SELECTIVE ACQUISITION TECHNIQUES FOR ENCULTURATION-BASED MELODIC PHRASE SEGMENTATION

Marcelo E. Rodríguez-López
Utrecht University
m.e.rodriguezlopez@uu.nl

Anja Volk
Utrecht University
a.volk@uu.nl

ABSTRACT

Automatic melody segmentation is an important yet unsolved problem in Music Information Retrieval. Research in the field of Music Cognition suggests that previous listening experience plays a considerable role in the perception of melodic segment structure. At present automatic melody segmenters that model listening experience commonly do so using unsupervised statistical learning with 'non-selective' information acquisition techniques, i.e. the learners gather and store information indiscriminately into memory.

In this paper we investigate techniques for 'selective' information acquisition, i.e. our learning model uses a goal-oriented approach to select what to store in memory. We test the usefulness of the segmentations produced using selective acquisition learning in a melody classification experiment involving melodies of different cultures. Our results show that the segments produced by our selective learner segmenters substantially improve classification accuracy when compared to segments produced by a non-selective learner segmenter, two local segmentation methods, and two naïve baselines.

1. INTRODUCTION

Motivation: In Music Information Retrieval (MIR), melody segmentation refers to the task of dividing a melody into smaller units, such as figures, phrases, or sections. Given that melody is an aspect of music shared by almost all cultures in the world, and that melodies are known to be memorable, many MIR systems base their functionality in melody processing. Automatic melody segmentation is hence an important preprocessing step for MIR tasks involving searching, browsing, visualising, and summarising music collections.

Scope: Research in automatic melody segmentation has been conducted by subdividing the segmentation problem into a number of subtasks, the most traditional one being segment boundary detection, i.e. automatically locating the time instants separating contiguous segments. In this paper we focus on detecting the boundaries of segments resembling the musicological concept of *subphrase*. The musical factors influencing the perception of melodic segment boundaries are diverse [4, 8]. In this paper we focus on modelling factors related to previous listening experience and melodic expectation [1, 9, 23, 25, 27].

Terminology: We use the term 'phrase' to refer to a sequence of notes lasting roughly from 6 notes to 8 bars. We use the term 'figure' to refer to a relatively short sequence of notes, lasting roughly from 2-6 notes. We use the term 'subphrase' to refer to melodic figures in the context of phrases, i.e. as the constituent parts of a melodic phrase.

Assumptions: Our main assumption is that human listeners exposed to melodies of a given culture acquire a vocabulary of melodic figures through 'incidental' learning, [1] and that this acquired melodic vocabulary aids the segmentation of phrases into subphrases. [2] We refer to such a listener as 'enculturated'.

Problem statement: At present, automatic melody segmenters that model previous listening experience usually do so by storing information indiscriminately into memory. We argue that selective (rather than indiscriminate) information acquisition is necessary to simulate enculturation. We hence propose and investigate two techniques for selective acquisition in the context of phrase segmentation: one in which an artificial learner selects the subphrases that give it the 'clearest' possible 'understanding' of a phrase, and another in which the learner attempts to use subphrases it 'knows well' to expand its melodic vocabulary. To compare the segmentations produced by enculturated segmenters using selective and non-selective acquisition techniques, we perform a melody classification experiment involving melodies of different cultures, where the segments are used as classification features.

Paper contributions: We have three main contributions. First, the proposed techniques for selective acquisition are, to the best of our knowledge, novel in the context of melody segmentation. Second, we focus on subphrase level segmentation, which is a neglected area in music segmentation research. Third, our results show that the segments produced by our selective learning segmenters substantially improve classification accuracy when compared to segments produced by using a non-selective learning

[1] We use the term incidental to mean that the listener does not have an explicit learning intention.

[2] Refer to [15, 16, 24] for experimental work in music cognition and cognitive neuroscience that supports our assumption.

segmenter, two local segmentation methods [5, 6], and two naïve baselines.

Paper summary: The remainder of this paper is organised as follows: §2 reviews related work, §3 describes our selective acquisition learning model , §4 describes our proposed enculturated segmenter, §5 describes the classification experiment and presents results, §6 discusses the evaluation results, and finally, §7 summarises our conclusions and outlines possibilities of future work.

