

The learning rate (LR) was configured at 1×10^{-5}, with a warmup phase of 5 steps to stabilize the training process. The LR scheduling was handled automatically by the trainer. The choice of LR was based on experiences and experiments that fine-tuning with high LR can lead to significant overfitting and performance drop. The training was conducted over 50 epochs, with a maximum generation length of 225 tokens. The optimization algorithm used was the AdamW implementation from the Hugging Face library. The model was evaluated at the end of each epoch, with Word Error Rate (WER) as the evaluation metric. Based on this metric, early stopping was incorporated into the training process with a patience value of 5. These hyperparameters were carefully selected to balance computational efficiency and model performance, ensuring optimal training conditions for our noise-augmented lyric transcription model.

5. RESULTS

5.1 Fine tuning results

Our fine-tuning approach yielded successful results for the proposed task. For Signal-to-Noise Ratio (SNR) values greater than -3, the model demonstrated a significant performance improvement over the baseline model. At an SNR of -3, the model's generalization ability was highly dependent on the training split and initialization. However, for SNR values less than -3, the model failed to generalize well on the validation set and tended to overfit the training data. These outcomes are considered highly sufficient, given that an SNR value of -3 indicates that the noise is twice as strong as the signal itself. Our results are depicted in Figure 1.

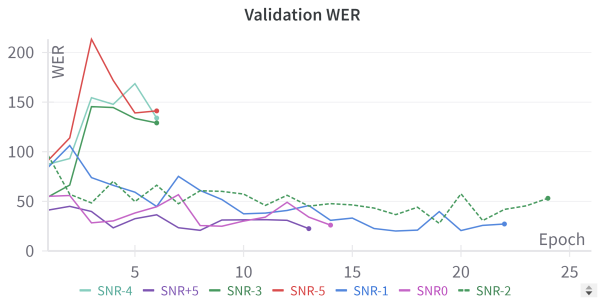
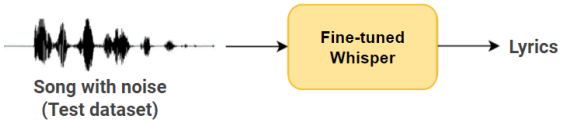
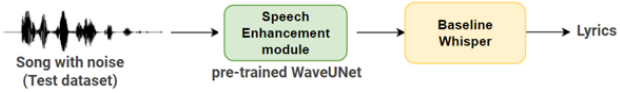


Figure 1: Validation WER with different SNRs. SNR $<$ -2: Stable fine-tuning. SNR = -3: Unstable, dependent on initialization. SNR $<$ -3: Overfitting, poor generalization



(a) Noise-augmented test data is inputted to fine-tuned Whisper



(b) Noise-augmented test data is preprocessed by speech enhancement Module, and then inputted to baseline Whisper

Figure 2: Pipeline of two types of lyrics transcription system in noisy environment

After gathering multiple models, we selected the one trained at the lowest SNR value with an acceptable Word Error Rate (WER) (SNR = -2) for testing with our recordings. We chose two samples: a less noisy song recorded with a phone in a café and a poor-quality concert recording from YouTube.

The predictions for these samples are presented in Table 2 and Table 3 in A. In the first example, the fine-tuned model closely transcribed the lyrics compared to the baseline model. Although in the second example, the transcription by the fine-tuned model still deviates significantly from the original lyrics, it represents a notable improvement over the baseline model, which failed to produce coherent English words.

	Our fine-tuned Whisper model	Baseline Whisper with SE module
WER (%)	18.05	63.45

Table 1: Performance comparison of our fine-tuned model with baseline whisper with speech enhancement module on test dataset

5.2 Comparison with Speech Enhancement Module

In this section, we present the results of our lyric transcription system in a noisy environment, comparing the performance of our fine-tuned Whisper model against a baseline Whisper model equipped with a speech enhancement module.

