Automatic Vocal Melody Alignment to Lyrics in Consumer Available Music

Jonathan Chung
Department of Electrical and Computer Engineering
University of Toronto
jono.chung@mail.utoronto.ca

Guowei (David) Lu
Department of Computer Science
University of Toronto
guowei.lu@mail.utoronto.ca

Ka Ho (Jeff) Wu
Department of Computer Science
University of Toronto
jeffkahow892@gmail.com

Abstract

We described a method to vocals melody to lyrics in consumer available music. Vocal extraction, noise reduction and forced alignment are the major components of the model. The vocal extraction component was completed using robust principle component analysis and the noise was reduced by estimating a binary mask for vocal and noise. Phones were detected using an acoustic model and they are aligned to the the lyrics to obtain timestamps. In our tests, a 70% overlap between the automatically aligned lyrics and manually aligned lyrics was found in some songs.

1 Introduction

The vocal melodies can be aligned to lyrics professionally by directly transcribing from the official music scores. However, official music scores are not readily available and are not easily interpreted by laymen. Lyrics that are aligned to the music is desirable for consumers in applications such as Karaoke or fan made music videos that are posted on video sharing websites. Music video showcased for Karaoke are professionally aligned word by word however, in fan made videos, the lyrics are typically displayed line by line. The discrepancies between the two suggest that aligning lyrics (word by word) to vocal melodies requires significant efforts to complete.

A fully automatic technique that include speech recognition and lyrics alignment is generally not required as song lyrics are easily interpretable and widely distributed. Therefore, a speech recogniser to detect words at each timestamp is over-complicating the problem. The problem with aligning speech to a list of predefined words may be more simple. However, when aligning vocal in music to lyrics: the background music, irregularities in vocals (i.e., some words are sung longer or shorter compared to average speech) and incomprehensible pronunciations often create issues in correctly aligning the vocals in music to lyrics. Typically, systems that align vocal in music to lyrics uses the following approach to maximise the accuracy: (1) vocal extraction, (2) vocal manipulation and/or signal processing then (3) forced alignment.

We introduce a method to automatically align vocal melodies in consumer available music to the corresponding lyrics. We take a song as an input and extract the vocal using robust principle component analysis (RPCA) [1]. The extracted vocal is denoised by using a graph cut algorithm to create a binary mask for noise and vocal components. The denoised vocal is placed into an HMM acoustic model for phone recognition, and this is followed by a forced alignment. Figure 1 describes the model:

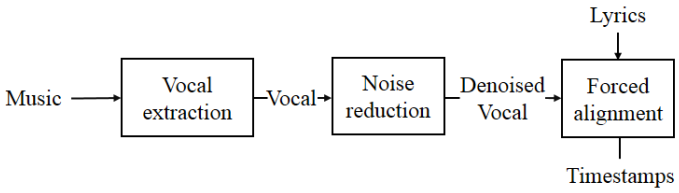

Figure 1: Overview of the model.

This paper is organised in the following manner. Firstly, the related work are described. This is followed by describing the major components of the system: vocal extraction, noise reduction and forced alignment. An experiment was conducted to test the system and the results are discussed. The paper is concluded with a summary and the future directions.

2 Related work

Previously described techniques of automatically aligning vocals in music to lyrics is performed by: vocal (singing voice) extraction and manipulation followed by lyrics to vocal forced alignment. The technique described by Fujihara et al. [2] used the F0 frequencies (fundamental frequency) to detect vocal sections of the music. In the detected sections with vocals, the non-vocal components are suppressed by extracting the harmonic structure of vocal and resynthesizing it. Fujihara et al. used an HMM based forced alignment technique with a Japanese phones language model and aligned the detected sounds with the corresponding lyrics. Mesaros and Virtanen [3] described a technique similar to Fujihara et al. but specifically developed the forced alignment technique with an English language model. Mauch et al. [4] extended the technique developed by Fujihara et al. by evaluating the background musics chord at each word of the lyrics. In addition to the phones of the vocal, the chord pitch were also aligned.

