A Computational Approach to the Modeling and Employment of Cognitive Units of Folk Song Melodies using Audio Recordings

Peter van Kranenburg[1], George Tzanetakis[2]

[1]Meertens Institute, Netherlands; [2]University of Victoria, BC, Canada

ABSTRACT

We present a method to classify audio recordings of folk songs into tune families. For this, we segment both the query recording and the recordings in the collection. The segments can be used to relate recordings to each other by evaluating the recurrence of similar melodic patterns. We compare a segmentation that results in what can be considered cognitive units to a segmentation into segments of fixed length. It appears that the use of 'cognitive' segments results in higher classification accuracy.

1. BACKGROUND

Large collections of monophonic folk song recordings are interesting from a music cognition perspective since they represent musical performances of common people. Most people share a 'common core of musical knowledge' (Peretz 2006: section 2). Since recorded folk songs were sung from memory, knowledge about the process of remembering and reproducing melodies can be used to employ these recordings in the context of folk song research, music information retrieval or music cognition studies.

This study combines ideas and approaches from etnomusicology, music cognition and computer science. One of the research questions of etnomusicology is how melodies in an oral tradition relate to each other (Nettl 2005, chapter 9; Van Kranenburg et al. 2009a). Samuel Bayard (1950) developed the concept tune family to denote a group of melodies that share a common origin, which, in the most simple case, is a single tune. The idea that melodies from the same tune family are related by shared melodic motifs has a long history in folk song research. Nettl (2005: p. 117f) discusses the relative independence of shorter units of musical thought. These might 'wander' from melody to melody and from country to country (Tappert 1890). Marcello Keller (1988) explains the relations between Trentino folk music compositions by means of a repertoire of 'segments' that is used in the act of composing. To cope with specific relations between melodies he encountered in Irish folk music, James Cowdery (1984) extended Bronson's concept of tune family by including melodies that are related by sharing melodic material from the same 'pool of motives'. Finally, one of the conclusions from a previous study on the same corpus of songs that we use in this paper is that recurring motifs are more important than contour and rhythm for recognizing a song (Volk et al. 2008).

Understanding the way melodies change in oral transmission involves understanding of encoding of melodies in, and reproduction of melodies from human memory. Cognitive studies indicate that melodies are not reproduced note by note, but as a sequence of higher level musical units, or chunks (Miller 1956). Much research has been done to model these chunks (e.g., Lerdahl and Jackendoff 1983, Narmour 1992, Cambouropoulos 1998). All mentioned approaches use a symbolic transcription of the melody in the form of a musical score and try to group notes into musically meaningful segments in a bottom-up or top-down fashion. In the current study we take as our starting point the audio recording of a song performance rather than its transcription. Thus, we can use aspects of the performance that are lost during transcription into musical notation.

The computational methods we use enable a data-rich, empirical approach to the study of segmentation and similarity of melodies (Clarke and Cook 2004). The current study has been performend in the context of a music information retrieval project that has the aim to design a search engine for folk song melodies (Wiering et al. 2009).

The two main questions in this paper are whether recurrence of audio segments can be exploited to classify a folk song recording into the correct tune family, and whether the use of cognitively and musically meaningful audio segments yields better classification performance than the use of fixed-length audio segments.

Our classification method consists of four stages: pitch extraction, segmentation, selection of representative segments for each tune family, and classification using these representative segments. These four stages are described in the next sections. The main idea for segmentation we employ in this paper is to take breathing and pauses during singing as segment boundaries, which can be conceived as chunk boundaries. Thus, segmentation results in musically and cognitively meaningful units. We do not assume a one-to-one relation between these breathing and pause boundaries at the one hand and chunk boundaries at the other hand, but we do assume a strong relationship.

Contribution: We widen the scope of automatic folk song classification by using audio recordings rather than symbolic data. To our knowledge, this is the first study in which aspects of folk song performance (breathing and pauses) are used to mark segment boundaries, and this is the first computational study that models a tune family by its most representative recurring segments.

2. DATA

We use a corpus of annotated songs from the Dutch collection 'Onder de groene linde', hosted by the Meertens Institute in Amsterdam (Wiering et al. 2009). A subset of 360 symbolically encoded strophes from 347 songs in 26 tune families was carefully selected by documentalists to be representative for the whole collection. This subset consists of tune families with considerable variation between the individual song instances. We use the 305 songs that are available as audio recordings.

3. PITCH EXTRACTION

Pitch extraction is necessary. We use the YIN algorithm (Cheveigne and Kawahara 2002), with time frames of 1024 samples and a YIN-threshold of 0.7, along with a newly developed post-processing filter. Our post-processing filter uses the dependencies between subsequent time frames to correct remaining errors. For each time frame, the filter replaces the detected pitch with the median of that pitch and the 5 preceding and 5 following pitches, which smoothes the contour.

