SAPIENZA
UNIVERSITÀ DI ROMA

université
Paris Ouest
Nanterre La Défense

Salience concept in auditory domain with regard to music cognition

Dottorato di ricerca in Psicologia Cognitiva,
Psicofisiologia e Personalità
XXV Ciclo

ED 139, Connaissance, Langage et Modélisation

PhD candidate
Maurizio Giorgio

Tutor
Marta Olivetti Belardinelli

Tutor
Michel Imberty

A/A 2012/2013

SAPIENZA
UNIVERSITÀ DI ROMA

université
Paris Ouest
Nanterre La Défense

Salience concept in auditory domain with regard to music cognition

Dottorato di ricerca in Psicologia Cognitiva, Psicofisiologia e Personalità
XXV Ciclo

ED 139, Connaissance, Langage et Modélisation

PhD candidate
Maurizio Giorgio

Tutor
Marta Olivetti Belardinelli

Tutor
Michel Imberty

Δεν τόσο πολύ χρειάζεται τη βοήθεια των φίλων, λόγω της βεβαιότητας της βοήθειά τους.

Επίκουρο

Table of Contents

Table of figures

Table of figures: Appendix A

Table of figures: Appendix B

1 Introduction

Consider the melody *a* in fig.1.1.

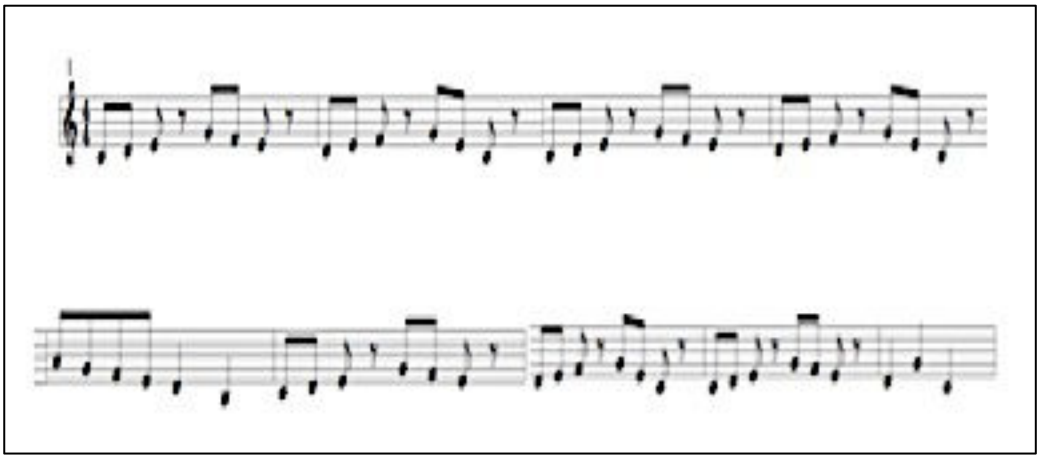


Fig. 1 How do we process music?

How does a listener organize the information contained in this short sequence of sounds? Does every pitch have the same importance in the listener's understanding of the piece? Is there any mechanism that allows the listener to simplify the memorization of this melody? Exploring the scientific literature it is possible to find several indications that allow to hypothesize the mental mechanisms involved in this task, with respect to a theoretic background that has been developed in the course of the last and the current century with regard to the emerging issues on music psychology and on general psychology too. Currently, it is possible to find in literature a conspicuous number of computational models, which try to reproduce the way a listener group and categorize simple auditory stimuli; the central concept of these models is, doubtless, the idea of "salience". Let's give a first, preliminary overview with respect to the above-mentioned melody: each element, that is, each pitch holds a series of specific auditory features, i.e., frequency, duration, intensity, timber; at the same time, each element forms with the surrounding elements precise relations based the specific features they have. While the specific features can be view as continuous (and sometimes complex) variables, the relations between tones move on an often culturally determined background. Coming back to the melody *a*, it is easy to observe that it presents a well defined temporal structure with the tones having only two relative possibilities in duration, respectively defined as “chrome” and “crotchet”. Furthermore, the melody is composed by a combination of only seven precise pitches that belong, moreover, to a precise tonality (C major). This multiple-level grammar, deeply inquired by Lerdahl and Jackendoff (1983) may allow the listener to easy categorize and memorize the elements of the melody: once an element have been chosen such a landmark, the ones surrounding it can be represented on the basis of the relations they have with it. In other words, it means that not all the pitches in the melody have exactly the same importance for the listeners. The most important elements are, exactly, the “salient element”, and they are though to assume their salience from their possess of “salient features”. An element may thus be salient if being louder or longer in duration with respect to the neighbor tones. A tone can also be salient if having a different timbre or if its frequency is quite different from the remaining tones. Furthermore, global features of the piece can make the listener assigns greater importance to a tone; this is the case of the accents provided by the tonal constraints, by the melodic structure (contour or structure of intervals between pitches), or by rhythmic and metric mechanisms. The mechanisms listed above have always a common point: the detection of a difference between elements or groups of elements. These differences can only arise on a musical surface that presents a certain degree of regularity or, in other words, which is composed by elements (or groups of elements) that can be perceived by the listeners as equal or, at least, similar. In the opinion of many authors, the alternation of similarity and differences in a composition has been thought to be the main reference that allows the listener to elaborate a representation of the piece. If this point finds a conspicuous degree of agreement in literature, it equally opens a series of questions that do not have yet a precise answer. In this work, we exactly focus on two of these questions:

1) Is the structure of a piece sufficient in itself to allow the mental representation of the piece or, otherwise, there is the need of implying an active role of the listener through some top-down kind of elaboration?
2) Assuming that the elaboration of the piece moves from the detection of similarities and differences, what are the structural levels of the piece that can be used by the listener to act this process?

In order to deepen the knowledge on this problems, we present two experiments that inquire the segmentation of atonal pieces taking into account the role of performance, duration and learning. On the basis of the existing literature, we also consider the possible effects of expertise, listeners' gender and musical aptitude by controlling these variables.

The two chapters that anticipate the presentation of the experiments have been thought to collect and summarize the current knowledge on music segmentation and grouping.

The Chapter II, that follows this introduction, presents the main modern theorizations related to the elaboration of the representation of music. In a first sub-chapter, we present the recent models of "auditory map of salience", computational models of the elaboration of auditory stimuli, while in the latter section we describe the main models of music segmentation developed through behavioral studies. The main aim of this chapter is the attempt to understand whether an integration of these different approaches is possible and to underline the points that still need to be deepened for a better comprehension of the elaboration of music.

In the Chapter III the focus is instead posed on the concept of salience referred to musical stimuli. As it will be shown, it is impossible unify the several viewpoints existing in literature in a unique, well-defined frame. In order to clarify the common assumptions we choose to divide the research contributions into two broad areas:

1) Studies focusing on "local salience". These works are often focused on single features of the musical surface and use short musical stimuli that are created by the searchers. Even if the choice of composing the stimuli allows a better control of the many variables implied in such kind of research object, these studies leave opened some doubt about the possibility of generalizing the results to the perception and elaboration of real music, where the different types of variables are contemporary present and tied to each other.
2) Studies focusing on "global salience". These works move from the idea that the salience of an element only arises from its relationships with the surrounding elements. These works find an historical reference in the Gestalt theory and are more often realized using real music (mainly classical western music). If it allows a better generalization of the results, it presents at the same time more limits in controlling the whole sample of variables implied.

From this first sight the problem of the cognitive elaboration of music seems to be very complex and articulated. In a real situation of musical listening, however, this complexity can only be increased. At the two sides of a piece, indeed, there are two big variables that are often forgotten in literature: the performer and the listener. Every musicians, as well as every listeners, have their specific personality, motor patterns, perceptual skills and past experience; even if supposing some common, universal mechanisms in the production and fruition of music, it is unlikely that the inter-individuals differences do not product any differences in these processes. While some variables belonging to the listener (e.g., expertise) have been often included in the studies on segmentation, this is not the case of the other extremity of the continuum, the performer. The second part of the second chapter, thus, exactly focuses on these problems with a deepening of the experimental knowledge on musical performance, listener's expertise and musical aptitude. Since the temporal dimension is the one the more influenced by the performer, a specific paragraph has been exactly dedicated to "duration and other temporal variables".

After the introductive chapters, it is here presented an experimental section composed by two behavioral studies on the segmentation of music. The experiments aim to inquire the role of performance and texture in the segmenting of a musical composition during the listening. A second question concerns whether the structure perceived by the listeners depends on processes developing simultaneously with the listening or on an a posteriori synthesis. Finally, we take into account the role of expertise and musical aptitude. For each experiment thirty subjects were asked to attentively listen to two versions of an atonal composition, identify the architecture underlying the piece and mark the boundaries between different segments by pressing the spacebar. The order of presentation of the two versions was balanced. In the first experiment the two performances differ in duration and in many dynamic aspects. In the second experiment, the two performances only differ in duration. Both for the first and the second experiment, results show a good number of coinciding segmentations in the two performances irrespective of the order of presentation. Musicians operate a lower number of segmentations than not-musicians, even if many of the chosen boundaries remain the same. No effect of musical aptitude has been found. The results are discussed with respect to the two question posed above. In the final discussion we suggest a possible integration of the theoretical approaches discussed in the first chapter with a need for the inclusion of inhibitory top-down mechanisms in the elaboration of musical maps of salience.

2 Theoretical background

2.1 Saliency maps

In recent years, an increasing interest can be remarked around the concept of “saliency map”. Unfortunately, the psychological research on perceptual processes has almost always focused on visual perception, with the consequence of remaining unexplored many of the problems peculiarly related to the other modalities.

Authorization requested
The image will be attached as soon as possible

Fig. 2 Itti and Koch’s map of salience.

The problem of building a map of salience for explaining and reproducing the selection and categorization of sensorial stimuli does not represent an exception. In effect, the most famous and cited map of salience, conceived by Itti and Koch (2000) (fig. 2.1), has been developed exactly with regard to the visual domain.

Their aim was that of identifying the basic features of the visual field necessary and sufficient to explain, through a minimum number of simple steps in the elaboration of the external stimuli, the way the visual system could assign to these stimuli the quality of figure and/or ground. One important feature of this model is the concept of "winner takes all", with this meaning that only the most salient stimuli can accede to the top down attentive system, while an inhibition of return stops the same possibility to the remaining stimuli. Even if the work of Itti and Koch pertains to the visual domain, it is useful to cite some other features of their model because, as we are going to see, this saliency map represents a cardinal point in the forwarding elaboration of auditory maps of salience.

In the Authors' opinion, in every single moment three basic features of the visual field are analyzed separately: color, intensity and orientation. For each of these variables, the system provides a detection of the differences between center and surround, followed by a normalization that generates several "features maps" for each of these basic features. Each of these groups of maps is then re-conducted to a "conspicuity map" with a process that includes across-scales combinations and a new normalization. The linear combinations of the conspicuity maps, finally, produce the saliency map. At this point the object resulting the most salient captures the whole attention of the viewer and inhibits the processing of the other stimuli. It is important to denote that, in this model, the attentive processes only can start successively to the completion of the saliency map, while no choice is given to the subjects in the attribution of saliency scores to the stimuli. This point will be analyzed with regard to the auditory perceptual models exposed in the next paragraphs, and will be discussed at the end of this chapter. To conclude this introduction, it must be said that this review does not take into account the models focusing only on the issue of auditory spatial orientation (see i.e.: Hang and Hu 2010) in order to avoid an surplus of information that could preclude more than help the comprehension of the central theme of this chapter, that is, the role of goal-driven and stimulus-driven attention in the uprising and modulation of auditory saliency maps.

The present chapter is structured into two main sub-sections. The first paragraph provides a rapid overview of the main models of auditory map of salience, putting mainly in evidence how the different authors face the problem of stimulus-driven/goal-directed attention. The second paragraph presents three different models of melodic segmentation.

2.2 Auditory maps of salience

2.2.1 Kayser's map

The first auditory saliency map here presented is that of Kayser (et al., 2005) (fig 2.2). For the authors, “saliency describes the potential influence of a stimulus on our perception and behavior” (pp. 1943), while “the salient stimuli are those which are more likely to attract our attention or which will be easier to detect” (idem). Concerning the maps of saliency, then, the authors claim that “in addition to being a theoretical concept describing properties of sensory stimuli, *they* serves as a basis for understanding the cortical representations and mechanisms which implement this weighting of sensory stimuli” (italics by me). Kayser’s model, as claimed by the own authors, is quite similar to the model proposed by Itti and Koch but it is obviously conceived to explain the choice of the stimuli in the auditory scene and not in the visual one. Three auditory features, intensity, frequency contrast and temporal contrast, take the place of the three basic visual features identified by Itti and Koch. With respect to the visual model, the authors insert a preliminary analysis of the frequencies, which allow the formation of the auditory image. The basic features are hence extracted from this image and are automatically organized in maps of features that, through a linear combination, flow to compose the saliency map. The Kayser's model has been tested in two correlational studies realized by the same authors and presented in the same work (Kayser 2005). The first study compares the prediction obtained by the model with explicit responses given by human subjects. Listeners were exposed to two complex auditory scenes and had to explicitly indicate the salient objects they heard. Results show that the model can well predict the people choice but only when they identify something salient in the auditory scene. Nonetheless, the model predicts the choice better than a model only based on intensity. The second experiment proposed concerns a stimulus detection task in monkeys. Authors used six different stimuli, three more salient and three less salient (on the basis of the author's model), inserted on a background noise played by two speakers placed at opposing sides of the animal's head. Monkeys turned their head toward the source of salient stimuli in a high percentage of trials both for the less and the more salient stimuli but with a significant level only for the more salient stimuli.

The model developed by Kayser does not consider any active role of the listener. On the opposite, it states that the features of the auditory stimulus are sufficient in themselves to obtain a saliency map.

Authorization requested
The image will be attached as soon as possible

Fig. 3 Kayser’s map of salience

2.2.2 Kalinli's map

The saliency map proposed by Kalinly and Narayanan (2009) (fig 2.3) moves from biological data, and takes into account the saliency of speech. Following the author’s explication, we can divide the model into three different steps.

The first one, the “multiscale feature map generator”, mimics the processing stages in the early and central auditory systems toward the realization of an auditory spectrum with time and frequency axes that the authors call “scene”.

Authorization requested
The image will be attached as soon as possible

Fig. 4 Kalinli’s map of salience

The scene is then analyzed to extract a set of multi-scale features that the authors identify on the basis of the information processing stages in the central auditory system. These features include intensity, frequency contrast, temporal contrast, orientation and pitch (contour). The term "orientation" is used to mean the slope in frequency. Some more words are necessary on the introduction of pitch contour: in effect, the authors allow and promote the possibility that top-down attentive mechanisms are active and influent on the selection of salient stimuli already in these early processes. Going on in the exposition of the model, center-surround differences are evaluated resulting in the feature maps. This process aims to mimic the properties of local cortical inhibition. The second step, called “gift features”, clearly recalls the fuzzy-trace theory proposed by Brainerd and Reyna (i.e.: Reyna and Brainerd, 1995), where the gist traces are defined in opposition to the verbatim traces, the first being fuzzy representations and the second detailed representations of a past event. Gist traces can be distinguished in perceptual gist and conceptual gist, where “perceptual gist refers to the representation of a scene built during perception, and conceptual gist includes the semantic information inferred from a scene and stored in memory” (Kalinly and Narayanan, 2009). The authors explicitly focus on perceptual gist, thought as a relatively low-dimensional acoustic scene representation, which describes the overall properties of a scene at low resolution. Every map of features produces a gist vector through an extraction process. These vectors are hence combined in a cumulative gist vector. The authors also consider a “principal component analysis to reduce redundancy. The third and last step is the “Task-Dependent Biasing of Acoustic Cues”. (Fig 2.4)

This step has been introduced to explicate the role of top down modulation on the target search, that is, implying the possibility of a central influence on the attribution or at least modulation of the saliency scores. The modulation proposed by Kalinly (et al, 2009) is explicable in terms of enhancement/inhibition of the response of neurons respectively tuned/detuned on the features of the target stimuli. Fundamental in this model is the presence of a “learner” that exactly provides the guidelines of the top-down modulation. To conclude their paper, the authors present a series of experiments that aim to investigate the validity and the limits of their own model. To conclude this paragraph, it must be specified that this model does not take into account the role of the spatial position of the stimulus.

