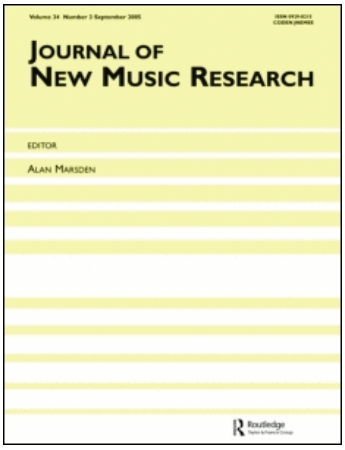
This article was downloaded by: *[Aalborg University]*
On: *11 May 2011*
Access details: *Access Details: [subscription number 912902580]*
Publisher *Routledge*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of New Music Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713817838

Cognition-based Segmentation for Music Information Retrieval Systems

Frans Wiering[a]; Justin de Nooijer[b]; Anja Volk[a]; Hermi J. M. Tabachneck-Schijf[a]
[a] Utrecht University, The Netherlands [b] Fortis ASR, Utrecht, The Netherlands

Online publication date: 30 November 2009

To cite this Article Wiering, Frans , de Nooijer, Justin , Volk, Anja and Tabachneck-Schijf, Hermi J. M.(2009) 'Cognition-based Segmentation for Music Information Retrieval Systems', Journal of New Music Research, 38: 2, 139 — 154
To link to this Article: DOI: 10.1080/09298210903171145
URL: http://dx.doi.org/10.1080/09298210903171145

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Journal of New Music Research
2009, Vol. 38, No. 2, pp. 139–154

Routledge
Taylor & Francis Group

Cognition-based Segmentation for Music Information Retrieval Systems

Frans Wiering[1], Justin de Nooijer[2], Anja Volk[1], and Hermi J.M. Tabachneck-Schijf[1]

[1]Utrecht University, The Netherlands; [2]Fortis ASR, Utrecht, The Netherlands

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Abstract

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 computational models 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.

We conducted two experiments, each with twenty experts and twenty novices. In the melody segmentation experiment, we found a high agreement between the participants. Evaluating algorithm output against participant data, we conclude that human output cannot be distinguished from three of the segmentation algorithms (Grouper, IDyOM and LBDM). For voice separation—which we evaluated by means of a melody identification task—the situation is different, as the combined results of two algorithms (Skyline and SSA) were shown to agree best with experimental results, and differences were found between novice and expert performance. Several other model selection criteria besides performance are discussed in conclusion.

1. Introduction

This paper evaluates computational models of melody segmentation and voice separation 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 the 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 overlapping chunks of 6–9 notes in order to make the system robust against melodic variation, tempo and pitch fluctuation. However, this also results in a considerable increase of database size; in addition, 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 also process a continuous stream of music into more manageable units. We mention two processes: the ability to perceive groups of successive tones as coherent melodic phrases (melody segmentation) and the ability to segregate the sound into a number of simultaneous streams (Bregman, 1990, ch. 5), typically one or two melodies and accompaniment (voice separation). For both of these tasks, several algorithmic models have been proposed in the symbolic domain.

As humans are the users of MIR systems, it seems reasonable to assume that an MIR system would provide better search results when melody segmentation and

Correspondence: Frans Wiering, Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands. E-mail: frans.wiering@cs.uu.nl

DOI: 10.1080/09298210903171145 © 2009 Taylor & Francis

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

voice separation are done by human cognition-based methods than when a primarily computational approach to segmentation is employed. Since a number of melody segmentation and voice separation algorithms have been developed that claim to model the human cognition of this task, the question is then which one(s) to choose. Various authors (listed in Section 3.3) have compared melody segmentation algorithms and one study has explored voice separation methods, but no firm conclusions can be drawn from these publications, since the results are often in conflict with each other and the methods used are not consistent (see Section 3.3).

This paper describes a systematic evaluation of prominent computational methods for music segmentation against human performance, with the aim of answering 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?

In order to be able to answer the above questions, we conducted two experiments in which participants were asked to carry out melody segmentation and voice separation on musical samples. The specific voice separation task we studied was melody identification, the separation of the principal melody from the accompanying melodies or chords. We developed novel methods that do not require any formal music training, so that both experts and novices could participate in the experiments. Thus, we are more likely to approach the actual target audience of an MIR system, which does not consist of musical experts only.

There have been several earlier evaluations of melody segmentation algorithms; this one however seems to be the first larger one which is not performed by the author of an algorithm that is part of the experiment. Concerning voice separation algorithms, very few evaluations have been published before. Neither the melody segmentation algorithm evaluations nor the voice separation algorithm evaluations described previously in the literature involve rigorous testing against human performance.

Contribution. We will show that for melody segmentation there is sufficient agreement among participants and that three algorithms cannot be distinguished from human segmentation. Based on the actual segmentations these algorithms produce, two of them can be considered more suitable for implementation. For voice separation, there are differences between types of users, and the optimum performance may be reached using the results of two algorithms and customizing the MIR system for experts and novices.

In Section 2, we discuss model selection in computational musicology. The algorithms that are used in the experiments are described in Section 3. This includes a summary of evaluation results from the literature. Section 4 provides a description of the experiments. Conclusions and suggestions for future research are presented in Section 5.

2. Model selection

The increasing number of computational models in music research calls for methods to evaluate and compare competing models. However, models for similar musicological tasks may have been developed in very different contexts, and thus be very different in nature. Therefore, it is often difficult to find a common perspective from which to compare them. In attempting to compare 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 an evaluation method for rhythmic-metrical models that is based on the number of 'correct' answers—that is, inferred by competent listeners—obtained from a specific corpus. However, his solution is only applicable to symbolic metrical models and not to the audio-based models discussed by Gouyon and Dixon (2005). Honing (2006) on the other hand states that in music cognition the goodness of fit between a computational model and the empirical data is not a sufficient criterion for the validity of the model, and he suggests the degree of surprise in the predictions of the model[1] as an important additional criterion. 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 amount of correct results produced.

