2016-6

Segmentation of Folk Songs with a Probabilistic Model

Ciril Bohak
University of Ljubljana, ciril.bohak@fri.uni-lj.si

Matija Marolt
University of Ljubljana, matija.marolt@fri.uni-lj.si

SEGMENTATION OF FOLK SONGS WITH A PROBABILISTIC MODEL

Ciril Bohak, Matija Marolt
University of Ljubljana,
Faculty of Computer and Information Science
{ciril.bohak,matija.marolt}@fri.uni-lj.si

1. INTRODUCTION

Structure is an important aspect of music. Musical structure can be recognized in different musical modalities such as rhythm, melody, harmony or lyrics and plays a crucial role in our appreciation of music.

In recent years many researchers have addressed the problem of music segmentation, mainly for popular and classical music. Some of the more recent approaches are Mauch et al. (2009), Foote (2000), Serrà et al. (2012) and McFee & Ellis (2014). Last three are included in the music structure analysis framework MSAF Nieto & Bello (2015). None of the mentioned approaches however, addresses the specifics of folk music.

While commercial music is performed by professional performers and recorded with professional equipment in suitable recording conditions, this is usually not true for folk music field recordings, which are recorded in everyday environments and contain music performed by amateur performers. Thus, recordings may contain high levels of background noise, equipment induced noise (e.g. hum) and reverb, as well as performer mistakes such as inaccurate pitches, false starts, forgotten melody/lyrics or pitch drift throughout the performance.

One of the most recent approaches which addressed folk music specifics was presented by Müller et al. (2013). The approach was designed for solo singing and was evaluated on a collection of Dutch folk music by Müller et al. (2010).

In our paper, we present a novel folk music segmentation method, which also addresses folk music specifics and is designed to work well with a variety of ensemble types (solo, choir, instrumental and mixtures).

2. METHOD

The proposed method processes the input audio recording in several steps and returns a list of segment boundaries. The method assumes that songs consist of similar repetitions of a single part (stanza).

2.1 Feature extraction

The method averages the input audio to a single channel and normalizes it. To find repetitions in a melodic/harmonic space, we use harmonic chroma features to represent the contents of recordings, more specifically we use 24-dimensional HPCP features presented in Gómez (2006).

2.2 Finding similarity

Our aim is to find segment boundaries that separate repetitions of a segment in a song. We do not know how long individual repetitions are, how many repetitions there are in a song nor how similar they are. To bootstrap the segment finding process, we randomly select a number of 10 second long parts in a song and calculate their distances to the entire song. We use dynamic time warping (DTW) to calculate the distances, as it can tolerate tempo variations well, the technique was already presented by Müller et al. (2009).

Besides rhythm and tempo variations, we also have to take into the account pitch drifting, which occurs when intonation of performers changes upwards or downwards over the course of a song. Ignoring pitch drift would result in inaccurate distance curves and thus poor segmentation. We thus calculate several distance curves for each selected segment, where we shift the intonation of the selected part before calculating the distance. As drifting occurs gradually, we obtain the final distance curve by minimizing distances across all curves, and at the same time restricting the number of intonation changes over the course of a song. An example of an obtained pitch drift curve is presented in Figure 1 (a).

The process results in a series of distance curves, describing the distance of each randomly selected part to the entire song, where tempo and intonation variances are taken into consideration. An example is given in Figure 1 (b). Local minima in these curves represent repetitions of a chosen part in the song. We then remove the self-similar parts of the distance curves, and the resulting curves are shown in Figure 1 (c).

2.3 Alignment and length

The set of distance curves (Figure 1 (c)) is not time aligned, since the parts used for their calculations were randomly chosen. To perform alignment, we select a reference distance curve, which is the one that has the highest correlation (is the most similar) to all other curves, thus we may say that it is very representative of the song. Alignment is performed by time-shifting each curve according to its closest local minimum to the part the reference curve was calculated for. From aligned curves we calculate the average distance curve shown in Figure 1 (e).

We also calculate the approximate segment length from

the average distance curve with auto-correlation, as shown in Figure 1 (f).

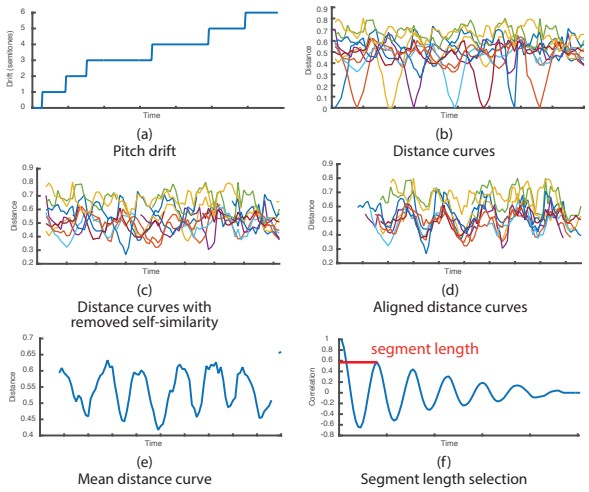

Figure 1: Segmentation steps.