Authorization requested
The image will be attached as soon as possible

Fig. 5 Kalinli's model: the Learner

2.2.3 De Coensel's map

Going on in the exposition of the existing models of auditory saliency map, it is here presented the point of view proposed by de Coensel and Botteldooren (2009, 2010) (Fig. 2.5). Before explaining it in details, it is useful an overview of some important points that characterize this map. Conceived as a computational model, this map has been built in line with the ASA (Auditory Scene Analysis), the peculiar approach to the study of auditory perception proposed by Bregman (1990). Coensel's map can hence be view as a CASA (Computational Auditory Scene Analysis) model. To explain the ASA point of view, we cite the words of the same authors: "In a simplifying manner, ASA is often regarded as a two-stage analysis-synthesis process. In the first stage (segmentation), the acoustic signal is decomposed into a collection of time-frequency (T-F) segments. In the second stage (grouping), segments that are likely to have arisen from the same environmental source are combined into auditory streams" (de Coensel and Botteldooren, 2010, pp. 2).

Authorization requested
The image will be attached as soon as possible

Fig. 6 De Coensel's map of salience

As for the map of Kalinli, this model takes into account the integration of bottom up and top down processes but with a later intervention of the second ones into the perceptual process. Another analogy with the former model stays in the exclusion of the sound location when providing the set of auditory variables considered. The explicit reference to the ASA theory is instead observable in the consideration that the authors give to the problem of the streams segregation, not considered in the former models exposed. De Coensel and Botteldooren divide their model into four steps that try in a certain measure to tie their computational description to the biological course of auditory information. During the first step, the "Peripheral auditory processing", the whole sound wave, described as the set of the environmental sounds in a given moment, is filtered and then rectified, integrated and compressed logarithmically into a spectrogram. This step mimics the processing of the sound through the tympanum and the acoustic nerve. During the next step (Auditory saliency map) the spectrogram is newly filtered on the basis of intensity, spectral contrast and temporal contrast. The output, a set of raw maps, is then converted in a set of feature maps (center-surrounds differences and normalization) and hence in a set of conspicuity maps (across-scale combination and normalization). The relation among the conspicuity maps, mainly of inhibitory nature, generates the saliency map. The second step aims to mimic the properties of the auditory cortex. The third and forth steps proposed by the authors represent the most original point of their work. The third step (stream-specific saliency scores) walks in parallel with the second one; the authors consider that the whole sound wave is composed by different simultaneous sounds. As for the whole wave, each stimulus is filtered and compressed to produce a specific spectrogram and a specific time-frequency mask (T-F Mask). The authors propose two different ways to calculate the T-F Mask, the first based on the ratio with the whole wave, the second conceived on a binary system. For each stream can be now calculated a saliency score; saliency combines additively across frequency channels. The last step, the "Auditory attention switching", is the lonely directly linked to attentive phenomena. Significant saliency scores activate the bottom-up attentive system. The selection of the stimuli that can be allowed to arrive in the working memory occurs on the basis of an inhibition of return mechanism. The "winner takes all" competition includes not only the auditory streams but also stimuli coming from different modalities. The intervention of top-down attention, strictly related with the WM system, shows itself in the modification of the choice of the competitors that have *already* overpass the filter represented by the bottom-up attention. In other words, de Coensel and Botteldooren do not consider the possibility that top down processes could participate in the attribution of saliency scores to the stimuli. Anyway, their model is mainly though to explain the formation and choice of auditory streams while it does not inquire directly the emergence of salient features inside a single stream.

2.2.4 Duangudom's map

The model elaborated by Duangudom and Andersen (2007; fig 2.6) explicitly moves from, and is thought relatively to, the visual map proposed by Itty and Koch (2000) and the auditory map provided by Kayser (Kayser et al., 2005).

Authorization requested
The image will be attached as soon as possible

Fig. 7 Duangudom's map of salience

Importantly, the authors acquiesce to the existence of two different perspectives on salience, that is, top down and bottom-up saliency, even if they describe top-down salience as the property of a sound to violate the expectations based on models of auditory scenes that have been previously learned: "from a top-down perspective, humans use previously learned models to understand complex auditory scenes and salient sounds are sounds that violate these models" (Duangudom and Andersen, 2007, pp. 1206). This idea (or, at least, this formulation) should need to be extended with a clear definition of the boundaries between the two kinds of salience described by the authors, in order to avoid some confusion on the same model that they propose. Overlooking this point, Duangudom's model focuses on bottom-up salience and proposes a general architecture for an auditory map of salience. Although at first sight this schema could seem very similar to Kayser's one, there are important differences that deserve to be argued. In the former models that have been shown it was always present the step called "Center-surround differences", which works as a dynamic filter that allows the transitions of the only stimuli having salient features. In other words, the quality "salience" organizes itself in the neighborhood of a central value; the filter stops every element having a value that falls out of a range built around the central point. Duangudom (ibidem) proposes a mechanism of inhibition active already inside the features maps, with the salient range inhibiting automatically the out-of-range elements. This active functioning aims to mimic the analogue neuronal organization in the primary auditory cortex as described by Wang and Shamma (1995). The authors place a second inhibition process before the linear combination of the feature maps in a saliency maps, and provides a competition between different sets of features. The authors provide different examples of the application of their algorithm and present an experiment with the scope to compare their model to human explicit answers. The task was based on the comparison within couples of auditory scenes relatively to the higher or lower salience perceived by the subjects. The results showed a good correlation between the answers given by the subjects and the computational model.

2.2.5 Coath's map

The next and last model that is going to be presented arises from a perspective, which is quite different from the one of the former models because of the focus posed only on the spectral dimension of the sound. In a first exposition Coath (and Denham 2005) focuses on speech classification and categorization; in a following work (Coath et al., 2009) the authors try to extend their algorithm to include the identification of the positions of perceptual onsets in musical stimuli. We present this model following its evolution for a better comprehension.

Authorization requested
The image will be attached as soon as possible

Fig. 8 Coath's model

Coath's model (fig. 2.7) is based "on the idea of 'spectro-temporal response fields' (*STRF*, italics by me), and uses convolution to measure the degree of similarity through time between the features detection and the stimulus" (Coath and Denham 2005, pp.22).

Already from this first statement, it is possible to observe the biological perspective of the authors. Nonetheless, Coath does not refers to the common first-order isomorphism across stimulus and neuronal response but prefers a second-order isomorphic mapping which provides for a conversion of the stimulus in a low-dimensional space. Importantly, this conversion must preserve the similarities between the stimuli. Any sound can then be positioned in the response space spanned by an ensemble of STRFs. Coath's model consists of six processing stages. The first and second steps, the "spectral decomposition" and the "transient extraction" mimic the behavior of the cochlea and of the sub-cortical auditory system. The third step, "convolution using a bank of STRFs", provides the detection of similarities between each incoming pattern and each STRF participating in a common ensemble. The next two steps, "event detection" and "mapping to response space" provide the process of segmentation of the stimulus. The word segmentation is here used in line with the ASA and pertains (although not explicitly) to the fragmentation and re-combination of a streaming. The most important point in the authors' view is that "the stimulus-ensemble-driven event detection results in a method of segmentation where auditory events are marked by coherence in the response of the ensemble, and not wholly by properties of the stimulus" (Coath and Denham 2005, pp.25). This peculiar view oversteps the classical distinction between bottom up and top down attentive processes and can be conceived as a "bridge" which provides an original perspective to approach this problem. Remaining in the field of a pure conjecture, one could indeed conceive the higher level processing as an extension of these basic mechanisms. Obviously, this view presents the limit to exclude the problem of consciousness. The sixth and last step consists in the classification of the patterns, which is realized through the mutual information between the class of the stimulus and the class of the response. In a more recent work, Coath (et al., 2009) uses his model to explain also the insurance of tactus and rhythm.

2.3 Theories on segmentation

2.3.1 Overview

The segmentation of a musical composition can be view as the other side of the more cited and studied theme of the musical grouping. While the latter word refers to the perceptual property of assembling different elements in a more complex perceptual unit, the word segmentation concerns the ability of dividing a complex stimulus into different, simpler elements. Music is a temporal phenomena, thus the segmentation of a composition during the listening necessarily refers to the division of the piece in a series of shorter, temporally located excerpts on the basis of laws and rules that several scientific researches have tried to identify. Many authors have posed and pose their attention on single features of the musical surface that allow and/or lead to the identification of precise boundaries between two segments in a piece. A clear example is provided by the Gap and Run rules proposed by Garner and Gottwald (1968), with ties the segmentation of a looping rhythm to the differences in duration within the tones and the rests of the phrase. Other authors put in evidence the role of different kinds of features in the mechanisms of musical grouping/segmentation: melodic accents (e.g., Deutsch 1999; Pfordresher 2003), harmonic or tonal accents (e.g., Boltz 1991; Pearce and Wiggins 2006), or accents deriving from an integration of two or more musical features (e.g., Jones 1987; Bigand and Pineau 1996). Regardless of the different targets, these contributions find a common point in the choice of posing the attention on local features of the musical stimulus. On the other side, there is a consistent number of works that try to analyze the problem of segmentation from an holistic perspective. These latter researches have been developed along a continuum that starts with the Generative Theory of Tonal Music (Lerdahl and Jackendoff, 1983) and that is currently in progress both in a more or less linear development of the original ideas (e.g., Ockelford, 2009) and in an attempt of linking these ideas with more recent trends in the cognitive sciences (e.g., Cambouropolus, 2001; 2003). However, the first experiments on musical segmentation based on the detection, during the listening, of the boundaries between chunks in atonal music can be found in the works of Deliège (1987) and Imberty (1987). The theoretic hypotheses deriving from this line of research have been described in the next paragraphs.

2.3.2 The Cue Abstraction Hypothesis

Since the ’80 of the last century, Deliège proposed a simple but effective theory, the *Cue Abstraction Hypothesis* (Deliège 1987; Deliège 1989; Deliège and El Ahmadi 1990; Mèlen and Deliège 1995; Deliège 1996; Deliège et al. 1997; Deliège 2001). In listening to a piece some elements stand out, becoming themselves a guideline to the processes of categorization and representation of the musical surface. These elements, that Deliège calls *cues*, allow the listener to form in his mind a plan of the piece, by dividing it in chunks on multiple hierarchic levels. The cues represent prototypes of redundant elements; music then, would be early analyzed with reference to similarities and differences between the chunks. Deliège's model is explicitly inspired by the Gestalt viewpoint that the author tries to overstep with the reduction of several perceptual principles (proximity, similarity, common fate, closure, and good continuation) in a simpler perspective: "… principles arose directly from the Gestalt principles of which they are an extension, or more precisely, a generalization insofar as the psychological mechanisms concerned (previously analyzed in terms of proximity, similarity, common fate, closure, and good continuation) are henceforth divided into only two categories: similarity proposes that small differences between elements within a unit are minimized, and difference assumes that the contrast between elements adjacent to a boundary is emphasized" (Deliège, 2001, pp.235-236). Thus the main concepts involved in Deliége's theorization are those of "same" and "different". The term "same" does not refer mono-directionally to the exact repetition of musical structures that represents only the extreme degree of similarity between chunks. This concept, indeed, finds a reference in Lerdahl and Jackendoff's "A generative theory of tonal music" (1983) and its subsequent extension to the atonal music by Lerdahl (1989). Using the four parameters described by Lerdahl and Jackendoff, atonal music can be categorized on the basis of the "prolongational structures", structures of tension and distension. This is due to the weak possibilities of grounding the understanding of the piece on metrical or tonal structures that are not well identifiable.

The theory by Deliége derives from her studies on atonal pieces that are mainly monotonic, thus it is possible to hypothesize that the cues have been extracted on the basis of physical (timbre, intensity), melodic (contour and intervals but not tonal constraints nor harmony), rhythmic (but not metric). The not-independence of these parameters with each other can give account of the difficulty of separating them in a theoretic model. For this reason, with the term "cue" we identify a not better-specified musical element that can be perceived as salient for a single feature as well as for the interaction of different features. These elements, formed by small groups of tones that are temporally close, are automatically chosen as labels for longer chunks. Different chunks can be thus categorized with reference to a same cue that belongs to them both. During the listening of a composition, the early and semi-automatic storage of such labels in memory allows the comparison between the musical elements heard in a given temporal moment and the ones that have been already listened. Through the definition of a restricted number of categories, the listener can represent the piece as a series of variations of the very basilar (redundant) elements of the piece-self. The fig. 2.8 illustrates the steps in elaboration theorized by Deliège and col. (In El Ahmadi and Deliège, 1990).

Authorization requested
The image will be attached as soon as possible

Fig. 9 Deliège's Cue Abstraction model

Deliège's conception of cues is strictly related to the model of categorization elaborated by Eleanor Rosch (Rosch 1978), who distinguishes two levels in categorization: the horizontal level, based on relations of similarity and contrast between elements; and the vertical level, related to hierarchic levels. Music can be view as a temporal pattern of auditory events; melody, instead, can be described as a pattern of tones and durations (e.g., Jones, 1987; Bigand and Pineau, 1996; Schwarzer, 1997). Since melody represents a special type of serial pattern (its serial nature is here strictly connected with the temporal dimension), melody has been expected to preserve the some of the properties described by Frank Restle about the learning of the serial patterns (Jones 1987). The very basic feature of these mechanisms is the possibility of including every group of elements in a larger group, on the basis of the recognizing of repetitions, mirrorings or transpositions (Restle, 1970; 1973). The final result is a tree-shaped structure as in fig. 2.9.

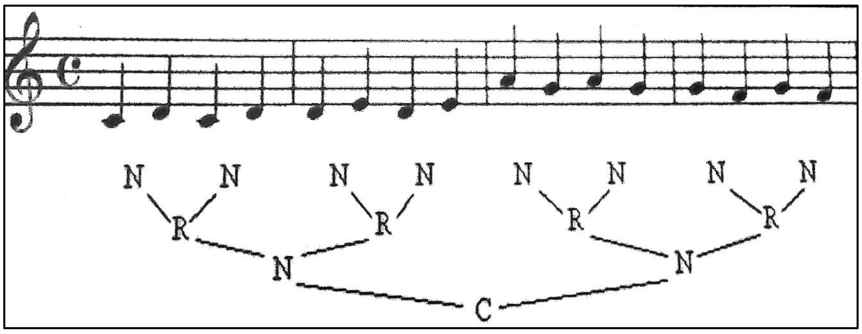


Fig. 10 Hierarchic levels in music

Many authors (Jones, 1987; Pfordresher, 2003; Ockelford, 2006), in effect, exactly try to transform the relationships between tones in numerical ratios, with the result of converting the musical stimulus in a serial pattern of rules. Regardless of the issues arising from this practice (e.g., Do these rules pertain to an early, pure auditory elaboration or only to a further, later step?), the melodic patterns contain another important source of information that cannot be forgotten: the tensional dimension. Imberty (1987, 2000) and Deliège (and Ahmadi, 1990), as well as other authors interested in music segmentation (e.g., Clarke and Krumahnsl, 1990; Lamont and Dibben, 2001), find that the tensional dimension is the main variables involved in the elaboration of atonal music. A problem with the concept of "musical tension" is related to the difficulty in understanding what are the musical elements that provide, by themselves or through some interaction with each other, the tensional structure of a composition. Unfortunately Deliège does not directly deepen this problem. It will be better discussed into the second chapter.

