INTEGRATING SEGMENTATION AND SIMILARITY IN MELODIC ANALYSIS

Tillman Weyde
Research Department of Music and Media Technology
University of Osnabrück
Osnabrück, Germany
e-mail: tweyde@uos.de

ABSTRACT

The recognition of melodic structure depends on both the segmentation into structural units, the melodic motifs, and relations of motifs which are mainly determined by similarity. Existing models and studies of segmentation and motivic similarity cover only certain aspects and do not provide a comprehensive or coherent theory.

In this paper an *Integrated Segmentation and Similarity Model* (ISSM) for melodic analysis is introduced. The ISSM yields an *interpretation* similar to a paradigmatic analysis for a given melody. An interpretation comprises a segmentation, assignments of related motifs and notes, and detailed information on the differences of assigned motifs and notes. The ISSM is based on generating and rating interpretations to find the most adequate one. For this rating a neuro-fuzzy-system is used, which combines knowledge with learning from data.

The ISSM is an extension of a system for rhythm analysis. This paper covers the model structure and the features relevant for melodic and motivic analysis. Melodic segmentation and similarity ratings are described and results of a small experiment which show that the ISSM can learn structural interpretations from data and that integrating similarity improves segmentation performance of the model.

1. INTRODUCTION

Recognizing the structure of a melody is an essential part of musical listening. Although a model of melodic structure is needed for analytical research as well as for practical applications, no generally accepted theory of melodic structures has yet been developed. Neither is it generally clear how melodic structures can be found, nor is it agreed on exactly what melodic structures are.

Yet there are two aspects that are essential for most theories of melodic structure: segmentation and similarity. Segmentation describes structural units: perceptual groups or musically speaking melodic motifs, the building blocks of melodic structure. Motif relations are mainly determined by similarity. The importance of motifs for melody has been stressed by many theorists. E.g. Riemann said that understanding a melody depends essentially on the recognition of the motif division intended by the composer [1, p. 15]. The role of similarity has also been recognized early by theorists like Koch who called for variety and uniformity by repeating musical parts and putting them into a new but similar form [2, p. 55].

Segmentation is influenced by factors like the *Gestalt* principle of proximity and properties of auditory perception like the maximum number of elements in one segment. But segmentation is also influenced by the similarity relations of motifs within a melody which vice versa depend on segmentation. In their *Generative Theory of Tonal Music*[3] Lerdahl and Jackendoff refer to this influence in their *Grouping Preference Rules* by defining parallelism as a criterion for segmentation. For a computational model of melodic structure both aspects should be integrated into one coherent model.

There are some computer based models for the segmentation of melodies like the *Temporal Gestalt Perception* (TGP) model by Tenney und Polansky [4] and the *Local Boundary Detection Model* (LBDM) by Cambouropoulos [5]. Similarity of motifs has also been computer modeled. One approach is to use geometrically motivated distance metrics for motifs like in TGP or in Mathematical Music Theory (MaMuTh). The problem with metrics is the handling motifs of differing lengths which is solved in MaMuTh by using motif topologies [6]. Another approach is to use string matching methods which calculate the editing distance of motifs by counting insertions and deletions of notes. These models can handle motifs of different length elegantly but they generally do not take into account the gradual differences of note onset time, duration, pitch, and loudness. Lately Smith, McNab, and Witten have presented a system which weights editing operations by differences in pitch and duration [7].

The *Integrated Segmentation and Similarity Model* (ISSM) presented in this paper combines segmentation and similarity relations in a model for the recognition of melodic structure. The goal is to determine a structural *interpretation* of a melody, i.e. a segmentation into motifs and assignment of each motif to one other motif like it is shown in figure 1 (see also [8]). This approach is similar to paradigmatic analysis (see [9]) in that the assignments represent how motifs are interpreted by a listener as being identical or similar to previous motifs. The choice of an adequate interpretation for a given melody is a difficult task. Although music theory and empirical studies have determined some influential factors for the choice of interpretations, no coherent or comprehensive theory has yet been established. So the aim of

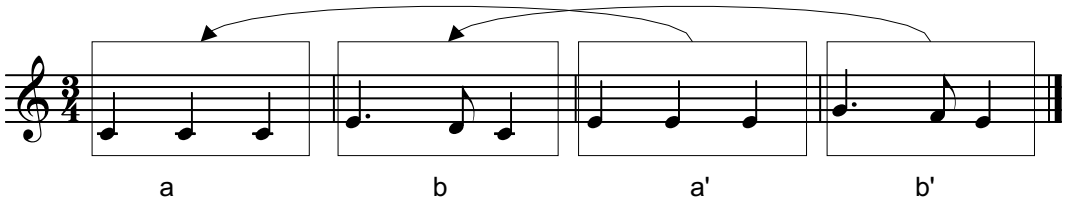

Figure 1: Interpretation of a simple melody.

this model is to integrate findings from music theory and existing studies with system optimization by experimental data.

