CALCULATING SIMILARITY OF FOLK SONG VARIANTS WITH MELODY-BASED FEATURES

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

ABSTRACT

As folk songs live largely through oral transmission, there usually is no standard form of a song - each performance of a folk song may be unique. Different interpretations of the same song are called song variants, all variants of a song belong to the same variant type. In the paper, we explore how various melody-based features relate to folk song variants. Specifically, we explore whether we can derive a melodic similarity measure that would correlate to variant types in the sense that it would measure songs belonging to the same variant type as more similar, in contrast to songs from different variant types. The measure would be useful for folk song retrieval based on variant types, classification of unknown tunes, as well as a measure of similarity between variant types. We experimented with a number of melodic features calculated from symbolic representations of folk song melodies and combined them into a melody-based folk song similarity measure. We evaluated the measure on the task of classifying an unknown melody into a set of existing variant types. We show that the proposed measure gives the correct variant type in the top 10 list for 68% of queries in our data set.

1. INTRODUCTION

With the rapid growth of digitization and appearance of digital libraries, folk song archives are (slowly but surely) entering the digital age. More and more folk song and music archives are being digitized, while most new data are already being collected in digital form.

Folk music is music that lives in oral tradition. It was composed by everyday people, and has in most cases never been written down or at least never published. It was mostly passed on to the next generation verbally and not in written form. Until folk music researchers started to put together folk music collections containing transcriptions, lyrics and other metadata, melodies were never put down in scores or any other symbolic representation. Several folk song collections are widely available; probably the most well known of them is the Essen Folksong Database [1] that includes 20.000 songs, mostly from Germany, Poland and China and minor collections from some other (mostly European) countries. The digital archive of Finnish Folk Tunes [2] is also a well known collection, containing approximately 9.000 folk tunes that were published as a collection of books between 1898 and 1933 and were digitized in 2002-2003. Some other collections are: The American Folk Song Collection [3], Australian Folk Songs [4], etc. We conducted our researh on songs from the the Ethnomuse archive [5], which contains folk music and dance collections of the Institute of Ethnomusicology, Scientific Research Centre of Slovene Academy of Sciences and Arts. The archive is especially suitable for our purpose, because it contains classifications of songs into *variant types*, tune families and genres.

Because folk songs live largely through oral transmission, there usually is no standard form of a song. As songs are passed through generations, they undergo an evolutionary process, parts change, they may be dropped and other parts may be added. Lyrics, as well as melodies get changed in the process. Each performance of a folk song may be unique and interpretations of the same song represent song *variants*. All *variants* of a song belong to the same *variant type*.

In this paper, we explore how measures extracted from folk song melodies relate to folk song variants. Specifically, we explore whether we can derive a melodic similarity measure that would correlate to variant types in the sense that it would measure songs belonging to the same variant type as more similar, in contrast to songs from different variant types.

The use of music information retrieval tools in folk music research was very limited until recently; a good overview can be found in the technical report of the Witchcraft project [6] as well as in [7, 8].

Our research is focused on developing an algorithm that calculates the similarity of two songs. In the following works several different approaches of a calculating the similarity measure are described. In [9] a method for melodic similarity of songs is presented; in [10], a method is described which uses each extracted statistical feature for training of separate self-organising map (SOM). All of the maps are later on used for training of a Supermap, on which melodies with similar features are located closer together; in [11] a comparison of different computational approaches to rhythmic and melodic similarity is made to find the features that characterise similarity of Dutch folk songs. A rhythmic similarity measure of folk songs is presented in [12]. Another similarity measure that uses pitch stability among a group of aligned folk songs is described in [13]. Which songs are similar, or how much they are alike, is not a precise problem. Not even humans always agree on whether two songs are similar or not, or which two songs are most alike. The study of how much experts agree on a manual annotation method for melodic similarity and the study of melody feature sets is described in [14].

In our paper we are proposing a system that uses simple melody-based features for classification of songs into variant types. While most of the previously mentioned papers describe methods for calculating rhythmic or melodic similarities in collections or finding features that are relevant in calculations of such similarities, our goal is to create a retrieval system for melodies, that will help us classify new unknown songs into already defined variant types.

