TECHNICAL REPORT: LYRICS TRANSCRIPTION FOR MIREX 2021

Zhen Yang[1], *Qichen Han*[1], *Xiang Li*[1], *Dong Liu*[1], *Peng Li*[2]

[1]NetEase (Hangzhou) Network Co., Ltd., China
[2]NetEase Cloud Music, China

{yangzhen1, hanqichen, hzlixiang, hzliudong, hzlipeng}@corp.netease.com

Abstract

This technical report describes our submitted systems for the lyrics transcription challenge of MIREX 2021. The task of lyrics transcription aims to identify the words from sung utterances, in the same way as in automatic speech recognition (ASR). We built the recognition system using traditional hybrid DNN-HMM method with Kaldi toolkit. Specially, our system takes as input original audios without the need for singing vocal extraction. For the language model (LM), we collected a large number of textual lyrics and trained a 5-gram in-domain LM. Experimental results have shown that our systems achieved large improvements over the prior work in the lyrics transcription task.

1. Introduction

Automatic lyrics transcription (ALT) aims to identify the words from sung utterances [1], in the same way as in automatic speech recognition (ASR). ALT is quite useful in a wide range of application scenarios, such as lyrics generation, security review, music copyright protection and etc. In the past few years, the performance of ASR systems has been dramatically improved. However, due to the complex and variable characteristic of singing audios, ALT still remains an open question. We found that the difficulties in ALT mainly came from two aspects. Firstly, singing voices are often fully mixed with music accompaniments, which are quite complicated with various rhythms and genres. Secondly, many pronunciations in singing voices are changed. For example, some vowels are often assigned with longer duration than that in speech while some consonants are omitted.

Similarly to ASR, ALT systems can be built using traditional hybrid methods [2, 3] as well as end-to-end ones [1, 4]. In the hybrid architectures, acoustic models are typically used together with some in-domain LMs to generate better results. A new architecture called MSTRE-Net is proposed in [2] where multiple streams of layers are trained in parallel to improve the recognition accuracy. Since pronunciations in singing voice are often changed in various ways, a genre and duration-based modified pronunciation lexicon is employed in [5] to match the different singing styles. As end-to-end models usually demand more training data, authors in [4] developed a voice-to-singing (V2S) module to convert natural speech to singing voice using a vocoder based speech synthesizer. The converted singing voices are then used as training data for an end-to-end ALT model. The following parts of this report will describe our submitted systems for the automatic lyrics transcription task of MIREX 2021.

2. AUTOMATIC LYRICS TRANSCRIPTION

Our ALT system is built using DNN-HMM architecture with Kaldi toolkit [6]. Specifically, we take advantage of Kaldi CHAIN [7] as the acoustic model together with a 5-gram language model (LM).

2.1. Data Preparation

Typically, sing vocals are firstly extracted from original audios using tools like Spleeter [8] before training or testing in lyrics transcription. This may eliminate the complex influences caused by music accompaniments. On the other hand, vocal extraction will also bring side effects where some vital acoustic information is lost and make it inconvenient to use in the production environments. As a result, our system takes as input original audios and no extra processing is required for recognition.

We used the LibriSpeech [9], DAMP [10] and DALI [11] datasets as the training data. Additionally, to further improve the system performance, we also collected large number of music audios and textual lyrics from Cloud Music to train acoustic and language models. Songs related to testsets are carefully checked and filtered out from all training sets for a reasonable evaluation.

2.2. The Proposed Method

For all of our experiments, we follow the training procedure of traditional hybrid ASR system. For training GMM-HMM acoustic models, 39-MFCC features including the deltas and delta-deltas are computed from singing audios. When training CHAIN models, we used 40-MFCC features with frame rate 10 ms and window length 25 ms, respectively. Firstly, we trained a basic GMM-HMM system to generate the alignments for CHAIN models. To accommodate the duration of vowels and consonants, we modified the training lexicon accordingly. We also randomly masked out some of the words in the training transcripts and inserted special $<NOISE>$ label at the beginning and end of a utterance, which is similar to the prior work of ZWZL1 in MIREX 2020 lyrics-to-audio alignment task. In this way, we improved the robustness of the system significantly.

When training LM, we explore the impact of different LMs. we trained the standard 5-grams LM with Kneser-Ney smoothing using SRILM toolkit [12] and the cutoff value is set to 1-1-1-2-3. We also tried to interpolate the in-domain song LM and the general LM, but there was no significant gain in the transcription task.

3. EXPERIMENTS

Table 1 shows the word error rate (WER) on the open-source Jamendo dataset [1]. We train several models on different sizes of datasets to explore the effects of our proposed method and the gains brought by large size of datasets. When training the model YHLLL3 only on open source datasets (i.e., LibriSpeech, DAMP and DALI), we follow the traditional GMM-HMM-CHAIN procedure and train a normal speech recognition model on LibriSpeech only. Then we finetune the aforementioned model using DAMP and DALI for another 30 epochs. The LM for this model is trained using transcriptions of DAMP and DALI as described earlier except that the cutoff value is set to 1-1-1-1-1.