A manual examination of all detected pitch curves reveals some main causes for bad pitch extraction: tape recorder hum, accompaniment, polyphonic singing, singing in octaves by male and female voices, and heavy noise in very old recordings. It seems that improvements of the pitch detection are achievable.

4. SEGMENTATION

There are time frames for which the YIN algorithm cannot detect a pitch. We assume that regions with a lot of these 'pitch-less' frames correspond to pauses in singing or to breathing. In addition to failing to detect a pitch, another indication of pause is a low energy of the signal. Our main idea for segmentation is to use these pitch-less regions as segment boundaries. This results in melodic segments in which a continuous flow of melody is present.

Since a short sequence of pitch-less time frames could also indicate a consonant like 'h' or a 'z', we set a lower limit to the length of the pitch-less regions to be considered as segment boundaries. After a small test on some representative examples, it appears that a good value for the minimal number of adjacent pitch-less time frames is 10, and that the median of the root-mean-square values of the time frames in the candidate boundary region should be smaller than 0.012.

As stated in the introduction, we assume that segment boundaries correspond to chunk boundaries. Therefore, relatively long segments are likely to be caused by under- segmentation. For that reason, and to decrease computation time, segments longer than 360 time frames (8.4 s) are removed from the data set. This leaves a data set with 5254 segments from 260 songs in 26 tune families. The threshold of 360 is somewhat arbitrary, but the exact value is not very important. It is unlikely that we throw away too many valid segments and the remaining set contains enough useful segments for the classification experiment.

The segmented recordings are available at: http://give-lab.cs.uu.nl/music/icmpc2010/segments

5. SIMILARITY MEASURE

To measure the similarity of two segments we use a variant of the Smith-Waterman local alignment algorithm (Smith and Waterman 1981). This algorithm finds the longest approximate common subsequence of two sequences of symbols along with an alignment of the matching parts and a score indicating the quality of the alignment. This score is the sum of the alignment scores of the individual symbols. If we consider two sequences **x**: $x_1,...,x_i,...,x_n$, and **y**: $y_1,...,y_j,...,y_m$, then symbol x_i can either be aligned with a symbol from sequence **y** or with a gap. Both operations have a score, the substitution score and the gap score. The gap score is mostly expressed as penalty, i.e. a negative score. The local alignment with the highest score is found by filling a matrix D recursively according to:

$$D(i,j) = \max\begin{cases} D(i-1,j-1) + S(x_i,y_j) \\ D(i-1,j) - \gamma \\ D(i,j-1) - \gamma \\ 0 \end{cases},$$

where $S(x_i,y_j)$ is the substitution scoring function, γ is the gap penalty, $D(i,0) = 0$ for $0 < i \leqslant n$, and $D(0, j) = 0$ for $0 < j \leqslant m$. $D(i, j)$ contains the score of the optimal local alignment up to x_i and y_j. The optimal local alignment can be found by starting at the cell with the highest value, which is the score of the alignment, and tracing back to the first cell with value zero. The standard dynamic programming algorithm has both time and space complexity $O(nm)$.

An audio segment is represented as a sequence of pitches for the consecutive time frames. The pitches are represented in continuous midi encoding, in which the middle c is represented by value 60.0, c# by 61.0, d by 62.0, and so on. By allowing fractional pitches we have a one-to-one correspondence to the frequencies, and a linear scale in the pitch domain.

The substitution scoring function, which returns values in the interval [−1,1], is defined as:

$$S(x_i,y_j) = \begin{cases} 1 - \dfrac{\text{interval}(x_i,y_j)}{7.0} & \text{if interval}(x_i,y_j) \leq 7.0 \\ -1 & \text{otherwise} \end{cases},$$

where $\text{interval}(x_i,y_j)=|p(x_i)-p(y_j)|\text{mod}12$, with $p(x_i)$ the pitch of symbol x_i. A perfect fifth has value 7 in midi-encoding. Thus all intervals up to a perfect fifth get a positive substitution score and all larger intervals are considered a bad match. This substitution score function was successful in a previous experiment on symbolic data (Van Kranenburg et al. 2009b). We use an extension of the algorithm proposed by Gotoh (1982), which employs an affine gap penalty function without loss of efficiency. In this approach, the extension of a gap gets a lower penalty than its opening. This prevents gaps from being scattered all over the alignment. We use 0.8 as gap opening penalty and 0.2 as gap extension penalty.