2. MODEL DESCRIPTION

The general scheme of the ISSM in application mode is to generate all possible interpretations and rate their quality to find the most adequate one. The quality of an interpretation depends on segmentation and motivic structure. The quality of motivic structure depends mainly on the similarity of motifs while the quality of segmentation depends on temporal and tonal distances combined with motivic relations. The rating of the interpretations is done on the basis of feature values. They are calculated from the interpretations and rate properties of the segmentation and the motivic structure. Segmentation features are for instance the length (number of notes) and the duration (temporal extent) of motifs and pitch intervals at motif boundaries. Features for the similarity of assigned motifs depend on similarity of pitch, tempo, loudness, and contour.

Rating all possible interpretations is computationally very inefficient because the number of possible interpretations grows exponentially with melody length. So only currently only a limited context of up to 10 notes is used and some optimizations are employed (see [10]). The most effective way to limit the search space is to use perceptually motivated constraints. These constraints correspond to the GTTM's *Grouping Well-Formedness Rules* and prevent the generation of implausible interpretations or filter them out before they are rated.

The calculation of an overall rating of an interpretation from the features is done by a neural net which is based on fuzzy rules (see [11]) and extended with a list processing feature (see [10]). Other adaptive mapping systems could also be used but need to be fitted to the list processing. The reason to use an adaptive system here is that the strength and interaction of the influential factors is generally not known and poses the greatest problem of modeling. The module structure of the ISSM is shown in figure 2.

3. INTERPRETATION RATINGS

The rating of interpretations is essential for the output of the ISSM. The features and rules related to rhythm and time are described in [12] and [10]. In this paper the features which introduced for melodic segmentation and motif similarity are described.

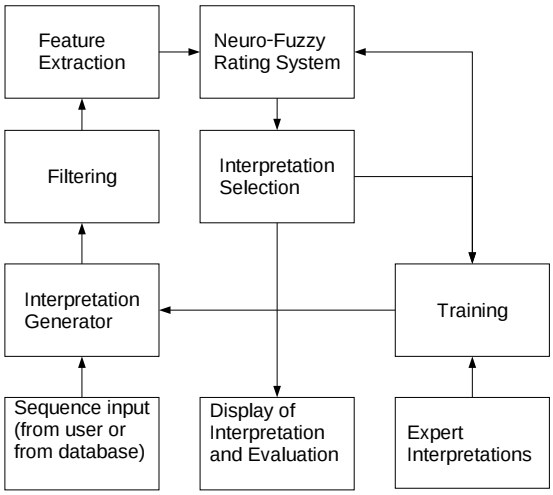

Figure 2: Modules of the ISSM.

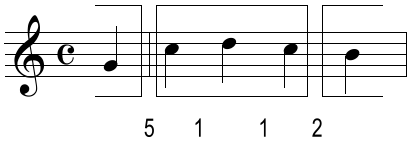

Figure 3: Inner and outer intervals of a motif. The inner intervals are both of distance 1 semitone, the outer intervals have an average distance of 3.5 semitones.

3.1. Segmentation

For segmentation the ratios of the average distance of the inner and outer intervals are calculated for each motif. The inner intervals are those between adjacent notes within a motif, the outer intervals are those between the adjacent notes of different motifs as illustrated in figure 3. Additionally for the outer intervals the minimal distance of interval notes in the circle of fifth is calculated. These features correspond to the Gestalt law of (tonal) proximity. The ISSM has a segmentation-only mode in which just these ratings and the rhythmic segmentation ratings are used.

3.2. Similarity

For motif similarity feature values for transposition, pitch differences, contour similarity and correctness are calculated. This is

based on the assignment of each note of a motif to a note of the assigned motif. Assigned notes are used for calculating transposition, pitch difference and contour similarity while the notes which are not assigned contribute to a correctness feature. For both rhythm and pitch the principle followed here is to separate local from global deviations similar to the MaMuTh approach. For pitch the global deviation is the transposition of a motif. The transposition of a motif a' compared to its assigned motif a is determined by the most frequent interval of assigned notes, in case of ambiguities the smallest interval and in the case of two smallest intervals the ascending one is chosen. This transposition value is subtracted from the pitch differences between the individual notes as is shown in figure 4. The local pitch deviation for a motif is calculated as the average difference of the notes.