The current paper compares computational models of melody segmentation and voice separation to human models in a specific research context. Our aim is to find the best candidate for incorporation into an MIR system. However, there is no undisputed ground truth available from which the 'correct' human segmentation could be determined. Therefore, in order to evaluate the model, we will compare the results produced by the models to those obtained in a human decision process. However, our main focus here is not the 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 requested by Honing (2006). Moreover, we will suggest additional criteria that are important within the context of MIR.

[1]A model is considered surprising if the predicted outcomes are a small fraction of the plausible outcomes.

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

3. Segmentation algorithms

Numerous algorithms have been proposed that claim to model the human melody segmentation and voice separation tasks. The following overview describes only those approaches that we tested for this research. Approaches that we were unable to include in the experiments because no program code was available, or because they needed a separate training corpus, or because they came to our attention only after the experiments, include Ferrand et al. (2003), Frankland and Cohen (2004), Juhász (2004), Ke et al. (2004) and Cambouropoulos (2006) for melody segmentation, and Szeto and Wong (2003), Kirlin and Utgoff (2005), Karydis et al. (2007) and Jordanous (2008) for voice separation. Rafailidis et al. (2008) present a wholly different viewpoint, creating by means of a single method 'stream segments' that are both horizontally and vertically separated from each other. Therefore it is difficult to compare the performance of their method to that of the ones discussed below.

3.1 Melody segmentation

The most important properties of the six melody segmentation methods studied here are shown in Table 1.

3.1.1 Temporal Gestalt units

Temporal Gestalt units (TGUs; Tenney & Polansky, 1980) model the Gestalt principles of proximity and similarity. The general idea is that a small distance in a musical dimension implies continuity, and that a large distance implies a border. 'Measures of change' are defined in four dimensions: pitch, duration, dynamics and timbre. The distance between two events is the weighted sum of these measures. For this experiment, absolute pitch interval (API, in semitones) and inter-onset interval (IOI, in eighth notes) are used. They receive equal weight. A boundary between so-called 'clangs' is constructed where a local maximum in the distance occurs. Each clang is then characterized by its onset time and average pitch. These values are submitted to the same procedure to create segment borders. For our experiment Eerola and Toiviainen's (2004) implementation was used. We analyzed both clang and segment borders.

3.1.2 Local boundary detection model

The local boundary detection model (LBDM; Cambouropoulos, 1998, 2001) employs three features: API, IOI and offset-to-onset interval (OOI). For each interval between two successive events, the boundary strength for each of the three features is calculated (details in Cambouropoulos, 1998, 2001). Then their weighted sum is calculated ($w_{\mathrm{API}} = 0.25$; $w_{\mathrm{IOI}} = 0.50$; $w_{\mathrm{OOI}} = 0.25$ in our experiment) and normalized in the range [0, 1]. We used Eerola and Toiviainen's (2004) implementation.

From these boundary strengths we derived boundaries in two different ways. The first is to create a boundary when the LBDM lies above a certain threshold (values 0.2, 0.3, 0.4 and 0.5 were used in the experiment, abbreviated LBDM2 to LBDM5). The second is to define a boundary as a local maximum. We used several different context sizes for this, specified by the number of notes before and after the candidate boundary. In order to create a boundary, all of the context notes must have a lower LBDM value than the boundary note. We tested contexts with an equal number of notes before and after the boundary (labelled LBDMmax2-2 to LBDMmax6-6) and contexts with one more note before the boundary than after it (labelled LBDMmax3-2 to LBDMmax5-4).

3.1.3 Grouper

Grouper (GRP) employs temporal information only (Temperley, 2001; implementation from Melisma by

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

Name	Method	Features	Input	Output	Processing	Parameters
TGUs	Model-driven	API, IOI; other features can be added	MIDI	Text, Graphic	Sequential	Weighting
LBDM	Model-driven	API, IOI, OOI	MIDI	Text, Graphic	Sequential	Weighting
GRP	Model-driven	IOI, OOI, Metre	MIDI Note list, Beat list	Text, Graphic	Non-sequential	Optimal length, length penalty, parallelism penalties
MSM	Model-driven	Pitch, IOI, OOI, Metre	MIDI	Graphic	Non-sequential	(None)
DOP	Data and model-driven	Pitch, duration	MIDI	Text, Graphic	Non-sequential?	(Unspecified)
IDyOM	Data-driven	Pitch, duration, onset, key	Humdrum	Humdrum	Sequential	(Unspecified)

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Sleator & Temperley, n.d.). Three preference rules are defined. The Gap Rule states that boundaries occur preferably at large IOIs or OOIs. The Phrase Rule sets the preferred phrase length at a certain number of notes. The Metrical Parallelism Rule states that successive groups preferably begin at the parallel points in the metrical structure. Grouper calculates a gap score for each pair of successive notes by taking the sum of IOI and OOI. Phrases receive a bonus for the gap scores at their boundary, but a penalty for the deviation from the ideal length, and for beginning on a different metrical position than the preceding group. The optimal segmentation is the one that has the lowest sum of penalties of all possible solutions.

We varied the optimum phrase length, using values of 6, 8 and 10 (labelled Grouper6, Grouper8 and Grouper 10). The metrical structure of a melody can be determined using the Metre tool from Melisma, but Grouper can also be applied without this information. In our experiment we have done both. Variant methods using metre information are labelled Grouper6mtr, Grouper8mtr and Grouper10mtr. For other settings we used Melisma's defaults as their individual and combined effects are not easy to interpret.

3.1.4 Melodic similarity model

The melodic similarity model (MSM), 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. It is not entirely clear how these principles interact in the implementation. The model provides a so-called section analysis as the result. Segmentation is hierarchical. We evaluated the lowest two levels of the hierarchy separately. Moreover, segmentations can be made using either a start-oriented or an end-oriented interpretation. These have been included in the evaluation as well.