A last observation on the studies discussed in this paragraph concerns the experimental paradigm adopted. Since these studies focus on the segmentation of atonal music, supposed to be hard to understand especially for the naïve listeners, the procedure provides for a first "not experimental" listening in order to allow a better familiarization with the stimulus. The consequence is that both the first and the second experimental sessions are realized when the listener has already heard the whole piece. In this condition it is impossible to understand the boundary (hypothesizing its existence) between a stimulus-driven and a goal-directed segmentation.

2.3.3 Ockelford's Zygonic model

The Zygonic theory, developed by Ockelford in the last decades, represents a possible evolution of Deliège's Cue abstraction hypothesis. The theory is here described with reference to Ockelford (2006), where the author tries to summarize the main points of his viewpoint. The basic concept, which provides also the name to the theory, has been described with the term "zygon". This Greek word indicates the yoke, a device used for attaching two animals to the same plow. The metaphor, thus, under means a mechanism that allows to tie two musical groups and to perceive them as "variations" of the same element (fig. 2.10). Already from this first sight it is possible to notice the strong relation, explicated by the author himself, with the same-different relations that constitute the basis of Deliège's hypothesis. The detection of similarity between chunks arises from the detection of "interperspective relationships" (fig. 2.11), that is, relationships based on mechanisms quite similar to those described by Restle (1970) and extended to music by Jones (1987). A melody can be repeated, transposed or temporally shifted on the metric structure of the piece and it can equally provide a sense of continuity in the listener's perception. The detection of such kinds of relationships between elements is automatic and active during the whole listening. Furthermore, this kind of elaboration has a dynamic nature, with this meaning that the reference points (what we should call cues in Deliège's theory) do not remain exactly the same during the listening. On the opposite, they develop and this development allows the listener to create expectations on the ongoing music. The result of this process is the elaboration, in the listener's mind, of a sense of auditory identity of the piece that can be perceived as a continuum regardless of the number of variations it contains. The dynamic nature of the labels gives account of the possibility to perceive as similar two very different musical chunks if the latter one is obtained from the first through an even long series of little, consecutive variations. This is what exactly happens in a precise musical genre, the "variazioni su tema". Another consequence of this model is that every musical piece, regardless of the previous experiences of the listener, produces by itself a familiarization with its rules that allows the listener to read it in a specific way. In classic experiments on segmentation (i.e., El Ahmadi and Deliège, 1990), the expertise of the listeners seems to interact with the number of segmentations they produce but not on the identification of the main boundaries between segments, that is, experts and naïve subjects have the same ability of identifying the basic structure of the piece. Ockelford's theory can explain these results since it implies that the perceived structure of a composition exactly arises from the way the piece develops. The more the composer speaks, the better the comprehension of the listener.

Turning into a more psychological perspective, Ockelford does not take a strong position on the cognitive level of elaboration of the interperspective relationships. Reasoning on the reference to a sub-conscious mechanism, it is possible to hypothesize that the author attributes to the detection of such features an essentially bottom-up nature, but the missing of an explicit reference does not allow to discuss adequately this point.

Authorization requested
The image will be attached as soon as possible

Fig. 11 Zygonic relationships in Ockelford's model

Authorization requested
The image will be attached as soon as possible

Fig. 12 Interperspective relationships

2.3.4 Cambouropolus' model

The Local Boundary Detection model (LBDM) is a computational model developed by Cambouropoulus from the second half of the nineties. Following the words of the author himself, the LBDM "calculates boundary strength values for each interval of a melodic surface according to the strength of local discontinuities; peaks in the resulting sequence of boundary strengths are taken to be potential local boundaries" (Cambouropoulos, 2001). The first version of the model, introduced in two works dated 1996 and 1997 (Cambouropoulos, 1996; 1997), has been developed during the following years, also through the introduction of new variables (e.g.: the degree of change between time/pitch intervals; Cambouropoulos 2001). The main principles at the basis of the model, however, remain the same during this development and can be described as two basilar rules:

1. Change rule: The strength of the boundary between two intervals is proportional to the degree of change (difference) between them. As for the second rule, the term interval can describe either a temporal interval or a pitch interval.
2. Proximity rule: If two consecutive intervals are different, the boundary introduced in the larger interval is proportionally stronger.

We can thus observe that the two rules proposed by Cambouropoulos move from a coherent viewpoint that allows two considerations: first, both the Change and the Proximity rule have been thought to be applied to the relationships (intervals) between elements, not to the elements selves. The strongly relational nature of this model does not take into account any absolute parameter (e.g., pitch, timber, etc.); thus, if we consider two melodies with the second one obtained through a transposition of the first one in a different tonality, we will obtain exactly the same values for each of the internal boundaries. Moreover, the same result will be obtained with a change in the global tempo of the piece. On one hand, several researches show that listeners can recognize a melody even if transposed (e.g., Dowling and Harwood, 1986). On the other hand, this is not absolutely true for changes in tempo (Monahan and Hirsch, 1990; Handel 1992). As showed by Handel (1993) proportionally scaled rhythms can often be not recognized as being identical by the listeners. The problem of tempo is not the only limit of the LBDM model that, as observed by the author himself (Cambouropoulos 2001, p. 2): "…not include harmonic profiles of melodies and also it does not take into account melodic similarity which is paramount for establishing important groups of notes". Shifting to the second consideration, we notice that both the rules arise from the detection of differences (the opposite of similarity) between time/pitch relationships. With respect to the models presented in the former paragraphs, however, the concept of similarity has not been applied to musical chunks. To avoid the limits arising from this absence, Cambouropoulos introduced in a later work (2003) the PAT (Pattern Boundary Detection Model).

Even if this model still deserves to be improved with the inclusion of some important variables, it has been shown to be already able to identify successfully a large number of local boundaries in one evaluation test that compares the LBDM with the punctuation rules developed by Friberg (et al., 1998). With respect to melodic similarity, Cambouropoulos (2003) proposes an integration of the LBDM with a new computational model, the "Pattern Boundary Detection model (PAT)", described as below:

"The pattern extraction procedure is applied to one (or more) parametric sequences of the melodic surface as required. No pattern is disregarded but each pattern (both the beginning and ending of pattern) contributes to each possible boundary of the melodic sequence by a value that is proportional to its Selection Function value. That is, for each point in the melodic surface all the patterns are found that have one of their edges falling on that point and all their Selection Function values are summed. This way a pattern boundary strength profile is created (normalized from 0-1). It is hypothesized that points in the surface in which local maxima appear are more likely to be perceived as boundaries because of musical similarity”. (Cambouropoulos 2003, p. 2). Similarity, hence, comes back to be one of the main actors in the cognitive elaboration of a composition.

3 From salient elements to the role of performance.

3.1 Atonal music, grouping and tempo

"A generative theory of tonal music", the book published by Fred Lerdahl and Ray Jackendoff in 1983, represents beyond any doubts one of the most important reference points for the understanding and the investigation of the cognitive processing of music. Inspired by the theory of Noam Chomsky (e.g., Chomsky 1968/1972), the authors tried to extend the concept of universal grammar from the spoken language to the tonal music. In line with a hierarchic idea, Lerdahl and Jackendoff propose four basic mechanisms, innate, that allow the listeners to understand the plan of a piece and to divide it according to different grouping levels. At a very basic level, the musical elements can be grouped in agreement with simple rules, such as proximity, symmetry or parallelism, that overlap in a certain measure the gestalt rules described in a further paragraph. Importantly, the authors underline that "only contiguous sequences can constitute a group" (Lerdahl & Jackendoff 1983, pp. 37); This unavoidable feature of grouping emphasizes the temporal nature of music and leads to conceive temporal proximity as the basic constraint for the listener's understanding of the piece. The simple rules above mentioned, however, belong to the first level of analysis proposed in the book at issue, the "Grouping structure", that joined with the "Metrical structure" constitutes the bottom-up level of elaboration of the musical surface. The perception of a meter is strictly related to the detection of three different kinds of accents: phenomenal, structural and metrical, defined as follows:

"By phenomenal accent we mean any event at the musical surface that gives emphasis or stress to a moment in the musical flow… By structural accent we mean an accent caused by the melodic/harmonic points of gravity in a phrase or section-especially by the cadence, the goal of tonal motion. By metrical accent we mean any beat that is relatively strong in its metrical context." (Idem, pp.17)

If, following Meyer (1956), we think to an accent as "something marked for consciousness" (hence a salient element), we can see that in Lerdahl and Jackendoff's theorization this salience can also be determined by cultural elements (structural accent), unless we conceive tonality as an innate and universal feature of music. Anyway, the stating for an influence of culture in the elaboration of a representation of the musical piece does not necessarily imply the need for top-down processing. This need belongs, on the opposite, to the further two levels of elaboration proposed by the authors: "time-span reduction" and "prolongational reduction". While time-span structure provides the perception of stability, prolongational structure is related to the contextual salience of tones or groups of tones. Moreover, in atonal music prolongational structures have a main role. A segmentation paradigm applied on atonal pieces, then, represents the best way to inquire these ideas (Imberty 2000). If we give a main role to prolongational structure, then we have to forecast that atomic elements, like long rests or melodic pivots, only have a secondary role in the choice of the boundaries between melodic segments. In effect, Lerdahl (1986) proposes three kinds of prolongation: strong prolongation (an event repeats), weak prolongation (an event repeats in altered form) and progression (connection with a new different event). This point puts the emphasis on same-different relation between elements as the main feature involved in the segmentation of atonal music. It is useful to specify that Lerdahl mainly refers to the dimension of tension/distension, which is supposed to be not necessarily related to structural elements. Nevertheless, Boltz (1998) finds that the perception of tempo can be influenced by the density of changes in direction of the melodic contour (peaks), as well as by the incompatibility between rhythmic accents and melodic peaks, that should carry the listener to be unable to perceive the musical surface in an unambiguous way. The perception of temporal tension/distension thus can be influenced by at least two orders of structural factors, i.e., rhythmic and melodic features of the musical surface. These results overlap some previous findings that show how the speed of melodic rhythm can influence the perception of tempo regardless of the actual beat rhythm of the piece (Kuhn 1987, Duke 1989). Geringen (et al., 2006), moreover, finds that the perception of tempo depends on the performer's style: playing a melody with *staccato* mode produces in the listener the impression of an increasing tempo, while the *legato* mode does not. Then, what does it happen when two musicians perform the same piece with different prolongational structures? One may suppose that the temporal dimension perceived by a listener will be related to both the structural nature of the piece and to the changes in tempo due to the specific of the musicians' interpretations. Tempo modulation, then, can produce both real changes in tempo and a different reading of the rhythmic and melodic accents structures that, agreeing with Boltz (1998), will produce in turn a further effect on the subject's perception of tempo. Dèliege and Ahmadi (1990) compared the structure of an atonal piece as perceived by the listeners with the structure emerging by the analysis of the score done by two composers. The results underlined a very similar sub-division of the piece into six different segments on the basis of "dynamics". Having the composers analyzed the piece only on the basis of the scores, these dynamics were essentially tied with the melodic and rhythmic features of the piece (that is, structural elements), while any effect seemed to be produced by the interpretation of the performer (what we call here "dynamic elements"). The natural question we have to pose, thus, can be expressed as follows: where is the limit between the piece itself and the outcome provided by the performer's interpretation? Two paragraphs of this chapter try to deepen this problem with the focus posed, respectively, on musical performance and on the temporal variables involved in music. A last paragraph faces the issues of expertise and musical intelligence in relation to music segmentation.

3.2 Global salience: The Gestalt psychology of music

The Gestalt psychology represents a well known psychological movement that has been developed in the beginning of the 20th century. The formal birth of this School is usually tied to the works of the Berlin School (with Wertheimer, Koffka, Köhler, Metzger as main proponents), which gave the start to a long series of studies aiming to illustrate and sometimes to explain the principles of perception. Among the forerunners of Gestalt psychology, a special role is certainly played by C. von Ehrenfels, whose writings represent indeed the basis for the term "Gestaltqualität" (Eichert et al., 1997) which describes the particular property of certain stimuli that remain invariant in perception even if physically varied. Von Ehrenfels developed this concept with respect to musical melodies, which can be indeed recognized even when transposed in a different tonality. The accent is thus posed on the relationships between the tones composing the melody more than on the tones themselves, hence highlighting that the perception of a complex stimulus cannot be exhaustively explained with an atomic associationism.

Even if melody could be view as the object that marks the origin of the ideas at the basis of the Gestalt Theory, the latter psychologists of the Berlin School, as well as their followers, elaborated a great corpus of phenomenological laws concerning visual but not auditory perception. For a strange fate, only in a second moment the Gestalt principles have been tested with regard to auditory issues. Nonetheless, when exploring the literature from the sixties of the last century it seems not to exist any insurmountable differences between the Gestalt rules identified with regard to visual perception and their equivalent in the auditory domain.

This similarity leads to suppose that the principles proposed by the theorists of Gestalt are a-modal, that is, not specific for sensorial channels. Anyway, regardless of the overlapping between different modalities, it exists a quite strong difference between visual and auditory perception that cannot be ignored: the temporal dimension. Music is, indeed, a temporal phenomenon, since it can emerge only grouping successive elements in a limited segment. It is hence evident that, from a cognitive point of view, an explication of the Gestalt principles in the auditory domain may necessarily move from the listener's comparison between element that are temporally dislocated with each other. This necessarily implies that, in order to verify some Gestalt principle such as similarity, the listener needs to compare the information received in a certain moment with the one previously heard. On the other hand, the search for an explication of the Gestalt principles is far from the original point of view of the Gestalt theorists. Following Tenney (Tenney & Polansky, 1980), "...the principles, as stated, were not "operational," but merely descriptive. That is, although they were able to tell us something about TGs (Temporal grouping, N/A) whose boundaries were already determined, they could say nothing about the process by which that determination was made. They described the results of that process, but not its mechanism". To overstep this impasse, it seems to be useful a quick overview on the auditory grouping principles inspired to the Gestalt that can be found in the scientific literature and on their experimental basis.

Getting a look to the literature focusing on the Gestalt principles in the auditory domain, it is possible to meet a good number of experiments that provide evidences for the possibility to apply some of the originally visual-derived laws in the perception of music. Already at the beginning of the eighties, Diana Deutsch (1982) presents a review concerning the "organizational processes in music" where she discusses the applicability in music perception of four Gestalt principles, i.e., proximity, similarity, good continuation and common fate. The new sub-paragraphs focus on three Gestalt rules: proximity, good continuation and common fate. Similarity has not been included in this paragraph since the theme has been deepened in the former chapter.

