THE SHEFFIELD UNIVERSITY SYSTEM FOR THE MIREX 2020:LYRICS TRANSCRIPTION TASK

Gerardo Roa Dabike **Jon Barker**
The University of Sheffield
{groadabike1,j.p.barker}@sheffield.ac.uk

ABSTRACT

This extended abstract describes the system we submitted to the MIREX 2020 Lyrics Transcription task. The system consists of two modules: a source separation front-end and an ASR back-end. The first module separates the vocal from a polyphonic song by utilising a convolutional time-domain audio separation network (ConvTasNet). The second module transcribes the lyrics from the separated vocal by using a factored-layer time-delay neural network (fTDNN) acoustic model and a 4-gram language model. Both the separation and the ASR modules are trained on a large open-source singing corpora, namely, Smule DAMP-VSEP and Smule DAMP-MVP. Using a separation module audio pre-processing reduced the transcription error by roughly 11% absolute WER for polyphonic songs compared with transcriptions without vocal separation. However, the best WER achieved was 52.06%, very high compared to WERs as low as 19.60% that we achieved previously for unaccompanied song [16].

1. INTRODUCTION

The task of automatic speech recognition (ASR) task is that of identifying and transcribing words directly from an audio signal, whether the signal is a single speaker, multiple speakers or speech in a noisy environment. We use the term lyric transcription (LT) when the audio signal corresponds to *sung* speech, where, in general, the voice may be unaccompanied (i.e., acapella) or in the presence of background musical accompaniment.

Existing acapella singing LT systems are typically based on successful approaches for spoken speech. In particular, these systems utilise the same acoustic features motivated by the similarities between sung and spoken speech, i.e., they share the same production systems and convey semantic information in the same way [6, 9, 16, 18].

However, several differences between the sung and spoken speech styles make LT a more difficult task. First, sung speech possesses a larger pitch range and higher pitch average than spoken speech [8]. Second, in spoken speech, the pitch can vary freely up to 12 semitones within a syllable [12] but, in sung speech, these variations are more discrete with changes no greater than two semitones [19]. Third, the duration of the sung speech syllables is larger than in spoken speech, which can lead to phone insertion and substitution errors [5]. Finally, singers can employ vibrato singing, a frequency modulation of periodic pitch variations of between 5.5 and 7.5 Hz [17]. These differences are the result of the fact that, in song, artistic interpretation tends to have greater importance than speech intelligibility.

The transcription task becomes even more complicated when the singing is in the presence of background instrumental accompaniment (polyphonic music). This presents a source separation problem which is made particularly challenging by the fact that singing is often highly correlated with the background, resulting in a frequency overlap and synchronised frequency and amplitude modulations.

There are basically two main ways to tackle this problem. First, train models to transcribed the lyrics directly from the polyphonic music audio. This approach has been employed in lyrics to audio alignment task. In these kinds of problems, unlike LT, the lyrics are known, and the objective is to align the phonetic units with the knowing lyrics [4]. The second approach relies on an audio source separation front-end to enhance the vocal source before passing it to a transcription back-end [9], trained on either enhanced or isolated sung speech.

Following the second approach, in this work, we present an extension of the solo-singing sung speech ASR system presented in [16]. We incorporate a source separation front-end consisting of a convolutional time-domain audio separation network (ConvTasNet) trained on the DAMP-VSEP corpus [2].

2. DATASETS

For training and evaluation, we use two open-source DAMP singing corpora sourced from the Smule [1] karaoke application, and the polyphonic musical corpus DALI [10].

First, we utilised the Smule Digital Archive of Mobile Performances - Vocal Separation (DAMP-VSEP) dataset [2] to train the audio source separation module. For each song, this dataset provides the vocal segments, the corresponding backtracking and a mixture of the two. It is composed of 41000 segments, each of 30-seconds, from recordings made in 155 countries and 36 different languages,

[1] www.smule.com

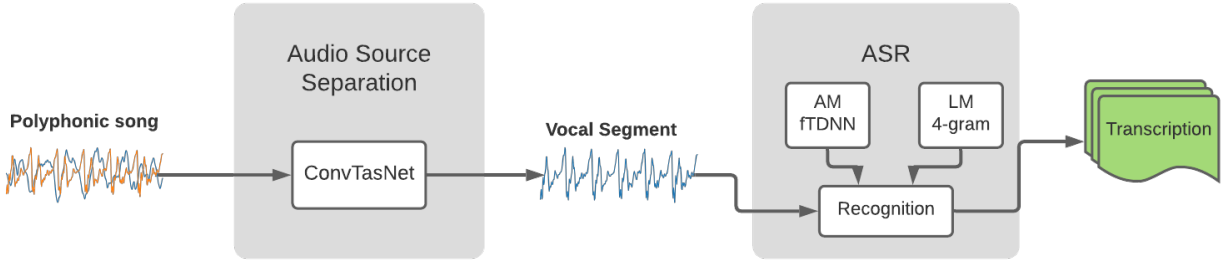

Figure 1. Lyrics transcription pipeline system.

