LYRICS-TO-AUDIO ALIGNMENT USING SINGING-ADAPTED ACOUSTIC MODELS AND NON-VOCAL SUPPRESSION

Chitralekha Gupta[1,2] **Bidisha Sharma**[2] **Haizhou Li**[3] **Ye Wang**[1,2]

[1] NUS Graduate School for Integrative Sciences and Engineering, [2] School of Computing,
[3] Electrical and Computer Engineering Dept., National University of Singapore, Singapore

chitralekha@u.nus.edu,{s.bidisha, haizhou.li, dcswangy}@nus.edu.sg

ABSTRACT

In this work, we describe an algorithm that we have submitted for the MIREX 2018 task of Automatic Lyrics-to-Audio Alignment. The goal is to automatically detect word boundaries in English pop music, given the mixed singing audio (singing voice + musical accompaniment) and lyrics as inputs. The key component of the this submission is the singing-adapted acoustic models with lexicon-based duration modeling. As singing voice differs from speech, we have adapted speech models to singing voice. Moreover, to account for the long duration vowels in singing, we have modified the lexicon with longer duration vowel pronunciation variants. In this algorithm, we also apply a singing-vocal detection method to suppress the non-vocal sections before forced-aligning with the singing-adapted models.

1. APPROACH OVERVIEW

In automatic speech recognition (ASR) tasks, word or phone-level segmentation is obtained by forced-aligning the transcription to the speech audio using acoustic models trained with speech data. In this MIREX task, we apply the same idea to align lyrics to music audio. However we introduce several changes to handle the differences between the speech and the singing vocals with background music.

Although singing and speech share the same vocal production machinery, they are different in their timbre, pitch, and duration. To address these differences, we adapt speech trained acoustic models to singing voice. Adaptation of speech models for singing was previously attempted by Mesaros et al. [4] who applied the speaker adaptation techniques to transform speech recognizer to singing voice recognizer with a small singing dataset. We apply the same speaker adaptive training (SAT) method, but now with a large, automatically cleaned and annotated solo-singing dataset [3, 8] to adapt speech models to singing voice.

One major difference between speech and singing voice is in the duration of vowels. The vowels in singing could be longer in duration than spoken vowels, because they are dictated by the melodic and rhythmic attributes of the song. Longer duration of vowels can be viewed as a type of pronunciation variation. Therefore we modify the lexicon to model the duration dynamics of vowels in singing. We adopt the strategy of optional repetition (up to 4 times, set empirically) of the vowels so as to allow longer duration of the vowels [2]. For example, the word *sleep* will have the following lexicon variants: [S L IY IY IY IY P], [S L IY IY IY P], [S L IY IY P], [S L IY P]. Such variants are created with respect to every vowel in the word, and the ASR selects the closest matching variant at the time of forced-alignment. We expect that this method will result in improvement in alignment as reported in [2].

The presence of background music is another major difference between speech and singing vocals+music audio and it may lead to increase in misalignment. One solution could be to extract singing voice from the background music, and then to apply the singing-adapted ASR for alignment. We chose a state-of-the-art algorithm to extract the singing voice [1]. However the extracted singing vocals were noisy resulting in distorted MFCCs. So the alignment was not successful when applied over extracted singing vocal. We observed that singing voice is loud and dominant over the background music and occupies a different range of frequencies than the overlying music in many popular English songs. Thus applying forced-alignment directly on the original songs gives a reasonably good alignment.

Many pop songs also have long interludes of music in between stanzas without singing vocals. Such periods of music are problematic for our acoustic models that are trained on solo-singing vocals. To solve this problem, we detected the instrumental segments, and replaced them with silence.

2. ALGORITHM: NON-VOCAL SUPPRESSION AND GMM-HMM (SAT) MODELS FOR ALIGNMENT

Long musical interludes cause errors in alignment because our models are not trained for music. Therefore, we attempted to detect the non-vocal (i.e. instrumental) regions of the song first and replace them with silence, so that those regions are detected as silence when forced-aligned with the ASR. To achieve this, we first extracted the singing vocals from the background music as described in [1]. In the extracted vocal we observed that the energy of the segments with only background music is suppressed to some

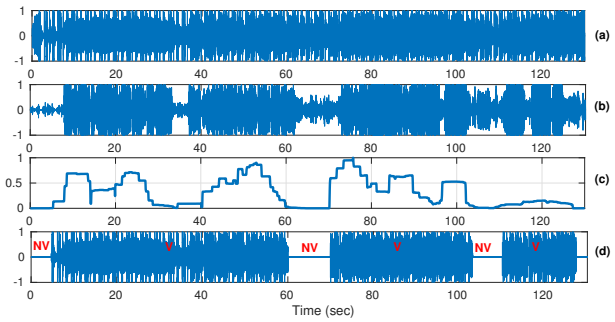

Figure 1. Non-vocal suppression method (a)original audio, (b)vocal extracted signal, (c)sub-band energy contour obtained from (b), (d)audio signal after suppressing non-vocal segments.

