Automatic Lyric Transcription

Victor Shan

November 25, 2023

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1 Introduction

- What is Automatic Lyric Transcription?
- Why is it important?
- What are the challenges? (Why can't we use ASR?)
- What are the applications?
- What is the expected output?

2 Datasets

- What are the requirements for a dataset?
- Why are there so few datasets?
- Challenges of creating new datasets?
- What are the existing datasets?

2.1 Dataset Augmentation

2.1.1 SpecAugment

2.1.2 Transforming Existing Datasets

2.1.3 Student-Teacher Approach

2.2 Existing Datasets

2.2.1 TIMIT

2.2.2 LibriSpeech

2.2.3 JamendoLyrics

Remember to record the performance of the state of the art models on this dataset

2.2.4 Children's Songs Dataset

2.2.5 MUSDB18

2.2.6 NUS Dataset

2.2.7 Other Datasets

- DALI
- DAMP! (Stanford - Smule dataset)
-

3 Related Works

3.1 Early HMM-Based Acoustic Models

3.2 Music Informed Models

3.3 wav2vec 2.0 and Transfer Learning

3.4 Whisper Word-Level Alignment

4 Method

5 Results

5.1 WER

5.2 PER

6 Discussion

7 Future Work

7.1 Adversarial Training

8 Conclusion

Scholarly Articles

1. [Ou et al., 2022]
2. [Baevski et al., 2020]
3. [Panayotov et al., 2015]
4. [Radford et al., 2022]
5. [Sharma et al., 2019]
6. [Nishikimi et al., 2021]
7. [Yong et al., 2023]
8. [Hansen, 2012]
9. [Shenoy et al., 2005]
10. [Nishikimi et al., 2020]
11. [Bain et al., 2023]
12. [Choi et al., 2020]
13. [Meseguer-Brocal et al., 2018]
14. [Duan et al., 2013]
15. [Stoller et al., 2019]
16. [Stöter et al., 2018]
17. [Rouard et al., 2023]
18. [CMU, 2023]
19. [Park et al., 2019]
20. [Wang et al., 2023]

Other Sources

1. [Raissi, 2021]
2. [Hannun, 2017]

References

[Baevski et al., 2020] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations.

[Bain et al., 2023] Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

[Choi et al., 2020] Choi, S., Kim, W., Park, S., Yong, S., and Nam, J. (2020). Children's song dataset for singing voice research. In *KAIST*.

[CMU, 2023] CMU (2023). Cmu pronouncing dictionary.

[Duan et al., 2013] Duan, Z., Fang, H., Li, B., Sim, K. C., and Wang, Y. (2013). The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9.

[Hannun, 2017] Hannun, A. (2017). Sequence modeling with ctc. *Distill*. https://distill.pub/2017/ctc.

[Hansen, 2012] Hansen, J. K. (2012). Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients.

[Meseguer-Brocal et al., 2018] Meseguer-Brocal, G., Cohen-Hadria, A., and Peeters, G. (2018). Dali: a large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. In ISMIR, editor, *19th International Society for Music Information Retrieval Conference*.

[Nishikimi et al., 2020] Nishikimi, R., Nakamura, E., Goto, M., Itoyama, K., and Yoshii, K. (2020). Bayesian singing transcription based on a hierarchical generative model of keys, musical notes, and f0 trajectories. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1678–1691.

[Nishikimi et al., 2021] Nishikimi, R., Nakamura, E., Goto, M., and Yoshii, K. (2021). Audio-to-score singing transcription based on a crnn-hsmm hybrid model. *APSIPA Transactions on Signal and Information Processing*, 10(1):–.

[Ou et al., 2022] Ou, L., Gu, X., and Wang, Y. (2022). Transfer learning of wav2vec 2.0 for automatic lyric transcription.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

[Park et al., 2019] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

[Radford et al., 2022] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

[Raissi, 2021] Raissi, M. (2021). Mel-spectrogram and mfccs.

[Rouard et al., 2023] Rouard, S., Massa, F., and Défossez, A. (2023). Hybrid transformers for music source separation. In *ICASSP 23*.

[Sharma et al., 2019] Sharma, B., Gupta, C., Li, H., and Wang, Y. (2019). Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 396–400.

[Shenoy et al., 2005] Shenoy, A., Wu, Y., and Wang, Y. (2005). Singing voice detection for karaoke application. In *Visual Communications and Image Processing*.

[Stoller et al., 2019] Stoller, D., Durand, S., and Ewert, S. (2019). End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model.

[Stöter et al., 2018] Stöter, F.-R., Liutkus, A., and Ito, N. (2018). The 2018 signal separation evaluation campaign.

[Wang et al., 2023] Wang, Y., Wei, W., and Wang, Y. (2023). Phonation mode detection in singing: A singer adapted model. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[Yong et al., 2023] Yong, S., Su, L., and Nam, J. (2023). A phoneme-informed neural network model for note-level singing transcription. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.