2. SIMILARITY MEASURE

The main hypotesis of our paper is: *It is possible to classify folk song melodies into correct variant types based on statistical features of their melodies alone.* To either accept or reject our hypothesis, we first have to answer the following questions: *What kind of data do we have at our disposal and how much of it? Which features are we going to use and how to choose them? Which statistical methods should we use and how to choose them?*

The goal is to train a classifier that will classify individual variants into variant types. For this we created pairs of songs from Ethnomuse archive. *A positive example* is a pair of songs that are from the same variant type; a *negative example* is a pair of songs from different variant types.

We selected 650 folk songs belonging to 40 different variant types from the dataset. The scores for these melodies are available in Sibelius format, which we converted to MIDI. The set was split into two subsets: a *learning set* of 600 and an *independent test set* of 50 songs. The learning set was again split into two subsets: an *attribute selection learning set*, and an *attribute selection test set*. For the attribute selection learning set we only used songs from variant types with more then 7 variants. From variant types with more then 10 songs, we only used 10 randomly selected ones. The attribute selection test set was put together from 100 randomly selected songs from the learning set. For the purpose of training the classifier we created pairs of all songs in each of the attribute selection sets. For these sets we calculated percentages of positive and negative examples. The attribute selection learning set consists of 12.7% positive examples and 87.3% negative examples; attribute selection test set consists of 15,48% positive examples and 84.52% of negative examples.

For each melody, we calculated a set of 94 melody-based features with the help of the MIDI Toolbox [15]. We analyzed whether these features can be used to compare pairs of melodies and decide whether they belong to the same variant type or not. Because the task involves pairs of melodies, we created pairs of songs, for which we calculated compounded attributes as the quotient and absolute difference of individual features. All of the calculated attribute values were normalised and the SMO attribute selection method [16] used on the attribute selection sets to rank the attributes. We found the following attributes to be useful for variant type classification (details on the attributes can be found in [17]):

complebm: the measure is an expectancy based model of melodic complexity based on optimal combination of pitch and rhythm-related components calibrated in relation to the Essen Folksong Collection, where higher value means higher complexity.

entropy: relative entropy of note distribution in a matrix representation of note events.

meteraccent: measure of phenomenal accent synchrony. Meter accent is defined as:

$$meterAccent = mean(mh \cdot ma \cdot du) * (-1) \quad (1)$$

where vector *Metric hierarchy* (mh) indicates the locations of notes in the metric hierarchy (meter is calculated as an autocorrelation-based estimate - for further information see [17]); *Melodic accent* (ma) assigns melodic accents according to the possible melodic contours arising in 3-pitch windows. One can say that the melodic accent will be greater in places where the pitch changes. *Duration accent* (du) is defined in [19]

***gradus*:** degree of melodiousness (mean of Gradus suavitatis) was defined by Euler [18]. Gradus suavitatis bases on prime factorisation of note frequency ratios decreased by 1 and summed together with 1:

$$gradus\_suavitatis = 1 + \sum_{p_1 \in P} (p_i - 1) \quad (2)$$

where P is set of all prime factors of frequency ratio and note frequency ratios are acquired from nominator and denominator matrices. Degree of melodiousness (gradus) is mean value of Gradus suavitatis for all note intervals:

$$gradus = mean(\sum_{n_i \in N} (n_i)) \quad (3)$$

where N is set of all note intervals.

compltrans: Simonton's melodic originality score based on 2nd order pitch- class distribution of classical music derived from a set of music themes.

The selected features were used to train a logistic regression (LR) model. For each pair of melodies, the model outputs values between 0 and 1 for each instance; the closer the calculated value is to 1, the more probable it is that the pair of melodies belongs to the same variant type and vice versa, the closer the value is to 0, the lower the chance that the selected pair of songs is from same variant type. For the calculated values we had to set the threshold, that determines when the songs of a selected pair are from same variant type, and when not. The threshold was set so that the F-Measure reached the maximum on the *attribute selection learning set*. For later evaluation we have also calculated the F-Measure on the *attribute selection test set*.

For comparison we have also used the data to build the same model with SVM Regression (SVM) as well as Multilayer perceptron (MP). The Table 1 shows that there are only small differences between different machine learning models and that all the models are better than random classifier (RC) in all measures.

Table 1. F-Measure, precision and recall values of different models, for the attribute selection test set

Method	F-Measure	Precision	Recall
LR	0.2837	0.9888	0.1656
SVM	0.3396	0.8750	0.2107
MP	0.3237	0.8661	0.1990
RC	0.2649	0.8503	0.1600

3. EVALUATION AND DISCUSSION

3.1 Testing the model

To evaluate the logistic regression based similarity measure on a realistic task, we set up a retrieval system that takes an unknown melody as a query and returns an ordered list of melodies that should belong to the same variant type as the query. The queries were chosen from the *independent test set* and were compared to songs in the *learning set* with the logistic regression classifier trained as described previously; its output was used to rank the results.