2. RELATED WORK

Previously proposed melody segmenters that model listening experience have mostly used non-selective learners. For instance, [23] presents a segmentation model with a long-term memory (LTM) component. To train the LTM model the prediction by partial match (PPM) algorithm [21] is used, which gathers and stores ngrams and ngram statistics indiscriminately into LTM. Much of the work carried out by the authors of [23] in melodic learning has focused mostly on dealing with melodic multidimensionality [19] and on the combination of short-term and long-term memory models [20], but not much attention has been paid to the construction of the LTM itself.

We base our approach on [11], where selective acquisition learning is used for motivic pattern extraction from a corpus of melodies. Our approach extends their work by proposing and testing different selective acquisition techniques, and by combining the learning approach proposed in [11] with characteristics of the 'feature selection' learning approach proposed in [3] for natural language processing. Moreover, we focus on using selective learning to create more powerful LTM models for melody segmentation. In the following section we describe our approach in detail.

3. ENCULTURATION VIA SELECTIVE ACQUISITION LEARNING

The goal of selective acquisition learning is to construct an enculturated LTM model. In this paper we model enculturation as a refinement process. That is, our learner takes two inputs: (1) a LTM model, which is simply a collection of melodic figures acquired during prior listening experience, and (2) a corpus of melodies of a given culture to which the learner is to be exposed. The output is a LTM model in which, ideally, only melodic figures characteristic of the culture to which the learner has been exposed are preserved. Our learning approach is summarised as pseudo code in Algorithm 1.

As shown in Algorithm 1, our learner 'listens' to each melody one phrase at a time, and decides which figures to store in LTM by evaluating different segmentations. That is, the learner stores in LTM only the figures that allow it to segment the phrase in an optimal way. This process is continued until the learner has acquired the melodic vocabulary that allows it to perform optimal segmentations. In the following sections we describe each part of the approach in more detail.

Input: LTM model, Phrase-segmented Melodic Corpus,
Output: LTM model

while *termination condition not met* **do**
  read melody from corpus;
  for *each phrase in melody* **do**
    Compute possible segmentations;
    Select the optimal segmentation;
    Store suphrases in LTM;
  Check termination condition;

Algorithm 1: Selective Acquisition Learning

3.1 Input/Output

3.1.1 Input: Melody Representation

Our learner takes as input melodies represented as a sequence of chromatic pitches, constrained to a range of two octaves. [3] Formally, we take $p = p_1 \dots p_N$ to be a sequence of pitch intervals, where each interval $p_i \in \mathcal{A} = \{-12, \dots, 0, \dots, +12\}$. In $\mathcal{A}$ each numerical value encodes the distance in semitones between two contiguous pitches, and the $\pm$ symbol encodes its orientation (ascending, descending).

3.1.2 Input: phrase segmented corpus

We assume input melodies are annotated with phrase boundaries, so that our learner can process melodies on a phrase by phrase basis, finding for each an optimal segmentation. We choose to process phrases based on cognitive constraints, as exhaustively evaluating multiple segmentations for a whole melody would break known limitations of human memory.

3.1.3 Input/Output: long term memory (LTM) model

We model LTM probabilistically using a Markov modelling strategy. Essentially this boils down to constructing a data structure to hold the number of times melodic figures up to 5 intervals appear in a corpus, and then use those counts to estimate probabilities (we go into more detail in §3.3). [4]

[3] In this paper our learner and segmenters take as input symbolic encodings of melodies, i.e. computer readable representations of scores transcribed by experts (see §5.1 for more details). Symbolically encoded melodies can be represented in a variety of ways, e.g. chromatic pitch, step-leap pitch intevals, inter onset intervals, and so on. In statistical learning this multi-dimensional attribute representation of melodic events can be tackled using *multiple viewpoint systems* [7, 19]. However, using multiple viewpoints comes at expense of a considerable increase in the complexity of the statistical model architecture, resulting in an increase in processing time and space requirements, as well as lower interpretability of the model. In this paper we favour using a single melodic representation to simplify the evaluation our of segmenters, which is important considering that we evaluate our segmenters indirectly, by means of a classification experiment (see §5).