3.2.1 Proximity

In the visual field, the law of proximity describes the tendency of close elements to be perceived as a whole. To apply this principle to the perception of music, it has to be considered that there are at least two kinds of proximity that deserve to be mentioned: frequency proximity and temporal proximity. Concerning the first one, the law can be described as follows: tones that are close in frequency will tend to be perceived as a group. This principle has been tested with different techniques. In two dichotic listening experiments, Deutsch (1975a, 1975b) invited the subjects to listen to two V-shape contemporary melodies, with the condition that when an ascending melody was in the right ear, a descending melody was in the left ear, and vice versa. When the listeners where asked to describe the melodies heard respectively in the two ears, they tended to place the higher tones in one ear and the lower ones in the other ear. Deutsch explain her data claiming that the subjects group the tones by proximity. A similar result has been obtained by Butler (1979) using two speakers placed on the left/right of the listeners instead of the headphones. Another evidence of the effect of frequency proximity is provided by Dowling (1973) in four experiments that inquire the recognition of well-known but overlapped melodies. The performance of the listeners improves when the two melodies are played in two different frequency ranges, while it decreases when using the same range. In a successive study, Dowling and Hollombe (1977) find that the difficulty in recognizing a well-known melody increases when the tones of the melody are placed in different octaves; this result can be view as a further evidence for the principle of proximity if considering the experimental proofs for the generalization of the octave (Deutsch, 1972). In a further work, Deutsch (1978) asked the subjects to hear short sequences of tones and to decide whether the first and the last pitches are equal or different. The results were more accurate when the interpolated sequences were composed of smaller melodic intervals. Finally, Deutsch (1991) shows that the preference for grouping pitches that are close in frequency is active also with simultaneous tones. In her experiment, the subjects had to listen to two consecutive couples of tones, with the second ones respectively spaced one semitone from the first two. In this condition, the listeners tend to perceive two different melodies organized in accordance with the law of proximity.

As told at the beginning of this paragraph, there is another kind of proximity that has to be considered, since it is one of the strongest organizational principles in music perception: temporal proximity.

Starting from 1962, Garner (Garner 1962, Royer & Garner 1966, Garner & Gottwald 1968, Royer & Garner 1970, Garner 1974) elaborated an experimental paradigm with the aim of inquiring the perceptual organization of serial patterns. The technique consists in the presentation of sequences of single tones or couples of tones played in loop. Subjects are asked to chose where they think that the pattern begins. Following the results of his experiments, Garner proposes two principles at the basis of serial grouping: the Gap principle and the Run principle. The first one asserts that the beginning of the pattern is perceived on the tone following the largest rest, while the second one states that, if there is not a longer rest, the beginning of the pattern is perceived on the first tone of the longer group. Both the Gap and the Run principles, hence, underline that the listeners organize tones that are closer in time in a common group. A further phenomenon that allows to understand the strong effect of temporal proximity on serial grouping is that of the perceived accent elicited precisely by longer rests. Several works show indeed that the presence of a rest can make the listeners perceive an accent on the tones that precede (Fraisse 1982, Drake et al. 1991, Drake & Palmer 1993, Pfordresher 2003) or follow (Jones 1987, Jones & Pfordresher 1997, Tekman 2002, Pfordresher 2003) the rest, or in both cases (Povel & Essens 1985). The evidences for these perceived accents seem to mark the emergence of new groups of tones as well as the end of a former group, giving to temporal proximity a very important role in the mechanisms of musical grouping and segmentation.

3.2.2 Good continuation and common fate

The terms "good continuation" and "common fate" indicates well-known Gestalt concepts. Developed with regard to visual perception, the principles can be expressed as follows:

1. Common fate: the elements involved in a similar movement are perceived as being part of the same greater element.
2. Good continuation: elements that are contiguous in space are grouped together if having the same direction.

With respect to the perception of music, these principles need to be shifted into the auditory domain; the natural consequence is that of taking into account the temporal nature of music: the grouping of different elements into a specific chunk only can be developed when the listener knows the several elements to group. The question opened by this point concerns whether the listener can form a representation of the object already during the listening or only when he has heard the whole sequence of tones in the piece.

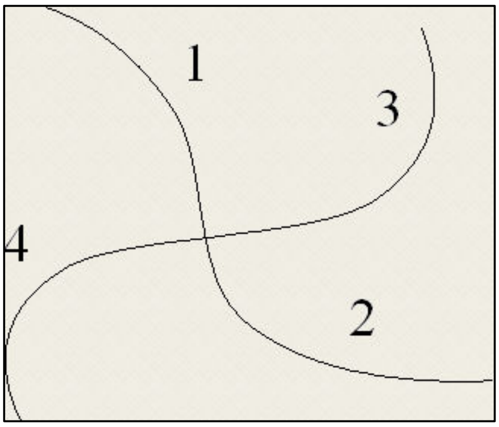


Fig. 13 Common fate in visual perception

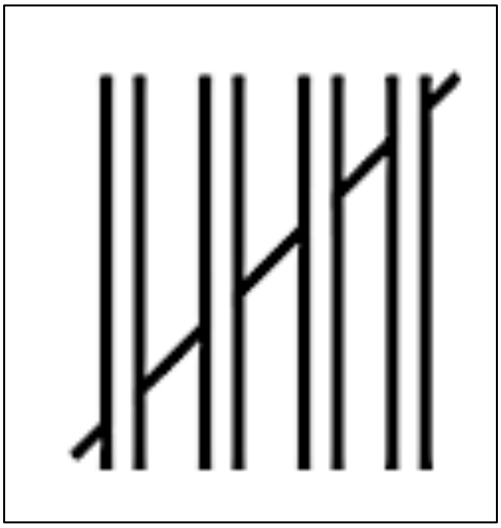


Fig. 14 Good continuation

A key-concept related to this question is that of melodic expectancy, strictly connected with the law of good continuation. Before coming back to this point, we should define the two laws with regard to the perception of musical stimuli:

1. Common fate; to define this rule we use the words of Deutsch (1979, p. 400): “when a sound mixture is presented such that one ear receives one portion and the other ear receives a different portion, it is unclear which elements of the total spectrum should be assigned to one source and which to another, or, indeed,

whether two sources rather than one are involved. If the onsets of these two signals are strictly synchronous, then, by the 'law of common fate,' this is evidence that a single source is involved."

2. Good continuation: This principle, described by Bregman and Dannenbring (1973) "would lead listeners to group notes together that form a coherent line, and a logical melody. This would work in conjunction with the preference for linear, stepwise or small-interval motion…" (Stainsby and Cross; 1982, p.54).

The first principle, hence, can be related with the localization of sounds and it becomes important when the listener is exposed to auditory stimuli produced by different sources. Using the differentiation of Bregman (1990), the law of common fate clearly concerns the problem of "simultaneous streaming" and it is thus poorly related to the temporal segmentation of melodic sequences (the "sequential streaming" of Bregman; idem). On the opposite, the principle of good continuation exactly concerns the issues of temporal grouping and segmentation in music. This topic deserves thus to be deepened soon after a brief parenthesis: the gestalt principles have been often discussed because of their descriptive nature that, even if underlying perceptual mechanisms that can be easily reproduced trough experiments, do not give any explications on the biologic substrates that produce these behaviors (Tenney & Polansky, 1980). Exploring this issue is not a goal of this work, since this review only tries to make some order into the broad range of concepts related to auditory salience and to the dynamics of musical grouping and segmentation. Nonetheless, the avoiding of this clarification could lead to some misunderstandings.

Evidences for the principle of good continuation applied to musical stimuli have been provided by a conspicuous number of researches already from the seventies of the last century. Studies as such of Divenyi & Hirsh (1975) or Van Noorden (1975) demonstrate that sequences of tones that follow a unidirectional pitch change are more effectively processed than those where pitch changes reverse direction. Good continuation in music is hence related to the direction of melodic contour. From an opposite perspective, it means that a change in direction of the melodic contour should be perceived as a boundary (discontinuity) between two groups of tones. Interestingly, many works put in evidence exactly the arising of a perceived accent in the area that surrounds the contour peaks (Fraisse, 1982; Povel & Essens, 1985; Jones, 1987; Drake et al., 1991; Drake & Palmer, 1993; Jones & Pfordresher, 1997; Tekman, 2002; Pfordresher, 2003); This link allows to notice the strong interconnection between local and global salience, whose difference is maybe only due to the choice of different viewpoints in inquiring this problem. However, as observed by Deutsch (1981, pp. 516-517), "if the pitch contour is repeatedly presented, the listener will tend to form groupings on the basis of this identity of contour". Dowling (1977) finds that the recognizability of a melody is increased when melodic contour remains intact at least in fragments of the original melody, while the other tones are shifted on higher or lower octaves. In a later work Dowling (1978) claims that melodic contour is treated by the listener as a perceptual attribute of music, independent either of the number of steps along a direction or the size of the intervals between tones. With regard to the auditory maps of salience, we can find the same consideration in the model proposed by Kalinly (2009).

If the repetition of a contour is related to the grouping and segmentation of musical sequences, then the principle of good continuation should work at different hierarchic levels; furthermore, at a higher level it may be strongly tied with the detection of similarity between chunks that becomes, one more time, the basilar condition of melodic segmentation.

At the beginning of this paragraph we suggested a question related to the applicability of good continuation to ongoing sequences of tones, with the listener ignoring the direction that the melody is going to take. This issue concerns the formation of melodic expectations. According to Koffka (1935, p.437): "A melody is a whole, organized in time ... The earlier notes of the melody have an effect upon the later ones, because they have started a process which demands a definite continuation. A melody ...[is] ... not analogous to beads on a string, but ...[is]... a continuous process .... These events have their own shape which demands a proper continuation”.

In literature, the formation of expectations has been showed to depend both on innate principles (Narmour, 1990) and on the cultural background of the listeners (Unyk and Carlson, 1987; Krumhansl, 1997; Tillmann et al., 2006; Tillmann and Lebrun-Guillaud, 2006). With respect to the expectations generated by melodic patterns, Schellenberg (1997; et al., 2002) describes two basilar principles obtained through a development of Narmour’s theorization (1990):

1. Registral direction: “small implicative intervals lead to expectancies for similarity in pitch direction, specifically that the next (realized) interval will continue the direction of the melody (upward followed by upward, downward followed by downward, or lateral followed by lateral).” (Schellenberg et al., 2002, p. 513).
2. Intervallic difference: “small implicative intervals generate expectancies for realized intervals that are similar in size, whereas large implicative intervals create expectancies for smaller realized intervals. Similarity in size depends on whether pitch direction changes or remains constant.” (Ibidem).

Again, the two principles clearly recall the laws of, respectively, good continuation and proximity; furthermore, the latter is explicitly linked with the principle of similarity (between relationships).

Coming back to the issue concerning the roles of innate features and cultural elements in the formation of expectancies, Bharucha (1994) distinguishes two different kinds of expectations: schematic and veridical. If the listener hears a piece a lot of time, then he learns to avoid expectations that will be violate, with the consequence of a decreasing emotional interest; If it does not happen, it is due to the schematic expectations that are automatic, culturally generic and can be modified only over years of experience in assimilating a musical genre. Mac Mullen and Saffran (2004) relate the development of expectancies to the statistical learning of the cultural features of music. It is possible because of the link between statistical expectations and emotional responses that can elicit basic neurobiological responses (Huron 2004).

3.3 Performance

When two different players perform a piece, it is quite difficult that they join the same, identical result. In our experience it is easy to focus on our ability to recognize the style of the performer we like, as he posed a signature on the particular way to make the notes flow from his musical instrument. The importance of including performance in the field of study about music cognition is well underlined by Sloboda (2000, p. 398): "Music performance has considerable intrinsic interest as an example of a highly complex perceptuomotor skill, and it has been used as a window onto a better understanding of motor programming and control. In addition, just as the study of speech has shed important light on our understanding of language representation, so the study of performance has also increased our understanding of the organization of musical cognition. Measurement of the microstructure of performance (e.g. timing, prosody, errors) is crucial in both psycholinguistics and music psychology".

Exploring performance psychology, thus, a first question one can pose may concern whether the unique way to perform a piece by a musician derives from a conscious choice or, otherwise, it is related to some implicit feature belonging to the performer. The role of the performer in the peculiar meaning that a musical piece assumes strongly arises from the "embodied cognition" point of view (e.g., Iyer, 2002). This current states that cognition is an activity that is structured by the body situated in its environment. Hence "cognition depends upon experiences based in having a body with sensorimotor capacities" (idem, p.389). Furthermore, "in the embodied viewpoint, the mind is no longer seen as passively reflective of the outside world, but rather as an active constructor of its own reality". Such considerations open a question that deserves to be explored: when stating about an active mind, are we speaking about an involvement of top-down mechanisms in perceiving a piece? The answer, at least in Iyers' viewpoint, seems at a first sight to be negative: "according to the embodiment hypothesis, cognitive structures emerge from reinforced intermodal sensorimotor coupling. In this view, short-time rhythm cognition might include physical sensation, visual entrainment, and sonic reinforcement, unmediated by a symbolic representation" (ibidem, p.396). On the other hand, the same author explains that the evidences "from neuroscience allows for postulating shared mechanisms for low-level control of embodied action and higher-level cognition (ibidem, p.389). Hence, even if focusing on the sensorimotor level the author does not exclude a possible role of goal-directed mechanisms in the perception of a musical piece. Anyway, coming back to the problem of performance, this work strengthens the idea that performance is strictly tied to peculiar features of the performer that are quite independents from a voluntary control. In literature we can find several studies demonstrating that each musician, when playing, produces a series of systematic micro-variations that are in a certain measure independent from a voluntary control. This is true for live performances of musical rhythms (Bengtsson & Gabrielsson, 1983), and it is also true for metronomic playing, that shows to have the same timing patterns as expressive playing, but to a lesser extent (Repp, 1999). The performer's personality also emerges if playing a simple musical scale (MacKenzie & Van Eerd, 1990). Regardless of these variations being produced by perceptual distortions (Drake, 1993) or by the greater difficulty in the movement necessary to some specific note transitions (Engel et al., 1997), it seems in each case clear that it is almost impossible for a musician avoiding to give to the piece a specific prompt related to his own psychophysic features. Of particular interest is the work of Van Vugt (et al., 2013); the authors asked to eight professional pianists to play the C-major scale, twice, with the second being played fifteen minutes after the first one. Musicians were asked to avoid expressive intentions. Then, the authors compared the deviations in timing both between the two performance of each pianist and among different players. The results show that the scales played by the same musicians are more similar in timing with respect to the scales played by different pianists. Finally, a simple classifier was used to test the possibility of recognizing the pianists on the basis of the deviations observed. The accuracy was that of 100%, and the result is moreover interesting if considering that listeners could not detect some of these differences.

If playing an instrument automatically generates deviations from the "pure piece" that are specific for each performer, a second question may concern the musical variables involved in such deviations. Sloboda (2000) argues that two different and separated components are involved in skilled musical performance: a technical component, "related to the mechanics of producing fluent coordinated outputs" (p. 398) and an expressive component that is "derived from intentional variations in performance parameters chosen by the performer to influence cognitive and aesthetic outcomes for the listener. The main expressive parameters available to performers are those of timing (both in note-onset and note-offset), loudness, pitch and timbre (sound quality). The precise parameters vary from instrument to instrument… Expressivity is also related to knowledge of musical genres” (ibidem). Agreeing with this vision, Palmer 1997, p.118 specify that performance expression concerns "the small and large variations in timing, dynamics, timbre, and pitch that form the microstructure of performance and differentiate it from another performance of the same music”. An experimental basis to this idea has been provided by Gabrielsson and Juslin (1996). The authors asked to nine professional musicians using various instruments to give different emotional interpretation to short melodies. The expressive intention had a marked effect on timing, tempo, dynamic (loudness) and spectrum. If considering professional musicians, hence, one can suppose that differences in performance provided from technical skills should tend to decrease, keeping unchanged only the temporal micro-deviations above mentioned. On the other side, deviations due to expressivity have been shown to increase in parallel with the increasing of musical experience (Sloboda, 1983; Repp, 1997; Geringer & Johnson, 2007). Such variations, anyway, are generally limited and led by the information provided by the musical score. Sloboda (1983, experiment 1) asked to six pianists of varying levels of experience to give repeated performances of a note sequence presented for sight-performance under two conditions. The only difference between the two conditions concerned the location of metrical accents. Results show that this variation produced differences in expressive interpretation and a great agreement between subjects concerning the position, nature and direction of expressive variation. The more experienced players, however, made greater use of expressive variation than did the less experienced players.