3.1.5 Data oriented parsing

Data oriented parsing (DOP) is a memory-based approach that uses the grouping structure of previously encountered pieces to determine the segmentation of a new piece (Bod, 2002). The frequencies of the fragments in a corpus are used to determine the new analysis. For this, a Markov grammar of the corpus is created. In the experiments described by Bod, a history of four notes was used. This model is then extended by a melodic grammar that describes the phrase structure. The model was tested on a subset of the Essen Folksong Collection, using a training set of 5251 melodies and a test set of 1000 melodies.

For our experiment Eerola and Toiviainen's (2004) implementation was used. Boundary scores are calculated between 0 and 1. We used the same methods to determine the actual boundaries as for LBDM. The absolute thresholds used are 0.5, 0.55, 0.6, 0.65 and 0.7 (abbreviated DOP5 to DOP7); the same contexts were used as for LBDM (abbreviated DOPmax2-2 to DOPmax6-6 and DOPmax3-2 to DOPmax5-4).

3.1.6 Information dynamics of music

Information dynamics of music (IDyOM) creates boundaries at points of expectancy violation and predictive uncertainty (Pearce & 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.

3.2 Voice separation

The most important properties of the five voice separation methods studied here are shown in Table 2.

3.2.1 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.

3.2.2 Nearest neighbour

This nearest neighbour algorithm (NN) 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.).

3.2.3 Streamer

Streamer (Temperley, 2001; implementation by Sleator & Temperley, n.d.) represents a polyphonic piece in quantized piano-roll notation. Melodies are formed by finding connections between notes that satisfy the well-formedness and preference rules. The well-formedness rules state that notes of a melody must be temporally connected, that within a melody only one note can occur at a time, that melodies do not cross, and that all notes must be entirely included in a melody. Preference rules state that large melodic leaps should be evaded, that the number of voices and the number of rests within

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

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	Metre (from Melisma) or MIDI	Text, Graphic	Non-sequential	Maximum number of voices, maximum number of 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

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

a melody should be minimized, and that inclusion of a note in more than one melody should be avoided. Penalties for violating the preference rules act as parameters of the model. The preferred voice separation solution is the one with the lowest total penalty.

3.2.4 Voice separation analyzer

The voice separation analyzer (VoSA), developed by Chew and Wu (2004), divides a polyphonic stream 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.

3.2.5 Stream separation algorithm

The stream separation algorithm (SSA) by Madsen and Widmer (2006) was inspired by Streamer but is 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. In this respect the model is stricter than Streamer. 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.

3.3 Previous evaluations

Table 3 provides an overview of tests and claims about the performance of melody segmentation algorithms that have been published in the literature. The tests and claims are summarized in Table 4. 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 3. 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 quantified Generative Theory of Tonal Music (GTTM; Frankland & Cohen, 2004), the Melodic Density Segmentation Model (MSDM; Ferrand et al., 2003) and the Pattern Boundary Detection Model (PAT; Cambouropoulos, 2006).

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
Bruderer and McKinney (2008)	LBDM > GTTM > Grouper
Claims	
Bod (2002)	DOP > Gestalt
Meredith (2002)	LBDM > Grouper

No previous experiments are known to us that compare the output of voice separation algorithms to human performance. Different variants of the Skyline algorithm were compared by Uitdenbogerd and Zobel (1998) and by Isikhan and Ozcan (2008). Jordanous (2008) aggregates performance evaluations from several published methods that used the fugues from J.S. Bach's *Das wohltemperierte Clavier* as a ground truth.

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

4. Experiments

4.1 Melody segmentation

In this experiment human performance in melody segmentation—a process sometimes referred to as 'chunking' or 'grouping'—was studied. The segmentations that humans generated in the experiment were compared to the segmentations of the same melodies by several prominent algorithms. In this section, we discuss the experiment's setup and summarize the results (see also Nooijer, 2007; Nooijer et al., 2008b).

4.1.1 Comparison against other methods

Melody segmentation tasks have been carried out in previous research using very different setups, often accessible only to music experts. For instance, the experiments reported by Thom et al. (2002) and Ahlbäck (2004) both involved the score of the piece. This required the participants to have formal music training in order to be able to read notation. Furthermore, Thom et al. (2002) used specific music terminology by asking to indicate the beginning of a phrase or sub-phrase.

A number of studies used only audio stimuli. For instance, Palmer and Krumhansl (1987) presented listeners with predetermined segments that were rated as to how complete the phrase sounded. Hence, no free choice of determining a segment was given. In the experiment by Koniari et al. (2001) children were asked to press a key at a segment boundary while listening to the piece three times. In a similar setup described in Spiro and Klebanov (2006), participants listened three times to a piece and identified phrase starts by pressing a key. Spiro and Klebanov developed a method to determine from the recorded keystrokes the actual segmentation intended by the participants, since listeners' real-time responses involve latency or may even contain errors.

In the design of our melody segmentation task we combined the real-time assignments of boundaries with the possibility to adjust the responses during repeated listening. Thus, participants could indicate precisely where they perceived a boundary.

4.1.2 Participants

Forty persons participated in this and the following experiment. Based on the years of formal musical education we divided participants into two categories (novices and experts) using cluster analysis. Each category contained twenty participants. Table 5 presents information on the participants' placement in the expert or novice category. Involving the additional data from this table in the cluster analysis did not result in a different categorization of experts and novices (details in Nooijer, 2007).

4.1.3 Materials and tools

The 10 melodies were randomly gathered from MIDI files on the Internet (using a crawler), such that the musical material is not biased towards certain criteria that might favour one of the algorithms when manually selected. Samples which did not have a recognizable melody were removed. For the selected melodies the accompaniment was removed. The melodies were subsequently converted into WAVE files, all using the same timbre. Each melody is between 25 to 30 s long, which is enough to be able to distinguish several segments. The number of melodies was set at 10 in order to keep the duration of the experiment reasonable. Obviously this influences the extent to which our conclusions can be generalized. The melodies, though mostly popular, differ considerably from each other. One example is shown in Figure 1. All melodies are accessible at http://give-lab.cs.uu.nl/music/.