Table 1. Description of the training datasets.

Module	Dataset	Songs	Size
Source Separation	DAMP-VSEP	9243	77 hrs.
Acoustic Model	DSing30	4234	149 hrs.

Table 2. Description of the system evaluation sets.

Dataset	Utt	Size
DAMP-VSEP	416	0.5 hrs.
DALI	515	0.4 hrs.

by 6456 singers with 11,494 song arrangements. We constructed a subset of the corpus by filtering all songs classified as English and rearranging the duets ensembles into two single performances. In the corpus, more than one performance of the same song may be available. For each performance, a copy of the background is provided. However, no information is provided that indicates that these copies are in fact, the same background. Therefore, to avoid overlap in the training and development test sets, we detected and grouped identical backgrounds by using their MD5 checksum. This resulted in about 2100 distinct backgrounds. Then, we cross-correlated the distinct backgrounds to group all perceptually similar ones. This last step was necessary because several non-identical backgrounds are slightly different versions of the same recording, e.g., they are time-shifted versions or has different volume level. Then, given the correlation matrix, we recursively group the backgrounds with a correlation greater than certain threshold. We tested several thresholds values and selected a value equal to 0.9831, which was a value where, after human evaluation, the clustering error was minimised. This process resulted in 1364 clusters of perceptually distinct backgrounds. Finally, the development and evaluation sets were constructed by selecting, and equally distributing, 200 backgrounds from the clusters with a single element. To add an extra precaution of avoiding overlapping with the train set, we chose the 200 backgrounds were its higher correlation with any other background is lower than 0.9440, which is the minimum threshold to complete the 200 backgrounds needed. This process resulted in 9243 performances for training, 100 for validation and 100 for evaluation. Further, the evaluation set was humanly aligned and transcribed at the utterance level, enabling it to be used as a system evaluation set.

For the ASR module, we utilised the DSing30 [16] for training a solo-singing AM model. DSing30 is the largest training set offered by the DSing dataset [16]. DSing corresponds a pre-processed dataset composed by 4,460 English karaoke performances from the larger multi-language karaoke performances Smule Multilingual Vocal Performance 300x30x2 (DAMP-MVP) dataset [1]. For details of the construction of DSing, please refer to [16].

Finally, we selected ten songs from the ground truth of DALI dataset [10] to be used for system evaluation. DALI is a collection of polyphonic songs sourced from YouTube with synchronised audio, lyrics and notes. Due to DALI relies on the availability of the videos on YouTube. At the time of writing, 91 out of the 105 ground truth songs remain online.

Table 1 presents a summarise of the source separation and AM model training datasets.

3. SYSTEM FRAMEWORK

The LT system presented is composed of two independent modules connected in a pipeline; the audio source separation and ASR module. Figure 1 shows a diagram of the transcription system.

For the audio source separation module, we train a ConvTasNet [7] model by using the Asteroid PyTorch-based audio source separation toolkit [11]. The model was trained on the DAMP-VSEP dataset for 100 epoch using four GPUs, learning rate of 0.0003 and the Adam optimization algorithm.

For the ASR module, we extended the sung speech recognition system described in [16], built using the Kaldi ASR toolkit [14], by expanding the MFCC feature vector with vocal source features (VSF). The VSF features are conformed by four pitch features, two jitter parameters, one shimmer parameter and harmonic to noise ratio. Using these expanded features, we trained a factorised time-delay neural network (TDNN-F) [13] AM with a lattice-free maximum mutual information (LF-MMI) loss function [15]. We employed a two frames context vector consistent of 40 MFCC and 8 VSF, plus 100 i-Vectors [3]. The model was trained on the solo-singing DSing30 corpus.

Table 3. Transcription performances decoding with the 4-gram LM. *Ref vocal* refers to the isolated vocal segment, *Sep vocal* refers to the separated vocal resulting from the separation module and *Mixture* refer to the polyphonic song.

Eval set	Audio conditions	WER
DAMP-VSEP	Ref vocal	24.07
	Mixture	63.98
	Sep vocal	**52.06**
DALI	Mixture	87.81
	Sep vocal	**75.91**

The language model (LM) employed is a 3-gram MaxEnt LM trained on an in-domain lyrics corpus sourced from Lyrics Wiki website [2] , and a 4-gram model for final rescoring trained on the same corpus.