3 Vocal Extraction

Instrumental extraction is a common problem and can be completed with simple signal processing techniques (see appendix) on the other hand, vocal extraction is a difficult problem. The repetitive nature (less variable) of the instrumental components (in contrast the variable nature of the vocal components) is leveraged to perform vocal extraction. The robust principle component analysis (RPCA) [5] is used to separate an input matrix into a low-rank and a sparse matrix. Previously Huang et al. showed that the RPCA can extract vocal successfully [1] as the vocal components exhibit large variability (sparse matrix). Firstly, the music is represented in the frequency domain using the short-time Fourier transform (STFT). The RPCA was used to separate the input in the frequency domain into a sparse and low-rank matrices. This was followed by taking the inverse STFT of the sparse matrix as the vocal component of the input signal (Figure 2).

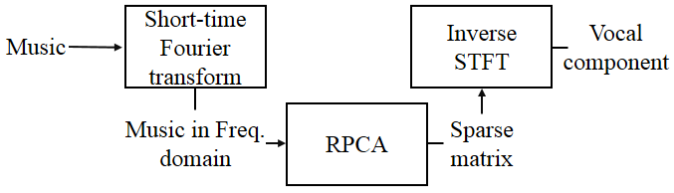

Figure 2: RPCA for vocal extraction.

4 Vocal Noise Reduction

The extracted vocal is not perfect and remnants of the instrumental components is often observed in the segments without vocals (i.e., labelled in Figure 4). Therefore our model reduces the effects of the instrumental that is found in the vocal components by using graph cut algorithm. The residues often causes errors in phone detection, therefore causes overall alignment errors.

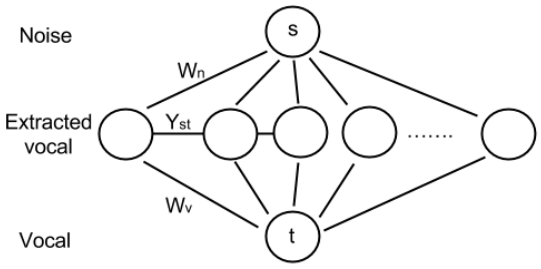

Figure 3: Model of noise reduction in the extracted vocal.

The system minimises the noise given by:

$$E(L) = \sum_{s \in S} D_s(L_s) + \sum_{(s,t) \in N} Y_{st}(L_s, L_t) \tag{1}$$

where S are all the samples in the extracted vocal and N are the connections between the neighbouring samples. Y_{st} is a constant throughout all the neighbours and D_s is defined as:

$$D_s(L_s) = \frac{1}{1 + exp[-k(L_s - L_o)]} \tag{2}$$

where k and L_o are parameters determined by the nature of vocal extraction through RPCA.

Generally, the residues appear in small patches, have a significantly smaller amplitude compared to the vocals and are in isolation compared to other residues (see Figure 4). The maximum flow will generate a binary mask that predicts whether a sample is the vocal or noise.

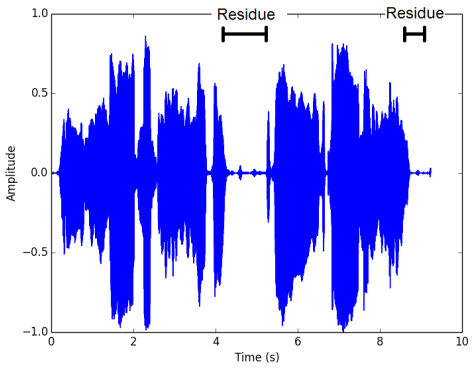

Figure 4: Vocal extraction of a segment for "Blowin' in the Wind" by Bob Dylan and the easily identified residues of the instrumental are indicated.

5 Vocal Aligner

In this section, we describe an HMM acoustic model for vocal recognition and timestamp prediction using forced alignment. The Prosodylab-aligner implementation was used to achieve this [6]. The CMU sphinx vocal aligner was also studied but yielded unsuccessful results (see appendix).

