Cognition-based segmentation for music information retrieval systems

Justin de Nooijer
Fortis ASR, Utrecht, Netherlands
justindenooijer@gmail.com

Frans Wiering, Anja Volk, Hermi J.M. Tabachneck-Schijf
Department of Information and Computing Sciences, Utrecht University, Netherlands
{frans.wiering; volk; h.schijf}@cs.uu.nl - http://www.cs.uu.nl

Proceedings of the fourth Conference on Interdisciplinary Musicology (CIM08)
Thessaloniki, Greece, 2-6 July 2008, http://web.auth.gr/cim08/

Background in computer science. This paper investigates the generic problem of model selection in the specific context of Music Information Retrieval (MIR). In MIR research, similarity measures are developed for ranking musical items with respect to their relevance to a user's musical query. The application of such similarity measures in MIR systems typically requires musical works to be divided into more manageable units. This involves two tasks: melody segmentation and voice separation. For both of these tasks, several solutions have been proposed in the symbolic domain. It seems reasonable to assume that those solutions that are most in accordance with human performance will result in the best ranking of retrieval output. As a first step towards this goal, this paper describes the evaluation of ten prominent methods against human performance.

Background in cognition. Human listeners generally possess two functions that allow them to process a continuous stream of music into understandable units: the ability to perceive multiple, successive tones as one coherent melodic phrase (melody segmentation) and the ability to differentiate melody notes from harmony notes (voice separation). Algorithms for mimicking these human functions have been developed from two perspectives: model-driven, taking Gestalt principles as a starting-point; and data-driven, inferring rules by learning from large amounts of data. One method excepted, this research focuses on model-driven approaches.

Aims. The aim of the research is to answer two questions: (1) is there enough agreement in human segmentation perception to function as a basis for measuring algorithmic performance? and (2) which algorithm best models human segmentation? These questions are investigated both for melody segmentation and for voice separation.

Main contribution. We conducted two experiments, each with twenty experts and twenty novices. In the first experiment, participants were asked to segment popular melodies. We found a high degree of intraclass agreement between the segmentation results of both novices and experts, but, surprisingly also for the combined class, justifying use of the results as basis for algorithm benchmarking. Evaluating algorithm output against participant data, we conclude that human output is closest to three of the chunking algorithms (Grouper, Information Dynamics and LBDM).

In the second experiment, participants judged which of several melody variants sounds most like the original by exhaustive pair-wise comparisons. According to the combined participants, the combined output of two algorithms (Skyline and SSA) shows the highest resemblance to the actual melody. Small differences were found between experts and novices.

Implications. We conclude that Grouper, Information Dynamics and LBDM are the best candidate algorithms for implementing melody segmentation in MIR systems. There is no reason to segment melodies in more than one way in order to accommodate different user groups. For voice separation, the situation is different, as the combined results of two algorithms were shown to agree best with experimental results, and differences were found between novice and expert performance. There is no immediate answer as to how voice separation should be implemented in a MIR system.

This paper investigates the generic problem of model selection in the specific context of Music Information Retrieval (MIR). In MIR research, strategies are developed for enabling automatic access to music collections. MIR systems enable music industry, music professionals and end users to search large quantities of musical audio or encoded scores (Casey et al. 2008). Three main components can be discerned in such systems: user interface, database, and similarity measure. The task of the last is to compare the user's query to the items in the database and to return a ranked list of search results.

The effective application of similarity measures typically requires musical works to be divided into more manageable units. Typke et al. (2007) provide an example of this. In their system, melodies are divided into a large number of overlapping chunks of 6-9 notes in order to make the system tolerant against melodic variation, tempo and pitch fluctuation. This results in a considerable increase of database size, and accidental matches of items that are not logical units from a musical perspective may negatively affect the ranking. Therefore it is important to look into alternative approaches to segmentation.

Human listeners generally possess two functions that allow them to process a continuous stream of music into smaller segments: the ability to perceive multiple, successive tones as one coherent melodic phrase (melody segmentation) and the ability to differentiate melody notes from accompaniment (voice separation). For both of these tasks, several algorithmic models have been proposed in the symbolic domain.

It seems reasonable to assume that a MIR system performs better when melody segmentation and voice separation are done by cognition-based methods than when a primarily computational approach to segmentation is employed. The question is then which one(s) to choose. Various authors have compared melody segmentation algorithms, but with inconclusive results. We are not aware of any previous evaluation of different voice separation algorithms.