The segmentation was done using Sony's Sound Forge. The interface displays the wave form of the audio file being played. A cursor shows the present location in the audio. While the song is playing, users can place markers, which can be moved around. This property of the program was exploited in the experiment, to allow the participants to fine-tune the placement of their markers. Figure 2 displays a screen capture of Sound Forge. The cursor (a vertical line) is shown towards the

Table 4. Summarized test and claims from Table 3.

	>	MSM	>	Grouper	>	
DOP	>	MDSM			>	LBDM
	>	Grouper	<	GTTM	<	
		PAT			>	

Table 5. Participant data from questionnaire.

	Experts	Novices	Combined
Average age	27	24	25.5
Participants per gender	♀: 6 ♂: 14	♀: 9 ♂: 11	♀: 15 ♂: 25
Average music education (in years)	11.3	1.25	6.75
Frequency of listening to music	Daily	Daily	Daily
Frequency of visiting concerts	Regularly	Sometimes	Sometimes
Preferred genre	Popular music	Popular music	Popular music

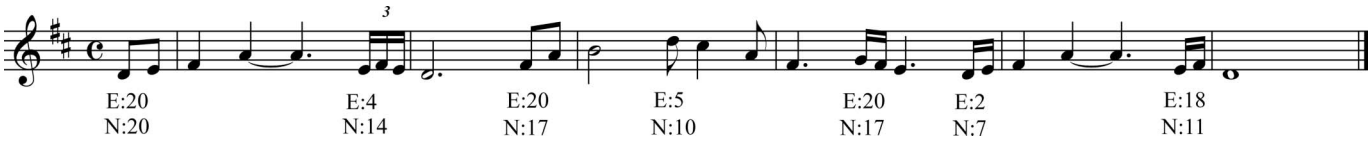


Fig. 1. Example melody used in the melody segmentation experiment. Human segmentation results are shown below the staff at the position of the segment beginning. (E: experts; N: novices).

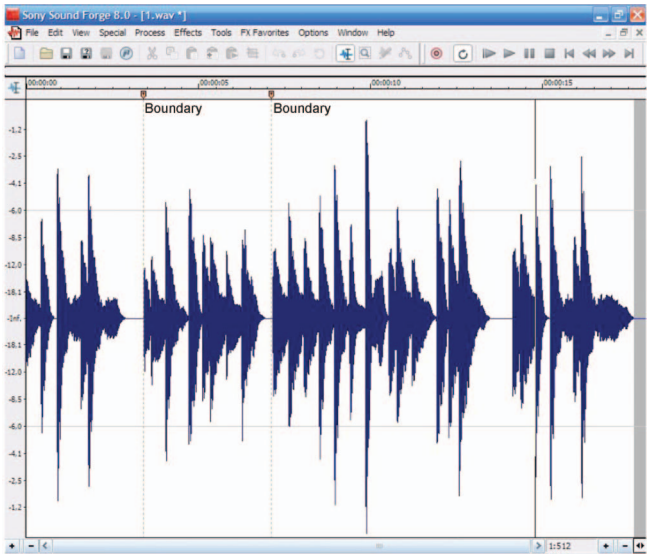


Fig. 2. A screenshot of Sound Forge playing a tune.

Table 6. Melody segmentation experiment scenario.

Time	Participant	Experimenter
Before experiment		
	Sit down at table	
		Laptop, instructions and drinks on table
Experiment		
0:00–0:10	Reads instructions	
0:10–0:20	Practices segmentation task	
0:20–1:20 (approx.)	Performs segmentation task	
1:20–1:30	Fills out questionnaire	
After experiment		
		Thanks participant for cooperation
		Packs up laptop and instructions

right side of the screen. Two markers, labelled 'Boundary' in the illustration, are displayed as dotted vertical lines towards the left side of the screen. We removed all task-irrelevant elements (a VU-metre, etc.) from Sound Forge's interface and maximized the program window, to ensure that on-screen distractions were minimized during the experiment.

4.1.4 Design and procedure

Participants were asked to divide a melody into smaller units by placing markers at locations where a segment ends. The participants received an instruction sheet that contained information on their tasks. Before executing the actual task, participants practiced on three well-known tunes. The experimenter did not answer any questions concerning the execution of the actual task during the experiment. Experiments followed a strict scenario, to ensure consistency. The scenario is displayed in Table 6.

It is more likely that participants will recognize the ending of a segment instead of the beginning of a new segment, because of the experience of closure (Snyder, 2000). Thus, asking participants to mark endings of segments relied more on their intuition than asking them to recognize the beginning of a new segment as in Spiro and Klebanov (2006). Hence, participants were asked to place markers at locations where a segment ends. A marker placed at a position where a segment ends automatically initiates the start of a new segment, which in turn is closed by the following end-marker.

After finishing the experiment, participants completed a short questionnaire related to their level of musical knowledge.

4.1.5 Results and discussion

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 inter-assessor (or intraclass) coherence. The results show that within both groups of participants, the coherence is very high—respectively $\alpha = 0.9675$ for novices and $\alpha = 0.9902$ for experts. The coherence between all participants of both groups is also high, $\alpha = 0.9864$. Thus, we conclude that the segmentation

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

results of novices and experts do not differ significantly, and that there is enough coherence between participants to function as a basis for algorithm benchmarking. This is an interesting observation, since multiple authors state that segmentation is a highly ambiguous task (Thom et al., 2002; Ahlbäck, 2004). The material used might be one reason for the difference. Previous researches have often used classical music. The popular melodies in our experiment may contain clearer cues about segment boundaries. It should also be mentioned that Thom et al. (2002) compare averaged F-scores between the participants in order to illustrate the ambiguity of the task; however, they do not measure whether these scores differ significantly from each other.