[4] The input LTM model can also be computed by sampling from known parametric distributions, e.g. in [2] the LTM model is constructed sampling from a Dirichlet distribution. However, by using corpus statistics we can assess how different (and perhaps more suitable) are the segmentations produced by one of the learners in respect to the others when exposed to the same melodies, which is a better way to try to prove or disprove our hypothesis.

3.2 Computing Possible Segmentations

Ideally, our learner should evaluate all possible segmentations of a phrase. However, processing time is exponential on the number of notes in the phrase, so in practice evaluating all segmentations is unfeasible. Thus, we use the algorithm proposed in [17] to efficiently compute a constrained space of possible segmentations. The algorithm takes as input the minimum and maximum length of subphrases, as well as the minimum and maximum number of subphrases. As we mentioned previously we have limited subphrases to be sequences of 1-5 intervals in length. We also limit phrases to be composed of at most 6 subphrases (by doing so we are able to cope with phrases of a maximum length of 30 intervals).

3.3 Select the optimal segmentation

Below we present two techniques to select an optimal segmentation. One in which the learner selects subphrases that give it the 'clearest' possible understanding of a phrase, and another in which the learner uses subphrases it 'knows well' to increase its vocabulary.

3.3.1 Common and Complete Figures

Melodic figures that aid segmentation should be 'characteristic' of a melodic culture. One way to measure how characteristic figures are is by searching for 'common' figures in a corpus representative of a melodic culture. However, common figures are mainly of short duration, and normally less specific and informative than figures of larger duration (see [28]). There is hence a trade-off between how common a figure is and how specific to a given tradition it can be. [5] Thus, we need a way to automatically determine how long do the figures we are after need to be, so that we search for the longest possible common figures instead of only the most common ones. One way to do so is by attempting to determine if a given figure is somehow 'complete' on its own, or if its part of a larger figure. Our search then would be for figures that are common, yet large enough so as to be perceptually complete. According to melodic expectation theory [14, 27], the perceptual completeness of a melodic figure is inversely proportional to the degree by which it stimulates expectation. In other words, melodic figures for which is hard to predict what comes next are perceived as more complete than those for which is easy to predict what comes next.

Using information theory we can attempt to jointly quantify the commonness and completeness of a figure. If from within a phrase of length T we take a figure $w = p_i \dots p_j$, with $i, j \in [1:T]$, we can compute its conditional entropy h as

$$h(x|w) = P(w) \sum_{x \in \mathcal{A}} P(x|w) log\left(P(x|w)\right) \qquad (1)$$

where x is used to symbolise melodic events that can follow w, and P denotes probability. In Eq. 1 the first term $P(\cdot)$ will be high for common figures in a corpus, and the second term $\sum P(\cdot) log\left(P(\cdot)\right)$ will be high if it is hard to predict what comes after w. Hence, h will be high for figures that are common and complete in an information theoretic sense.

The values of probabilities $P(\cdot)$ can be estimated from the counts of w and the concatenation wx in a given melodic corpus: $P(w) \sim N(w)/N_T$ and $P(x|w) \sim N(wx)/N(w)$, where $N(\cdot)$ denotes counts, and N_T denotes the total number of counts for figures of length equal to w in the corpus.

3.3.2 Monitoring LTM

Using conditional entropy we can monitor the state of our LTM before and after a new melodic figure is listened to. So, first, the total entropy for figures w of the same size is

$$H^o = - \sum_{w \in \mathcal{A}^*} P(w) \sum_{x \in \mathcal{A}} P(x|w) log P(x|w) \qquad (2)$$

where we use $\mathcal{A}^*$ to denote the space of all figures of size o with attribute space $\mathcal{A}$. In our LTM $o = \{1, \dots, 5\}$ and hence its total entropy is

$$H = H^1 + \dots + H^5 \qquad (3)$$

and then we can define ΔH as

$$\Delta H = H_{\text{after listening to } w} - H_{\text{before listening to } w} \qquad (4)$$

which allows us to monitor the evolution of our LTM.