Aim and contribution. This paper describes the first systematic evaluation of prominent computational methods for music segmentation against human performance, with the aim to answer two questions:

1. Is there enough agreement in human segmentation perception to function as a basis for measuring algorithm performance?
2. Which algorithm best models human segmentation?

These questions are investigated both for melody segmentation and for voice separation. We will show that for the former there is sufficient agreement and that there are three methods that perform better than the others. For the latter, there are differences between types of users, and the optimum performance may be reached using the results of two algorithms.

Organisation. First we discuss model selection in computational musicology. Then the algorithms that are used in the experiments are briefly described. This includes a summary of evaluation results from the literature. Next the experiments are described. Conclusions and perspectives for future research conclude this paper.

Model selection

The increasing number of computational models in music research calls for methods to evaluate and compare competing models. At present, there are no general strategies available for such comparisons. Since models dedicated to similar musicological questions are often very different in nature, it is difficult to find a common perspective that all models should be reconceived upon. For instance, in an attempt to compare different automatic rhythm description systems, Gouyon and Dixon (2005) concluded that 'there are no precise problem definitions or evaluation criteria, because rhythm description systems have been built for diverse applications using diverse data sets'. Temperley (2004) suggested such an evaluation system for rhythmic-metric models, which is based on the number of 'correct' answers obtained from a specific corpus. However, his system is only applicable to symbolic metrical models and hence not applicable to the audio-based models discussed by Gouyon and Dixon (2005). Honing (2006) on the contrary states that in the specific case of a computational model in music cognition the goodness of fit with the empirical data is not sufficient in order to test the validity of a model. Hence, the number of correct answers of the system should not be the only criterion of an evaluation system. He suggests to consider the degree of surprise in the predictions of the model as an important factor of the evaluation. Similarly, Volk (2005) proposes that surprising results of a computational model might lead to interesting insights into the investigated phenomenon and could therefore be as important as the number of results correctly (in accordance with the empirical data) produced.

In contrast to these very general considerations as to how to select between competing computational models, the current paper compares models on melody segmentation and voice separation in a specific research context. Our aim is to find the best candidate for incorporation into a MIR-system. In contrast to Temperley's evaluation model that derives the 'correct' answer directly from the music notation, there is no formal theory available from which the 'correct' segmentation could be determined. Hence, we compare the results produced by the models to those obtained in a human decision process in order to evaluate the model. However, our main focus here is not a general psychological validation of these models. By measuring the fit between the model and the empirical observation we provide the starting point for a more general verification of these models, as pointed out by Honing (2006). Moreover, we suggest specific model selection criteria that are important within the context of MIR.

Segmentation algorithms

Numerous algorithms have been proposed for both melody segmentation and voice separation tasks. The following overview describes only those approaches that we tested for this research. De Nooijer (2007) provides a fuller survey of known approaches.

Melody segmentation

Five melody segmentation methods were studied in our experiment. Their most important properties are shown in Table 1.

Temporal Gestalt units (TGU's). TGU's were introduced by Tenney and Polansky (1980); for the experiment Eerola and Toiviainen's implementation was used (2004). This model employs several 'measures of change': absolute pitch interval (API, in semitones) and inter-onset interval (IOI, in eighth notes) are used for this experiment. The distance between two events is the weighted sum of these measures. In our experiment, they receive equal weight. A boundary between so-called 'clangs' is constructed where a local maximum in the distance occurs. Each clang is then characterised by its onset time and average pitch. These values are submitted to the same procedure to create segment borders.

Local Boundary Detection Model (LBDM). LBDM (Cambouropoulos 1998, 2001) employs three parameters: API, IOI and offset-to-onset interval (OOI). These values are normalised and their weighted sum is calculated (w_{API} = 0.25; w_{IOI} = 0.50; w_{OOI} = 0.25 in our experiment). A degree of change is then calculated for each pair of successive intervals. The boundary strength of an interval is determined from its weight and its change degree to the preceding and following intervals. For details of the calculations see Cambouropoulos (1998, 2001). A boundary is created when this value lies above a certain threshold (values 0.4, 0.5 and 0.6 were used in the experiment, abbreviated LBDM4, LBDM5 and LBDM6).