Some observations can be made concerning our raw data. The judgments of experts generally show a higher overall similarity amongst participants than those of the novice participants. Spiro and Klebanov (2006) also noticed this, and attribute this phenomenon to the fact that novices and experts have different knowledge available to process the tunes. Specifically, some tunes have stronger cues (for example, more apparent rests, larger pitch intervals or recurring rhythmical patterns) than others. Novices have to rely solely on these cues, while experts can also rely on higher hierarchical structures, learned during their education. Thus, tunes with stronger local cues tend to show more consistency in segmentation, while the results for tunes with weaker cues tend to be more diverse amongst novice participants.

We noticed that novices have the tendency to discern a larger number of boundaries than the experts, sometimes creating phrases of just six notes. Upon further analysis, the novices whose results display this effect often place boundaries at locations where the TGU algorithm marks a clang boundary. A clang is a lower level musical element consisting of a small number of notes. Clangs together form sequences, or segments (Tenney & Polansky, 1980). This observation confirms the assumption that education forces experts to think of a piece in hierarchical structures (Levitin, 2006; Spiro & Klebanov, 2006), while non-experts base their decisions on local cues in the melody.

For our second research question we evaluated the algorithms in two ways, first by measuring their performance and then by establishing whether or not the algorithmic segmentations can be distinguished from those of the participants.[2] We measured the performance of the algorithms using the F1-measure. The F1-measure is commonly used to give a single-number value for retrieval performance tasks. It has also been applied to the evaluation of segmentation algorithms, for example by Thom et al. (2002). Given the number of true positives tp, the number of false positives fp and the number of false negatives fn, one first calculates the precision $P = tp/(tp + fp)$ and the recall $R = tp/(fp + fn)$. The F1-measure is the harmonic mean of P and R, so $F1 = 2PR/(P + R)$ (Manning et al., 2008). Please note that the F1-measure provides no information about the statistical significance of the differences in performance.

In order to be able to calculate the F1-measure, we need to convert the human segmentation judgement results into binary boundary decisions. We did so by requiring the percentage of the participants that agreed upon this boundary to be above 60. This yields 27 boundaries in total. Setting the cutoff level at a lower value seems inappropriate, since one would expect a search engine to use boundaries that the majority of the audience agrees to. Below reasons are given why this value cannot be set much higher. The performance evaluation results are shown in Table 7. For LBDM and DOP, we show only the best two variants, plus the ones for which $F1 \geq 0.50$. LBDM6 and all DOP variants except DOP6 and DOPmax2-2 are therefore eliminated.

The best-performing method is thus Grouper8, with three other variants of Grouper closely following. Several LBDM variants also perform well. The ones using absolute cutoff values perform considerably worse than most variants using local maxima. All other methods perform worse than Grouper and LBDM. The best-performing methods tend to have the largest number of true positives, but note that even these display many false positives and/or false negatives. LBDM2 is exceptional in that it produces, despite its low rank, the same number of true positives as Grouper8: its bad overall performance is caused by the high amount of false positives. A similar situation exists for TGUclang. The bad performance of all DOP variants may be related to the fact that this measure was trained on phrase information from the Essen Folksong Collection, whereas the participants in the experiment tend to distinguish boundaries both at the phrase level and at the level below this.

We used the Wilcoxon Signed-Rank Test to determine whether or not there are significant differences in segmentation performance between algorithms and the groups of participants (novices, experts, all). The results are presented in Table 7. When differences must be considered non-significant ($p > 0.05$), the hypothesis that human and algorithmic segmentations are different cannot be rejected. This can mean two things: either the method is an adequate model of human segmentation, or there are not enough data to reject the hypothesis. Non-significance thus constitutes only a weak proof of the algorithm's adequacy.

Table 7 shows that algorithms with a good F1-measure tend not to be distinguishable from the human segmentations. In particular, this is true for the

[2]This evaluation of the melody segmentation algorithms replaces the preliminary evaluation that was presented in Nooijer et al. (2008a,b).

Table 7. Segmentation evaluation. Left: F1-measure results; right: Wilcoxon Signed-Rank Test results. For the latter, non-significant values ($p > 0.05$) are printed in bold.

	Performance evaluation						Wilcoxon Signed-Rank Test		
Method	F1-measure 60%	True Positives	False Positives	False Negatives	Precision	Recall	Novices	Experts	All
Grouper8	0.7302	23	13	4	0.6389	0.8519	**0.692**	**0.142**	**0.590**
Grouper10	0.7143	20	9	7	0.6897	0.7407	**0.113**	**0.898**	**0.189**
Grouper10mtr	0.6667	18	9	9	0.6667	0.6667	0.040	**0.564**	**0.065**
Grouper8mtr	0.6557	20	14	7	0.5882	0.7407	**0.798**	**0.320**	**0.965**
LBDMmax4-3	0.6557	20	14	7	0.5882	0.7407	**0.994**	**0.272**	**0.860**
LBDMmax4-4	0.6552	19	12	8	0.6129	0.7037	**0.436**	**0.678**	**0.541**
LBDMmax3-3	0.6349	20	16	7	0.5556	0.7407	**0.610**	**0.142**	**0.468**
LBDMmax5-5	0.6250	15	6	12	0.7143	0.5556	0.001	0.038	0.001
LBDMmax3-2	0.6087	21	21	6	0.5000	0.7778	**0.074**	0.013	0.049
LBDMmax5-4	0.6038	16	10	11	0.6154	0.5926	0.041	**0.463**	**0.065**
Grouper6	0.6027	22	24	5	0.4783	0.8148	0.006	0.002	0.004
LBDM4	0.5769	15	10	12	0.6000	0.5556	0.027	**0.325**	0.048
Grouper6mtr	0.5714	20	23	7	0.4651	0.7407	**0.069**	0.008	0.047
LBDM3	0.5667	17	16	10	0.5152	0.6296	**0.847**	**0.466**	**0.983**
LBDM5	0.5417	13	8	14	0.6190	0.4815	0.002	**0.060**	0.002
LBDMmax2-2	0.5316	21	31	6	0.4038	0.7778	0	0	0
LBDMmax6-6	0.5000	10	3	17	0.7692	0.3704	0	0	0
LBDM2	0.5000	23	42	4	0.3538	0.8519	0	0	0
MSMend-low	0.4507	16	28	11	0.3636	0.5926	0	0.041	0
MSMend-high	0.4444	10	8	17	0.5556	0.3704	**0.065**	0.009	0.046
IDyOM	0.4242	14	25	13	0.3590	0.5185	**0.362**	**0.068**	**0.311**
TGUseg	0.4167	10	11	17	0.4762	0.3704	0.011	**0.092**	0.009
MSMstart-high	0.3830	10	8	17	0.5556	0.3704	0.048	0.009	0.040
TGUclang	0.3654	19	58	8	0.2468	0.7037	0	0	0
LBDM6	0.2941	5	2	22	0.7143	0.1852	0	0	0
MSMstart-low	0.2778	10	35	17	0.2222	0.3704	0.007	**0.102**	0.005
DOP6	0.2597	10	40	17	0.2000	0.3704	0.010	0.001	0.008
DOPmax2-2	0.2250	9	44	18	0.1698	0.3333	0.002	0	0.001