Table 2. Ranks of first correct hits according to the proposed similarity measure.

Rank	Number of correct hits
1st	5
2nd	10
3rd or 4th	6
5th - 20th	18
21st - 30th	4
31st or worse	7

The ranked list of hits contains 600 songs; the *correct hits* are those belonging to the same variant type as the query song. For our model, the majority of first correct hits are ranked 30th or better (in 43 of 50 cases). The overall worst first correct hit is on 422nd place. Table 2 shows where first correct hits were ranked for the entire test set. For our model 68% of first correct hits lie within the top 10; a random classifier would reach 41%, so this is a significant improvement. The average rank of the first correct hit is 20.60th place, but if we exclude the most divergent results, the average rank of the first correct hit is 7.21th place for all but the worst 7 songs.

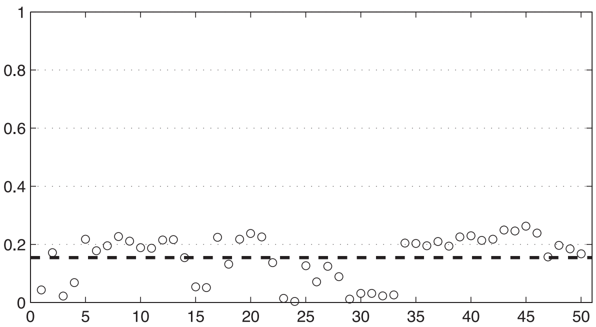
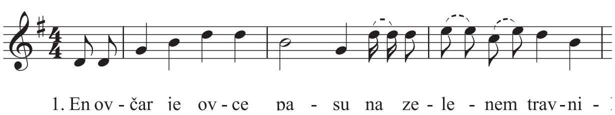


Figure 1. 11 point precision averages of test set items and their mean value.

Another measure frequently used for MIR system evaluation is 11 point precision average. This measure is calculated as the average precision at recall levels 0.0, 0.1, . . . , 0.9, 1.0. For our test set the calculated value of 11 point precision average for songs from the independent test set is 0.1544. In Figure 1 the circles represent 11 point precision average measure of each of the test set items. The dashed line indicates the mean value for all the test set items. Most of the worst cases (those under the mean line in Figure 1) are either from variant types with less then 7 variants or variants from bigger variant types that derogate the most.

3.2 Case study

The variant type with the most songs in our data set contains 163 songs. The best first correct hit for a query song from this variant type is 2nd, while the worst first correct hit is in the 31st place. 11 point precision average for this variant type is 0.2096, which is close to 11 point precision of the best query song - 0.2494. Following is the comparison of the best first correct hit and worst first correct hit examples for this variant type.



(a) query song

(b) first correct hit song (2nd)

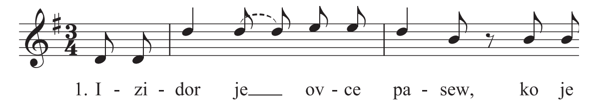
Figure 2. Example of a good result (first correct hit at 2nd place) for the same variant type as in Figure 3.

Figure 2 shows an example, where the first correct hit was on the 2nd place; the query and the correct hit are shown. The reason why this result is ranked so good (it was ranked 2nd) is because not only *complebm* (the values, 5.1332 of query and 5.2631 of first hit song, are quite similar), *meteraccent* (pitch in both, query and first correct hit song, is not monotonic) and *gradus* (both songs have quite high melodiousness) values are very similar with the values for query song, but also other two features (*entropy* and *compltrans*) are very similar with values for query song; which is not true for the previous example.

Figure 3 shows the worst first correct hit example. The query song, its first and last correct hits and the first hit song on the ranked list returned by our system are given. The main reason why the song in Figure 3(c) was ranked so low, is because of the major differences in *complebm* (the query song value is 5.0880, the first correct hit song value is 4.8136 and the last correct hit song value is 3.9494), *meteraccent* and *gradus* melodic features in comparison to the query (Figure 3(a)); on the other hand *complebm* and *compltrans* values of the first hit song (Figure 3(d)), are closer to the query song values, then the first correct hit song values.