Grouper (GRP). This method is based on metric information only (Temperley 2001; implementation by Sleator and Temperley n.d.). A gap score for each interval is calculated by taking the sum of IOI and OOI. Candidate boundaries are scores above a certain threshold. Candidate groups are given a penalty for their deviation from the ideal length, and for beginning on a different metrical position than the preceding group. The optimal segmentation is the one, which has the lowest sum of penalties of all possible solutions.

Melodic Similarity Model (MSM). This model, proposed by Ahlbäck (2004) combines bottom-up Gestalt-oriented principles such as similarity, proximity and good continuation, with a top-down analysis involving melodic parallelism and structure. The model provides a so-called section analysis as the result.

Name	Method	Features	Input	Output	Processing	Parameters
TGU's	Model-driven	API, IOI; other features can be added	MIDI	Graphic	Sequential	Weighing
LBDM	Model-driven	API, IOI, OOI	MIDI	Graphic	Sequential	Threshold, weights
GRP	Model-driven	IOI, OOI, Meter	Meter (from Melisma)	Text	Non-sequential	Threshold, ideal length, length penalty, metrical penalties
MSM	Model-driven	Pitch, IOI, OOI, Metric	MIDI	Graphic	Non-sequential	(None)
ID	Data-driven	Pitch, duration, onset, key	Humdrum	Humdrum	Sequential	(Unspecified)

Table 1. Properties of the selected melody segmentation algorithms. Abbreviations are explained in the main text.

Information Dynamics (ID). This model creates boundaries at points of expectancy violation and predictive uncertainty (Pearce and Wiggins 2006; Potter et al. 2007). The assumption is that listeners perceive a boundary where the context fails to inform about forthcoming events. The model has a long-term and a short-term memory model. The former is an *n*-gram model that was trained on c. 900 tonal melodies. The latter has no prior knowledge but learns from the current piece. Boundaries are calculated using the entropy of events as a measure for the uncertainty of the model's expectation.

Voice separation

Five voice separation methods were studied in our experiment. Their most important properties are shown in Table 2.

Skyline. For each onset time in a polyphonic piece, this algorithm determines the highest sounding note (Clausen n.d.). All other notes are discarded. Thus, a monophonic melody is created consisting solely of the highest-pitched notes. Uitdenbogerd and Zobel (1998) describe several variants that deal with MIDI-specific issues.

Nearest Neighbour (NN). This algorithm creates a set of melodies out of a polyphonic piece by joining each note to the immediately preceding note that is closest to it in pitch (Clausen n.d.).

Streamer. This model (Temperley 2001; implementation by Sleator and Temperley n.d.) represents a polyphonic piece in quantised piano-roll notation. Melodies are formed by finding connections between notes that satisfy four wellformedness rules and five preference rules. Penalties for violating the latter act as parameters of the model.

Voice Separation Analyzer (VoSA). In this method, developed by Chew and Wu (2004), a polyphonic stream is divided into short monophonic fragments. These are grouped in simultaneously sounding *contigs*. Within a contig, the number of fragments is constant at any time. Fragments within a contig satisfy a number of requirements that derive from the pitch proximity principle and the avoidance of stream-crossing principle. The contigs with the maximum number of voice fragments are used as starting points for iteratively connecting contigs into larger units.

Name	Features	Input	Output	Processing	Parameters
Skyline	Pitch, onset, duration	MIDI	MIDI	Sequential	None
NN	Pitch, OOI	MIDI	MIDI	Sequential	None
Streamer	Pitch, onset, duration	Meter (from Melisma)	Text	Non-sequential	Max. voices, max. collisions, penalties for violating preference rules
VoSA	Pitch, onset, duration	MIDI	MIDI, Graphic	Non-sequential	None
SSA	Pitch, onset	MIDI	MIDI	Sequential	Penalties for starting notes, ending notes, inserting rests and leap size

Table 2. Properties of the selected voice separation algorithms. Abbreviations are explained in the main text.

Stream Separation Algorithm (SSA). Madsen and Widmer (2006) present an algorithm inspired by Streamer but particularly intended for online use. It considers groups of notes that begin approximately at the same time. Sustained notes are not included. Groups are processed sequentially. In assigning notes to a voice the following requirements apply: (1) each note must be assigned to exactly one voice and (2) overlapping notes are not allowed. Also, leaps, number of voices and number of rests within a voice must be minimized. Voice crossing is not prohibited but is expensive as it generally involves multiple leaps.