For models YHLLL1 and YHLLL2, we use an extra large internal dataset collected from Cloud Music to train both the acoustic and language models. The only difference is that YHLLL1 is trained for 4 epochs while YHLLL2 is trained for 10 epochs. For the model YHLLL4, we combined transcriptions of DAMP and DALI with our internal datasets and train a new LM. The other parts are kept the same as YHLLL2. Finally, to evaluate the effectiveness of large LMs, we replace the LM in YHLLL4 with the one used in our ALT systems in the real production environments, which is referred as YHLLL5. The LM used in YHLLL5 was trained on a full textual lyrics corpus covering nearly all the common songs, including those in testsets.

Table 1: *Word error rate for 5 models on Jamendo dataset.*

	Jamendo
Librispeech + DAMP and DALI finetuning (YHLLL3)	25.91
+ internal data (YHLLL1)	23.50
+ extra 6 epochs (YHLLL2)	21.17
+ DAMP and DALI transcriptions in LM (YHLLL4)	20.12
+ LMs in production environments (YHLLL5)[1]	7.06

This is just a demonstration of how we use ALT in our production environments (i.e., covering as many songs as possible in LMs), which may not be a valid case for quantitative evaluation. Thus, we did not submit this model for any further evaluation. It is obvious that covering as many songs as possible in LMs is an easy and effective way to improve the recognition accuracy. Table 2 shows a detailed WER for YHLLL4.

4. CONCLUSIONS

We provide an approach for automatic lyric transcription with original music audios and no data pre-processing is needed. We build the recognition system using traditional hybrid DNN-HMM method with Kaldi toolkit. Data augmentation in training transcriptions is employed to improve the system robustness. Experimental results have shown that our system benefits a lot from large datasets as well as adapted training strategies.

5. References

[1] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 181–185.

[2] E. Demirel, S. Ahlbäck, and S. Dixon, "Mstre-net: Multistreaming acoustic modeling for automatic lyrics transcription," *arXiv preprint arXiv:2108.02625*, 2021.

[3] ——, "Low resource audio-to-lyrics alignment from polyphonic music recordings," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 586–590.

[4] S. Basak, S. Agarwal, S. Ganapathy, and N. Takahashi, "End-to-end lyrics recognition with voice to singing style transfer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 266–270.

[5] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics alignment and transcription in polyphonic music: Does background music help?" in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 496–500.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[1] not submitted

Table 2: *Word error rate (WER) of YHLLL4 on Jamendo dataset.*

Track	**WER**	**Insertions**	**Deletions**	**Substitutions**	**TotalWords**
Avercage_-_Embers	67.72	0	120	8	189
Color_Out_-_Falling_Star	12.56	7	8	13	223
Cortez_-_Feel__Stripped_	8.78	7	17	7	353
Explosive_Ear_Candy_-_Like_The_Sun	5.66	1	4	10	265
HILA_-_Give_Me_the_Same	12.11	0	31	8	322
JASON_MILLER_-_CROWD_PLEASER	15.77	11	13	58	520
Kinematic_-_Peyote	18.37	1	17	9	147
Lower_Loveday_-_Is_It_Right_	8.96	2	9	8	212
LUNABLIND_-_Vision__Radio_Edit_	8.68	10	4	11	288
Moon_I_Mean_-_Wrong_Concept	54.68	0	122	24	267
Pure_Mids_-_The_Leader	43.86	0	46	4	114
Quentin_Hannappe_-_Keep_On	10.86	1	7	11	175
Ridgway_-_Fire_Inside	39.80	8	83	30	304
Rxbyn_-_Bad_Side	15.45	6	51	11	440
Slingshot_Miracle_-_Whistler	18.52	1	20	9	162
Songwriterz_-_Back_In_Time	20.00	1	33	13	235
The.madpix.project_-_One_Way_Street	14.21	2	8	16	183
The_Rinn_-_Voices__2017_Version_	27.72	3	37	16	202
Tom_Orlando_-_The_One__feat._Tina_G_	20.61	5	61	36	495
Wordsmith_-_The_Statement	14.29	17	15	51	581
Total	20.12	83	706	353	5677

[7] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[8] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020, deezer Research. [Online]. Available: https://doi.org/10.21105/joss.02154

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[10] I. Smule, "DAMP-MVP: Digital Archive of Mobile Performances - Smule Multilingual Vocal Performance 300x30x2," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.2747436

[11] 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 *19th International Society for Music Information Retrieval Conference*, ISMIR, Ed., September 2018.

[12] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.