(a) query song

(b) first correct hit song (31st)



(c) last correct hit song (592nd)

(d) first hit song (1st)

Figure 3. Example of a bad result (first correct hit at 31st place) for the variant type with the most examples.

4. CONCLUSIONS AND FUTURE WORK

As we show, there is some correspondence between simple statistical measures calculated on folk song melodies and the classification of folk songs into variant types. While results are far from very good and such a basic approach cannot be used to build a fully automatic variant type classification system, the obtained similarity measure is good enough to create a retrieval system for melodies of an unknown variant type that will give us list of a few (in our case 10) variant types, that will contain the correct type with high probability (in our case 68%). We also plan to combine the obtained melodic similarity measure with lyrics-based similarity measures and to use it for visualization of folk song melodies in the Ethnomuse archive.

5. ACKNOWLEDGMENT

Authors would like to thank anonymous reviewers for their constructive comments, suggestions and for pointing out errors; they helped make this article better.

The work was partly supported by the Slovenian Government-Founded R&D project EthnoCatalogue: creating semantic descriptions of Slovene folk song and music.

6. REFERENCES

[1] Helmut Schaffrath. The Essen Folksong Collection. D. Huron (ed.), Stanford, CA, 1995. computer database.

[2] Finnish Folk Tunes. University of Jyvskyl, 2004. URL: esavelmat.jyu.fi.

[3] The American Folk Song Collection, 2004. URL: kodaly.hnu.edu/home.cfm.

[4] Australian Folk Songs, 1994. URL: http://folkstream.com/.

[5] Grega Strle and Matija Marolt. Conceptualizing the Ethnomuse: Application of CIDOC CRM and FRBR. *Proceedings of CIDOC2007*, 2007.

[6] Peter van Kranenburg, Jörg Garbers, Anja Volk, Frans Wiering, Louis P. Grijp, and Remco C. Veltkamp. Towards integration of music information retrieval and folk song research. Technical Report UU-CS-2007-016, Department of Information and Computing Sciences, Utrecht University, 2007.

[7] Petri Toiviainen and Tuomas Eerola. Visualization in comparative music research. In *COMPSTAT 2006 - Proceedings in Computational Statistics*, pages 209–221, 2006.

[8] Peter van Kranenburg, Jörg Garbers, Anja Volk, Frans Wiering, Louis P. Grijp, and Remco C. Veltkamp. Towards integration of mir and folk song research. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pages 505–508, Vienna, Austria, September 2007. Österreichische Computer Gesellschaft.

[9] Rainer Typke. *Music Retrieval based on Melodic Similarity*. PhD thesis, Utrecht University, Netherlands, February 2007.

[10] Petri Toiviainen and Tuomas Eerola. Method for comparative analysis of folk music based on musical feature extraction and neural networks. In *In III International Conference on Cognitive Musicology*, pages 41–45, 2001.

[11] Anja Volk, Jörg Garbers, Peter van Kranenburg, Frans Wiering, Louis P. Grijp, and Remco C. Veltkamp. Comparing Computational Approaches to Rhythmic

and Melodic Similarity In Folksong Research. In *Proc. MCM 2007*, 2007.

[12] Anja Volk, Jörg Garbers, Peter van Kranenburg, Frans Wiering, Remco C. Veltkamp, and Louis P. Grijp. Applying rhythmic similarity based on inner metric analysis to folksong research. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pages 293–296, Vienna, Austria, September 2007. Österreichische Computer Gesellschaft.

[13] Jörg Garbers, Peter van Kranenburg, Anja Volk, Frans Wiering, Remco C. Veltkamp, and Louis P. Grijp. Using pitch stability among a group of aligned query melodies to retrieve unidentified variant melodies. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pages 451–456, Vienna, Austria, September 2007. Österreichische Computer Gesellschaft.

[14] Anja Volk, Peter van Kranenburg, Jörg Garbers, Frans Wiering, Remco C. Veltkamp, and Louis P. Grijp. A manual annotation method for melodic similarity and the study of melody feature set. *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, September 2008.

[15] Tuomas Eerola and Petri Toiviainen. Mir in matlab: The midi toolbox. In *ISMIR*, 2004.

[16] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[17] Tuomas Eerola and Petri Toiviainen. *MIDI Toolbox: MATLAB Tools for Music Research*. University of Jyväskylä, Jyväskylä, Finland, 2004.

[18] Leonhard Euler. *Tentamen novae theoriae musicae*. 1739.

[19] Richard Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11(4):409–464, 1994.