Previous evaluations

Figure 1 provides an overview of tests and claims about the performance of melody segmentation algorithms that have been published in the literature. For voice separation algorithms, the only comparison we know of is one of variants of the Skyline algorithm (Uitdenbogerd and Zobel 1998).

Test results

Cambouropoulos (2006)	PAT > LBDM
Thom et al. (2002)	Grouper > LBDM
Ferrand et al. (2003)	MDSM > LBDM
Ahlbäck (2004)	MSM > LBDM
Ahlbäck (2004)	MSM > Grouper

Claims

Bod (2002)	DOP > Gestalt
Meredith (2002)	LBDM > Grouper

Figure 1. Tests and claims about the performance of melody segmentation algorithms. The operator '>' should be read as 'performs better than.' Algorithms not discussed in the main text are Data Oriented Parsing (DOP, Bod 2002), the Melodic Density Segmentation Model (MSDM, Ferrand et al. 2003) and the Pattern Boundary Detection Model (PAT, Cambouropoulos 2006).

The tests and claims are summarised in Figure 2. Most of these were published by the authors of the algorithms, and generally a comparison is made to only one other algorithm. There is no consistency between experiments in method, criteria and circumstances. Therefore it is impossible to create a reliable overview out of these data, as is evident from the different judgments of the relative merits of LBDM and Grouper. Instead, we offer a systematic and independent evaluation of all algorithms that were available to us.

<table>
<tr><td>DOP</td><td>&gt;</td><td>MSM</td><td>&gt;</td><td>Grouper</td><td>&gt;</td><td rowspan="4">LBDM</td></tr>
<tr><td rowspan="2"></td><td rowspan="2"></td><td colspan="3">MDSM</td><td>&gt;</td></tr>
<tr><td colspan="3">Grouper</td><td>&lt;</td></tr>
<tr><td></td><td></td><td colspan="3">PAT</td><td>&gt;</td></tr>
</table>

Figure 2. Summarized test and claims from Figure 1. Claims are shaded in grey.

Experiments

Melody segmentation

Method. In the melody segmentation experiment, 40 participants were asked to chunk 10 melodies. The group consisted of 20 novices and 20 experts (conservatory students, (semi-)professional musicians, etc.) We performed cluster analysis to determine the division of participants between groups. Because of the inclusion of novices, we avoided using any musical terminology in our instructions, as this might give the experts an advantage over the novices. For example, we used the term 'sentence' instead of 'phrase,' as the term phrase may indicate a piece of a certain length to those with formal music education, while anyone with or without musical education has a similar conception of the term sentence.

Participants were instructed to place segment boundaries at those locations where they heard a melodic sentence ending, which automatically implies the start of a new sentence. For this purpose, they used the Sony Sound Forge program, which allowed them to place a boundary while the melody was playing. The interface also allowed the participants to correct small errors caused by latency, as it displays a visual representation of the audio file.

Materials. The 10 melodies were selected from a collection of popular music in MIDI format that was gathered from the Internet by a crawler. Each melody is between 25 to 30 seconds long, which is enough to be able to distinguish several segments.

Results. The data gathered from the experiment were converted to note lists, in which marker placement for each participant is indicated by either a 1 (marker placed) or a 0 (no marker placed). From these note lists, we first determined the degree of similarity for inter-participant segmentation results, using Fleiss and Cohen's (1973) test for interassessor (or intraclass) coherence. The results show that within both groups of participants, the coherence is very high—respectively α = .9675 for novices and α = .9902 for experts. The coherence between all participants of both groups is also high, α = .9864. Thus, we can state that there is enough overlap between segmentation amongst participants to function as a benchmark for algorithms.

We compared participant output to the following algorithms: TGU's, Grouper, MSM, LBDM (in three variants, LBDM4, LBDM5 and LBDM6, as described above) and ID. We used the Wilcoxon signed rank test to determine whether or not there are significant differences in segmentation results between algorithms, and between each algorithm and the groups of participants. The results are presented in Table 3. Significant differences are marked bold.