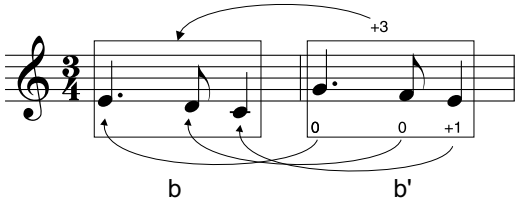

Figure 4: Global and local pitch differences of motifs a and b.

The contour rating is based on melodic difference vectors. These are based on the vector of each inner time and pitch interval in the pitch-time space as shown in figure 5. Each interval vector in a motif is subtracted from the corresponding interval in the assigned motif. An example is shown in figure 6. The average euclidian length of the difference vector defines the contour distance. In the example shown the difference vector for the first interval $b_1 - b'_1$ has the length 0 and the second one $b_2 - b'_2$ has the length 1 (assuming 1 semitone corresponds to 1 unit) and the average is 0.5.

This brings about the question of scaling the pitch and time dimensions as in all models using vector metrics (TGP, LBDM, MaMuTh). The hypothesis employed here is that the size of changes relative to the size of intervals is important and that this relation is local to the motif. So the scaling factor s of time relative to pitch is calculated as the ratio of the average time and pitch intervals:

$$s = \frac{\text{average pitch interval}}{\text{average onset time interval}} \tag{1}$$

The contour difference value cd of two assigned Motifs M, N is calculated as

$$cd = \frac{1}{l-1} \sum_{i=1}^{l-1} \sqrt{(\Delta m_i^p - \Delta n_i^p)^2 + (s(\Delta m_i^t - \Delta n_i^t))^2} \tag{2}$$

where m_i, n_i are the ith assigned notes of M and N with n^p as the pitch and n^t as the onset time and Δm_i^x is the difference $m_{i+1}^x - m_i^x$.

These ratings take only the assigned notes into account. The notes which are not assigned – and thus interpreted as insertions or deletions – contribute to an correctness value, similar to an inverted editing distance. The correctness rating is calculated by adding the differences of the extra notes from the average of their

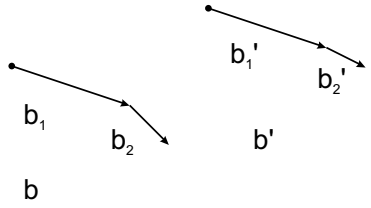

Figure 5: Interval vectors of motifs b and b'.

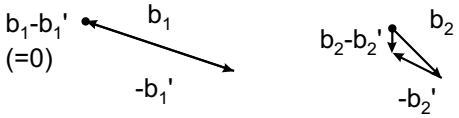

Figure 6: Interval difference vectors of motifs b and b'.

surrounding notes in the motif. This means that an inserted note makes less difference if its pitch lies between that of the surrounding notes.

3.3. Interpretations

The individual ratings for segmentation and similarity described above and those for rhythm need to be combined to an overall rating of interpretations. This can be done by mappings like linear combinations or neural nets. In the ISSM a neuro-fuzzy-system is used which consists of a neural net whose structure is defined by fuzzy rules. In this type of net each connection of neurons corresponds to a fuzzy rule. This allows to integrate prior knowledge with learning from data by defining the rules and training the corresponding net.

4. LEARNING FROM DATA

The ISSM learns from examples of melodic interpretations which it uses in an iterative training scheme. A neural net (and other adaptive systems) can be trained by relative samples containing two interpretations of which one is to be rated (see [13]). Iterative training then generates relative samples whenever the system chooses an interpretation that differs from the one provided by the expert (see [10]). The learning process changes the weights in the neural net which in the case of a fuzzy-logical neural net corresponds to the fuzzy truth values of fuzzy rules which can have a meaningful interpretation (see [11]). In the case of linear nets this corresponds to linear regression while in neural nets with error backpropagation, sigmoid activation function, and weight decay can be interpreted as a maximum likelihood estimate of the weights given the data. Yet a maximum found by backpropagation is not guaranteed to be global, for this a full Bayesian model would be needed (see [14]).