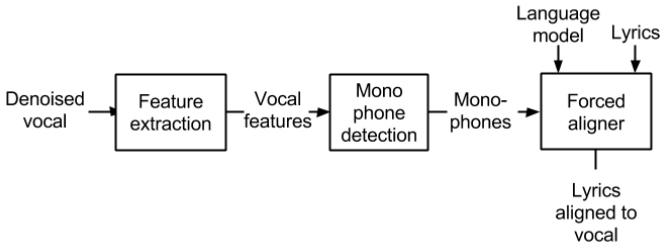

Figure 5: Model for vocal aligner.

The acoustic model is responsible for detecting monophones (also known as phonemes), which are the most basic unit of sound like "AA", "AO", from the denoised vocals. The denoised vocal is separated into ten millisecond samples and 39 Mel frequency cepstral coefficients were calculated from it. The coefficients were then placed into a model consisting of Gaussian mixtures to estimate the monophones.

The monophones are placed into the forced aligner. The forced aligner requires the song lyrics and a language model that contains a pronunciation dictionary. A pronunciation dictionary was built to map the words of the lyrics to their pronunciation (e.g., Hello = HH AH L OW). The new pronunciation dictionary is then merged with the standard US English pronunciation dictionary. The detected phones are then aligned to the phones corresponding to the lyrics (Figure 6).

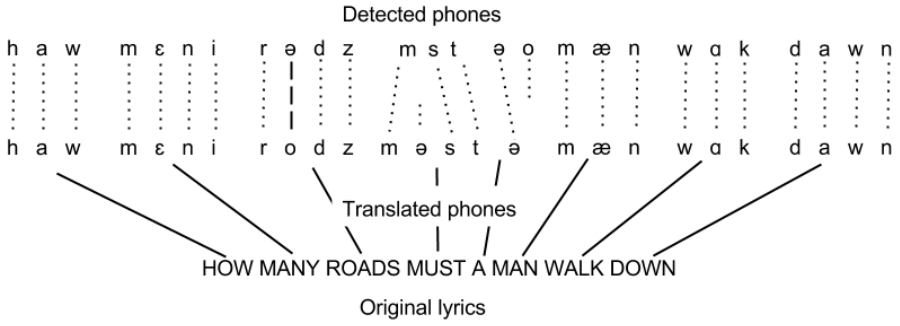

Figure 6: An example of forced alignment. The example showed instances of errors such as: ROAD has a substitution error, MUST has a deletion error and A has an insertion error.

6 Experiments

6.1 Evaluation

The system performance was evaluated using 10 English song segments. Each segment was sampled at 16khz and is 4 to 18 seconds long. The lyrics of the song segments were obtained online. We manually labelled the approximate timestamps for the start and finish of each word based on the original songs. The autoaligned timestamps are then compared with the manually aligned expected timestamps. We used two metrics:

- Percentage of overlap: An overlap is defined as the period of time in which the start and end time in both the autoaligned words and manually labelled words are intersecting. The percentage of overlap is the sum of the overlap of all words over the sum of the duration of all the expected words. This metric was used in [2].
- Percentage of correct words: A correct word occurs when the difference between the start time and end time in the autoaligned and manually aligned is less than 300 ms.

6.2 Results

As an example, a song segment from "Blowin' in the Wind" by Bob Dylan was used to illustrate each component of the model (Figure 7).

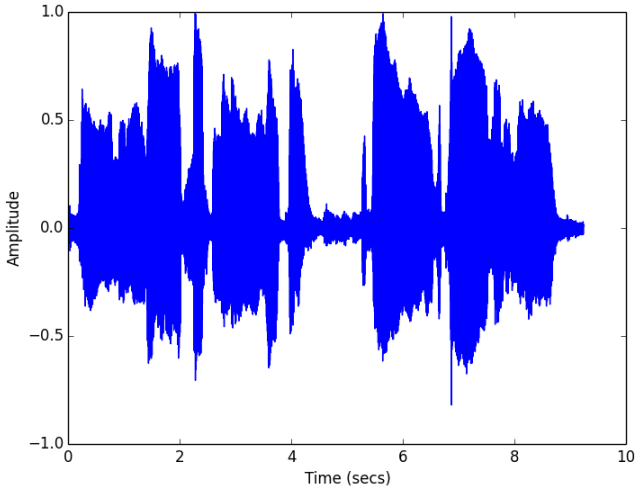

Figure 7: Original signal of the song segment for "Blowin' in the Wind" by Bob dylan.