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

top-7. However, there are some notable exceptions. Therefore it makes sense to compare the two evaluation measures. The Wilcoxon Signed-Rank Test examines the distribution of the differences between two measurements of the same population. Here the population consists of all possible boundaries. For the human measurement, the boundary score s_h is a rational number between 0 and 1, for an algorithm, the boundary score s_a is either 0 or 1. The Wilcoxon score is minimal if the positive and the negative differences have the same distribution. This situation can be approximated as follows. If $s_h < 0.5$, s_a is set to 0; if $s_h > 0.5$, s_a is set to 1. For $s_h = 0.5$, half of the cases are set to 0 and the other half to 1. In this manner, for each participant group an optimum segmentation was constructed and evaluated using the F1-measure at 60%, 70% and 80% cutoff levels. The results are shown in Table 8. These numbers are the maximum scores that can be attained using the setup of this article, and thus give an indication of the amount of algorithmic improvement that is still possible.

Table 8. F1-measures for segmentations optimized for Wilcoxon Signed-Rank Test. For each of the user groups (novices, experts and all participants) an optimal version was created.

Cutoff	F1-60 %	F1-70 %	F1-80 %
Opt-novices	0.8136	0.7857	0.7692
Opt-experts	0.9286	0.9057	0.8163
Opt-all	0.9643	0.9057	0.8163

Conversely, from the point of view of the F1-measure, the best possible results come from an algorithm that creates the exact boundaries at the different cutoff levels. If we evaluate these optimum boundary sets against the human results using the Wilcoxon Test, we get the results shown in Table 9. In all cases the result is significantly different from human segmentation, except for the 60% optimum boundary set when compared to the experts. It

Table 9. Wilcoxon Signed-Rank Test for segmentation optimized for different cutoff levels, compared to user groups. Non-significant values ($p > 0.05$) are printed in bold.

	Novices	Experts	All
Opt-60%	0.002	**0.101**	0.017
Opt-70%	0.000	0.002	0.000
Opt-80%	0.000	0.000	0.000

therefore makes less sense to use the higher cutoff levels in the evaluation. On the other hand, setting the boundary lower than 60% is not desirable from the point of view of search engine design.

The above shows that the two evaluation measures have different properties and can be considered complementary. Therefore it is unproblematic that methods that cannot be distinguished from humans can yet have a low F1-measure. Notably, this is the case for LBDM3 and IDyOM. What these have in common with the top-7 is that they find more than half of the true positives and that the number of false positives is at most twice the number of false negatives, and at least about the same number. It follows from the description above that in the Wilcoxon Test, false positives and false negatives tend to cancel each other out. However, on inspection of the raw data it appears that often the false positives coincide with boundaries perceived by a substantial minority of humans. The actual contribution of the false positives to the Wilcoxon score is thus smaller than that of the false negatives; hence there can be more of them. This also explains why LBDMmax5-5, LBDMmax3-2 and Grouper6 differ significantly from human segmentation performance.

There are some minor but interesting differences between the three participant groups. These suggest that experts' segmentations agree better with methods that provide fewer boundaries and thus longer segments (Grouper10mtr, LBDMmax5-4, LBDM4, LBDM5 and TGUseg), whereas novices seem to prefer more boundaries and shorter segments (LBDMmax3-2 and Grouper6mtr). However, for MSM the opposite seems true.

4.1.6 Evaluation

We 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). These results are promising for the implementation of an MIR system for popular melodies: they suggest that since there are no significant differences in human segmentation, one accurate automatic segmentation of the tunes in the corpus of an MIR system may suffice. In other words, it does not appear to be necessary to offer different segmentation options to, for example, people with different levels of music education.

We can therefore give a positive answer to our first research question. Since there is a high intraclass agreement within experts and within novices, as well as in the combined group, their combined melody segmentation results can function as a basis for algorithm benchmarking.

Comparing human and algorithmic segmentations, we conclude that the results of Grouper8 and Grouper10, LBDM in several variants, and IDyOM do not differ significantly from humans. Considering the F1-measure as well, the best candidates for implementation seem to be Grouper and LBDM. For Grouper, the optimum phrase length seems to be 8; applying first the metre tool to the data does not result in a measurable gain. The best choice for LBDM seems to be to use an approach that uses local optima. LBDMmax4-3 is then the best candidate, though this measure is apparently fairly robust against different sizes of the local context. Given the widely diverging results for different absolute cutof levels of LBDM, this approach seems more hazardous.

4.2 Voice separation

In the following experiment, human performance of a specific voice separation task, the separation of the melody from accompanying voices, was studied and compared to algorithmic performance of the same task. This specific task is often referred to as 'melody identification' or 'melody extraction'. In this section, we present the setup and results of this experiment (see also Nooijer, 2007; Nooijer et al., 2008b).

4.2.1 Method

For the voice separation experiment, the same 40 participants as in the melody segmentation experiment were presented with a short musical sample and were then asked to determine through exhaustive 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. What has been said about the limited number of samples in the melody segmentation experiment (see Section 4.1.3) applies here as well. One sample is shown in Figure 3; all samples are accessible at http://give-lab.cs.uu.nl/music/.

