# Low Resource Training for Automatic Lyric Transcription using WAV2VEC2

**David Scaperoth**
scaperoth@berkeley.edu

## Abstract

Lyrics and captions for videos with singing (and singing with musical accompaniment) have been in production in industry for many years. Today, polyphonic Automatic Lyric Transcription (ALT) and related tasks are still considered challenging. This paper investigates automatic lyric transcription (ALT) for a specific use case where fine-tuning a pre-trained model with a small dataset (or low resource) may be useful (2 - 10 hours of data). Results are published for a fine-tuned variant of WAV2VEC2 (Baevski et al., 2020) on a polyphonic dataset. These results are compared with previous work done by (Ou et al., 2022) on a different unaccompanied singing dataset. The paper's finding show that low resource training on WAV2VEC2 benefits heavily from the multi-task approach mentioned in (Ou et al., 2022). This paper also shows new results for a low resource fine-tuning of WAV2VEC2 trained on polyphonic music that was not previously reported.

## 1 Introduction

This paper investigates the use of supervised machine learning for ALT and motivated by a specific use-case around live band karaoke. In live band karaoke a band plays known songs with various modifications depending on style, number of band members, and skill set of the team. In one use case, a band plays to a click track to ensure that a prompt is able to stay in sync with the songs, etc. However, in some cases, it would be preferable to have a way to ensure that either prompt software is dynamically 'keeping in sync' with the band. ALT software that can not only keep up real time, but also handle complex environments where the tempo changes, the instruments are off key potentially or that the band chooses to take the song in a direction that shortens or lengthens segments of the song.

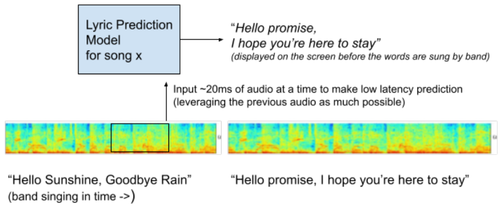
Lyric Prediction Model for song x
"Hello promise,
I hope you're here to stay"
(displayed on the screen before the words are sung by band)
Input ~20ms of audio at a time to make low latency prediction
(leveraging the previous audio as much possible)
"Hello Sunshine, Goodbye Rain"
(band singing in time ->)
"Hello promise, I hope you're here to stay"

Figure 1: Automatic Lyric Transcription Concept for Live Karaoke

Aside from venues that support the live band karaoke concept, there are also many evangelical church services which operate in a similar way with a volunteer 'keeping up' with the band on stage to display words for the audience to sing to. Because generalizing this task is challenging, a low resource training on several hours of 'similar' songs may improve models attempting to automate this type of work.

(Ou et al., 2022) mentions that when using his WAV2VEC2 architecture, the model may not need a great deal of data to achieve close to state of the art (SOTA) performance. Figure 2 shows Ou's plot and the dramatic drop in WER around 2hrs of training on the Dsing30 dataset (Roa Dabike and Barker, 2019). This range is central to the experiment conducted in this paper.

The paper contains the following contribution:

- Report results for the first time on an experiment where WAV2VEC2 is fine-tuned on a sub-set of the DALI dataset (∼10 hours) and compare the results with similar previous work.

Note that this is not intending to create a SOTA system, but simply improve from an existing pre-trained solution and attempt to 'get close' to SOTA while giving options to refine the model over time on small sets of data.

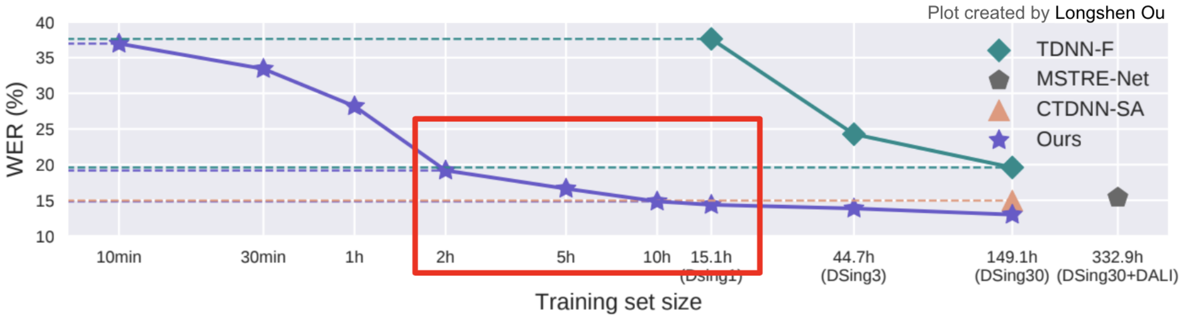

Figure 2: Low Resource Training for ALT. Red box is the target range of training hours for this experimentation. Plot created by Longshen Ou et. al.