Since the score of an alignment depends on the length of the alignment, we normalize by dividing the alignment score by the score of the query segment with itself. Thus, an exact match that is embedded in a longer segment results in the maximal score (which is 1.0). Alignment with a short segment that has an exact match embedded in the query segment, results in a lower score. This makes our approach robust against under-segmentation as well as over-segmentation. As long as we have enough correctly detected segments, we will find related segments that are embedded in longer segments, but we will not find segments that are considerably shorter and that possibly match with many unrelated segments, so called hubs.

Since the songs are sung at various pitch heights, the alignment needs to be transposition-invariant. The tentative solution we use for this is to add a constant to the pitches of one of the segments such that the means of the pitches are the same for both segments.

The normalized scores are converted to distances by taking one minus the normalized score. This results in distances within the interval [0, 1]. Figure 1 shows an example of an aligment.

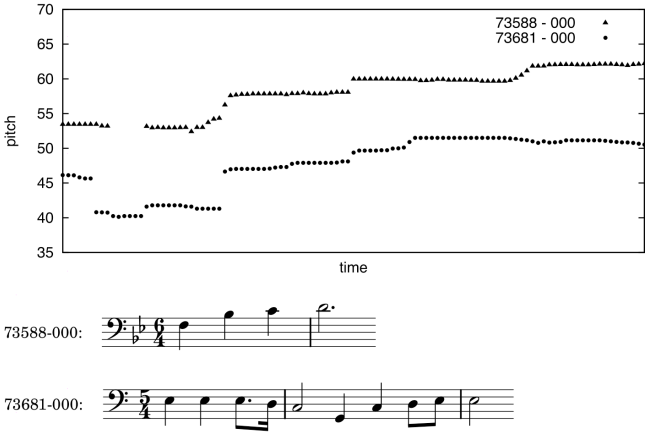

Figure 1: Example alignment of two segments (identified by: song id - segment). The alignment of the matching parts of the pitch curves as well as the symbolical transcriptions of the audio segments are shown at original pitch. 73588-000 matches with the second half of 73681-000. Apparently, gaps are needed at the beginning of 73588-000.

6. SELECTING REPRESENTATIVE SEGMENTS

As discussed in section 1, shared melodic patterns are important to recognize relations between tunes. Therefore, it seems a good approach to search for similar melodic segments among the songs that belong to the same tune family. We use an automatic selection procedure. For each tune family, we select the two segments that have the largest number of similar segments within the tune family, but that are not similar to each other. The selection procedure is as follows. For each segment all other segments in the dataset are ordered by distance according to the local alignment score. For a particular tune family, the segment that has the largest number of segments from the same tune family in the top 100 of the ranking list is selected as the first representative segment. To find the second representative segment the same criterion is applied with the additional constraint that the distance to or from the first selected segment is greater than 0.35. The histogram of all distances reaches its peak around 0.35. Therefore, this seems a safe value not to get a similar second segment. Thus, we find two dissimilar representative segments for each tune family. The threshold of 100 was established by inspecting the ranking lists manually. There is no segment for which the 100 nearest neighbors are all from the same tune family.

For some selected segments, the most common tune family among the 100 nearest neighbors is another tune family. These are removed from the set of representative segments. For six tune families no representative segment could be found at all. The numbers of songs in these families are 9, 7, 5, 4, 4, and 3. The small size of most of these families seems the cause for the failure to find representative segments. After removing these tune families, 228 recordings from 20 tune families remain.

Furthermore, there are nine tune families for which only one representative segment could be found.

7. CLASSIFICATION EXPERIMENT

In a classification experiment, we use the selected representative segments to find the tune family of a query recording. The aim of the experiment is both to evaluate whether the method presented in this paper is able to recognize a song at all, and to show the improvement of using 'cognitive' segmentation over fixed-length segmentation, in which the recordings are split into segments of 4.3 seconds, the average length of the 'cognitive' segments.

The procedure is as follows. We take the distances from all selected representative segments to all segments of the query song. After sorting, the tune family that is most common among the first n segments (the n nearest neighbors) is the tune family that is assigned to the query recording. It appears that 3 is a good value for n.

We cannot assume that the distribution of the distances to a particular segment is the same for each segment. Especially the variation in the minimal distance is problematic. To cope with this problem, for each representative segment, the distance from the first nearest neighbor to the representative segment is subtracted from all distances. The result of this linear shift is that all segments that are close to any of the representative segments are at the top of the sorted list.

When using segments of fixed-length, 95 of the 224 (42.4%) recordings are classified into the correct tune family. This result is positively biased because the songs that contain the selected representative segments are among the classified songs. If we disregard these songs, 62 out of 191 songs (32.5%) are correctly classified. Using the same 228 recordings, with the 'cognitive' segmentation, the respective numbers of correctly classified songs are 121 out of 228 (53.1%) and 92 out of 199 (46.2%), which is considerably better. If we take into account that there are 20 tune families, these are a quite good success rates.