3.3.3 Selection Technique 1

We have now the necessary information to formulate our first selection technique. Since common and complete figures are expected to have high entropy, a 'good' phrase segmentation among a group of possible segmentations is that segmentation with the highest average ΔH. That is, if we have a space of possible segmentations $\mathcal{S}$, the average ΔH of a candidate segmentation $s = w_1, \dots, w_m$ is

$$\phi(s) = \frac{\Delta H(w_1) + \dots + \Delta H(w_m)}{m} \qquad (5)$$

and hence our first selection technique is

$$s^* = \underset{s \in \mathcal{S}}{\operatorname{argmax}}\, \phi(s) \qquad (6)$$

Where s^* denotes the segmentation with maximal score. Note that, to ensure convergence, the leaner stores in LTM only the subphrases in s^* for which ΔH is positive.

One problem with our first technique is that it makes our learner very conservative. The melodic figures stored are characteristic of the corpus as a whole. Hence, the technique operates under the assumption that the corpus is stylistically homogeneous. For most cultural traditions the assumption of complete stylistic homogeneity is too strong (it is likely that certain figures are important but only characteristic of subsets of the corpus).

[5] In natural language this is also a commonly found problem, 'content' or informative words (e.g. nouns) tend to be of greater length that 'non-content' words (e.g. determinants).

Collection Name Abbreviation	Subset Name	Cultural Origin of Sample	Encoding	Number of Melodies	Average Melody Size in Notes	Number of Phrases	Average Phrase Size in Notes
MTC	FS	Dutch	**kern	4120	52.3 (22.5)	19935	9.1 (2.5)
EFSC	CHINA	Chinese	**kern	2201	62.8 (41.2)	11046	12.5 (4.7)
OHFT	-	Hungarian	EsAC	2323	38.6 (12.0)	9308	9.6 (3.2)

Table 1. Melodic Corpora. Numbers in parenthesis correspond to standard deviation.

3.3.4 Selection Technique 2

Our second technique aims to relax the assumption of homogeneity and stimulate the learner to expand its vocabulary. More importantly, it aims to reveal segmentations in which one or more subphrases are common and complete, and others are representative of the melody, yet relatively rare in the corpus. For a figure w the latter idea can be quantified as

$$\rho(w) = -P_{melody}(w) * log\left(P_{corpus}(w)\right) \tag{7}$$

with $P_{melody}(w) \sim M(w)/M_T$ and $P_{corpus}(w) \sim N(w)/N_T$, where M denotes counts of w in the melody, M_T is used to indicate the total number of counts of figures of size equal to w in the melody/corpus, and N denotes counts of w in the corpus.

For a complete segmentation we take the average of ρ

$$\overline{\rho}(s) = \frac{\rho(w_1) + \cdots + \rho(w_m)}{m} \tag{8}$$

Finally, we combine the $\overline{\rho}$ and ϕ using a geometric mean: [6]

$$\lambda(s) = \sqrt{\phi(s) \cdot \overline{\rho}(s)} \tag{9}$$

and compute our second technique as

$$s^* = \underset{s \in \mathcal{S}}{\operatorname{argmax}}\, \lambda(s) \tag{10}$$

Where s^* denotes the segmentation with maximal score. Our leaner stores all subphrases of s^* in LTM.

3.4 Termination Condition

We keep track of the scores of s^* when processing the corpus, expecting that, as the learner reaches convergence, the score difference between subsequent instances s^* gets smaller and smaller. We hence assume convergence has been reached if $\Delta s^* < \epsilon$.

Since Eq. 10 encourages learning new vocabulary, convergence is slow and not guaranteed. Thus, in addition to $\Delta s^* < \epsilon$, we also set a maximum number of learning iterations as a second termination condition.

4. ENCULTURATED SEGMENTATION

Once the LTM model has been trained (via either selective or non-selective learning), our segmenter proceeds in a way similar to Algorithm 1. That is, it processes each melody a phrase at a time, for each phrase it computes a space of possible segmentations, and selects the best one. However, this time the selection of the best segmentation is made by computing

$$\overline{h}^* = \underset{s \in \mathcal{S}}{\operatorname{argmax}}\, \overline{h}(s) \tag{11}$$

where $\overline{h}(s) = \frac{h(w_1)+\cdots+h(w_m)}{m}$ and $h(w)$ is computed using Eq. 1.

5. EVALUATING SUBPHRASE SEGMENTATIONS

At present, freely available corpora annotated with subphrase boundaries do not exist. This implies we are unable to evaluate our segmenters in a traditional scenario (i.e. by comparing automatic segmentations to human-annotated segmentations). Hence, we opt for a 'use-case' evaluation scenario: test the output of our segmenters in a melody classification experiment.