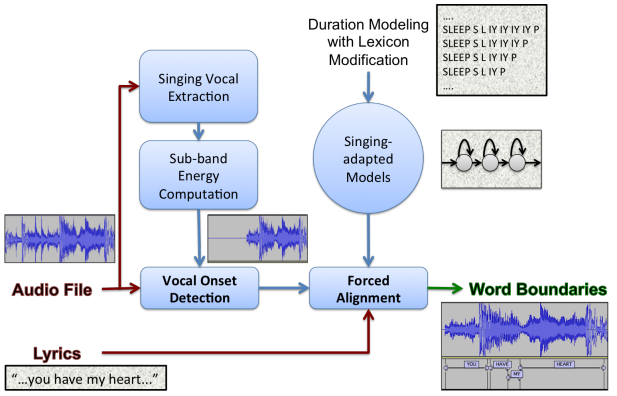

Figure 2. Framework of automatic lyrics-to-audio alignment.

extent. To analyze this, we divided the spectrum of each frame (framesize 25 ms, frameshift 5 ms) into four equal sub-bands. The energy corresponding to the second sub-band shows a prominent difference between the segments with vocals and without vocals. A threshold based on average second sub-band energy contour is set to classify the frames into vocal and non-vocal segments. The non-vocal segments which are of very long duration are more likely to increase the misalignment of the lyrics with the song. Therefore, with this algorithm these regions are replaced with silence in the original audio. Figure 1 shows the process of non-vocal suppression.

The overview of the complete framework is depicted in Figure 2. In this algorithm, the non-vocal suppressed audio is forced-aligned with the lyrics using singing-adapted speech models. The baseline ASR is a tri-phone GMM-HMM trained on Librispeech corpus [5] using MFCC features on Kaldi toolkit [6]. We use feature-space maximum likelihood linear regression [7] to compute transformations of the singing feature vectors. These transformations were applied at the time of training for the adaptation of the speech models to singing voice using solo-singing data, called SAT [3]. The duration modeling with lexicon modification was also applied at the time of training [2].

To make the Viterbi alignment algorithm operate over the long duration of songs, we set the alignment retry-beamwidth to a high value of 2000. We apply the flag for optional silence to handle intermittent pauses.

	Absolute Average Error	
Track	Non-Vocal Suppressed and SAT	MIREX 2017 best results
_umbrella_rihanna	0.59	2.65
_muse_guiding_light	4.23	28.46

Table 1. Average absolute error in word alignment, and percentage correct frames using non-vocal suppression and SAT model.

3. PRELIMINARY RESULTS

The Lyrics-to-Audio subtask-2 in MIREX 2018 has provided two example songs, along with their lyrics and the ground-truth word alignment files. The organizers have also provided the evaluation code that provides the absolute average error metric. The algorithm evaluation results on this data are given in Table 1. The results show improvement compared to last year's best performing system in MIREX for these two songs. Moreover, we observe a larger improvement for the song *_muse_guiding_light*. This is because this song is slow with long and multiple stretches of musical interludes. So the combination of non-vocal region suppression and vowel duration modeling in our algorithm results in improved word alignments.

4. REFERENCES

[1] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer, 2017.

[2] Chitralekha Gupta, Haizhou Li, and Ye Wang. Automatic pronunciation evaluation of singing. In *Interspeech 2018 (To appear)*, 2018.

[3] Chitralekha Gupta, Tong Rong, Haizhou Li, and Ye Wang. Semi-supervised lyrics and solo-singing alignment. In *ISMIR 2018 (To appear)*, 2018.

[4] Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):546047, 2010.

[5] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015.

[6] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[7] Daniel Povey and George Saon. Feature and model space speaker adaptation with full covariance gaussians. In *Ninth International Conference on Spoken Language Processing*, 2006.

[8] Smule. Digital Archive Mobile Performances (DAMP). `https://ccrma.stanford.edu/damp/`. [Online; accessed 15-March-2018].