		Algorithms						
		TGU	**GRP**	**MSM**	**LBDM4**	**LBDM5**	**LBDM6**	**ID**
Algorithms	**TGU**		.160	.239	.071	.732	.450	**.007**
	GRP	.160		**.007**	.683	.108	**.003**	.157
	MSM	.239	**.007**		**.003**	.117	.655	**.000**
	LBDM4	.071	.683	**.003**		**.003**	**.000**	.276
	LBDM5	.732	.108	.117	**.003**		**.014**	**.013**
	LBDM6	.450	**.003**	.655	**.000**	**.014**		**.000**
	ID	**.007**	.157	**.000**	.276	.276	**.000**	
Participants	**NOV**	**.017**	.085	**.000**	.420	**.001**	**.000**	.250
	EXP	.092	.985	**.002**	.643	.102	**.003**	.072
	ALL	**.014**	.128	**.000**	.568	**.003**	**.000**	.220

Table 3. P-values indicating differences between algorithms and participants. Significant scores are shown in bold print. Abbreviations: NOV=novices, EXP=expert, ALL=all participants.

Several conclusions can be drawn from these results. The output from the TGU algorithm does not differ significantly from any of the other algorithms except ID. The output of TGU's is most similar to the output of LBDM5 (and vice versa). The three different parameterizations of LBDM lead to three significantly different results. Hence, the parameters strongly influence the result of this model instead of creating slight variations within the results. The output from Grouper and LBDM5 differs from two other algorithms, and the output from MSM, LBDM4 and ID differs from three other algorithms. LBDM6 differs significantly from four algorithms. The most similar output to Grouper is that of LBDM4, LBDM6 is most similar to MSM. ID is the method that is differs most from the others. It most resembles LBDM4, but the converse is not true.

When compared to participants' output, ID, Grouper and LBDM4 show no significant differences with any group of participants, whereas the other algorithms differ significantly from at least two or three groups of participants.

Evaluation. We can conclude that, in our experiment segmentation does not appear to be an ambiguous task, contrary to what was concluded by Thom et al. (2002) and Ahlbäck (2004). The material used might be the cause of this: previous researches have often used classical music, but we chose to use popular melodies. These seem to contain clear cues about segment boundaries.

The results are promising for the implementation of a MIR system for popular melodies: since there are no significant differences in human segmentation, one accurate segmentation of the tunes in the corpus of a MIR system will suffice. In other words, it is not necessary to offer different segmentation options to, for example, people with different levels of music education.

Based on the test results, we can positively answer our first research question. Since there is a high intraclass agreement between experts and novices, their melody segmentation results can function as a basis for algorithm benchmarking.

In addition, we can conclude that the results of LBDM4, Information Dynamics and Grouper's neither differ significantly from the participants' results, nor from each other.

Hence, based on the results of this experiment, none of the three models can be selected as the best one. Thus, LBDM4, Information Dynamics and Grouper are plausible candidates for implementation.

Voice separation

Method. For the voice separation experiment, the same 40 participants were presented with a short musical sample and were then asked to determine through pair-wise comparison of two monophonic lines ('variants') which line best resembled the melody heard in the original musical piece. Participants were allowed to listen to the pairs as many times as they found necessary. This process was then repeated for all 8 samples.

Materials. Eight polyphonic musical samples were extracted from the same collection as in the first experiment. Each sample is approximately 10 seconds long, which is enough to be able to distinguish melody and harmony. Melody and harmony notes are of the same timbre, since algorithms do not take timbre into account for separation; therefore, participants could have an advantage when hearing different timbres.

For each sample, a set of melodic variants was created. The number of variants in a set ranges from 4 to 8, and was determined by the output of the algorithms. Each sample was processed using the voice separation algorithms. The resulting melodies were used as variants in the experiment. In cases where several algorithms produced the same variant, we have included this variant only once in the set. For algorithms that output multiple monophonic lines, the one that most accurately represents the melody was selected.

Furthermore, each set includes a manually extracted variant representing the melody as it was perceived by the first author, and two variants consisting of a randomly selected combination of harmony and melody notes from the original.

Results. Because of the complexity of the dataset, we applied our statistical measure, Cronbach's alpha for inter-rater coherence, only to the highest ranked variant for each set by each participant. The resulting values are α_{nov} = 0.8966, α_{exp} = 0.9389 and α_{all} = 0.9230. Since these values are all high (taken into account that a1 value of $\alpha > 0.70$ is considered acceptable), we conclude that the inter-rater coherency between all groups is high, and that coherence amongst experts is somewhat higher than coherence amongst novices. After having established a high inter-rater coherence, we calculated rankings based on the experts' results (Table 4).