The vocal extraction with RPCA is shown in Figure 8. The RPCA implementation used parameters $\mu_{fac} = 125$ and $\rho = 1.5$ where μ_{fac} governs the intiatiate conditions and ρ governs the update speed.

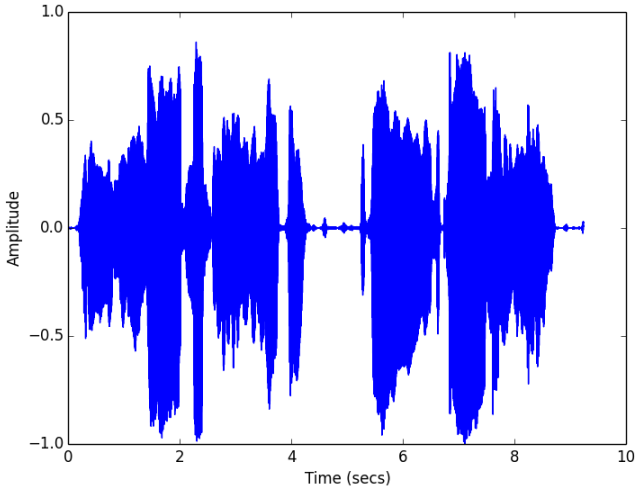

Figure 8: Results of vocal extraction with RPCA.

As shown in Figure 8, RPCA produced a reduction in the instrumental section (between 4.0 to 5.5 seconds and 9.0 to 9.5 seconds) however, residuals are still observed. Also, slight reduction of amplitude is observed throughout the course of the song. For vocal denoising the parameters were selected as follows: $Y_{st} = 20$, $L_o = 0.1$, $K = 5$ where Y_{st} determines the relationship between the neighbours, L_o and K determines the probability of given the initial value. The results of vocal denoising is shown in Figure 9.

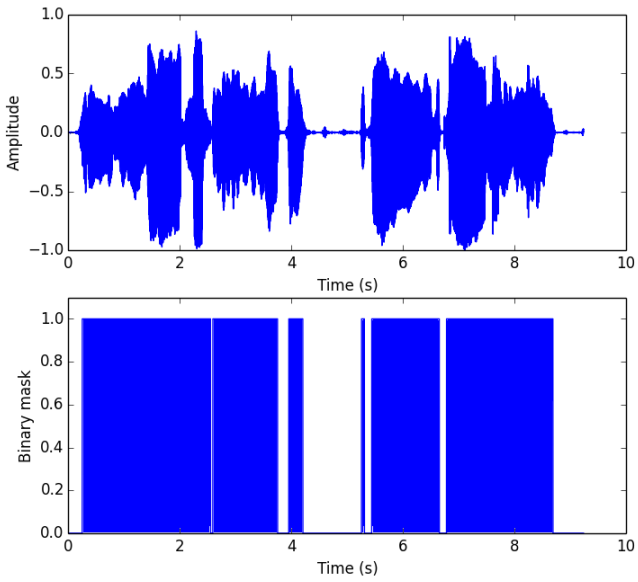

Figure 9: Results of the vocal denoising. The binary mask denote areas at which the model predicts the samples are vocal or noise.

As shown in Figure 9, the binary mask removed major components between 4.0 to 5.5 seconds and 9.0 to 9.5 seconds. The denoised vocal was placed into the aligner and the following results were obtained. The overall system performance for "Blowin' in the Wind" is illustrated in Figure 10.