## 2 Background

For many years, ASR and transcription algorithms heavily relied on Hidden Markov Model (HMM) based algorithms. Even today for low resource applications HMMs and hybrid HMM / deep learning systems are still in use.

Deep learning has made a heavy influence over polyphonic dataset tasks. For lyric alignment, the leading performers are largely transformer based models.

WAV2VEC2 is trained using novel self-supervised techniques that have demonstrated state of the art performance for many tasks surround speech and ASR. The model is pre-trained on 53,000 hours of unlabeled data where masking of audio representations is done much like BERT. The original work has been followed up with many fine tuned variants. One notable variant from the original authors involved fine tuning on WAV2VEC2 on 960hrs of Librispeech (pre-trained on the LibriVOX dataset (Baevski et al., 2020)).

WAV2VEC2 is especially attractive to support to study ASR applications because of the numerous contributions and tutorials that build upon its current capabilities and because the model is a 'barebones' model in which it can be evaluated and extended as an important open source contribution.

## 3 Methodology

(Ou et al., 2022) discusses the ability to achieve less than 20% WER's with only 15hrs of training data using a multi-task training regime. This paper attempts to explore the benefits of low resource fine tuning of WAV2VEC2 for ALT for a dataset (DALI) not reported by Ou and attempt to reproduce the DSing results previously found.

Additionally, we evaluate already fine-tuned models which have been designed to perform exceptionally well on speech data to understand our baseline for high performing pre-trained models.

### 3.1 Datasets and Pre-processing

| Split | Data | # Utt. | Total Dur. |
|---|---|---|---|
| Test | DSing$^{test}$ | 480 | 46 min |
| Test | DALI$^{test*}$ | 1,000 | 46 min |
| Dev | DSing$^{dev}$ | 482 | 41 min |
| Train | DSing1 | 8605 | 12.8 h |
| Train | DALI$^{train*}$ | 9186 | 8.0 h |

Table 1: Datasets used for experimentation, *refers to atypical split variants for that dataset.

This paper uses mainstream datasets for lyric transcription experiments (DALI and DAMP). For testing we evaluate against a curated versions of both DAMP Sing! 300x30x2 (Smu) called DSing and DALI v1 (Meseguer-Brocal et al., 2018). For DALI, the user must download ∼200GB of audio data and then using work from (Demirel et al., 2021b), the data is then reformatted to support baseline experiments using DALI$^{test}$. Then, the DALI data is then processed to change the data into utterances to avoid Out of Memory (OOM) issues when training with GPUs in a manner described by Ou. The authors from (Ou et al., 2022) did publish their source code for their experiment, but the dependencies were challenging resolve because of some major changes to SpeechBrain (the framework used by the researchers) since the experiment was conducted. DALI$^{test}$ was the only split that was created for this paper.

DSing consists of three training sets (DSing1, DSing3, DSing30), each with progressively more

| **Method** | **DSing$^{test}$** | **DALI$^{test}$** | **DALI$^{test*}$** |
|---|---|---|---|
| (Dabike and Barker, 2019) | 19.60 | 67.12 | - |
| (Demirel et al., 2020) | 14.96 | 76.72 | - |
| (Demirel et al., 2021b) | 15.38 | 42.11 | - |
| DSing30+DALI (Ou) | **12.99** | **30.85** | - |
| DALI$^{train*}$ (Scaperoth) | 39.4 | - | 66.6 |
| DSing1 (Scaperoth) | 41.4 | - | 83.0 |
| Whisper-Small-En (OpenAI) | 16.8 | - | 68.4 |
| WAV2VEC2 LARGE (Meta) | 49.1 | - | 91.0 |

Table 2: WER Performance across various ALT systems experimentation. Bold is SOTA.

data and supported languages. The DAMP dataset 300x30x2 is first downloaded, then leveraging a specific github repo published by (Roa Dabike and Barker, 2019), data splits are created. Once all the audio is labeled and correctly preprocessed, the author created Huggingface Dataset objects to support loading the data into Pytorch Dataloaders or Huggingface Trainer objects.

For the DSing dataset, we perform fine-tuning with DSing1. For DALI v1, we pre-process and divide up what is in literature referred to as DALI$^{test}$. For DALI test split, we sample 1000 utterances out of the 10186, leaving 9186 for fine tuning (around 8 hrs). Throughout the paper, we will refer to these splits as DALI$^{test*}$ and DALI$^{train*}$ to indicate that they are not the same as well known versions by similar names in prior research. This author was not able to obtain detailed information about the published data set variants for DALI and thus created one based on the size split similar to DSing$^{test}$. A summary of the dataset details is in Table 1.

For pre-processing, the paper largely follows (Ou et al., 2022) to support resampling to 16kHz, single channel audio, and for DALI we break apart the songs into utterances greater than 1 second and less than 28 seconds. When pre-processing the polyphonic data, these experiments do not perform source separation or train with data augmentation.