Rank	Sample 1	Rank	Sample 2	Rank	Sample 3	Rank	Sample 4	Rank	Sample 5	Rank	Sample 6	Rank	Sample 7	Rank	Sample 8
1	VoSA	1	*Author*	1	SSA	1	Skyline	1	*Rand1*	1	*Author*	1	*Author*	1	VoSA
2	*Rand2*		VoSA	2	*Author*	2	NN	2	*Author*		Madsen		Skyline		SSA
3	*Rand1*		Skyline	3	Streamer	3	*Rand1*	3	Skyline		Skyline		VoSA	2	Streamer
4	Streamer	2	SSA	4	VoSA	4	*Author*	4	Streamer	2	VoSA		Streamer	3	Skyline
5	SSA	3	Streamer	5	Skyline	5	VoSA	5	SSA	3	NN		SSA	4	*Rand1*
6	Skyline	4	NN	6	*Rand1*	6	SSA	6	VoSA	4	*Rand2*	2	*Rand2*	5	NN
	NN	5	*Rand1*	7	NN	7	Streamer	7	NN	5	Streamer	3	NN	6	*Author*
7	*Author*	6	*Rand2*	8	*Rand2*			8	*Rand2*	6	*Rand1*	4	*Rand1*	7	*Rand2*

Table 4. Expert rankings of the variants per sample. Here and in the following tables, variants generated by the algorithms are printed in Roman type; those created by the researchers are printed in italic. *Author* is the optimum variant; *Rand1* and *Rand2* are the variants that were created by randomly selecting notes from the sample.

When examining these rankings, one can see that the results differ rather much per sample. For example, when reviewing the rankings throughout all results of the SSA algorithm, one observes that it ranks in the top two for melodies 2, 3, 6, 7 and 8, but also that it occupies a mere fifth place for melody 1 and 5, and a sixth place for melody 4. A similar pattern occurs for VoSA and Streamer. Thus, in order to get a clearer view of which algorithm's variants are generally ranked higher, we determined how many times each algorithm was ranked at the first place. This tells us which algorithm is able to produce accurate results on most of the samples. The results are shown in Table 5.

Here, we see that three algorithms' variants are ranked at the first place on four occasions: VoSA, SSA and Skyline. Thus, we can assume that any of these three algorithms provide the most accurate results. However, each algorithm's results are only accurate for four of the eight samples we investigated.

The accumulated novice results are shown in Table 6. Here, too, the rankings of the algorithms differ considerably between samples.

The number of times each algorithm's variant was ranked at the first place by novices is shown in Table 5. Here, we see that the variants of algorithms SSA and Skyline each are ranked first three times. This result matches the experts' result; however, the novices chose VoSA's variants less often than the experts: the novices only ranked it highest in two cases. The first author's melody however was ranked highest for five out of eight samples.

Experts		**Novices**	
Variant	**# 1st**	**Variant**	**# 1st**
VoSA	4	*Author*	5
SSA	4	SSA	3
Skyline	4	Skyline	3
Author	3	VoSA	2
Rand1	1	*Rand1*	1
Streamer	1	Streamer	1
Rand2	0	*Rand2*	0
NN	0	NN	0

Table 5. Number of times a variant is ranked first place by experts (left) and novices (right).

Evaluation. Both novices and experts prefer the melodies generated by the SSA and Skyline algorithms. Experts additionally prefer the VoSA variants equally often, when solely considering variants generated by algorithms. However, novices prefer the melody hand-extracted by the first author to computer-extracted melodies.