To evaluate our fine-tuned model, we fed noise-augmented test data directly into it. For the baseline model, we pre-processed the noise-augmented test data using a speech enhancement module before transcribing the lyrics. The overall pipeline is illustrated in Figure 2.

As shown in Table 1, our fine-tuned model achieved a Word Error Rate (WER) of 18.05, while the baseline model achieved a WER of 63.45. These results demonstrate that the fine-tuned Whisper model significantly outperforms the baseline model in noisy environments, even without pre-processing. Additionally, the transcription results in Table 4 in A indicate that the baseline model's transcriptions were significantly different from the ground truth lyrics, whereas the fine-tuned model produced more accurate results.

The baseline model's poor performance is due to two main factors. First, the speech enhancement module was insufficient in reducing noise in highly noisy environments, leaving residual noise. Second, the enhancement process inadvertently removed critical linguistic information necessary for accurate lyric transcription due to over-suppression. This compromised the integrity of the lyrics, leading to high WER.

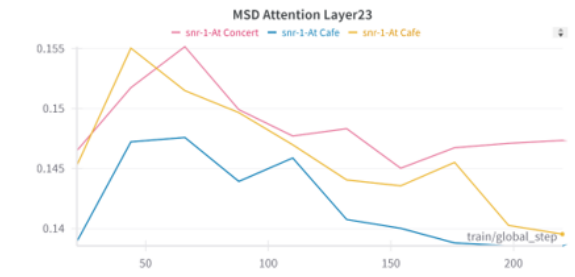
In contrast, the fine-tuning process effectively enhanced the Whisper model's capability to accurately recognize lyrics in the presence of noise. This underscores the importance of domain-specific training for improving transcription accuracy. The low WER and the qualitative similarity of the transcriptions to the ground truth confirm that the fine-tuned model is robust against noise interference.

5.3 Comparison with Activations in Different Layers

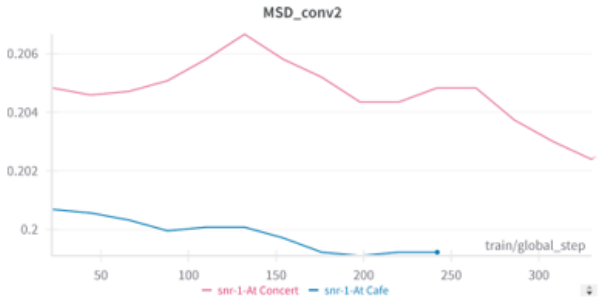
We analyzed the behavior of the Whisper Model by comparing its response to music in a clean environment versus music in a noisy environment. This comparison was conducted by computing the Mean Squared Difference (MSD) as explained above. While the majority of layers did not exhibit a significant difference, there was a notable decrease in MSD_{conv} values in the second-to-last convolutional layer and the $\text{MSD}_{\text{trans}}$ values in the last transformer layer, as shown in the Figure 3. This is anticipated because the final layers have the greatest impact on the model's output. These results suggest that the fine-tuned model responds similarly to noisy and clean data, enabling it to accurately transcribe lyrics from songs even in noisy environments.

6. LIMITATION AND FUTURE WORK

While our approach produced promising results, it presents several limitations and opportunities for future work.



(a) $\text{MSD}_{\text{trans}}$ in the last transformer layer



(b) MSD_{conv} in the last convolutional layer

Figure 3: Comparison of Mean Squared Difference (MSD) in activations between clean and noisy music inputs in the Whisper Model

To begin with, due to computational constraints, we were restricted to fine-tuning the medium Whisper model. Future research could explore utilizing the larger Whisper Large-v3 model, which might yield superior results due to its higher capacity.

Additionally, incorporating dynamically changing SNR during training could improve robustness. Training with a static low SNR value may present an overly challenging problem, but adjusting SNR based on evaluation metrics could enhance the model's adaptability to varying noise levels.

Furthermore, introducing a wider variety of noise types, such as concert noise, could further improve the model's robustness. This would enable the model to perform more effectively across a broader spectrum of real-world audio environments.