The classification task consists in predicting the cultural origin of each melody in a dataset of melodies, using subphrases as classification features. In this scenario 'good' segmentations should facilitate classification and thus result in high classification performance.

In the following subsections we describe the melodic corpora used for our classification experiment, the compared segmenters, the classifiers employed, and finally we list evaluation metrics and present results.

All segmenters and baselines were coded in Matlab. All source files as well as the train/test data listings are available at `http://www.projects.science.uu.nl/music/`.

5.1 Phrase Annotated Melodic Corpora

The melodic corpora used in our experiments is summarised in Table 1. The *Meertens Tune Collection* [7] (MTC) is a collection of Dutch folk songs. The *Essen Folk Song Collection* [8] (EFSC) is a collection of vocal folk songs from Eurasia. The *Old Hungarian Folksong Types* collection [9] (OHFT) is a collection of vocal folk songs from Hungary.

All corpora summarised in Table 1 have been annotated with phrase boundaries by expert Ethnomusicologists. [10] We cleaned the collections by removing all melodies with overly short and overly long phrases. We considered a

[6] Since $\phi(s)$ can in principle be negative, to compute λ we consider negative $\Delta H(w)$ values to be zero when computing $\phi(s)$ to avoid the possibility of negativity.

[7] `http://www.liederenbank.nl`

[8] `http://www.esac-data.org`

[9] We obtained the OHFT data directly from the author of [11].

[10] In the case of the EFSC-CHINA the origin of the phrase markings is uncertain. However, it is often assumed it corresponds to notated breath marks and/or to the phrase boundaries of lyrics. In the case of the MTC-FS phrase boundary markings where produced by two experts (which agreed on a single segmentation). The annotation process is detailed in [26]. In the case of the OHFT the phrase boundary marking process is detailed in [10, 12].

Segmenter	Parameter Setting	Segmentation Results (for the best parametric setting)								
		Mean Number of Subphrases per Phrase			Mean Number of Subphrases per Melody			Total Number of Unique Subphrases per Corpus		
		C	H	D	C	H	D	C	H	D
NS	LTM training: PPM-C, with exclusion, 1000 melodies of each culture.	5.0	4.8	4.8	27.7	16.7	23.4	1300	1091	1197
ST1	LTM training: convergence 10E-8 or 8000 phrases, 1000 melodies of each culture.	3.8	3.7	3.7	21.7	13.7	19.6	3437	2562	2204
ST2	LTM training: convergence 10E-8 or 8000 phrases, 1000 melodies of each culture.	3.6	3.5	3.5	23.2	15.4	21.3	3566	2743	2311
LBDM	detection threshold $\{0.2, \mathbf{0.4}, 0.6\}$, others: suggested setting in [5].	3.0	2.9	2.9	15.5	10.4	15.4	4497	3841	2999
PAT	detection threshold $\{0.2, \mathbf{0.4}, 0.6\}$, others: suggested setting in [6].	3.6	3.4	3.4	19.3	11.4	16.7	3603	3139	2810
FIXLEN	constant size $CS = 3$ intervals.	4.0	3.8	3.8	22.0	13.5	18.9	1371	1179	1474
RAND	constant size $RS \in [2-4]$ intervals.	4.1	3.9	3.9	22.2	13.5	19.2	2827	2413	2551

Table 2. Parameter settings and segmentation results. C - Chinese, H - Hungarian, D - Dutch. Text in bold indicates best performing parametric settings.

phrase to be overly short if it contains only one note or one interval. We considered a phrase to be overly long if it is longer than 30 notes in length.

5.2 Enculturated Segmenters

We evaluate three enculturated segmenters: NS, ST1, ST2. The NS segmenter uses a LTM model trained with non-selective acquisition (using the PPM-C algorithm [22]). The ST1 segmenter uses a LTM model trained with the selective acquisition technique 1, Eq. 6. The ST2 segmenter uses a LTM model trained with the selective acquisition technique 2, Eq. 10. A sample of 1000 melodies from each collection is used to train the LTM models. The parametric settings for each enculturated segmenter are specified in Table 2.