Variations in timing, loudness, timbre and other parameters, related to the performer's expressivity, are supposed to have the main aim to produce specific emotions in the listener. Several works inquired this issue. A first topic exactly concerns the ability of the musicians of communicating specific emotions, and the way to investigate this point usually consists in the study of the overlapping between the emotions explicitly proposed by the performers and the ones perceived by the listeners. Many works demonstrate that there is a great concordance between the emotional performer's intentions and the emotions felt by the listener (e.g.: Behrens & Green, 1993; Gabrielsson & Juslin, 1996; Juslin, 2000; Canazza, et al., 2001). Adachi and Trehub (1998) report that even children (4-12 y.o.) seem to be able to use some variables (e.g.: tempo, legato) to express emotions in music. The strong rule of certain variables in giving to a piece specific emotional connotations is clearly demonstrated by Juslin (1997): in a first experiment, the author systematically modified a group of variables (tempo, sound level, spectrum, articulation, attack, vibrato, and timing) in a synthesized musical piece to produce specific emotions in the listener. Subjects showed great agreement in the recognizing of these emotions. In a second experiment some of the previous variables (tempo, loudness, spectrum, attack, vibrato and timing) have been random varied, with the result that the emotions described by the subjects have almost no relations with all the parameters used. In a later work, Juslin and Madison (1999) try to understand the effect of timing on the emotional recognition by removing it from the piece; as a consequence, the listener ability to identify the right emotion becomes strongly weaker. When leaving unchanged timing but not other dynamic parameters, on the other side, the subjects show to remain able to recognize the emotional intents even if less than having the possibility of using other variables too. Timing seems to be thus one of the main variables involved in the expressivity of the performer, and it is not only related to emotional contents. As shown in the next paragraph, variations in timing also can influence the perception of a piece, as well as variations in other temporal dimensions (e.g.: tempo, duration).

3.4 Duration and other temporal variables

Following Clarke and Krumhansl (1990, p.220), "Theories of time perception are themselves divided rather clearly into two models: those based on the idea of an internal clock or pacemaker and those based on the idea that perceived duration depends on the amount of information processed or stored". The same authors provide some examples of the two viewpoints, the first represented by Treisman (1963), Luce (1972) or Kristofferson (1980), the latter by Fraisse (1963), Ornstein (1969) or Michon (1972). Even if these works seem at a first sight to be quite dated, the debate on the existence and the features of an internal clock is still far from a definitive solution. To try to reduce the confusion surrounding this problem, it can be useful a brief historical excursus about the concept of tempo in music. A good starting point is surely given by a short description of the temporal variables in music and their definitions. Tempo and duration are indeed only two of a larger group of concepts related to the temporal dimension of music. From a traditional perspective, in each musical piece we can distinguish the following components: rhythm, meter, tempo and duration. Furthermore, we can also define one more dimension given by the pattern of variations in tempo along the piece (accelerations and decelerations). Such pattern is often included in the dynamic dimension of music together with other variables that do not seem to pertain to the same temporal field (e.g., variations in intensity). Nonetheless, it exist some researches that exactly focus on this specific aspect of music perception, which deserves thus to be mentioned in this review.

Rhythm is a very basic dimension of the musical phenomena. It is usually conceived as the whole feeling of movement in time, including pulse, phrasing, harmony, and meter (Apel, 1972; Lerdahl & Jackendoff, 1983). Resuming the thought of Cooper and Meyer (1960), Large and Palmer (2002, p.3) explain that: "more commonly, however, rhythm refers to the temporal patterning of event durations in an auditory sequence. Beats are perceived pulses that mark equally spaced (subjectively isochronous) points in time, either in the form of sounded events or hypothetical (unsounded) time points. Beat perception is established by the presence of musical events; however, once a sense of beat has been established, it may continue in the mind of the listener even if the event train temporarily comes into conflict with the pulse series, or after the event train ceases". This broad definition leads to introduce another important variable, the musical meter. The meter can indeed be conceived as a perceived "hierarchy of stress, that is, a pattern of strong and weak beats" (Krumhansl, 2000; p.162). In this definition, as in other ones, the word "perceived" remarks a fundamental aspect of meter: it is a result of the listener's perceptual processing. Famous works of Povel and colleagues has provided a very strong proof to this thesis. Povel and Okkerman (1981) demonstrated that when listening to isochrone patterns of equal tones, the subjects perceive local accents, that is, differences in the loudness of these tones depending on the number of elements in each pattern. The accents were heard on single, isolated elements, on the latter of a pattern of two tones, on the first and last tone in a pattern of three or more elements. Such results evidence the natural human tendency of organizing the temporal information through the selection/construction of temporal cues that allow or are at least tied to the temporal segmentation of the stimulus. In 1984 Povel has developed a temporal grid model to explain the arising of a beat from a pattern of tones differently distanced in time: the best interval is the one that allows to insert more elements into the grid and to minimize the number of "empty spaces" on the points of the grid. In a later work, Povel and Essens (1985) improved the temporal grid model by theorizing an “internal clock model” that takes into account the distribution of the accents produced by grouping which becomes “the main determinant of the hypothesized internal clock” (ibidem, p.415). Before discussing the consequences of adopting or not an “internal clock” perspective, it is useful introduce the other variables involved in the perception of the temporal dimension of music and their relations with rhythm and meter.

The tempo variable, traditionally expressed at the beginning of a musical score, defines the effective duration of each note in the piece. While in the music notation the tempo is usually expressed in terms of beats per minute (BPM), in the literature concerning the study of musical cognition it is more frequently described as the number of milliseconds that characterize the duration of the temporal unit (beat). The tempo assigned to a piece by the performer defines, obviously, the duration of the piece itself. Furthermore, the initial tempo given to a composition can undergo several variations during the execution. While a part of these variations are suggested by the composer through more or less explicit formal dynamic indications (e.g.: accelerando, solenne, grave, con gioia…), another part is provided by the performers. Already from this first sight emerges the complexity of approaching the study of the whole sample of temporal variables implied in the perception of a musical composition. Thus, it is possible to understand the reasons that have carried the searchers to avoid the use of real music in the examination of music cognition, preferring instead ad hoc created stimuli that allow to investigate from time to time the specific features whose the searcher is interested in. Temporal variables are not mutually independent, and the modification of one of these can influence the perception of the other ones. We know, indeed, that differences in tempo can influence the perception of a rhythm (Povel 1977, Clarke 1982, Marshburn e Jones 1985, Handel 1993) and, on the other hand, that different rhythms can modify the perception of tempo (Wang et al. 1983 and 1984, Drake e Botte 1993, Boltz 1998). Wang (1984) asked the subjects to listen to patterns of tones that could be isochronous or could be formed by tones with different duration. A change in tempo occurs in a certain point of the patterns, and the subjects have to identify this point. The RTs were shorter for isochronous patterns, maybe due to the different difficulty in defining a meter or at least a beat for the two kinds of patterns. If these results only let glimpse interdependence between meter and tempo, this connection is better expressed in previous works by Oshinky and Handel (1978) or Handel and Oshinky (1981). In this latter study, the subjects listen to polyrhythmic stimuli consisting of two contrasting rhythms (e.g., 3 vs. 5) that can allow the perception of only one meter or can be opened to more interpretations. For the latter class of stimuli, the perception of a meter depended on the tempo of the pattern. To provide an example, let's consider a pattern formed by two sequences of, respectively, three and five tones. The sequences have the same total duration. When the pattern is longer than 2 sec, the meter arises from the five-elements sequence (each tone during around 400ms); shortening the duration under the threshold of 1.2 sec, however, the meter arises instead from the three-elements sequence (with each tone still during approximately 400ms). The constancy in duration within the tones used to extrapolate a beat and a meter, that is, the 400ms described by this specific example, seems to suggest the existence of an absolute temporal unit in human perception. In effect, scientific literature is not extraneous to this idea, and we can find studies focused on this topic already in the first part of the 20th century. Farnsworth (et al., 1934), in a work entitled "Absolute tempo", asked blindfolded subjects of finding and tap on a telegraph key the right tempo for the piano tunes they were listening to. The results suggested the existence of a controlling absolute tempo of about 120 BPM (500ms per tones). Later searches strengthen this idea even if with little modifications to this absolute tempo (Lund, 1939; Halpern, 1988). It must be observed that these works are based on the motor behaviors of the subjects, that is, tapping, and it is not surprising that their results are quite similar to those obtained by Fraisse (1982) on spontaneous tempo. Fraisse asked the subjects to simply produce a tempo by tapping at a comfortable rate. He observed that people were consistent in their preference for tempos with a beat period about 600ms. As he noticed, this tempo was quite similar to the rate at which people walk and, moreover, if they were asked to produce slow and fast tempos, the subjects tended to use durations related by 2:1 ratios. These data seem hence to indicate a possible motor source for rhythm regularity and they anticipate maybe the "embodied" viewpoint developed in more recent times.

However, the idea of an absolute tempo that allows the listener to interpret the temporal dimension of a musical piece has been weakened by later researches as, for example, the one realized by Lapidaki (et al.) in 1991: in this work, the subjects were invited to listen to the same three compositions in three different sessions. During each listening, the participants might vary the rate of the pieces till they think they have found the right tempo. The author varies the initial tempo of the performances (very slow or very fast) in the first two sessions, while in the third one each piece is played with a starting tempo corresponding to the mean of the "right tempos" chosen by each subject in the previous two sessions. The large differences of the judgments among the sessions suggest that, supposing that an absolute tempo existed, it should be quite flexible and/or be eventually strictly tied with the perceived tempo. Such doubt seems to propose again the debate, still existing, about the validity of the "internal clock models".

In 1978, Schulze tried to understand how a listener understands the regularity of an auditory pattern by comparing three different models:

a) The listener extracts the temporal regularity by comparing the duration of interonset intervals that are close in time.

b) The listener produces his own rhythmic pattern and compares it with the rhythmic pattern of the piece.

c) After listening to the early interonset intervals in the pattern, the listener produces an internal representation that is used to judge the next intervals and the temporal stability of the pattern.

While the first model has a local nature, the second and third have a global nature. In the exposed experiment, subjects had to decide whether an auditory sequence was irregular or regular. Shulze's results give reason to the second model and provide a basis to hypothesize an internal "time keeper" that can be viewed as a precursor of the internal clock hypothesis. Nonetheless, the study of Shulze presents two limits that reduce the generalizability of his results; first, the author uses monotonic pattern, thus only focused on the rhythmic aspect of music; second, listeners do know when they have to expect a temporal variation.

Povel and Essens (1985) proposed the existence of an "internal clock" mechanism. In this work, they expose and analyze three experiments aiming to understand the strong and weak points of some of the main coeval theories on tempo perception. The investigated viewpoints are:

a) Absolute clock: Every subject is provided with an absolute internal clock, which has a very short time unit (around 1 millisecond). The temporal intervals in a sequence are thus measured on the basis of this unit. One consequence of this theory is that every sequence having the same number of intervals may be equally well perceived and reproduced regardless of the durations of the intervals. As noted by the authors, experiments such those of Fraisse (1956), Povel (1981) or Sternberg (et al. 1982) explicitly disconfirm this idea.

b) A clock with a time unit derived from the sequence: the time unit selected for the clock is identical to the shortest interval in the sequence. This model allows a solution of the former problem but still implies some consequences disconfirmed by the experience.

Following this model, indeed, two sequences of four tones spaced respectively 200ms-200ms-400ms or 200ms-400ms-400ms should be equally well perceived and reproduced. On the opposite, Povel (1981) finds that, while the first sequence can be well reproduced by the subjects, the second one is usually poorly reproduced, hence disconfirming this model.

c) Hierarchical clock: To cite the authors, the "pulsed intervals", at least in simple music, "are not composed of very small durations, but are rather of a medium duration which can either be subdivided or concatenated" (Povel and Essens 1985, p. 414). In such kind of model the unit is flexible and "continuously adapting to the sequence under consideration" (ibidem). In the authors' viewpoint this model allows a better comprehension of temporal representation.

The internal clock proposed by Povel and Essens is supposed to have two levels that define, respectively, the unit of the clock and the subdivision of the unit. The clock, moreover, has two parameters: its unit (the interval between ticks) and its location (synchronization). The listener defines the unit on the basis of the accents distribution in the sequence. These accents are those described by Povel and Okkerman (1981) and here illustrated in the former part of this chapter. The reference both to the synchronization and to the accents distribution lets suppose that the perception and representation of the temporal structure of a musical piece may be strongly related with the identification of a meter. If it is true, what does it happen when the temporal distribution of the notes does not allow any fixed meter? The authors try to answer underlying the existence of another type of temporal representation "called figural coding by Bamberger (1978)", that "capitalizes on the perceptual grouping of events. In this latter grouping strategy, detailed information about the relative durations of intervals would seem to be left encoded" (Povel and Essens 1985, p.437). Going back to the definition of rhythm given by

Cooper and Meyer (the temporal patterning of event durations in an auditory sequence), the figural coding can be viewed as the representation of a rhythm without any possibility of creating a temporal grid. If the events in the sequence cannot be located on a temporal grid, then the listener should need to memorize a quite greater amount of information to represent the piece.

This consideration leads us to open a brief parenthesis: weak meters are frequently used in contemporary music and they are often joined with the absence of a tonal structure. Meter and tonality can be considered as two grids governed by rules that allow the representation and memorization of a lot of information in an economic way (experimental data are provided, for example, by the works of Restle 1970 and 1972 on the learning of serial patterns). Bharucha (et al., 2006) divide hierarchical representations into tonal hierarchies and event hierarchies. Tonal hierarchies organize tones within a key into stable and unstable pitches. Event hierarchies extend upward from smallest subdivisions of a beat, to beat level, then measure, phrase, period, and large-order forms. When the listener cannot use these grids, he does not have a framework to which tie the elements that he is hearing. Hence it is likely that the representation of the piece will be led by simple grouping mechanisms based on the detection of boundaries between chunks. Similarity among chunks (rhythmic, melodic, harmonic…), hence, can be a valid strategy for obtaining an enough precise representation of the piece and for reducing the quantity of elaborated information.