Rank	**Sample 1**	**Rank**	**Sample 2**	**Rank**	**Sample 3**	**Rank**	**Sample 4**	**Rank**	**Sample 5**	**Rank**	**Sample 6**	**Rank**	**Sample 7**	**Rank**	**Sample 8**
1	*Rand1*	1	*Author*	1	*Author*	1	Skyline	1	*Author*	1	*Author*	1	*Author*	1	SSA
2	*Rand2*		VoSA	2	NN	2	NN	2	SSA		Skyline		Skyline	2	*Author*
3	VoSA		Skyline	3	SSA	3	*Author*	3	*Rand1*		SSA		VoSA	3	Streamer
4	Streamer	2	Streamer	4	Streamer	4	SSA	4	Skyline	2	VoSA		Streamer		VoSA
5	SSA	3	SSA	5	Skyline	5	*Rand1*	5	Streamer	3	NN		SSA	4	Skyline
6	Skyline	4	NN	6	VoSA	6	VoSA	6	VoSA	4	*Rand2*	2	*Rand2*	5	NN
	NN	5	*Rand1*	7	*Rand1*	7	Streamer	7	*Rand2*	5	Streamer	3	NN	6	*Rand2*
7	*Author*	6	*Rand2*	8	*Rand2*			8	NN	6	*Rand1*	4	*Rand1*	7	*Rand1*

Table 6. Novice rankings of the variants per sample. Variants generated by the algorithms are printed in Roman type; those created by the researchers are printed in italic. *Author* is the optimum variant; *Rand1* and *Rand2* are the variants that were created by randomly selecting notes from the sample.

None of the algorithms is able to end up at the highest rank for more than half of the melodies. Therefore, we might consider offering multiple solutions. When we offer both SSA and Skyline as voice separation algorithms, they would together deliver the most resembling (or: highest ranked) melodies in five out of eight cases according to novices' rankings, and six out of eight cases according to experts' rankings. Other combinations of algorithms yield lower scores.

Conclusion and future work

The evaluation of computational models based on the measure of fit to empirical data we describe in this paper obtained different kinds of results for the melody segmentation and voice separation tasks. The results of the human melody segmentation experiment showed sufficient agreement among the participants to function as a basis for measuring algorithm performance. Three algorithms were close to the human performance of this task: Grouper, Information Dynamics and LBDM4. Among the voice separation algorithms, none of the models came close to human performance; therefore, combining the results of two algorithms, SSA and Skyline is suggested. Furthermore, novices and experts differed in their evaluation of the results of the voice separation algorithms. Hence, there is no immediate answer which voice separation model should be selected.

The evaluation of computational models in this paper does not aim at a general psychological validation of these models but is a first step for a model selection for a MIR system. In order to select the most appropriate model among the three highest scoring candidates for melody segmentation, two additional criteria need to be applied, namely the nature of the corpus and the requirements of an efficient implementation.

Potter et al. (2007) claim that the data-driven Information Dynamics model represents the 'typical human Western musical experience'. It would therefore seem to be generally applicable and furthermore not to require any further training. We do not know how musical features are weighed in this model. However, if we know certain properties of the repertoire, algorithms may be selected on the basis of this knowledge. In particular, if the corpus contains rhythmically strong tunes, Grouper might be the most appropriate algorithm, while for tunes with less differentiated rhythms the pitch-based LBDM4 is likely to produce better results. Also, as one can discern a trend for Information Dynamics' results to differ from Grouper's and LBDM4's, it seems likely that in a MIR system Information Dynamics may produce different, but equally good retrieval output as the others. In other words, one would expect the retrieval results to be complementary.

For an implementation in a MIR system the output as well as the input of the models have to be considered. Most algorithms offer a graphical representation as output, which allows a quick interpretation of the results by humans (see Tables 1 and 2). However, such output is not directly usable in MIR systems, which need the data generated by the segmentation method to be processed in a symbolic format. Some of the algorithms would therefore need to be re-implemented.

Another consideration for implementation is computational efficiency. We did not study the computational properties of the methods in detail, yet it seems likely that methods that process the data sequentially, such as Information Dynamics, LBDM and Skyline, possess a better time complexity than methods that process the music in several iterations, such as Grouper, Streamer and VoSA, especially if the number of iterations depends on the length and/or number of polyphonic parts of the music.

In order to test our underlying hypothesis, namely that a MIR system performs better when melody segmentation and voice separation are done by cognition-based methods, we need to do another series of experiments. For these, the best performing algorithms will be implemented in a MIR system. They will be employed to segment both the queries and the dataset. The retrieval performance of this system will be compared to that of another version of the MIR system, which employs the same similarity measure but segments the data with a method that is not cognition-based. The comparison of the results of both versions of the MIR-system will determine whether or not one system performs significantly better than the other.