5. RESULTS

The ISSM has been implemented and can be used for motivic analysis, melody comparison or in segmentation-only mode where no similarity information is used. It computes and graphically displays segmentation, motivic relations, and data on the individual differences of related motifs and notes (differences in pitch or timing, inserted or deleted notes or groups). An expert user can provide interpretations with the graphical user interface. Previous versions have been tested intensively with rhythms.

A small experiment has been conducted with 15 beginnings of song melodies (2–4 bars). They have been used as samples for training the ISSM together with structural interpretations and with segmentations. Learning of interpretations was successful for all 15 samples. This indicates that the ISSM can learn interpretations with the currently used features.

Using segmentation-only mode the system could learn correct interpretations only for 10 of the 15 samples. This confirms that the motif similarity information which is not used in segmentation-only mode is of importance for the segmentation. The weights of the trained net were higher for the similarity features than for the segmentation features which further corroborates the importance of motif relations for segmentation. Yet to draw conclusions on perception and cognition of musical structure in general, larger samples sets from a group of subjects under controlled conditions would be needed.

6. CONCLUSIONS

The ISSM is an integrated model for segmentation and structural interpretation based on similarity. An integrated model is necessary since both processes influence each other. The weighting and interaction of aspects concerning both segmentation and interpretation poses a problem for which an adaptive system that learns from data proves to be useful. The ISSM learns from examples to generate musically meaningful interpretations of melodies which can be useful for applications like music retrieval, music tutorials, and interactive music production tools. First results support the view that it is necessary to take similarity into account for modeling segmentation. Future work should include experiments with larger data sets to further explore the capabilities of the ISSM and to learn more about the perception and cognition of melodies.

References

[1] Hugo Riemann. *Musikalische Dynamik und Agogik*. Hamburg, Leipzig, St. Petersburg, 1884.

[2] Heinrich Christoph Koch. *Versuch einer Anleitung zur Composition*, volume III. Leipzig, 1793. Reprografischer Nachdruck, Hildesheim 1969, Olms.

[3] Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music*. The MIT Press, Cambridge, Mass., 1983.

[4] James Tenney and Larry Polansky. Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–41, 1980.

[5] Emilios Cambouropoulos. A formal theory for the discovery of local boundaries in a melodic surface. In *Proceedings of the III Journees de'Informatique Musicale*, 1996.

[6] Chantal Buteau and Guerino Mazzola. From contour similarity to motivic topologies. *Musicae Scientiae, European Society for Cognitive Sciences of Music (ESCOM)*, IV(2):125–149, Fall 2000.

[7] Lloyd A. Smith, Rodger McNab, and Ian H. Witten. Sequence-based melodic comparison: A dynamic-programming approach. In Walter B. Hewlett and Eleanor Selfridge-Field, editors, *Melodic Similarity*, volume 11 of *Computing in Musicology*, pages 101–117. The MIT Press, Cambridge, Mass., 1998.

[8] Tillman Weyde. Grouping, smilarity and the recognition of rhythmic structure. In *Proceedings of the International Computer Music Conference 2001*, Habana, Cuba, 2001.

[9] Jean-Jaques Nattiez. *Music and Discourse - Toward a Semiology of Music*. Princeton University Press, Princeton, New Jersey, 1990.

[10] Tillman Weyde and Klaus Dalinghaus. Recognition of musical rhythm patterns based on a neuro-fuzzy-system. In Chian H. Dagli, editor, *Proceedings of the ANNIE 2001 conference*, 2001.

[11] Detlef Nauck, Frank Klawonn, and Rudolf Kruse. *Neuronale Netze und Fuzzy-Systeme*. Computational Intelligence. Vieweg, Braunschweig, 2 edition, 1996.

[12] Tillman Weyde. Recognition of rhythmic structure with a neuro-fuzzy-system. In C. Woods, G. B. Luck, R. Brochard, F. Seddon, and J. A. Sloboda, editors, *Proceedings of the Sixth International Conference on Music Perception and Cognition*, pages 1467–77, Keele, Staffordshire, UK, 2000. Department of Psychology, Keele University. CD-ROM (pdf, html).

[13] Heinrich Braun, Johannes Feulner, and V. Ulrich. Learning strategies for solving the planning problem using backpropagation. In *Proceedings of NEURO-Nimes 91, 4th International Conference on Neural Networks and their Applications*, 1991.

[14] Christopher M. Bishop. *Neural networks for pattern recognition*. Clarendon Press, Oxford, 1997.