LYRICS ALIGNMENT SYSTEMS SED1 AND SED2 FOR MIREX 2019

Daniel Stoller
Queen Mary University of London
d.stoller@qmul.ac.uk

Simon Durand
Spotify
durand@spotify.com

Sebastian Ewert
Spotify
sewert@spotify.com

ABSTRACT

This article describes the systems “SED1” and “SED2” for the MIREX 2019 Automatic Lyrics Alignment Challenge. They are based on a modified Wave-U-Net architecture, which predicts character probabilities directly from raw audio using learnt multi-scale representations of the various signal components. There are no sub-modules whose interdependencies need to be optimized. Our training procedure is designed to work with weak, line-level annotations available in the real world, and achieves substantial performance improvements over all previous approaches evaluated at MIREX.

1. SYSTEM DESCRIPTION

The two submitted systems are exactly the ones presented in the paper “END-TO-END LYRICS ALIGNMENT FOR POLYPHONIC MUSIC USING AN AUDIO-TO-CHARACTER RECOGNITION MODEL” [2]. The paper is accessible at `https://arxiv.org/abs/1902.06797` and supplementary materials are available at `https://sigport.org/documents/end-end-lyrics-alignment-using-audio-character-recognition-model`.

More specifically, the SED1 system is the main system presented in the paper (denoted as “Ours”), and uses an adapted version of the Wave-U-Net [3] to predict characters directly from the raw waveform of the music input, which is trained using a CTC loss.

The SED2 system is functionally equivalent to the SED1 system, but instead of using the polyphonic music directly as input, the accompaniment is first removed by a pre-trained singing voice separation system. The system is also explained in the paper (see “Discussion” section), and achieves a further improvement in alignment accuracy.

Note that the MIREX results can slightly vary from the ones obtained in the paper when evaluating on the Mauch dataset [1], since MIREX maintains its own version of the Mauch dataset that can differ slightly from the ones used by the authors.

2. REFERENCES

[1] Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations. In *Proceedings of the Sound Music Computing Conference (SMC)*, pages 9–16, 2010.

[2] Daniel Stoller, Simon Durand, and Sebastian Ewert. End-to-end Lyrics Alignment for Polyphonic Music Using An Audio-to-Character Recognition Model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019. IEEE.

[3] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Source Separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, volume 19, pages 334–340, 2018.