LYRICS TO AUDIO ALIGNMENT FOR KARAOKE IN POP MUSIC

Georgi Dzhambazov[1,2]
Marius Miron[1] **Xavier Serra**[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Voice Magix, Barcelona, Spain
georgi.dzhambazov@upf.edu, info@voicemagix.com

ABSTRACT

In this paper we describe an algorithm for automatic lyrics-to-audio alignment. It has as a goal the automatic detection of word boundaries in multi-instrumental English pop songs. We rely on a phonetic recognizer based on hidden Markov models: a widely-used method for tracking phonemes in speech processing problems. Tracking lyrics in music audio is harder than tracking text in speech because, unlike speech, the singing voice is mixed with multiple instruments. To address this obstacle we apply a convolution neural networks-based method for singing voice separation. We present a prototype of a practical application based on the alignment method - the highliting of lyrics in a karaoke-like fashion.

1. APPROACH OVERVIEW

We adopt the classical approach of alignment of speech and text - phonetic recognizer that has been the predominant choice of lyrics-to-audio alignment research [3]. The lyrics of a song is expanded to a network of phonemes using the CMU pronunciation dictionary [1]. The phoneme network is a Hidden Markov Model (HMM), wherein each phoneme is modeled by a monophone model, trained on clean singing voice. The sequence of feature vectors, extracted from the audio, is aligned to the phonemes by finding the most likely path in the phoneme network by means of a forced alignment Viterbi decoding [2]. Singing voice detection (SVD), followed by segregation of the spectral content of singing voice, are performed as preprocessing steps.

1.1 Features

The features for training the phoneme models are the first 13 Mel Frequency Cepstral Coefficients (MFCCs), where the 0-th coefficient represents the signal's energy. Their deltas and deltas of deltas are appended to form a 39-dimensional feature vector. The MFCC follow the default htk type of high-frequency-preemphasis, mel-scale equation, DCT-type [6]. The htk feature parameter used is *MFCC_0_D_A_Z*.

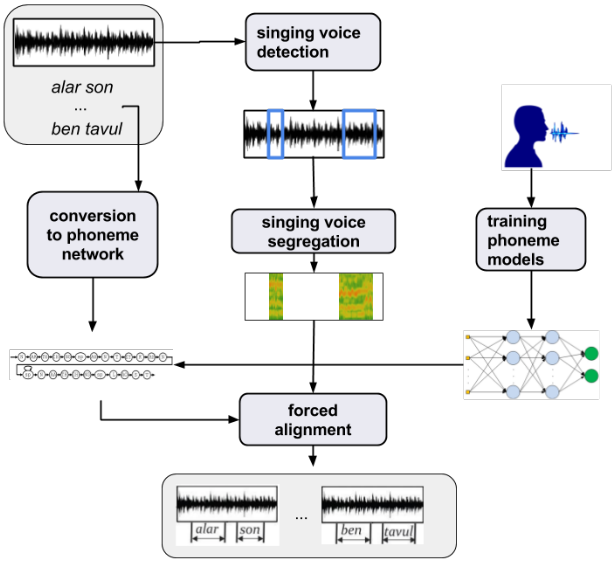
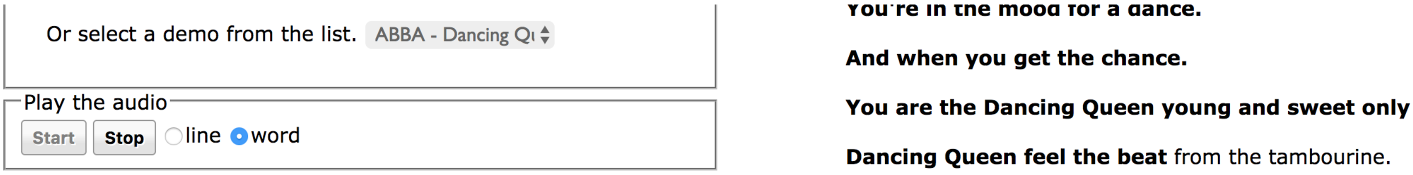