4. EXPERIMENTS RESULTS

For the evaluation of the system, we employ two different datasets: DAMP-VSEP and DALI. Table 2 summarises the size of these evaluation sets. First, we measure the performance for the isolated vocal (Ref vocal) and the the DAMP-VSEP mixture (Mixture) without applying the source separation pre-processing. We obtained WERs of 24.07% and 63.98% for the isolated vocal and the mixture, respectively. These results serve as an upper- and lower- performance band. Then, we activated the source separation module to separate the vocal (Sep vocal) from the DAMP-VSEP mixture, obtaining a WER of 52.06%, i.e. at 12% absolute decreased compared to the unprocessed mixed audio. Then, we repeated the experiments this time using the DALI mixture audio. Transcribing the lyrics directly from the DALI mixture audio led to a WER of 87.81% and introducing the source separation reduced the WER to 75.91%.

In both evaluations, the utilisation of the source separation front-end decreased the transcription error of polyphonic music in about 11%. However, the WER on the DALI data is significantly higher than that obtain with DAMP-VSEP. The acoustic difference between DALI and DAMP-VSEP may explain these disparate results. Most of the backgrounds of the latter are simpler mixtures, i.e., they are acoustic arrangements composed of a single instrument like piano or acoustic guitar. Also, in DAMP-VSEP, there are fewer types of instruments across the dataset, making it less variable than DALI.

Table 3 presents a summarise of the system performance in WER. The performances of the system using the source separation module before the ASR are presented in bold.

5. CONCLUSIONS

We presented a two-component lyric transcription system composed of a source separation front-end and an ASR back-end. The front-end module is a ConvTasNet source separation system trained on DAMP-VSEP dataset using Asteroid toolkit. The second module is an extension of the ASR system presented in [16], where we expanded the MFCC features with voice source features. The source separation component produced a 11% absolute WER improvement when evaluated on either DAMP-VSEP (Karaoke performances) or DALI (music from YouTube). However, there is plenty of space for improvement as shown by the DAMP-VSEP results, where the WER for the pre-mixed vocal track is 24.07%, but for the separated vocal it increases to 52.06%. These results suggests that large performance gains could be made through improvements to the separation front-end, or by retraining the acoustic model to reduce mismatch caused by residual noise in the separated signal. We are following both these directions in ongoing work.

[2] https://lyrics.fandom.com/wiki/LyricWiki

6. REFERENCES

[1] Smule, Inc. (2018). DAMP-MVP: Digital Archive of Mobile Performances - Smule Multilingual Vocal Performance 300x30x2 (Version 1.0.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.2747436.

[2] Smule, Inc. (2019). DAMP-VSEP: Smule Digital Archive of Mobile Performances - Vocal Separation (Version 1.0.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3553059.

[3] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[4] Chitralekha Gupta, Emre Yılmaz, and Haizhou Li. Acoustic Modeling for Automatic Lyrics-to-Audio Alignment. In *Proc. Interspeech*, 2019.

[5] Dairoku Kawai, Kazumasa Yamamoto, and Seiichi Nakagawa. Lyric recognition in monophonic singing using pitch-dependent DNN. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[6] Anna M Kruspe. Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[7] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27:1256–1266, Aug 2019.

[8] Julia Merrill and Pauline Larrouy-Maestri. Vocal features of song and speech: Insights from Schoenberg's Pierrot lunaire. *Frontiers in Psychology*, 8(JUL), 2017.

[9] Annamaria Mesaros and Tuomas Virtanen. Recognition of phonemes and words in singing. In *Proc. IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010.

[10] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. DALI: a large Dataset of synchronized Audio, LyrIcs and notes, automatically created using teacher-student machine learning paradigm. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[11] Manuel Pariente, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian-Robert Stöter, Mathieu Hu, Juan M. Martín-Doñas, David Ditter, Ariel Frank, Antoine Deleforge, and Emmanuel Vincent. Asteroid: the PyTorch-based audio source separation toolkit for researchers. In *Proc. Interspeech*, 2020.

[12] Aniruddh D. Patel, Meredith Wong, Jessica Foxton, Aliette Lochy, and Isabelle Peretz. Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Perception*, 25:357–368, 04 2008.

[13] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech*, 2018.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[15] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proc. Interspeech*, 2016.

[16] Gerardo Roa Dabike and Jon Barker. Automatic lyric transcription from Karaoke vocal tracks: Resources and a baseline system. In *Proc. Interspeech*, 2019.

[17] Johan Sundberg. *The science of the singing voice*. Northern Illinois University Press, 1987.

[18] Che-Ping Tsai, Yi-Lin Tuan, and Lin-Shan Lee. Transcribing lyrics from commercial song audio: The first step towards singing content processing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[19] Piet G. Vos and Jim M. Troost. Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception: An Interdisciplinary Journal*, 6(4):383–396, 1989.