4.2.2 Materials

Eight polyphonic musical samples in MIDI format were randomly gathered from the Internet in a similar way as in the first experiment. Each sample is approximately 10 s

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Fig. 3. Polyphonic sample used in the voice separation experiment, shown in piano roll notation.

long, which is enough to be able to distinguish melody and harmony. Since algorithms do not take timbre into account for separation, we assigned melody and harmony notes the same timbre; therefore, participants could not have an unintended advantage over the algorithms by using timbre information (Bregman, 1990, ch. 5). 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. When an algorithm outputs multiple voices containing monophonic lines—sometimes up to more than ten voices—we selected the one that, according to the second author, most accurately represents the melody for inclusion in the set of variants. If multiple monophonic lines contained similar amounts of the actual melody notes, we randomly selected only one of these.

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

4.2.3 Design and procedure

First we briefly describe a rejected version of the voice separation experiment. In this version, the piece was presented to the participants in the Cubase environment as piano roll notation. The task was to listen to the piece and to erase those notes from the piano roll notation that in the participant's opinion did not belong to the melody. Participants could listen to the piece as often as they wanted, and deletions could be reversed. The advantage of this setup was that each participant would provide precisely one solution for each voice separation task, whereas in the final experiment they produced a considerable number of judgments on the qualities of different alternatives. However, even for experienced musicians the erasing task proved to be too challenging. Therefore, this version of the experiment was rejected.

Table 10. Voice separation experiment scenario.

Time	Participant	Experimenter
Before experiment		
	Sit down at table	
		Laptop, instructions and drinks on table
Experiment		
0:00–0:10	Reads instructions	Starts Cubase, loads sound files
0:10–0:20	Practices separation task	Plays sound files
0:20–1:20		Plays sound files, notates scores
1:20–1:35	(break)	
1:35–2:35 (approx.)	Performs separation task	Plays sound files, notates scores
After experiment		
	Signs payment form	Pays and thanks participant for cooperation
		Packs up laptop, score sheets and instructions

The final version of the experiment was as follows. Participants were presented with a polyphonic piece. The polyphonic piece (the 'sample') was played, and then followed by a pair of melodic variants, for example, variant 1 and variant 2. Then, the participant had to decide which of the two variants he or she considered to be more similar to the melody heard in the polyphonic sample. The participant could listen to the sample as often as (s)he wanted before the next pair of variants was presented – for example, variant 1 and variant 3. This process was continued until each variant had been judged against each of the other variants and, with each pair, the original. The scenario of the experiment is shown in Table 10.

4.2.4 Results

Because of the complexity of the dataset, we applied our statistical measure, Cronbach's alpha for inter-rater

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Table 11. 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.

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*		SSA		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*

coherence, only to the highest ranked melodic variant for each polyphonic sample by each participant. The resulting values are $\alpha_{nov} = 0.8966$, $\alpha_{exp} = 0.9389$ and $\alpha_{all} = 0.9230$. Since these values are all high (taking into account that a 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. Next, we calculated rankings based on the experts' results (Table 11).

When examining these rankings, we observe that the results differ considerably per sample. For example, the SSA algorithm ranks in the top two for melodies 2, 3, 6, 7 and 8, but 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 the largest number of the samples. The results are shown in Table 12.

Here, we observe that three algorithms' variants are ranked at the first place on four occasions: VoSA, SSA and Skyline. Thus, any of these three algorithms provides the maximum number of accurate results. However, each algorithm is only this accurate for four of the eight samples we investigated.

The accumulated novice results are shown in Table 13. 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 12. 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. However, the second author's melody was in fact the winner for the novices, as it was ranked highest for five out of eight samples.

Table 12. Number of times a variant is ranked first place by experts (left) and novices (right).

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

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Table 13. 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.

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*

Experts chose the second author's melody only three times as the best matching. Despite this, the second author must be qualified as an expert using the criteria of section 4.1.2.

4.2.5 Evaluation

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

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.

5. Conclusion and future work

5.1 Method improvements

For the melody segmentation experiment, we chose Sound Forge as the user interface. Its main advantage was that it allowed participants to make up for any latency occurring between hearing a boundary and pressing the appropriate key to place a boundary. However, one could argue that Sound Forge's method of visualization might trigger visual cues about the musical piece. Participants could then, instead of relying solely on the supplied auditory information, use these visual cues to segment the musical piece. To eliminate this potential bias, one would need to design an interface that utilizes a custom, non-suggestive visualization method for displaying the audio signal. However, this cannot be done in Sound Forge, as all manipulations on the visual representation will automatically result in a change of the audio signal. An improved visualization should of course preserve the main advantage of Sound Forge, namely that it allows the participants to apply latency correction by using the visual representation.

The voice separation experiment has an important drawback in that in order to consider other new algorithms for comparison, the entire human experiment must be redone, including all the old and the new algorithms—combinatorially a daunting prospect. Devising a modified version of the experiment that does not suffer from this drawback is an important future goal. A possible solution would be a hybrid experiment,

in which pair wise judgements are used to eliminate those notes from the sample that do not belong in the sample's melody.

5.2 Selection of models

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 provide segmentations that could not be distinguished from human segmentations: Grouper (in several variants), IDyOM and LBDM (in several variants). Among the voice separation algorithms, none of the models came close to human performance of melody identification; 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 an 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, being data-driven, IDyOM 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 weighted 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 and rhythm-based LBDM is likely to produce better results. For LBDM a version using local maxima seems best suited. Both Grouper and LBDM can also be configured to suit a certain expected average segment length. IDyOM, though equally indistinguishable from humans, performs less well than Grouper and LBDM using the F1-measure. To what extent this behaviour depends on the size and nature of the sample studied here is unclear. In this respect, much insight can be gained from repeating the experiments with a much larger number of samples in different musical styles.

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 IDyOM, 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 different voices of the music.