Whereas the model proposed by Povel and Essens focuses mainly on the hierarchic nature of temporal representation, other authors give the accent to the internal experience of the listener. In this sense, a famous work is that of Parncutt (1994). The "pulse sensation" described by the author refers to the same phenomena inquired by Povel and Essens but differs from it with regard to both theoretic background and emphasised points. "*Internal clock* alludes to an underlying neurophysiological mechanism;

pulse sensation to the experience of the listener" (p. 411). Interestingly, "pulse sensations may be experienced during rhythmic perception (listening), rhythmic action (performance), or both". Parncutt refers here to the ecological psychology developed by Gibson (1972) but, at the same time, he anticipates the current embodied cognition viewpoint. However, the author presents two experiments where the subjects were asked to listen to cyclically repeating rhythmic patterns (monotonic) and to tap along with the underlying beat. Parncutt's data indicate that the subjects tend to use time-units that space within 550 and 800 ms. These numbers are quite similar (lightly slower) to the 120 bpm described by Fainsworth (et al., 1934) or by Fraisse (1982) but, just like in the former works, they are extrapolated by a motor behavior (tapping). Furthermore, the experiments of Parncutt have been re-analyzed by van Noorden and Moelants (1999) together with analogous works by Vos (1973) and Handel and Oshinsky (1981); the authors reach the conclusion of an existing preferred beat that is of 500-550 ms. This time unit should be related to a muscular specific resonance, where the resonance is defined as: "the increase in amplitude of oscillation in a physical system exposed to a periodic external force of which the driving frequency (or one of its component frequencies) is equal or very close to a natural frequency of the system" (p. 44). However, other studies explicitly contradict these data. Mc Kinney and Moelants (2006), for example, present a series of experiments to inquire the existence of a preferred tempo when perceiving musical stimuli. Subjects are asked to listen to musical excerpts varying in tempo, from very slow to very fast, and to tap the most salient pulse of these excerpts. The central idea of this research is that it does not exist a single preferred inter-beat interval traditionally reported to a tempo near to 120 bpm; results show indeed that the pulse is perceived at different tempo when changing the speed or, in other words, when changing the musical content. In authors' opinion the most likely explication is tied to the perception (or to the replication) of different metric levels by the subjects. About the idea of resonance, that is, of a pulse sensation as expressed by Parncutt (1994), Mc Kinney and Moelants argue that dynamic accents, as well as other types of accents not included in their study (e.g., melodic, durational) can have a role in the listener's perception of salient tempi. Another evidence against the "absolute time unit" can be derived by the experiments of Rankin and Large (2009); if we state that a subject will tend to categorize the temporal events on the basis of an internal pulse-unit, then the subtle variations around this unit are expected to be quite casual. On the opposite, Rankin and Large (2009) show that the temporal fluctuation around the unit moves in a systematic way during the performance (exp. 1) and, moreover, that this fluctuation can be forecasted by the listeners when they are asked to identify the beat underlying the piece. The ability to detect the pulse variations is stronger with longer time units (1/4 vs. 1/8). Studies by Penel and Drake (1998) or Repp (1999) allow then to exclude that this orderliness is due to the voluntary expressiveness of the performer, since they both find that the pattern of temporal variations during the performance does not depend on this parameter.

The last parameter considered in this chapter is strictly linked with the last researches presented. In real music, often, the tempo is not invariable. Especially in classical western music, we can assist to multiple changes in tempo during the ongoing of the compositions. Thinking, for example, on the Grieg's "In the Hall of the Mountain King" it is easy to notice that changes in tempo can also be very broad. Thus, taking into account the idea of an internal clock (regardless of the specific model), what does it happen when the former temporal unit cannot be used to interpreter the new tempo? An interesting work, in this sense, is that of Frañek and Mates (1997), which exactly investigates the theme of the perception of acceleration and deceleration of isochronous patterns. Subjects listen to a first pattern and have to reproduce it by tapping. Then a second pattern with different speed is played before the end of the first one. The subjects, in the space of 3-5 taps, have to pass to the new beat without stop their movement. The listeners do not know when the first pattern will finish. While for the deceleration trials the subjects obtain very high rating in accuracy, the deceleration seems to provide some more difficulties; furthermore, the listeners need longer time to reach the new tempo when the second pattern is faster than the first. Pouliot and Grondin (2005) find similar results but only for “musical” stimuli, that is, patterns involving more than a single repeated tone.

3.5 Expertise

After the theorization of the main mechanisms involved in the processing of tonal music into the “Generative Theory of Tonal Music” (Lerdahl and Jackendoff, 1983), Lerdahl (1989) tried to extend these mechanisms to the perception and categorization of atonal music. The need for differentiating the music on the basis of the presence/absence of a tonal structure cannot be related to a single, exhaustive reason but, on the other side, it is clear that the cultural background of the listeners may have here an important role. The rules that lead the composition of a musical piece, indeed, follow a set of parameters that are at least partially determined by the geographical position and the historical age of the composer. Every listener grows up in a cultural determined musical environment and, regardless of his consciousness about the redundancy of certain musical elements, he will use these elements to become able to forecast the development of a piece from the early tones he hears (e.g., Unyk and Carlson, 1987; Krumhansl, 1997; Tillmann et al., 2006; Tillmann and Lebrun-Guillaud, 2006). The exposition to a musical genre, thus, provides the listener with competencies that allow him to abstract the structure of a piece and memorize it. Obviously, this expertise can be improved by specific kinds of training that can be passive (listening) or active (playing an instrument) and that seem to produce many differences in cognitive processes not exclusively related to the fruition of music. The traditional way to inquire the effects of musical expertise moves from the comparison between musicians and musically naïve subjects in several kinds of tasks. The behavioral techniques have been strengthened during the last years with the parallel employment of the techniques of neuroimaging that put in evidence quantitative and sometimes qualitative differences in cerebral patterns of activation due to musical training. However, the issue remains complex, articulated and there is not yet a unified theory that is able to explain all the data collected until now. Our review, hence, will try to resume the current knowledge about musical expertise with the awareness of their fragmentation.

A first difference between musicians and non-musicians has been identified in the ability of discriminating pitches that are different but very close. Studies such those of Spiegel and Watson (1981) or of Kishon-Rabin (et al., 2001) demonstrate that the minimum interval between different pitches detectable by musicians is about half the size of that non-musicians need (1/3 the size in the study of Spiegel and Watson, 2001). This difference increases when considering professional musicians with over ten years of practice (Micheyl et al., 2006). The study presented by Micheyl and colleagues (2006) offers further evidences of the effect of training: the non-musicians, indeed, were asked to complete 2 h of training using an adaptive two-interval forced-choice procedure (exp.1), with the result of reducing the difference with the musicians' group from 1/6 to 1/4 of the detectable interval. In the second experiment the authors exposed the musically naïve subjects to a maximum of 14 hours of specific training. After 4-8 hours the results of non-musicians were similar to those of the musicians. In another recent work, Tervaniemi (et al., 2005) asked the subjects to identify deviant sounds in couples of tones spaced from 1 to 30 Hz (one of the two tones is always the C5:

528Hz). Musicians were able to detect the pitch changes faster and more accurately than non-musicians. Furthermore, the authors recorded the auditory event related potentials during the session and observed that The N2b and P3 responses have larger amplitude in musicians than in non-musicians. Nb2 and P3 responses are not generated when the listener's attention is directed away from stimuli and can be related with the updating of expectations (Näätänen et al., 2007); in other words, the ability to detect small differences between tones seem to depend on attentive processes and not necessarily on pre-attentive levels. This data show that also some of the simplest processes involved in the cognition of music cannot be well explained from a pure bottom-up perspective and puts in evidence the strong difference between passive and active listening. Obviously, the presence of processes that require the conscious elaboration of the stimuli does not reduce the effort of passive processes. Remaining into the field of EEG studies, Pantev (et al., 1998) finds larger N1m responses in musicians exposed to piano tones with respect to non-musicians. This effect is then instrument-specific, that is, related to the instrument played by the musician (Pantev et al., 2001). Larger P2 has been found by Shahin (2003) in pianists and violinists exposed respectively to piano and violin tones. Musicians also seem to be more able than non-musicians in the elaboration of tones with complex spectrum, as indicated by larger P2 and P2m responses (Shahin et al., 2005; Kuriki et al., 2006; Kuriki et al., 2007). The spectrum of a tone is strictly related with his timbre, the quality of the sound that allow, for example, to distinguish a musical instrument from another. Several studies focus on the discrimination of timber in musicians and non-musicians and demonstrate better accuracy due to expertise (e.g., Münzer et al., 2002; Chartrand and Belin, 2006). McAdams (et al., 1995) try to identify the parameters explaining the better performance of the musicians in timbre discrimination tasks Using simple tones, the authors vary the attack time, the spectral centroid and the spectral flux; they conclude that there are not differences in the way musicians and non-musicians elaborate the sound, but only in the consistency and precision of the answers. Musicians also seem to be more able in understanding the spatial position of an auditory source (Münte et al., 2001). The subjects of this research were conductors, thus it is not possible to generalize these results to all musicians. On the other side, this data underline a clear effect of expertise.

The big amount of data that demonstrates differences in the elaboration of auditory stimuli due to expertise gave the start, during the last decades, to a broad sample of experiments that inquire these difference trough the use of neuroimaging techniques. Such works are based on different paradigms and describe a very complex situation due, *in primis*, to the complexity of the processing of music. The number of involved cerebral areas is broad and comprehends "the corpus callosum (Schlaug et al., 1995a; Schmithorst et al., 2002), the Heschl gyrus gray matter volume (e.g., Gaser & Schlaug, 2003; Schneider et al., 2002), the planum temporale (Keenan et al., 2001; Luders et al., 2004; Schlaug et al., 1995b), the inferior frontal gyrus (Gaser & Schlaug, 2003; Luders et al., 2004), the primary motor cortex (Amunts et al., 1997) and the cerebellum (Hutchinsons et al., 2003)" (in Besson et al., 2007, p. 400).

Another topic related to the effect of musical expertise concerns the issue of dependence/independence among the processing of music and speech. Several studies demonstrate that the two examined competences, that is, music and speech, are not completely independent and that musical training can produce influences in the processing of speech. Some examples can be indicated in the works of Schön (et al., 2004), Thompson (et al., 2004) or Magne (et al., 2006), which put in evidence improvements in the processing of speech prosody due to musical expertise. In another recent study Marques (et al., 2007) finds that musicians are also more able in detecting pitch variations in foreign (unknown) languages. Furthermore, children and adults provided with

musical training show improved verbal memory (Chan et al., 1998; Ho et al, 2003; Fujiooka et al., 2006) and improved reading ability (Besson et al., 2007).
The effects of musical expertise have been taken into account also in studies on musical segmentation, which are usually interested in both the numbers of segmentations operated by expert and naïve subjects and in the differences between these groups in the choice of the boundaries between segments.

Concerning expertise, differences in the number of segmentations done by musicians and not musicians are remarked not only in classic segmentation studies but also in analyzing the performance of the same composition by more-less expert musicians (Ordoñana and Laucirica 2010). Koniari (et al., 2011) find differences in segmentation accuracy with regard to groups of 10-11 years old children with different degrees of training. Olivetti (1996) finds differences in musical aptitude among genders with naïve but not with expert subjects.

4 Experimental contributions

4.1 Experiment 1

4.1.1 Method

Participants

30 subjects with normal hearing, volunteers, aged 20-65 years (Mean age: 27,46; Sd: 8,86). 12 M, 18 F. All of the subjects were right-handed. None of the subjects was a professional musician.

Authorization requested
The image will be attached as soon as possible

Fig. 15 Antonio Vivaldi: secondo movimento of the "Inverno" from "Le quattro stagioni "

Stimuli

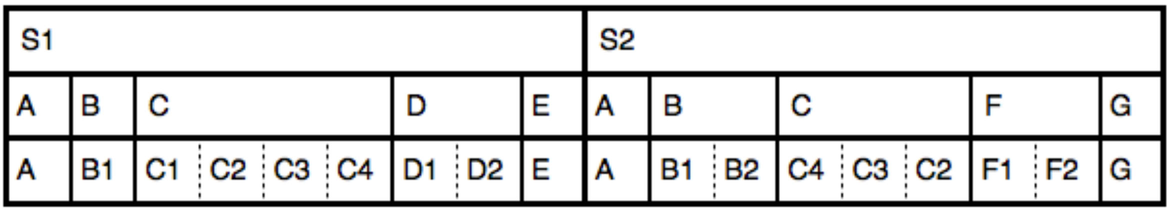
The piece used in this experiment is the "Sequenza VI per viola solo", composed by Luciano Berio in 1967. The reasons at the basis of this choice are quite different from the reasons that carried us to choose the piece used as stimulus for the tutorial session. The next two subparagraphs are thought to explain these choices.

The tutorial piece

The piece chosen for the tutorial is the second movement of the "Inverno" by Vivaldi (Largo) from the "Le quattro stagioni" (fig. 12).

This piece has been chosen because of its very clear melody and structure. On the basis of relations of proximity and similarity this piece can be divided, indeed, in accordance with three different hierarchic levels (table).

Table 1 Hierarchic analysis of the tutorial piece



S1									S2								
A	B	C				D		E	A	B		C			F		G
A	B1	C1	C2	C3	C4	D1	D2	E	A	B1	B2	C4	C3	C2	F1	F2	G

Thus, regardless of the level chosen, we can use this piece for having a measure of the comprehension of the task developed by the listener. We are here interested in understanding whether the boundaries marked by the subject follow the attempt to elaborate a representation of the structure of the piece or, on the opposite, these are only due to a casual choice.

The experimental piece

"Sequenza VI per viola solo", is an atonal composition written by Luciano Berio in 1967. This piece has been used by Dèliege and Ahnmadi in a work published in 1990 aiming to inquire the segmentation of atonal music by musicians and musically naïve subjects. Agreeing with the authors, a first feature that makes this piece a good stimulus is the presence of only an instrument that "avoids the problem of plunging the inexperienced subject into a too complex see of sound" (Dèliege & Ahnmadi 1990, pp. 23). "Berio", moreover, "has given invariance an important place in his writings: the effect of the relations of similarity induced the segmentations by the grouping cues should be clearly evident in the results" (ibidem). The authors also provide an analysis of the piece realized by two composers, Claude Ledoux and Gerhard Sporken, that identify six different sections in the Berio's composition, on the basis of intensity (thus dynamics). Using this piece, then, allows us to have both a musical analysis and experimental results that can be compared with our.

Procedure

Before the beginning of the experiment, the subjects were presented with instructions that explained them the task and included a simple metaphor aiming to make the task understandable even for musically naïve subjects. Thus, the subjects were asked to hear a tutorial piece and to follow the indications read in the instructions. This piece was presented through the laptop speakers, without earphones, in order to make the experimenter understand the real comprehension of the task obtained by the subjects. In the experimental tasks, the subjects were asked to attentively listen to each version of Berio's piece, to capture the piece structure and to press the spacebar in order to mark a boundary between two different parts. We used individual administrations of the instructions, the tutorial and the test. The order of presentation of the two performances was balanced across subjects. A graphic interface was realized using software Max/msp (Cycling74) in order to administrate the instructions, the tutorial piece and the test and to collect the data. Analyses were performed using R-packages.

Measures and analysis

Effect of performance

To understand the effect of performance we used two different, parallel types of statistical analyses: Monte Carlo (MC) simulation and cross-correlational analysis.

To evaluate whether the different performances influence the number of segmentations, we used a symmetric MC simulation with the hypothesis that, in a randomized situation, the answers given by the whole sample of subjects were uniformly distributed between the two classes (Performance A and B). The internal logic is not different from that of a binomial distribution. The MC simulation, however, allowed us to establish if the total number of answers was sufficient to obtain the hypothesized distribution. Furthermore, the formula $p=(r+1)/(n+1)$ provided by Davison & Hinkley (1997) gave us the possibility of calculating an empiric p-value.

The cross-correlational analysis, on the other side, was used to assess the effect of performance on the placement of the segmentations. The choice of this kind of correlational analysis with respect to the classic Spearman's method is due to the differences in temporal distension across the performances. To compare the distributions of segmentations, indeed, each performance was divided into one hundred consecutive temporal intervals having the same duration that is respectively 7.32 sec (Performance A) and 8.54 sec (Performance B). For each temporal cluster we measured the frequency of answers within the subject. Even if the different duration could make one think that the longer performance is played slower than the shorter one, this is not true at a local level, where the differences in acceleration/deceleration produce sometimes an inversion of the velocity ratio between the two. The cross-correlation analysis allows us to shift the temporal clusters of the two performances with respect to each other. Hence the vector of correlations obtained gives account of the degree of similarity of the two distributions of answers even if the performances miss a perfect overlapping.