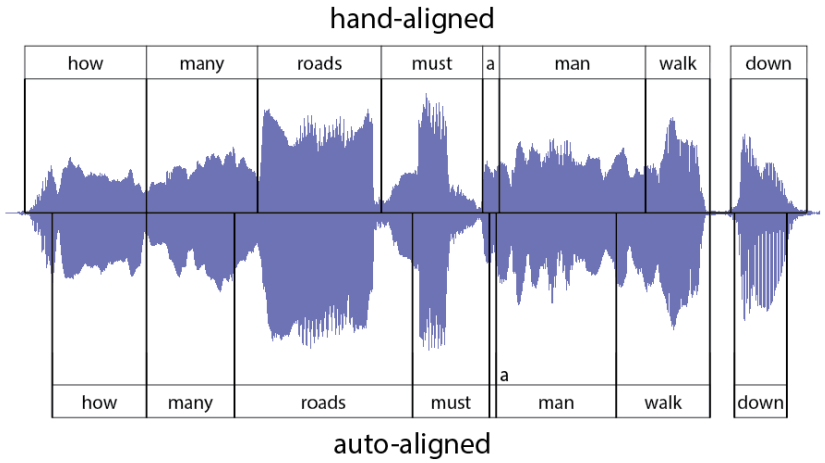

Figure 10: The original music of "Blowin' in the Wind", the lyrics above indicate the automatically aligned lyrics and the lyrics below indicate the manually aligned lyric.

Table 1: Results of the overall system performance

Segment ID	Overlap %	Correct %
1	70.7	71.4
2	12.0	7.0
3	39.5	50.0
4	45.4	50.0
5	45.4	57.1
6	62.2	73.3
7	29.5	33.3
8	26.1	35.0
9	50.0	50.0
10	26.9	26.9
	40.7 ± 17.8	45.4 ± 20.3

The results of all the song segments are shown in Table 1.

7 Discussion

As shown in Table 1, the average percentage of overlap and percentage of correct words identified are relatively low. However, the results vary dramatically between segments. For example, segment 1 produced 70.7% overlap and 71.4% correct words identified but segment 2 had 12.0% and 7.0% respectively (therefore the standard deviation is very high). Fujihara et al. [2] showed an average of 80% accuracy compared in alignment. However, as indicated by [3], aligning Japanese is easier than aligning English. The other related methods did not use the same metric to evaluated their results.

8 Conclusion and Future Directions

In this paper we described method for aligning lyrics in consumer available songs. The method consists of model that could be separated into 3 major components: vocal extraction, noise reduction and forced alignment. The vocal extraction component was completed using robust principle component analysis on the music in the frequency domain. Noise reduction was conducted to reduce the residue instrumental sounds in the extracted vocal. This was completed using graph cuts to estimate a binary mask for vocal and noise. Finally, phones were detected with an HMM based monophone model and were inputs for forced alignment to obtain timestamps for each word. The model was tested on multiple segment of songs and in some instances more than 70% overlap between the autoaligned lyrics and manual aligned lyrics were found.

Future developments emphasize two major points:

- Noise reduction is highly sensitive to the parameters: Methods to automatically find the optimal parameter can be obtained by training the parameters with songs and the corresponding songs with only vocals.
- Currently the segments are analysed independently and the repeating nature of the vocal melody is not used. In the future, the pitch of the melody at which each word is sung can be extracted [7]. This provides additional information for alignment (see Figure 11).

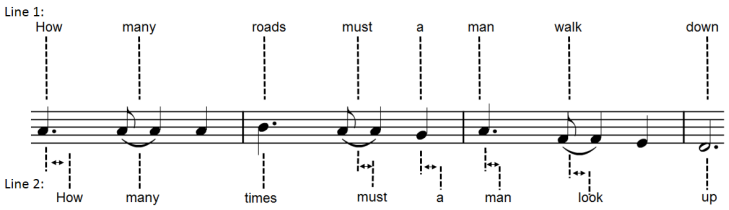

Figure 11: Use the vocal melody between separate lines to provide additional information for the alignment. The same logic could be applied to extracting the background chords [8] and beat estimation [9] for additional information.

