LYRICS TO AUDIO ALIGNMENT IN POLYPHONIC AUDIO

Georgi Dzhambazov
Marius Miron **Xavier Serra**
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{georgi.dzhambazov,marius.miron,xavier.serra}@upf.edu

ABSTRACT

In this paper we describe the two algorithms we submitted for the MIREX 2017 task of Automatic Lyrics-to-Audio Alignment. The task has as a goal the automatic detection of word boundaries in multi-instrumental English pop music.

We rely on a phonetic recognizer based on hidden Markov models (HMM): 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 propose the application of two separate methods for segregating the singing voice from the multi-instrumental mix. One of them is based on the detection of vocal harmonic partials, whereas the other extracts the vocal content by means of source separation.

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 pronounciation 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.

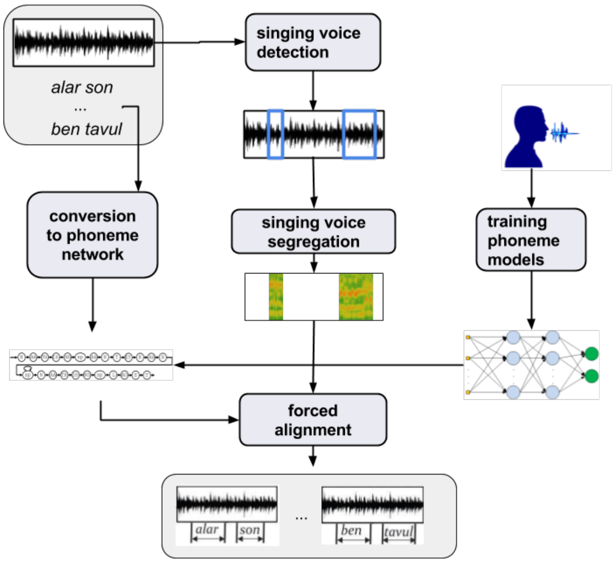

Figure 1. Overview of the steps for extracting the singing voice from the multi-instrumental mix and its alignment to the lyrics

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 feture vector. The MFCC follow the defalut htk type of high-frequency-preemphasis, mel-scale equation, DCT-type [7]. The htk feature parameter used is *MFCC_0_D_A_Z.*

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 multilayer perceptron feed-forward network, trained by Kruspe on material from clean singing [4].

For the a capella dataset the approach is applied as it is,

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

whereas for the multii-nstrumental ones, it is augmented with singing voice segregation strategies.

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. For the recognizer, the accompanying instruments essentially deteriorate the intelligibility of the phonemes. 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. We present two separate vocal segregation strategies.

2.1 Harmonic-partials-based

To filter the spectral peaks corresponding to the harmonic partials of the singing voice we utilize the harmonic model of [**?**] . The spectral peaks are computed at the expected location of harmonic partials at multiples of the fundamental frequency f_0. To extract the f_0 of the singing voice, we perform melody contour extraction of the predominant source in multi-instumental recordings [6]. We estimate a relatively large number of harmonics (30), in order to preserve the phonetic timbre as much as possible, and cut peaks at -65 db. The vocal part is obtained by resynthesis, whereby regions with no predominant voice detected remain silent.

2.2 Source-separation based

A recent source separation methods based on convolutional neural networks separates the signal into four parts - vocal, bass, drums and the rest [2]. The model is trained on the dataset *DSD100* [3] . Analysing the separated vocal part, we relized 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

Some song sections (e.g. intro, instrumental solos, interline breaks) contain no singing voice and an alignment should ideally be able to classify them as non-vocal segments. The melody contour detection algorithm [6] performs decent detection of vocal segments prior to melody contour detection, so we did not run SVD for the approach from section 2.1.

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.

Acknowledgements This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583) and partly by the Spanish Ministry of Economy and Competitiveness, through the "María de Maeztu" Programme for Centres/Units of Excellence in R&D" (MDM-2015-0502).

4. 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] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[7] 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/