Finally, integrating a multi-stage pipeline where vocals are first separated using source separation modules before applying noise reduction and transcription could potentially enhance performance. This layered approach might better handle the intricacies of noisy music recordings.

7. CONCLUSION

This research demonstrated that fine-tuning the Whisper-medium model with a noise-augmented dataset significantly improved performance for the Automatic Lyrics Transcription (ALT) task. The fine-tuned model showed notable enhancements over the baseline for SNR values above -3 dB, though it struggled with generalization at lower SNR levels. These results highlight the effectiveness of noise augmentation in training lyric transcription

models and suggest potential for further advancements by utilizing larger models, incorporating dynamically changing SNRs, and introducing a wider variety of noise types. Our findings provide a solid foundation for future work in developing more robust and accurate transcription systems, paving the way for reliable lyric transcription in real-world noisy environments.

8. AUTHOR CONTRIBUTIONS

Kwang Bin Lee analyzed the model's activations and led the overall report writing. Zsombor Banfi conducted the model fine-tuning process, performed performance analysis based on fine-tuning and SNR, and contributed to report writing. Hye Bin Ahn handled dataset preprocessing, analyzed the comparison between the baseline model with the SE module and the fine-tuned model, and contributed to report writing.

9. REFERENCES

[1] Emir Demirel, Daniel Stoller, and Simon Durand. Jamendolyrics multi-lang – an evaluation dataset for multilanguage lyrics research. `https://github.com/f90/jamendolyrics/`, 2020.

[2] Yuchen Hu, Nana Hou, Chen Chen, and Eng Siong Chng. Dual-path style learning for end-to-end noise-robust speech recognition. *arXiv preprint arXiv:2203.14838*, 2022.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, PMLR, pages 28492–28518, 2023.

[4] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus, 2015. arXiv:1510.08484v1.

[5] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

A. TRANSCRIPTION EXAMPLES

	Transcription
Fine-tuned Whisper	i didn't think you would understand me how could you ever leave me alone i don't want to do the same i don't want to hide but i don't want to be the same
Baseline Whisper	Oh, I didn't mean it. I'm going to make a I'm sorry. I don't want to fight. I don't want to fight.
Groud truth	I didn't think you'd understand me How could you ever even try? I don't wanna tiptoe, but I don't wanna hide But I don't wanna feed this monstrous fire

Table 2: Transcriptions of a less noisy sample by the fine-tuned and baseline Whisper models

	Transcription
Fine-tuned Whisper	i've got a secret i've got a sight i've got a feeling that you'll find out someday i understand how i hold it inside me i don't know how but i believe i'll do it out of love
Baseline Whisper	Yn y bwysigrwydd, yna'r peth dwi'n ei ddweud Mae'n ddim unrhyw fath i'w gael i'w gael Mae gen i rhywbeth a'i deall Rwy'n ei gynnal yn syth Yn y bwysigrwydd Mae'n
Groud truth	Well I've got a secret, I cannot say Blame modern movement to give it away You've got somethin', that I understand Hold it in tightly, call on command Leap of faith, do you doubt? Cut you in, I just cut you out

Table 3: Transcriptions of a very noisy sample by the fine-tuned and baseline Whisper models

	Transcription
Fine-tuned Whisper	every day and every week doch an manchen tagen ist es einfach schwer da wünschcht ich mir the new guy is on schedule man behind bars and thats minus the federal stone giant how you let me down but somehow deep down i still do love you now your out for night
Baseline Whisper with SE module	thing is night day en menschenipp das nichtuerschen man mir when new guy is on schedule behind cameras in this when the price stop giant to let it down you it down in just can not you but i mine of the
Groud truth	every day and every week doch an manchen tagen ist es einfach schwer da wünscht ich mir the new guy is on schedule man behind bars and thats minus the federal stone giant how you let me down but somehow deep down i still do love you now your out all night

Table 4: Transcriptions by the fine-tuned Whisper model vs. baseline model with SE on selected test utterances