Acknowledgements. Our gratitude goes out to the authors of the algorithms who have helped obtaining output data and the people who have participated in the experiments. We thank Bas de Haas for his comments on the draft version of this paper.

References

Ahlbäck, S. (2004). *Melody beyond notes: A study of melody cognition.* Göteborg: Göteborg University.

Bod, R. (2002). Memory-based models of melodic analysis: Challenging the Gestalt principles. *Journal of New Music Research* 31, 27-37.

Cambouropoulos, E. (1998). Musical parallelism and melodic segmentation. In: *Proc. of the XII Colloquium of Musical Informatics,* Gorizia, Italy, 111-114.

Cambouropoulos, E. (2001). The Local Boundary Detection Model (LBDM) and its application in the study of expressive timing. In: *Proc. of the International Computer Music Conference,* Havana, Cuba.

Cambouropoulos, E. (2006). Musical parallellism and melodic segmentation: A computational approach *Music Perception* 23 (3), 249-267.

Casey, M.A., Veltkamp, R.C., Goto, M., Leman, M., Rhodes, C., Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE* 96(4), 668-696.

Chew, E., Wu, X. (2004). Separating voices in polyphonic music: a contig mapping approach. In: *Proc. of Computer Music Modeling and Retrieval 2004,* Esbjerg, Denmark.

Clausen, M. (n.d.) *Melody extraction.* Retrieved 4-7-2006 from: www-mmdb.iai.uni-bonn.de/forschungprojekte/midilib/english/skydemo.html

Eerola, T., Toiviainen, P. (2004). *The MIDI Toolbox: MATLAB tools for music research.* Retrieved 31-5-2007 from www.jyu.fi/musica/miditoolbox/

Ferrand, M., Nelson, P., Wiggins, G. (2003). Memory and melodic density: a model for melody segmentation. In: *Proc. of the XIV Colloquium on Musical Informatics*, Firenze, Italy.

Fleiss, J. L., Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33, 613-619.

Gouyon, F., Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal* 29, 34-54.

Honing, H. (2006). Computational modeling of music cognition: A case study on model selection. *Music Perception* 23, 365-376.

Madsen, S. T., Widmer, G. (2006). Separating voices in MIDI. In: *Proc. of the 7th International Conference on Music Information Retrieval,* Victoria, Canada, 8-12.

Meredith, D. (2002). Review of David Temperley, *The cognition of basic musical structures*. *Musicae Scientiae* 6(2), 287-302.

Nooijer, J. de. (2007). *Cognition-based segmentation for music information retrieval systems.* Master's thesis, Utrecht University.

Pearce, M.T., Wiggins, G.A. (2006). The information dynamics of melodic boundary detection. In: *Proceedings of the Ninth International Conference on Music Perception and Cognition*, Bologna, 860-865.

Potter, K., Wiggins, G.A., Pearce, M.T. (2007). Towards greater objectivity in music theory: Information-dynamic analysis of minimalist music. *Musicae Scientiae* 11(2), 295-322.

Sleator, D. D. K., Temperley, D. (n.d.) *The Melisma Music Analyzer.* Retrieved 15-1-2007 from: www.link.cs.cmu.edu/music-analysis/

Temperley, D. (2001). *The cognition of basic musical structures.* Cambridge, MA: MIT Press.

Temperley, D. (2004). An evaluation system for metrical models. *Computer Music Journal* 28, 28-44.

Tenney J., Polansky, L. (1980). Temporal Gestalt perception in music. *Journal of Music Theory* 24, 205-241.

Thom, B., Spevak, C., Höthker, K. (2002). Melodic segmentation: Evaluating the performance of algorithms and musical experts. In: *Proc. of the International Computer Music Conference.* Göteborg, Sweden.

Typke, R., Wiering, F., Veltkamp, R.C. (2007) Transportation distances and human perception of melodic similarity. *Musicae Scientiae Discussion Forum* 4A, 153-181.

Uitdenbogerd, A. L., Zobel, J. (1998). Manipulation of music for melody matching. In: *Proc. of the ACM Multimedia Conference '98,* Bristol, UK., 235-240.

Volk, A. (2005). Coffee bean (and other) models about the metric structure of music. In: Sebastian Bab et al. (Eds.): *Models and Human Reasoning*, W&T Verlag, Berlin.