### 3.2 Experimental Setup

For fine tuning, the WAV2VEC2 model selected has already been pre-trained on 53k hours of speech data and fine-tuned on 960hrs of speech data. To fine tune the model further, CTC loss is used over 10 epochs with a batch size of 4 on a single A100. The CTC loss is evaluated for WER using beam search with a beam size of 512. The vocabulary used is 32. During training WAV2VEC2 feature encoder layers were frozen. The learning rate used was $1x10^-5$ with a "Reduce Learning Rate On Plateau" scheduler with 0.9 annealing factor, and a patience of 3. This was done in lieu of the NewBob technique used in SpeechBrain previously. Two fine tuning exercises were performed with the WAV2VEC2 model: (1) one model trained with DSing1 and (2) one model trained with DALI$^{train*}$. All fine tuning (even with DALI$^{train*}$) was evaluated using DSing$^{dev}$.

## 4 Results and Discussion

### 4.1 Baseline Results

To baseline the results for WER on DSing and DALI, four pre-trained models were evaluated that have no previous training on musical data: two variants of WAV2VEC2, two variants of Whisper by OpenAI (Radford et al., 2022). The first WAV2VEC2 model was the base model with no prior fine tuning, which performed so poorly it was not reported. Similarly, Whisper Tiny performed very poorly on DALI in particular and not reported on. The remaining two models and their results can be seen in Table 3.

Whisper Small performs surprisingly well against unaccompanied singing (i.e. the DSing dataset) and achieves near SOTA for DSing. This result has not been previously reported.

| **Model** | **DSing$^{test}$** | **DALI$^{test*}$** |
|---|---|---|
| Whisper-Small-En | 16.8 | 68.4 |
| WAV2VEC2 LARGE | 49.1 | 91.0 |

Table 3: WER performance compared to models not fine-tuned on singing data.

Compared to models that have not been trained on the distribution, ~10 hours of musical data can improve WER by 7 to 10 percentage points.

### 4.2 Fine-tuning on WAV2VEC2

We report results from fine-tuning the same WAV2VEC2 model on DSing with ∼10hrs of data and on DALI with 8hrs of data. The results in Table 4 including DALI have not been reported previously for WAV2VEC2, though they are still well below the state of the art. Some performance degradation is expected. (Ou et al., 2022) uses two techniques not implemented in our experiment: (1) multi-task learning with both CTC loss and Sequence to Sequence word generation (i.e. Language Modeling) (2) data augmentation for training. These results indicate performance is significantly improved through these missing techniques.

Table 4 shows results from experiments with different model variants of WAV2VEC2[1] fine-tuned on 15 hrs of DSing .

| **Method** | **DSing**$^{test}$ | **DALI**$^{test*}$ |
|---|---|---|
| Dsing30 (Ou) | 14.96 | - |
| DALI$^{train*}$ (Scaperoth) | 39.4 | 66.6 |
| DSing1 (Scaperoth) | 41.4 | 83.0 |

Table 4: WER performance of WAV2VEC2 variants fine-tuned on ∼10hrs of musical data.

## 5 Conclusions

This paper discusses experimental results related to fine-tuning WAV2VEC2 for the ALT task. Utilizing even relatively small amounts of singing data can make major improvements in WER. However, training using a simple CTC loss strategy was not nearly as effective as the the multi-task strategy taken by (Ou et al., 2022). As such, we do not meet the objective of recreating the results illustrated in Figure 2 with either unaccompanied and polyphonic datasets. However, we do improve over the WAV2VEC2 pre-trained model baseline. We published new results from Whisper that show that without further fine tuning the model can achieve near state of the art on unaccompanied music.

Future work should be considered to incorporate baselines for lyric alignment and measurements of real time factors for live-karaoke environments.

## References

Smule Sing! 300x30x2 Dataset. `https://ccrma.stanford.edu/damp/`. Accessed March 5, 2024.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Gerardo Roa Dabike and Jon Barker. 2019. Automatic lyric transcription from karaoke vocal tracks: Resources and a baseline system. In *Interspeech*, pages 579–583.

Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2020. Automatic lyrics transcription using dilated convolutional neural networks with self-attention. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2021a. Low resource audio-to-lyrics alignment from polyphonic music recordings. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590. IEEE.

Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2021b. Mstre-net: Multistreaming acoustic modeling for automatic lyrics transcription. *arXiv preprint arXiv:2108.02625*.

C. Gupta, E. Yılmaz, and H. Li. 2020. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? pages 496–500.

G. Meseguer-Brocal, A. Cohen-Hadria, and P. Geoffroy. 2018. Dali: A large dataset of synchronized audio lyrics and notes automatically created using teacher-student machine learning paradigm.

Longshen Ou, Xiangming Gu, and Ye Wang. 2022. Transfer learning of wav2vec 2.0 for automatic lyric transcription.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Gerardo Roa Dabike and Jon Barker. 2019. 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)*.

[1] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self