2.4 Segmentation

Segmentation is performed with a probabilistic framework similar to hidden Markov models. The model has a state for every possible segment beginning (placed at each second of a song). Segmentation is calculated as an optimal path through the model, defined by state and transition probabilities.

State probabilities are proportional to the likelihood of placing a segment boundary at a certain time. We assume that this likelihood is larger if the boundary is preceded by a region of low-amplitude: for singing, this often corresponds to breathing pauses, while for instrumental music this often corresponds to end of phrases. The longer this region is, the higher is the probability of a segment boundary.

Transition probabilities represent the probability of placing a segment boundary at certain time i if the previous was located at some other time j. We consider three restrictions in calculation of transition probabilities: (a) two segments beginning at times i and j should be similar; (b) the segments should be separated by approximately the estimated segment length and (c) only forward transitions are allowed.

To find an optimal path through states of this model, we use Viterbi algorithm, whereby we allow the starting state to occur within first 6 seconds of a song and enforce the ending in the last state. The resulting sequence of states represents the set of found segment boundaries, as the states are directly mapped to time.

The detailed description of the method and its individual steps can be found in Bohak & Marolt (2016).

3. EVALUATION AND RESULTS

We have evaluated the methods on a collection of folk music from the Ethnomuse presented in Strle & Marolt (2007) archive and part of the Dutch folk music collection presented in Müller et al. (2010). The EthnoMuse collection consists of different ensemble types: solo singing, two- and three-voice ensembles, choirs, instrumental and mixed singing and instrumental ensembles. We chose 206 songs of different types and recording quality for our collection with a total duration of 534 minutes. The collection was manually annotated, placing segment boundaries with $\pm$ 100 ms precision.

We calculated precision, recall and F1 measure values per song for each ensemble type and for the entire collection. The estimated segment boundary was considered as correct (true positive) if it was located within a $\pm$ 3 second window around an annotated boundary (the same window size as in MIREX evaluations).

The proposed approach significantly outperforms compared methods for non-instrumental music, while for instrumental it is comparable to the best performer. The overall results are presented in Table 1.

Results are also comparable with current state-of-the-art segmentation method for folk music presented in Müller et al. (2013), with an F1 measure of 0.87 on a collection of solo Dutch folk songs - our method achieves an F1 measure of 0.85 on the same collection.

Table 1: Evaluation results.

Method	P	R	F1
Mauch et al. (2009)	0.74	0.40	0.4
Foote (2000)	0.39	**0.81**	0.52
McFee & Ellis (2014)	0.41	0.59	0.48
Serrà et al. (2012)	0.41	0.56	0.47
Proposed method	**0.78**	0.80	**0.76**

4. CONCLUSION

We presented a novel approach to segmentation of folk music. The method takes into account folk music specifics and significantly outperforms current state-of-the-art segmentation methods for segmenting commercial music and is on pair with a state-of-the-art method for solo singing segmentation.

As part of our future work we can envision several improvements of the method, especially for segmentation of instrumental music. We also plan to further specialize the method for better performance with individual ensemble types, by first automatically detecting ensemble type and then choosing an appropriate set of method parameters. We also aim to extend the method for hierarchical musical structure discovery.

5. ACKNOWLEDGMENT

This work would not have been done without field recordings provided by the Institute of Ethnomusicology at Research Centre of Slovenian Academy of Sciences and Arts.

6. REFERENCES

Bohak, C. & Marolt, M. (2016). Probabilistic segmentation of folk music recordings. *Mathematical Problems in Engineering*, *2016, Article ID 8297987*.

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*. IEEE.

Gómez, E. (2006). *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra.

Mauch, M., Noland, K. C., & Dixon, S. (2009). Using Musical Structure to Enhance Automatic Chord Transcription. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, (pp. 231–236).

McFee, B. & Ellis, D. P. W. (2014). Analyzing Song Structure With Spectral Clustering. In *Proceedings of 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, (pp. 405–410).

Müller, M., Grosche, P., & Wiering, F. (2009). Robust Segmentation and Annotation of Folk Song Recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 735–740)., Kobe, Japan.

Müller, M., Grosche, P., & Wiering, F. (2010). Automated analysis of performance variations in folk song recordings. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, (pp. 247–256)., Philadelphia, Pennsylvania, USA.

Müller, M., Jiang, N., & Grosche, P. (2013). A Robust Fitness Measure for Capturing Repetitions in Music Recordings With Applications to Audio Thumbnailing. *IEEE Transactions on Audio, Speech & Language Processing*, *21*(3), 531–543.

Nieto, O. & Bello, J. P. (2015). Msaf: Music structure analytis framework. In *Proceedings of 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*.

Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2012). Unsupervised Detection of Music Boundaries by Time Series Structure Features. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, (pp. 1613–1619). AAAI Press.

Strle, G. & Marolt, M. (2007). Conceptualizing the ethnomuse: Application of cidoc crm and frbr. In *Proceedings of CIDOC2007*.