Figure 1. Overview of the steps for extracting the singing voice from the multi-instrumental mix and its alignment to the lyrics

1.2 Phoneme network

The phoneme network is a sequence of the phonemes. At the end of each line of the lyrics a silent pause model is appended to accommodate possible short non-vocal breaks. The HMM transition probabilities force a transition only to the following phoneme from this sequence. To represent the *observation probability* $P(y_k|x_k)$ of observing the MFCC feature vector y_k, generated by a phoneme x at time instant k, we utilize the softmax probabililty of the multi-layer perceptron feed-forward network, trained by Kruspe on material from clean singing [4].

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[2] We implemented the Viterbi decoding https://github.com/georgid/AlignmentDuration



You're in the mood for a dance.

And when you get the chance.

You are the Dancing Queen young and sweet only

Dancing Queen feel the beat from the tambourine.

Figure 2. Web interface for highliting words and lines with simultaneous audio playback

2. SINGING VOICE SEGREGATION

It is difficult to successfully track the phonemes in multi-instrumental music signals by using the models, trained solely on *a cappella* singing. Therefore we perform as a preprocessing step a segregation of the spectrum with origin in the singing voice and extract the MFCCs from it, as if it were a cappella singing.

A recent source separation method based on convolutional neural networks separates the signal into four parts - vocal, bass, drums and other instruments [2]. The model is trained on the dataset *DSD100* [3] . Analysing the separated vocal part, we realized it has a significant leak from background instruments, especially on regions with no singing voice present. For this reason, a SVD method is needed.

3. SINGING VOICE DETECTION

The SVD model is trained on the vocal part extracted with the source separation method for a dataset with annotations of segments with singing voice. We employed a subset of the *medleyDB* that contains singing voice (50 recordings, ˜3 hours) [1]. Then two separate GMMs with 20 components are fit on the MFCCs and their deltas - one that returns the probability of a frame k being vocal $P_{vocal}(k)$ and one being non-vocal $P_{non-vocal}(k)$. Then we compute a soft weight $V_{soft}(k)$:

$$V_{soft}(k) = \frac{P_{vocal}(k)}{P_{vocal}(k) + P_{non-vocal}(k)}$$

This strategy has been adopted from [5]. As non-vocal are considered segments that have a sequence of frames with V_{soft} above a threshold $\theta = 0.55$. Within the detected non-vocal segments the observation probability of the silent model is set to 1 and 0 for all the rest of the phonemes.

4. EXPERIMENTS

We performed alignment on the 30 pop songs from the dataset, used in MIREX 2017 task on Automatic Lyrics-to-audio Alignment [4] . We arrived at a median alignment error of 0.37 s for the Hansen's easier subdataset of 9 songs and 5.12 on the harder Mauch's dataset of 21 songs. Detailed results are at `http://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results`. A prototype of highlighting lyrics on the lyrics word-level and line-level is shown in Figure 2 [5] .

Acknowledgements We thank to Jeff Scott from *Gracenote* for the hints on the singing voice detection step. This work is supportedby the Spanish Ministry of Economy and Competitiveness, through the "María de Maeztu" Programme for Centres/Units of Excellence in R&D" (MDM-2015-0502).

5. REFERENCES

[1] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, volume 14, pages 155–160, 2014.

[2] 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.

[3] Hiromasa Fujihara and Masataka Goto. Lyrics-to-audio alignment and its application. *Dagstuhl Follow-Ups*, 3, 2012.

[4] Anna M Kruspe. Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing. In *Proceedings of 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, New York, NY, USA, 2016.

[5] Sang Won Lee and Jeffrey Scott. Word level lyrics-audio synchronization using separated vocals. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 646–650. IEEE, 2017.

[6] Steve J Young. *The HTK hidden Markov model toolkit: Design and philosophy*. 1993.

[3] `https://sisec.inria.fr/sisec-2016/2016-professionally-produced-music-recordings/`

[4] `www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment`

[5] A live demo can be seen at `https://georgid.github.io//lyrics-align/lyric.html`