Most, but not all, tune families show an improvement in the case of 'cognitive' segments.

8. DISCUSSION AND FUTURE WORK

The results clearly show that the recurrence of specific melodic patterns can be exploited to identify folk songs (i.e., to find the tune family to which they belong). We also conclude that 'cognitive' segments are more useful than fixed-length segments. This indicates a limitation of the n-gram approach that is widely used for similarity assessment or indexing of melodic material.

The segmentation we employ is entirely based on features of the audio recordings that are lost in the process of transcribing the songs into musical score. This shows that the focus of computational folk song research on symbolic musical data has to be widened. Integration of methods from both fields will lead to richer computational models of the concept of tune family.

The system is successful as a proof-of-concept. Since all phases clearly show many opportunities for improvement, we expect that the current results can be substantially improved. For example, the segmentation can be improved by using a proper breath detection algorithm instead of our simple model. The selection of representative segments could be improved by inferring the number of representative segments from the data rather than using a fixed number for all tune families. Thresholds were often defined by quick inspection. These could be determined in a more robust way. Probably these thresholds have different optimal values for different tune families.

This study offers many leads for further research that is relevant to ethnomusicology, computer science and music cognition. It would be interesting to evaluate the musical properties of the selected representative segments. Do they have occurrences in all tunes in the tune family? Are there types of representative segments? Investigating the false positives and negatives, might reveal relations between tune families that were unnoticed before. Also, a further study of the relation between the obtained segments and cognitive models of melodic chunks seems necessary. Finally, this research strongly indicates that musically and cognitively meaningful models are very important for Music Information Retrieval and other computational approaches to music, and therefore indicate that interdisciplinary collaboration between music scholars and computer scientists is of major importance.

9. REFERENCES

Bayard, S.P. (1950). Prolegomena to a Study of the Principal Melodic Families of British-American Folk Song. *Journal of American Folklore, 63* (247), 1-44.

Clarke, E., & Cook, N. (Eds.). (2004). *Empirical Musicology: Aims, Methods, Prospects*. Oxford: Oxford University Press.

Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. Ph.D. Thesis. University of Edinburgh.

Cowdery, J.R. (1984). A Fresh Look at the Concept of Tune Family, *Ethnomusicology, 28* (3), 495-504.

De Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111* (4), 1917-1930.

Gotoh, O. (1982). An Improved Algorithm for Matching Biological Sequences. *Journal of Molecular Biology, 162,* 705–708.

Keller, M.S. (1988). Segmental Procedures in the Transcription of Folk Songs in Trentino, *Sonus, 8* (2), 37-45.

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, Mass.: MIT Press.

Miller, G.A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review, 63*, 81-97.

Narmour, E. (1992). *The Analysis and Cognition of Basic Melodic Structures*. Chicago: University of Chicago Press.

Nettl, B. (2005). *The Study of Ethnomusicology*. 2nd edition. Urbana: University of Illinois Press.

Peretz, I. (2006). The Nature of Music from a Biological Perspective. *Cognition 100*, 1-32.

Smith, T.F., & Waterman, M.S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology 147*, 195–197.

Tappert, W. (1890). *Wandernde Melodien: Eine musikalische Studie*. Leipzig: List und Francke.

Van Kranenburg, P., & Garbers, J., & Volk, A., & Wiering, F., & Grijp, L.P., & Veltkamp, R.C. (2009a). Collaboration perspectives for folk Song research and music information retrieval: The indispensable role of computational musicology. *Journal of Interdisciplinary Music Studies*. doi: 10.4407/jims.2009.12.030.

Van Kranenburg, P., & Volk, A., & Wiering, F., & Veltkamp, R.C. (2009b). Proceedings of the 10th International Society for Music Information Retrieval Conference: *Musical Models for Folk-Song Melody Alignment.* Kobe: International Society for Music Information Retrieval.

Volk, A., & Van Kranenburg, P., & Garbers, J., & Wiering, F., & Veltkamp, R.C. & Grijp, L.P. (2008). *The Study of Melodic Similarity using Manual Annotation and Melody Feature Sets.* (Technical Report UU-CS-2008-013). Utrecht: Utrecht University.

Wiering, F., & Veltkamp, R.C., & Garbers, J., & Volk, A., & Van Kranenburg, P. & Grijp, L.P. (2009). Modelling Folksong Melodies. *Interdisciplinary Science Reviews, 34* (2-3), 154-171.