LYRICS-TO-AUDIO ALIGNMENT USING SINGING-ADAPTED ACOUSTIC MODELS

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

We describe two algorithms 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 two submissions 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 the first algorithm, we use the speaker adaptive trained (SAT) models to forced-align lyrics-to-audio. In the second algorithm we use a deep neural network (DNN) model trained on top of the SAT models for the forced-alignment.

1. APPROACH OVERVIEW

In automatic speech recognition (ASR) tasks, word or phone-level segmentation is obtained by forced-aligning the transcription to the speech 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. [5] who applied the speaker adaptation techniques to transform speech recognizer to singing voice recognizer with a small singing dataset. We apply the same SAT method, but now with a large, automatically cleaned and annotated solo-singing dataset [4, 10] 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 [3]. 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 [3].

The presence of background music is another major difference between speech and singing vocals+music audio. The background music may interfere with the singing voice if they lie in the similar frequency range. One solution could be to extract singing voice from the background music, and then to apply the solo-singing trained models for alignment. Singing voice extraction is an active research area, and we chose a state-of-the-art algorithm to extract the singing voice [1]. However the extracted singing vocals from the algorithm were noisy and resulted in distorted Mel frequency cepstral coefficients (MFCCs). Thus applying forced-alignment on such extracted singing vocals was not successful. 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 have long musical intro that is without singing vocals. Such periods of music are problematic for our acoustic models that are trained on solo-singing vocals. Therefore, we detected the instrumental segments in the beginning of the song, and replaced them with silence.

2. ALGORITHM A: GMM-HMM (SAT) MODELS FOR ALIGNMENT

The overview of the framework is depicted in Figure 1. In this algorithm, the input audio file is forced-aligned with the lyrics using singing-adapted speech models. The baseline speech acoustic model is a tri-phone Gaussian mixture model (GMM)-hidden Markov model (HMM) trained

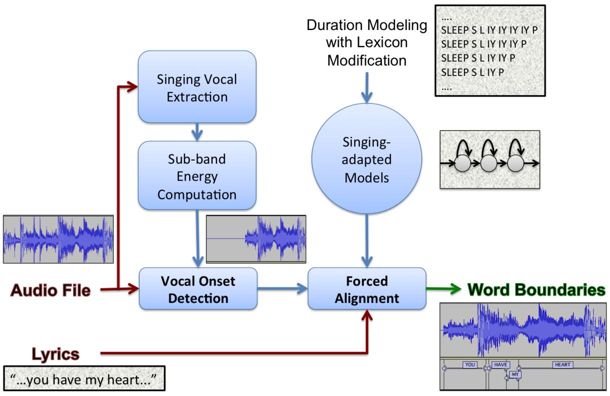


Figure 1. Framework of automatic lyrics-to-audio alignment.

on Librispeech corpus [6] using MFCC features on Kaldi toolkit [8]. We use feature-space maximum likelihood linear regression [9] to compute transformations of the singing feature vectors. These transformations were applied at the time of training for a semi-supervised adaptation of the speech models to singing voice using solo-singing data, called SAT [4]. The duration modeling with lexicon modification was also applied at the time of training [3].

To make the Viterbi alignment algorithm operate over the long duration of songs (∼4–5 minutes), we set the alignment retry-beamwidth to a high value of 2000. Also the flag for optional silence was on to handle the possibility of pauses. To avoid misalignment due to the presence of long duration musical intro, we apply an energy-based algorithm over the extracted vocals to detect the non-vocal part over the first few seconds of the song. These are then replaced with silence.

3. ALGORITHM B: SAT+DNN MODELS FOR ALIGNMENT

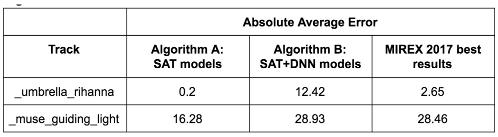
This algorithm is the same as algorithm A, except for the singing-adapted models. A DNN model [2] is trained on top of the SAT model with the same set of training data. During DNN training, temporal splicing is applied on each frame with left and right context window of 4. The SAT+DNN model has 3 hidden layers and 2,976 output targets.

DNN models are not good for alignment since the objective function they are trained with does not force them to produce good alignments. For forced-alignment, a GMM-based model is generally recommended [7]. So we expect that algorithm A should perform better than algorithm B.

4. 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 gives the absolute average error metric. We have evaluated both the algorithms A and B on this data, which is shown in Table 1.

Overall, algorithm A performs better than last year's best performing system for both the songs. The improvement could be due to the clean annotated singing data used for model adaptation, and the duration-based lexicon modification. Also, the performance of the SAT models is better than the SAT+DNN models, because the DNN models are not optimized for alignment, as discussed in Section 3.



	Absolute Average Error		
Track	Algorithm A: SAT models	Algorithm B: SAT+DNN models	MIREX 2017 best results
_umbrella_rihanna	0.2	12.42	2.65
_muse_guiding_light	16.28	28.93	28.46

Table 1. The average absolute error in word alignment, and the percentage correct frames using Algorithm A (SAT models) and Algorithm B (SAT+DNN).

5. 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] G. Hinton et al. Deep neural networks for acoustic modeling in speech recognition. In *IEEE Signal Processing Magazine*, volume 29, pages 82–97, 2012.

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

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

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

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

[7] Daniel Povey. Kaldi Google group for help. `https://groups.google.com/forum/#!topic/kaldi-help/cSAm5iXGhZo`. [Online; accessed 9-August-2018].

[8] 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.

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

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