5.3 Reference Segmenters and Baselines

We compared the performance of the enculturated segmenters to two local boundary detection segmenters (LBDM and PAT), and two naïve baseline segmenters (FIXLEN and RAND). The LBDM and PAT segmenters were selected for comparison because they have been used for subphrase level segmentation in the past [6, 18]. The LBDM segmenter [5] computes subphrase boundaries by detecting large pitch intervals and inter-onset-intervals. Intervals sizes are given a score by comparing them to immediately surrounding intervals (the larger the difference the higher the score). High scoring intervals are taken as subphrase ends. The PAT segmenter [6] computes subphrase boundaries by detecting and scoring repetitions of pitch interval sequences within each phrase. The starting points of high scoring repetitions are taken as subphrase starts. The FIXLEN baseline segments a phrase into subphrases of constant size. The RAND baseline segments a phrase into subphrases of randomly chosen sizes. The parametric settings for each of the reference and baseline segmenters are specified in Table 2.

5.4 Features and Classifiers

As mentioned above, in our experiment we are interested in evaluating the effectiveness of subphrases as classification features. To use subphrases in the most transparent way, we represent melodies as a 'bag-of-subphrases'. That is, we use a vector space model representation, [11] where each vector element is weighted using the common term frequency - inverse document frequency ($tf * idf$) heuristic [13]. We then use two simple and well known classifiers for the cultural origin prediction task: *k-means* and *k nearest neighbours* (kNN).

Segmenter	k-means (k=3)			kNN (k optimised)		
	$\overline{R}$	$\overline{P}$	$\overline{A}$	$\overline{R}$	$\overline{P}$	$\overline{A}$
NS	0.94	0.93	0.71	0.93	0.87	0.83
ST1	0.90	0.95	**0.74**	0.93	0.94	0.87*
ST2	0.92	0.93	0.71	0.92	0.96	**0.88**
LBDM	0.47	0.50	0.47	0.75	0.84	0.76
PAT	0.74	0.76	0.58	0.83	0.87	0.79
FIXLEN	0.88	0.89	0.67	0.86	0.90	0.83
RAND	0.84	0.84	0.63	0.88	0.85	0.78

Table 3. Clasification results: recall ($\overline{R}$), precision ($\overline{P}$), and accuracy ($\overline{A}$) averaged over 10-folds. Text in bold highlights the highest performances. Asterisks indicate performances that are not significantly different from the highest performances.

5.5 Test set, Performance Measures, and Results

We constructed a dataset of 3000 melodies by randomly sampling 1000 melodies from each corpus. (All melodies used to train the enculturated segmenters were excluded from the sample.) For each of the 3000 melodies, the classifiers are required to predict whether the melody is of Hungarian, Chinese, or Dutch origin. **Validation technique**: We used 10-fold cross validation to iteratively separate the melodic dataset into training and test sets. **Evaluation measures**: Given a N_{total} of melodies per fold to be classified, we use tp to indicate the number of true positives, fp the false positives, and fn the false negatives. With these statistics we measure measure classification performance using accuracy $A = \frac{N_{correct}}{N_{total}}$, precision $P = \frac{tp}{tp+fp}$ and recall $R = \frac{tp}{tp+fn}$. **Statistical testing**: We

[11] in a vector space model, melodies are represented as a vector of size $|V|$, where $|V|$ is the number of unique figures occurring in the corpus. If a figure occurs in the melody, its value in the vector is equal to the number of times it appears in the melody. The frequency of occurrence of each figure is then used as a feature for classification.

used an an ANOVA test ($\alpha = 0.01$) with Bonferroni correction to test the statistical significance of the differences in accuracy for each segmenter. **Setting and optimising classifier parameters**: The training sets were used to optimise the permutation labels of the k-means classifier and select the optimal number of nearest neighbours for the kNN classifier. The optimal number of nearest neighbours (selected from $k \in [1, 15]$) was set by optimizing cross-validated accuracy on the training data.

The results of our experiment are presented in Table 3. We discuss our results below.