The above-described analyses, however, do not allow to understand whether an eventual strong correlation across the performances can be read as a perfect overlapping across the musical elements that elicited the answer in the two versions of the piece. To inquire this point, a new MC analysis was performed, with each performance divided into twenty-four consecutive clusters during, respectively, 30.512 sec (A) and 34.519 sec (B). The number of clusters was chosen because it was the largest number guarantying a uniform distribution in the MC analysis with respect to the total number of answers we had. This technique allowed us to identify the local portions of the two performances that collected more segmentations. A consequential comparison of these clusters, with the frequencies of segmentations temporally mapped on the score sec per sec, allowed us to study the degree of similarity between the precise locations of the boundaries between segments marked by the subjects.

Temporal progression

In our hypothesis we forecasted that the exposition to the piece allows the listeners to categorize musical phrases in larger segments. We were hence interested both in a comparison between the numbers of segmentations done by the subjects during the first and the second listening, and in the study of the temporal progression of the number of segmentations during a single listening. Two MC analyses served this scope; to compare first and second listening we used the same method chosen for investigating the effect of performance. To assess the progression during the single listening we divided each performance into four segments with equal duration. Thus we performed a MC simulation to compare the distribution of the segmentations in the four clusters with regard to an hypothetic, uniform distribution. The analysis of the correlation between first and second listening was obtained by dividing the performances into one hundred equal segments.

Other variables

The MC method used for the above-described analyses was also useful to investigate two further variables: order of presentation of the two performances (AB vs. BA) and gender. Even in this case the formula by Davison & Hinkley (1997) helped us to define an empiric p-value.

4.1.2 Results

Two-classes MC simulations

A first MC simulation allowed us to inquire a casual distribution of 1721 answers in two mutually exclusive classes, and to establish the boundaries of the same. A comparison between our data and this simulation allowed us to investigate the effects of performance (A-B), order of presentation (AB-BA), and first-second listening (I-II) on the number of answers given by the subjects. An empiric p-value is calculated using the formula $p=(r+1)/(n+1)$(Davison & Hinkley, 1997) taking into account the two tailed distributions. Data show a strong decrease in the number of segmentations from the first to the second listening, with both N_I=999 and N_{II}=722 out of the range obtained in the MC simulation (p<.001, fig. 13).

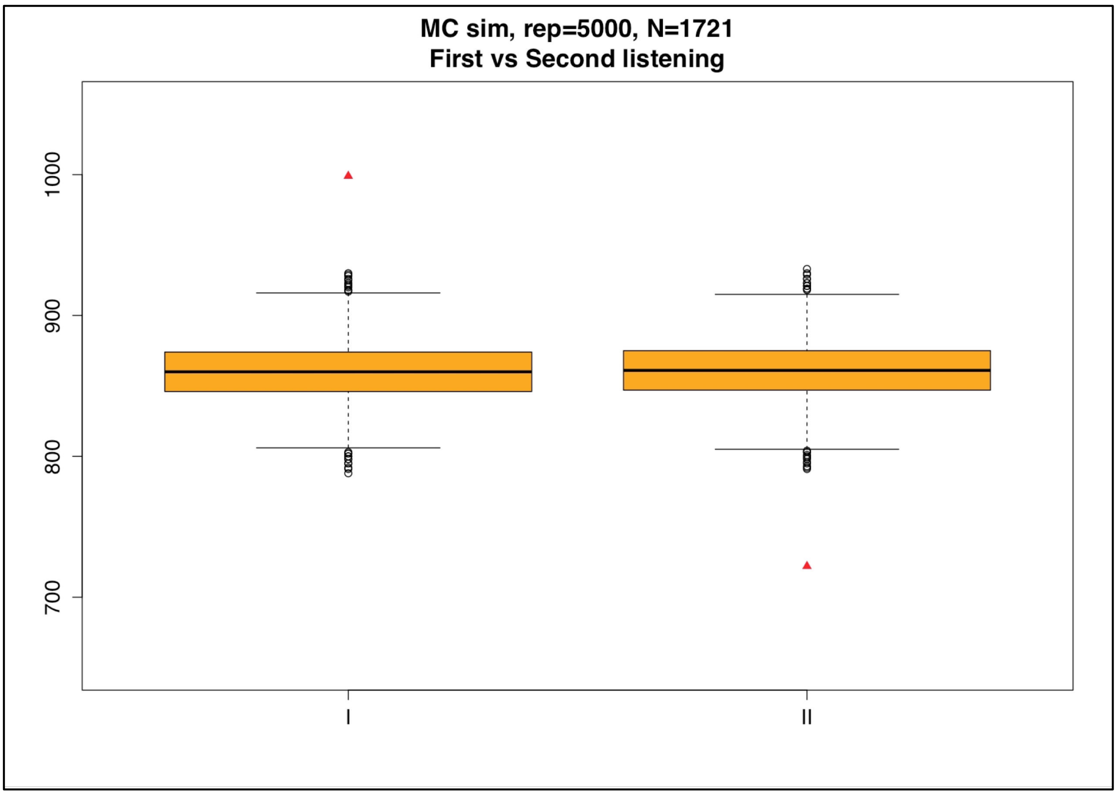


Fig. 16 First and second listening

MC simulation: comparison between the numbers of segmentations during the first and the second listening

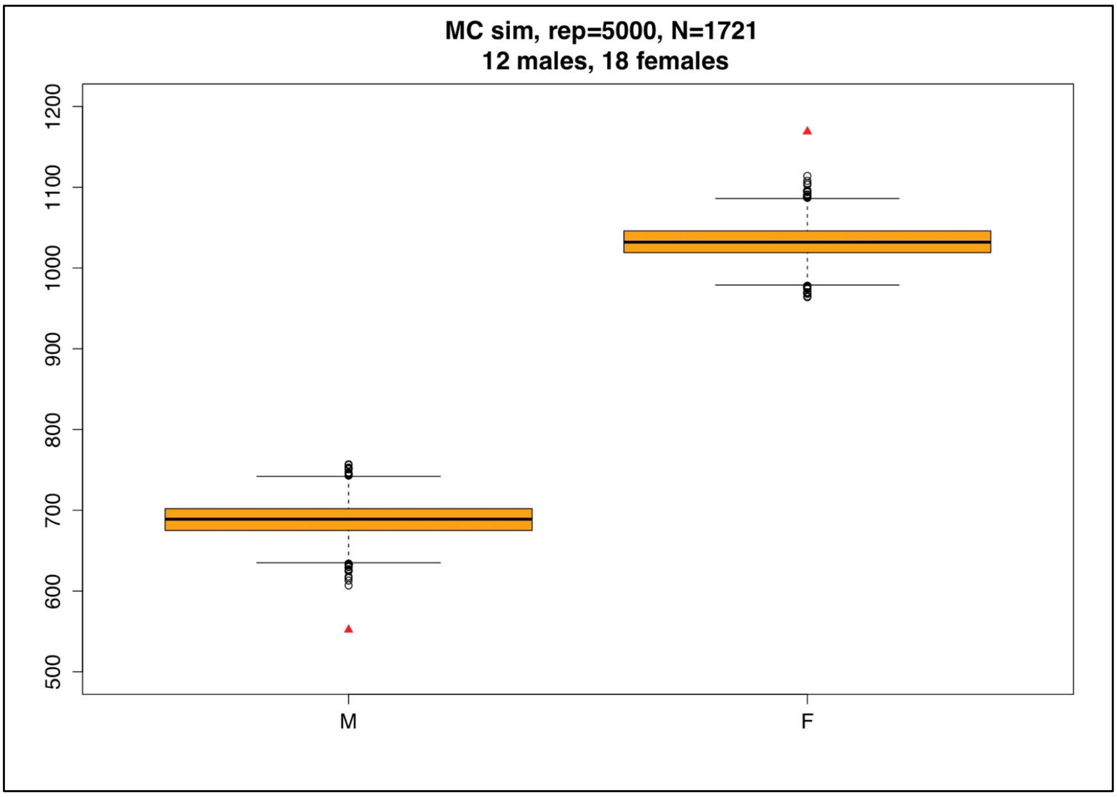


Fig. 17 Effect of gender

Gender seems to be another factor providing differences in the number of segmentations, with women indicating more segmentations than men (N_M=568, N_F=1153, p<.001; MC analysis of randomized distribution of 1721 answers in two classes with width equal to the numbers of male/female subjects; fig. 14). No effect of the order of presentation was found (N_{AB} = 877, N_{BA} = 844, p=.446).

Performances comparison

According with the limits of a MC analysis, each version of the Berio's *Sequenza VI* was divided in 24 classes of equal width (performance A = 30.512 sec, B = 34.519 sec). Then we used a new MC simulation to study the behavior of a randomized distribution of 852 (performance A, fig. 15) and 869 (performance B, fig. 16) answers in 24 classes. In a macro-analysis, we identified the classes collecting the highest numbers of answers (right tail of the distribution, p<.1).

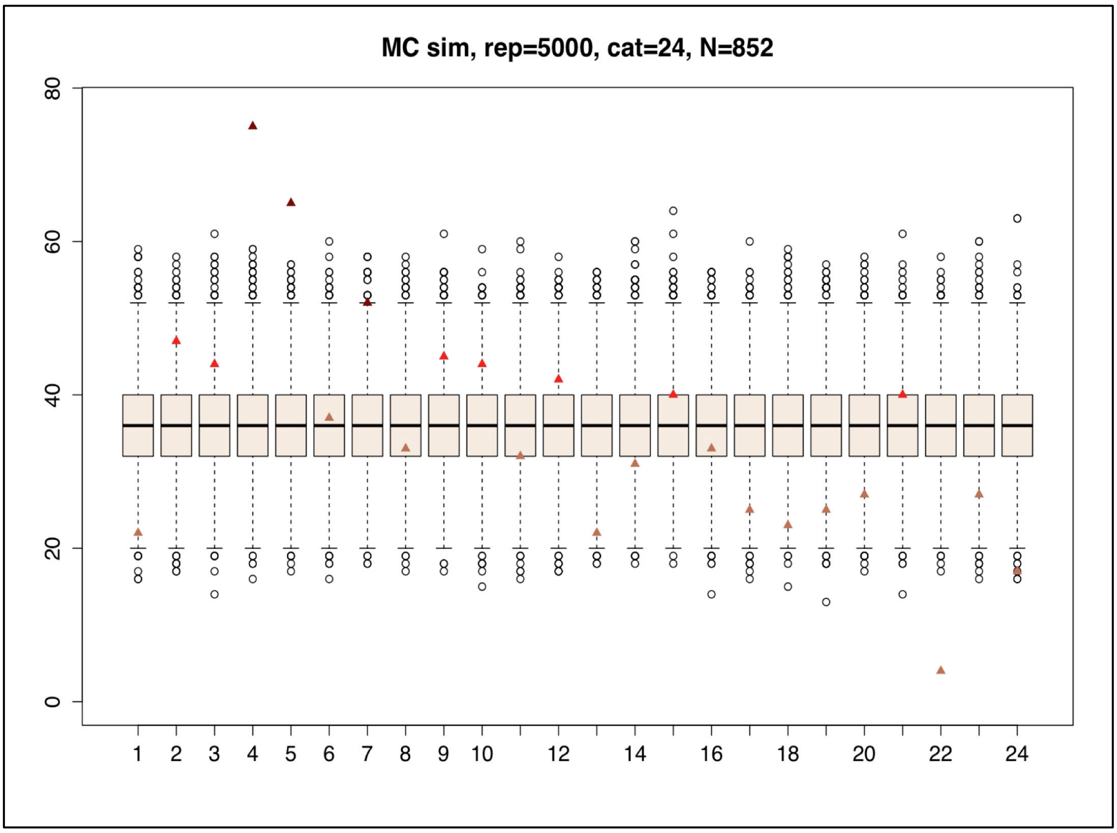


Fig. 18 Segmentation during the listening: Performance A

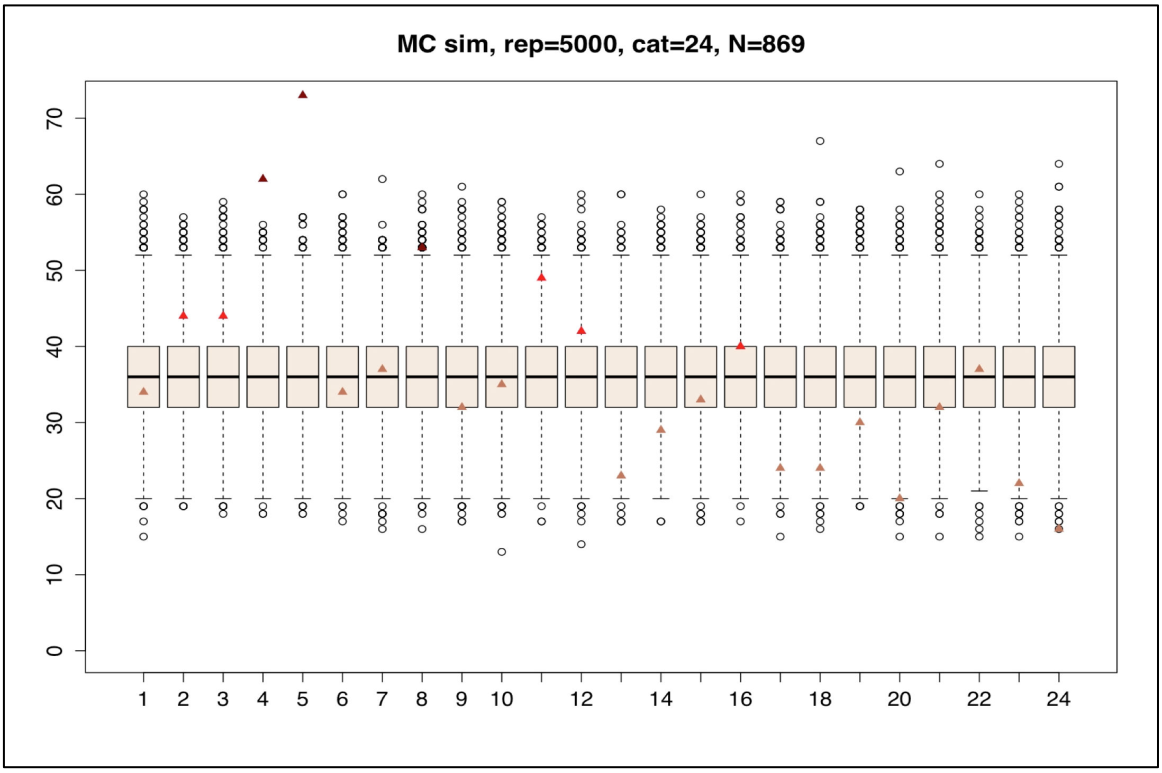


Fig. 19 Segmentation during the listening: Performance B

Analysis shows 3 main segmentation areas (MSA) for the performance A (classes 4, 5 and 7) and 3 for the performance B (classes 4, 5 and 8).

In order to inquire the possible overlapping of the segmentations in the two performances, we temporally mapped the score on the basis of the performances; then, we could mark the relevant segmentations obtained with the two versions of the piece in the overlapping areas of the MSA. Analysis shows 15 peaks in version A and 14 in version B, with 12 common pivots (82,75%) (Appendix A).