In order to test our underlying hypothesis, namely that an 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 an 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

We would like to thank the authors of the algorithms who have kindly answered our questions and requests for cooperation, in alphabetical order: Sven Ahlbäck, Emilios Cambouropoulos, Elaine Chew, Sren Madsen, Marcus Pearce and David Temperley. We also thank all the participants in the experiments for their cooperation. Bas de Haas and two anonymous reviewers gave valuable comments on earlier versions of this article.

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.

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Bruderer, M.J., & McKinney, M.F. (2008). Perceptual evaluation of models for music segmentation. In *Proceedings of the Fourth Conference on Interdisciplinary Musicology (CIM08)*, Thessaloniki, Greece. Retrieved December 5, 2008, from http://web.auth.gr/cim08/

Cambouropoulos, E. (1998). Musical parallelism and melodic segmentation. In *Proceedings of the Twelfth Colloquium of Musical Informatics*, Gorizia, Italy, pp. 111–114.

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

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Downloaded By: [Aalborg University] At: 13:07 11 May 2011

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation: A computational approach. *Music Perception*, *23*, 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 *Proceedings of Computer Music Modeling and Retrieval 2004*, Esbjerg, Denmark.

Clausen, M. (n.d.). *Melody Extraction*. Retrieved July 4, 2007, from www-mmdb.iai.uni-bonn.de/forschung projekte/midilib/english/skydemo.html

Eerola, T., & Toiviainen, P. (2004). *The MIDI Toolbox: MATLAB Tools for Music Research*. Retrieved May 31, 2007, from www.jyu.fi/musica/miditoolbox/

Ferrand, M., Nelson, P., & Wiggins, G. (2003). Memory and melodic density: A model for melody segmentation. In *Proceedings of the Fourteenth 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.

Frankland, B.W., & Cohen, A.J. (2004). Parsing of melody: Quantification and testing of the local grouping rules of Lerdahl and Jackendoff's *A generative theory of tonal music*. *Music Perception*, *21*, 499–543.

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.

Isikhan, C., & Ozcan, G. (2008). A survey of melody extraction techniques for music information retrieval. In *Proceedings of the Fourth Conference on Interdisciplinary Musicology (CIM08)*, Thessaloniki, Greece. Retrieved December 5, 2008, from http://web.auth.gr/cim08/

Jordanous, A. (2008). Voice separation in polyphonic music: A data-driven approach. In *Proceedings of the International Computer Music Conference*, Belfast, UK. Retrieved December 16, 2008, from http://www.informatics.sussex.ac.uk/users/akj20/papers/2008_ICMC_VoiceSeparation.pdf

Juhász, Z. (2004). Segmentation of Hungarian folk songs using an entropy-based learning system. *Journal of New Music Research*, *33*, 5–15.

Karydis, I., Nanopoulos, A., Papadopoulos, A.N., & Cambouropoulos, E. (2007). VISA: The voice integration/segregation algorithm. In *Proceedings of the Seventh International Conference on Music Information Retrieval*, Vienna, Austria, pp. 445–448.

Ke, W.-J., Chang, C.-W., & Jiau, I.C. (2004). Representative music fragments extraction by using segmentation techniques. In *Proceedings of the International Computer Symposium (ICS2004)*, Taipei, Taiwan, pp. 1156–1161.

Kirlin, P.B., & Utgoff, P.E. (2005). VoiSe: Learning to segregate voices in explicit and implicit polyphony. In *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, UK, pp. 552–557.

Koniari, D., Predazzer, S., & Melen, M. (2001). Categorization and schematization processes in music perception by children from 10 to 11 years. *Music Perception*, *18*, 297–324.

Levitin, D.J. (2006). *This is your Brain on Music: The Science of a Human Obsession*. New York: Dutton.

Madsen, S.T., & Widmer, G. (2006). Separating voices in MIDI. In *Proceedings of the Seventh International Conference on Music Information Retrieval*, Victoria, Canada, pp. 8–12.

Manning, C.D., Raghavan P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

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, Netherlands.

Nooijer, J. de, Wiering, F., Volk, A., & Tabachneck-Schijf, H.J.M. (2008a). Cognition-based segmentation for music information retrieval systems. In *Proceedings of the Fourth Conference on Interdisciplinary Musicology (CIM08)*, Thessaloniki, Greece. Retrieved December 5, 2008, from http://web.auth.gr/cim08/

Nooijer, J. de, Wiering, F., Volk, A., & Tabachneck-Schijf, H.J.M. (2008b). An experimental comparison of human and automatic music segmentation. In *Proceedings of the Ninth International Conference on Music Perception and Cognition*, Sapporo, Japan, pp. 399–407.

Palmer, C., & Krumhansl, C. (1987). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(1), 116–126.

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, Italy, pp. 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.

Rafailidis, D., Nanopoulos, A., Cambouropoulos, E., & Manolopoulos, Y. (2008). Detection of stream segments in symbolic musical data. In *Proceedings of the Ninth International Conference on Music Information Retrieval*, Philadelphia, USA, pp. 83–88.

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

Snyder, B. (2000). *Music and Memory, an Introduction*. Cambridge, MA: MIT Press.

Spiro, N., & Klebanov, B. (2006). A new method for assessing consistency or real-time identification of phrase-parts and its initial application. In *Proceedings of the Ninth International Conference on Music Perception and Cognition*, Bologna, Italy.

Szeto, W.M., & Wong, M.H. (2003). A stream segregation algorithm for polyphonic music databases. In *Proceedings of the Seventh International Database Engineering and Application Symposium*, Hong Kong.

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*(3), 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 *Proceedings 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 *Proceedings of the ACM Multimedia Conference '98*, Bristol, UK, pp. 235–240.

Volk, A. (2005). Coffee bean (and other) models about the metric structure of music. In S. Bab, J. Gulden, T. Noll & T. Wieczorek (Eds.), *Models and Human Reasoning* (pp. 317–325). Berlin: W&T Verlag.

Downloaded By: [Aalborg University] At: 13:07 11 May 2011