6. DISCUSSION

6.1 Selective vs. Non-Selective Learning Segmenters

Table 2 shows the NS segmenter produces relatively short segments, resulting in an average of ∼4.9 subphrases per phrase, and an average of ∼1196 unique subphrases over all three corpora. Conversely, the ST1-2 segmenters produce larger segments, resulting in an average of ∼3.6 subphrases per phrase, and an average of ∼2767 unique subphrases over all three corpora. Using the k-means classifier with subphrases computed using ST1 we obtain a (statistically significant) 3% $\overline{A}$ improvement over the NS segmenter, which seems to be driven by a 2% improvement in $\overline{P}$. Using the k-NN classifier with subphrases computed using both ST1 and ST2 we obtain (statistically significant) 3-4% $\overline{A}$ improvements over the NS segmenter, which are again in pair with 7-9% increases in $\overline{P}$. These results show the larger segments produced by the ST1-2 segmenters allow better discrimination between melodies of different cultural origin, suggesting that selective learning leads to better models of prior listening experience than non-selective learning.

6.2 Selective Learning Segmenters vs. Local Segmenters

Segmentation results in Table 2 show that local segmenters prefer larger segments than the ST1-2 segmenters. Also, the local segmenters produce an average of ∼3481 unique subphrases over all three corpora, which is 741 subphrases larger than the average of unique subphrases produces by the ST1-2 segmenters. Table 3 shows that $\overline{A}$ results using the segments produced by ST1-2 are >8% better than $\overline{A}$ results using the segments produced by LBDM and PAT. The $\overline{A}$ performance improvements are in line with relatively large improvements in both $\overline{P}$ and $\overline{R}$. These results show that the larger segments produced by the local segmenters leads to an increase in unique subphrases, and that these unique subphrases are not discriminative of cultural origin. The relatively large improvements in $\overline{A}$ of the ST1-2 segmenters over the local segmenters supports the hypothesis that enculturated listening might be of importance for the segmentation of melodic phrases.

6.3 Selective Learning Segmenters vs. Baselines

Table 2 shows the baseline segmenters produce relatively short segments (of 2 or 3 intervals), resulting in an average of ∼3.9 subphrases per phrase, and an average of ∼1969 unique subphrases over all three corpora. When using the k-means classifier we can observe significant and relatively large differences ($> 5\%$) between the $\overline{A}$ obtained using ST1-2 and those obtained using the baseline segmenters. These results show the larger segments produced by the ST1-2 segmenters allow better discrimination between melodies of different cultural origin than the shorter segments produced by the baseline segmenters, indicating once more the ST1-2 segmenters might be capturing important aspects of subphrase structure.

6.4 Scepticism

Any conclusions from our use case evaluation results are limited to classification schemes using 'bag-of-subphrases' representations of melodies. This representation limits the similarity assesment between any two subphrases to exact matches, which might be introducing an unwanted bias on the evaluation. To draw more definitive conclusions our experiment needs to be complemented with other use case studies.

7. CONCLUSIONS

In this paper we introduce techniques for selective acquisition learning in the context of melodic segmentation, specifically the segmentation of melodic phrases into subphrases. Our aim is to show that enculturated listening is important for the segmentation of melodic phrases, and that selective rather than indiscriminative acquisition techniques are better to model an enculturated segmenter. We present two selective acquisition techniques: one in which an artificial learner selects the subphrases that give it the 'clearest' possible understanding of a phrase, and another in which the learner attempts to use subphrases it 'knows well' to expand its melodic vocabulary.

To test the segmentations produced by enculturated segmenters using selective and non-selective acquisition techniques, we perform a melody classification experiment involving melodies of different cultures. Our results show that the segments produced by our selective learning segmenters substantially improve classification accuracy when compared to segments produced by using a non-selective learning segmenter, two local segmentation methods, and two naïve baselines.

In future work we plan to conduct experiments to test the sensitivity of our selection techniques to cross-learning. That is, cases in which the learners have prior knowledge of one melodic tradition and are required to adapt their knowledge to the particularities of a different melodic tradition. We also plan to extend the current approach so that it can process multiple attribute representations of a melody. To this end an integration between our approach and the multipleviewpoint formalism of [7, 19] is planned.

Acknowledgments: We thank Z. Juhász for sharing with us the OHFT dataset, and also to the anonymous reviewers for the useful comments. This work is supported by the Netherlands Organization for Scientific Research, NWO-VIDI grant 276-35-001 to A. Volk.