9 Appendix

9.1 Channel cancellation

In order to build a "Karaoke ready" track, the instrumental must be extracted from the music. Typically, simple filtering of the music can yield reasonable results (built-in function in Audacity). In addition, channel cancellation is also a plausible method that is a simple subtraction between the two stereo tracks. This is due to the nature of studio recording, the instrumental components are independently recorded with two tracks (stereo) whereas the vocal components recorded with one track (mono) and is repeated in both the stereo tracks. Therefore, a subtraction cancels the vocal components out leaving the instrumental components behind. The modified instrumental can be further smoothed and refined with the denoising technique we proposed.

9.2 CMU Sphinx forced alignment

Another possible method for forced alignment used the CMU Sphinx architecture. We used the lyrics of 100 original songs to build a N−Gram language model, which contains the probabilities of the words and word combinations, in addition to a pronunciation dictionary that maps every word in our lyrics to their corresponding phonetic representation (e.g., Hello $\rightarrow$ HH AH L OW). The language model was then merged with a standard English language model we found in the CMU archive. We built a mixture of experts by tuning the mixture weights to optimize recognition accuracy. Our dictionary was also merged with the standard English language dictionary CMU to build a wholesome dictionary of over 135000 words.

The combined language model and pronunciation dictionary was then used to construct an acoustic model responsible for recognizing words in the songs, as well as being an engine in our force aligner to detect phones and to align them with the expected ones from the lyrics.

Once the model is ready, we passed the denoised RPCA vocals into it to evaluate the model and the performance of the aligner that is built around it. Just like the vocal aligner we described in section 5 of the paper, the acoustic model is used to detect monophones in the input audio, and the detected phones are then aligned to the phones corresponding to the lyrics. For the average percentage of overlap, we obtained a result of 15.2%, while only getting 20.5% of the words correct. As we can see, the performance of the Sphinx aligner is not satisfactory in comparison to the one described earlier.

We decided to examine the data set carefully to decipher what elements affect our results negatively. For the segments that were perfectly recognized, the audio quality tends to be higher than average, and more importantly, the duration of each word is closer to the length of the same word spoken in conversation. In fact, during singing, some words are stretched for so long that our speech recognition system failed to adjust itself to properly account for them. For example in a song by the artist Muse, Our hopes and expectations, black holes and revelations, the line was stretched to full 15 seconds in the song, even though the same 7 words are much shorter in conversational speed.

The results of all song segments using the Sphinx aligner is shown in Table 2 and the results are found to be significantly weaker compared to the results in Table 1.

Table 2: Results of the Sphinx aligner performance

Segment ID	Overlap %	Correct %
1	13.3	21.4
2	10.9	14.3
3	14.2	14.3
4	7.65	18.8
5	11.1	28.6
6	19.6	20.0
7	31.9	38.9
8	22.2	30.0
9	18.6	11.5
10	2.87	3.85

References

[1] Huang, P. S., Chen, S. D., Smaragdis, P., & Hasegawa-Johnson, M. (2012, March). *Singing-voice separation from monaural recordings using robust principal component analysis.* In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 57-60). IEEE.

[2] Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T., & Okuno, H. G. (2006, December). *Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals. In Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on (pp. 257-264). IEEE.*

[3] Mesaros, A., & Virtanen, T. (2008, September). Automatic alignment of music audio and lyrics. In Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08).

[4] Mauch, M., Fujihara, H., & Goto, M. (2010). Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations. In Proceedings of the 7th Sound and Music Computing Conference (SMC) (pp. 9-16).

[5] Cands, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis?. Journal of the ACM (JACM), 58(3), 11.

[6] Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics, 39(3), 192-193.

[7] Chien, Y. R., Wang, H. M., & Jeng, S. K. *Vocal melody extraction based on an acoustic-phonetic model of pitch likelihood.*

[8] Jensen, K., & Andersen, T. H. (2003, October). Beat estimation on the beat. In Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on. (pp. 87-90). IEEE.

[9] Khadkevich, M., & Omologo, M. (2011, May). Time-frequency reassigned features for automatic chord recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 181-184). IEEE.