8. REFERENCES

[1] S. Abdallah, H. Ekeus, P. Foster, A. Robertson, and M. Plumbley. Cognitive music modelling: An information dynamics approach. In *3rd International Workshop on Cognitive Information Processing (CIP)*, pages 1–8. IEEE, 2012.

[2] S. Abdallah and M. Plumbley. Information dynamics: patterns of expectation and surprise in the perception of music. *Connection Science*, 21(2-3):89–117, 2009.

[3] A. Berger, V. Della Pietra, and S. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[4] M. Bruderer, M. Mckinney, and A. Kohlrausch. The perception of structural boundaries in melody lines of western popular music. *Musicae Scientiae*, 13(2):273–313, 2009.

[5] E. Cambouropoulos. The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC01)*, pages 232–235, 2001.

[6] E. Cambouropoulos. Musical parallelism and melodic segmentation. *Music Perception*, 23(3):249–268, 2006.

[7] D. Conklin and I. H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

[8] I. Deliège. Wagner alte weise: Une approche percepitve. *Musicae Scientiae*, 2(1 suppl):63–89, 1998.

[9] D. Huron. *Sweet anticipation: Music and the psychology of expectation*. MIT press, 2006.

[10] P Járdányi. Experiences and results in systematizing hungarian folk-songs. *Studia Musicologica*, pages 287–291, 1965.

[11] Z. Juhász. Segmentation of hungarian folk songs using an entropy-based learning system. *Journal of New Music Research*, 33(1):5–15, 2004.

[12] Z. Kodály and L. Vargyas. *Folk music of Hungary*. Da Capo Press, 1982.

[13] C. D. Manning, P. Raghavan, and H. Schütze. Scoring, term weighting and the vector space model. *Introduction to Information Retrieval*, 100, 2008.

[14] L. B. Meyer. Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, pages 412–424, 1957.

[15] S. J. Morrison, S. M. Demorest, E. H. Aylward, S. C . Cramer, and K. R. Maravilla. FMRI investigation of cross-cultural music comprehension. *Neuroimage*, 20(1):378–384, 2003.

[16] Y. Nan, T. R. Knösche, and A. D. Friederici. The perception of musical phrase structure: a cross-cultural erp study. *Brain research*, 1094(1):179–191, 2006.

[17] J.D. Opdyke. A unified approach to algorithms generating unrestricted and restricted integer compositions and integer partitions. *Journal of Mathematical Modelling and Algorithms*, 9(1):53–97, 2010.

[18] N. Orio and G. Neve. Experiments on segmentation techniques for music documents indexing. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 104–107, 2005.

[19] M. Pearce. *The construction and evaluation of statistical models of melodic structure in music perception and composition*. PhD thesis, Department of Computing, City University, 2005.

[20] M. Pearce, D. Conklin, and G. Wiggins. Methods for combining statistical models of music. In *Computer Music Modeling and Retrieval*, pages 295–312. Springer, 2005.

[21] M. Pearce and G. Wiggins. An empirical comparison of the performance of ppm variants on a prediction task with monophonic music. In *Artificial Intelligence and Creativity in Arts and Science Symposium*, 2003.

[22] M. Pearce and G. Wiggins. Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, 33(4):367–385, 2004.

[23] M. Pearce and G. Wiggins. The information dynamics of melodic boundary detection. In *Proceedings of the Ninth International Conference on Music Perception and Cognition*, pages 860–865, 2006.

[24] M. Rohrmeier, P. Rebuschat, and I. Cross. Incidental and online learning of melodic structure. *Consciousness and cognition*, 20(2):214–222, 2011.

[25] C. Thornton. Generation of folk song melodies using bayes transforms. *Journal of New Music Research*, 40(4):293–312, 2011.

[26] P. van Kranenburg, M. de Bruin, L. Grijp, and F. Wiering. The meertens tune collections. 2014.

[27] G. Wiggins and J. Forth. IDyOT: A computational theory of creativity as everyday reasoning from learned information. In *Computational Creativity Research: Towards Creative Machines*, pages 127–148. Springer, 2015.

[28] J. Wołkowicz, Z. Kulka, and V. Kešelj. N-gram-based approach to composer recognition. *Archives of Acoustics*, 33(1):43–55, 2008.