Correlational analyses

The cross-correlational analysis of the performances highlights a significant correlation in the zero/shift point (acf=0.251, two tails, p<.05), and other significant correlations with the LAGs surrounding the zero point (fig. 17,18). Nevertheless, it's important to note that the higher r-value is not relative to LAG 0, but to LAG -3, that is, with a mean delay of 97,55 sec between the two performances. These results are probably due to an imperfect overlapping between the corresponding clusters of the two performances, deriving from a different approach of the performers to the piece.

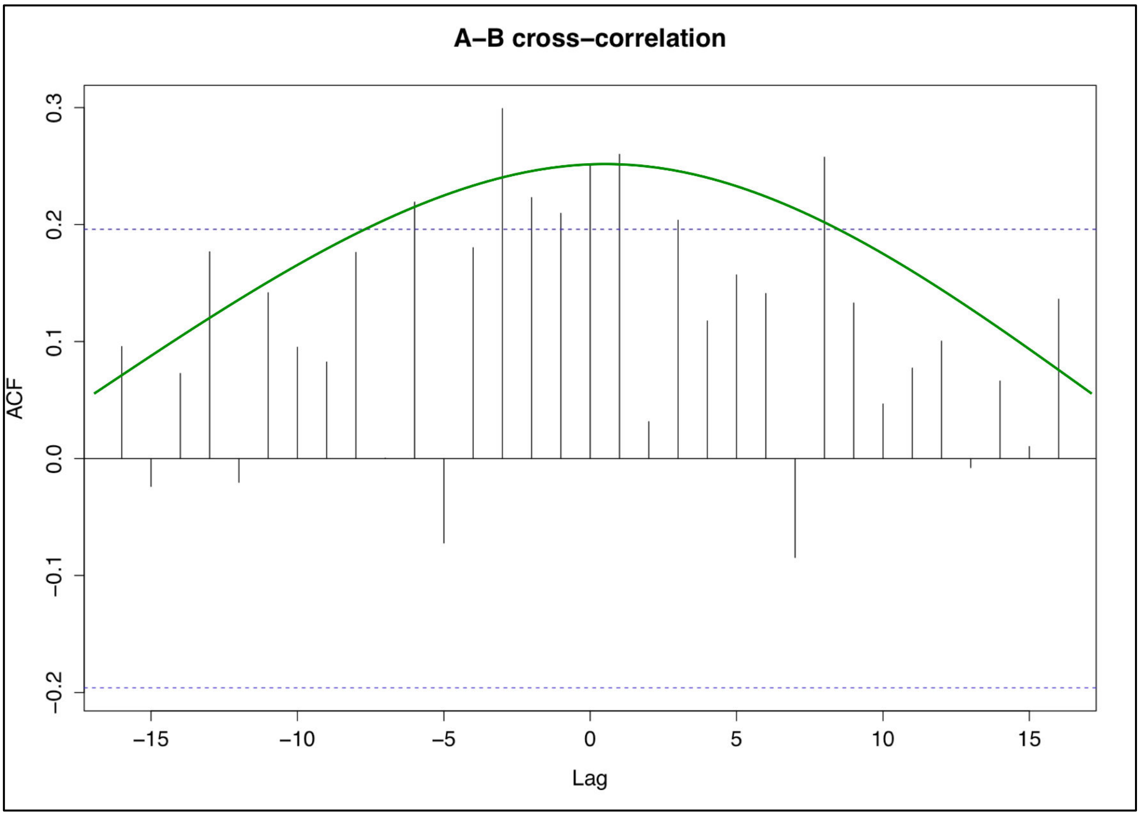


Fig. 20 Cross correlation analysis: comparison between the performances.

Concerning the cross-correlational analysis between first and second listening, we found a very strong correlation in LAG zero (acf=0.805, two tails, p<.001; fig. 19 shows correlation in LAG 0). It's important to specify that for these cross-correlational analyses we did not use the former 24 classes, but a more informative subdivision in 100 clusters of equal duration.

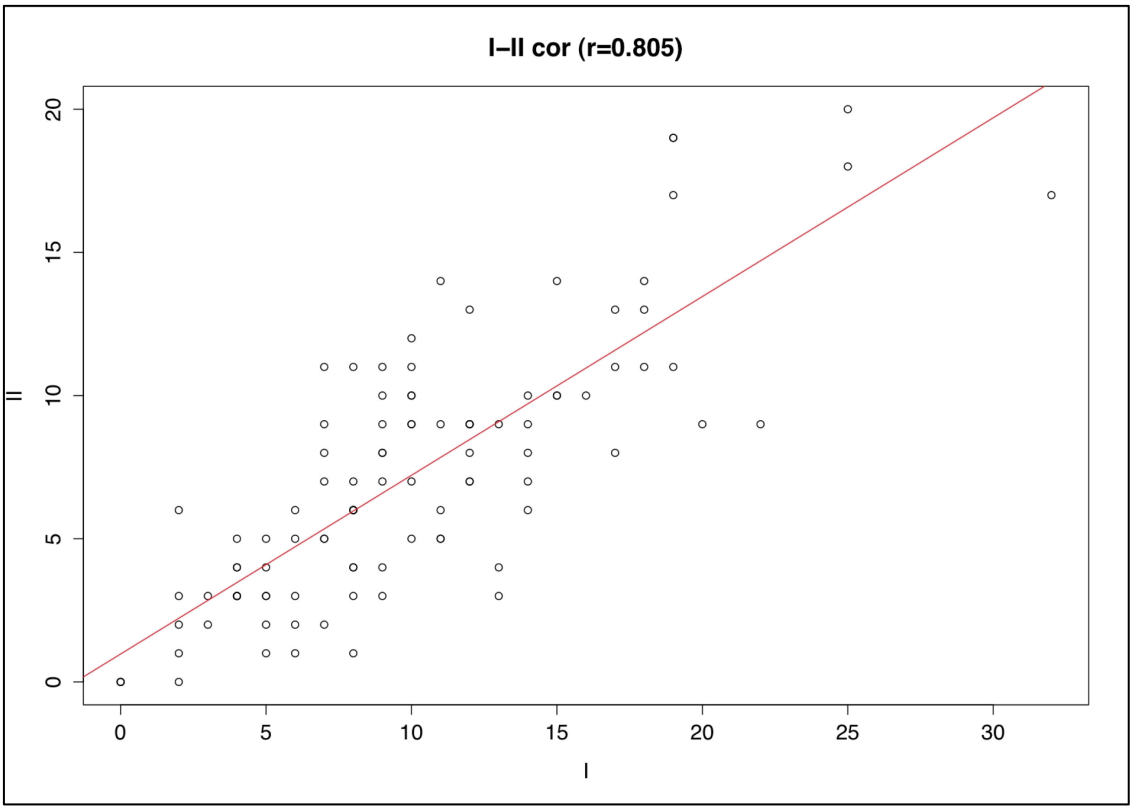


Fig. 21 Correlation analysis: first and second listening

Temporal distribution during the listening

A further MC analysis investigates the distribution of the segmentations during the listening of the piece. Each performance has been divided into four sections having the same duration. Then, each cluster represented in the figure contains the segmentations operated by the subjects in both the performances. Results clearly show that the number of answers of the listeners decreases during the listening.

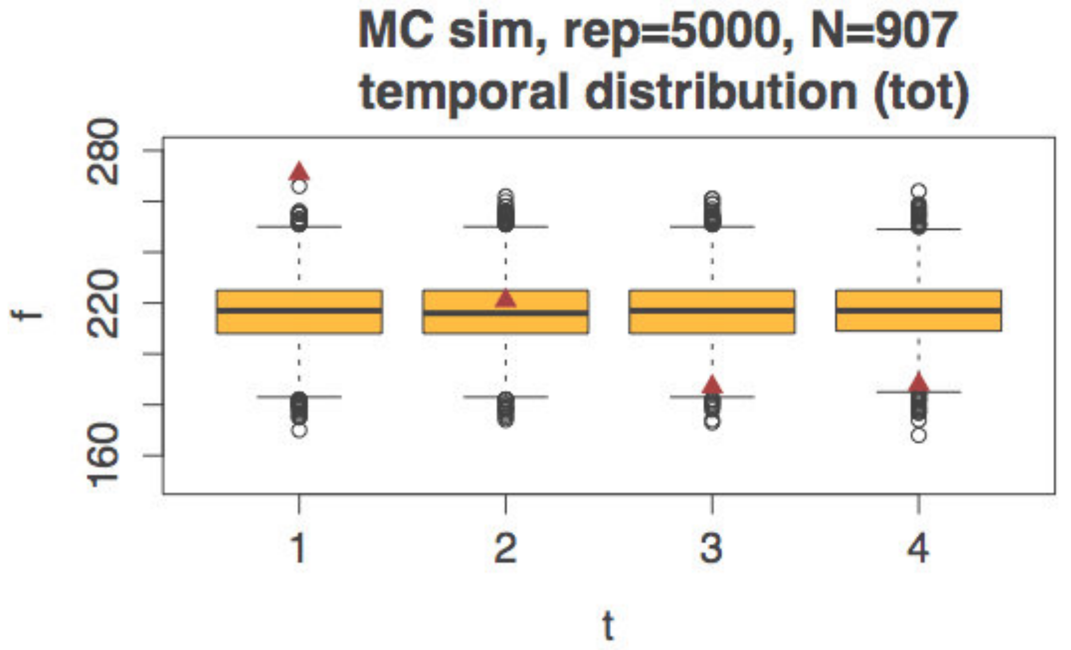


Fig. 22 Temporal distribution of segmentations during the listening

4.1.3 Discussion

Even if the cross-correlation analysis between the two performances could inspire some doubts on the prevalence of the texture in forming a representation of the plan of the pieces, these doubts are erased when looking at the placement of the answers. We can see, in effect, a good overlapping in segmentation placements either in the macro- and the micro-analysis. These data clearly suggest that salience of some groups of tones resides on a property of the same groups more than on the changes in dynamics provided by the performers. The value of cross correlation in LAG 0 (acf=0.251), moreover, is very far from the maximum correlation found with regard to the patterns of intensity (acfMAX=0.072). The prevalence of local accents then, can be excluded if looking at the MSAs: while the most of boundaries coincide with the surrounding of a rest, not every rest elicits an answer (App. A).

The very strong correlation existing between the first and the second listening says us that the knowledge of the piece has a very weak role in structuring the mental plan of the composition. The only difference found is related to the numbers of answers in the two hearings, that is, to the false alarms that decrease from the first to the second listening. From a psychological point of view, these results imply an important role of expectations. From a different perspective we could say that the salience of certain parts appears related to former portions of the score much more than to the ones yet to hear.

Last, we have to consider the difference found in the number of segmentations between males and females. In this test we did not take into account measures of expertise and aptitude, so we cannot exclude a bias produced by this lack. In this respect, Olivetti (1996) finds differences in musical aptitude among genders for naïve but not for expert subjects. Our data could be hence explained by the lack of expert subjects in the experiment. To solve this doubt, in the second experiment we chose to take into account both the musical aptitude and the expertise of the listeners.

4.2 Experiment 2

4.2.1 Method

Participants

30 subjects with normal hearing, volunteers, aged 23-57 years (Mean age: 34,16; Sd: 10,64), 15 males and 15 females; ten of them were professionals or semi-professionals musicians (2 singers, 4 pianists, 4 guitarists), 20 non-musicians. The mean of years of formal musical training for musicians is 5.4, with sd=4.351. All of the subjects were right-handed.

Stimuli

For the tutorial we used the same piece of the first experiment, the “Inverno” by Vivaldi (Largo) from “Le “quattro stagioni”, since it has proven to work well for its intended purpose. The experimental piece is instead the “Sequenza III per voce solo” by Luciano Berio (1965). This is an atonal composition for voice only that Berio wrote for his wife, the soprano Cathy Berberian. As described by Imberty (2005), the composition has a well understandable structure, with two alternated kinds of elements that can be described as “spoken” and "sung". At the beginning of the piece, spoken parts are formed by a mixture of short syllables without any sense and short English words while sung parts generally consist of long vowels. After a while, it is possible to notice an inversion of these features, with syllables used in the sung parts while the spoken parts become vowels or simple sounds that reproduce several human vocal emissions (e.g., laughters, coughs, whispers, etc.). Since one of our aims is to inquire the effect of duration, we chose two performances of this piece realized by the same singer, the above cited Cathy Berberian (1966, 1969). The performances differ in duration (8.48 vs. 6.55 min.) but have almost equal timbre and dynamics. Given that it is reasonably impossible to find two performances that have exactly the same temporal distensions, the choice of using recordings by the same singer seems to be the best way to face this problem avoiding the use of ad hoc created stimuli.

Procedure

The method is the same of the former experiment, except for the use of the “Wing Standardised Test of Musical Intelligence” in the Italian adaptation provided by Olivetti Belardinelli (1993, 1995). The Wing test was administered before giving the instructions to the subjects. We thought the aptitude subtests (1,2,3) are sufficient for our goals; the trials concerning the musical taste (4 to 7) were not administered. As provided by the Wing test, we also asked to the subjects some explicit information about their musical competences.

Measures and analysis

Effect of duration

To investigate the effect of duration we compared the two performances following the same methods as in the former experiment. MC simulations aim both to inquire whether there are differences in the number of segmentations across the performances and to understand the degree of overlapping of the main segmentation areas. Cross correlation has been used to analyze the similarity/difference of the distribution of segmentations between the two versions.

Temporal progression

Given the results obtained in the first experiment, we compared also in this study the numbers of segmentations in the first and in the second listening; we also analyzed the temporal progression of the frequencies of answers within the composition, by dividing it in four parts with equal duration. The statistical methods are the same as in the former experiment.

Musical intelligence and expertise

Differently from the first experiment, we decided this time to take into account the effects of both musical intelligence and expertise. The Wing Standardised Test of Musical Intelligence served to the first scope. To analyze the effect of expertise on the Wing scores, we performed one-way Anova for the three sub-tests and for the total score between groups (musicians - not musicians). A multivariate analysis of variance provided a measure of the influence of the three sub-tests on the differences in the total score produced by expertise (Test Pillai). MC simulations served then to clarify the effect (or the lack of effects) of, respectively, musical intelligence and expertise on the number of segmentations.

Gender and serial order

The results concerning the variable “gender” that we obtained in the first study were hypothesized to be the effect of a bias produced by the poor musical expertise of the sample. To verify this hypothesis we compared the number of segmentations marked by, respectively, male and female subjects in this new experiment. We also analyzed the differences related to gender within the two groups of musicians and musically naïve subjects. The method is the same as in the first experiment. Another symmetric MC simulation was realized to assess the role of the serial order of presentation of the two performances (AB vs. BA)

4.2.2 Results

Two-classes MC simulations

A first group of symmetric MC simulations allowed us to inquire a casual distribution of 907 answers in two mutually exclusive classes with the aim of inquiring the effects of duration (A vs. B), order of presentation (AB vs. BA), first and second listening (I vs. II) and gender (M vs. F). Another MC simulation with two classes with different width (Mus=10, NoMus=20) explored the effect of expertise. Results show an effect of the all variables except for the gender (N_A= 499, N_B= 408, p<.01; N_{AB}= 524, N_{BA}= 383, p<.001; N_I= 532, N_{II}= 375, p<.001; N_M= 435, N_F= 472, p=.23; N_{MUS}= 226, N_{NOMUS}= 681, p<.001; fig. 20, 21, 22, 23). The empiric p-value is still calculated with the formula *p=(r+1)/(n+1)*(Davison & Hinkley, 1997) and takes into account the two tailed distributions.

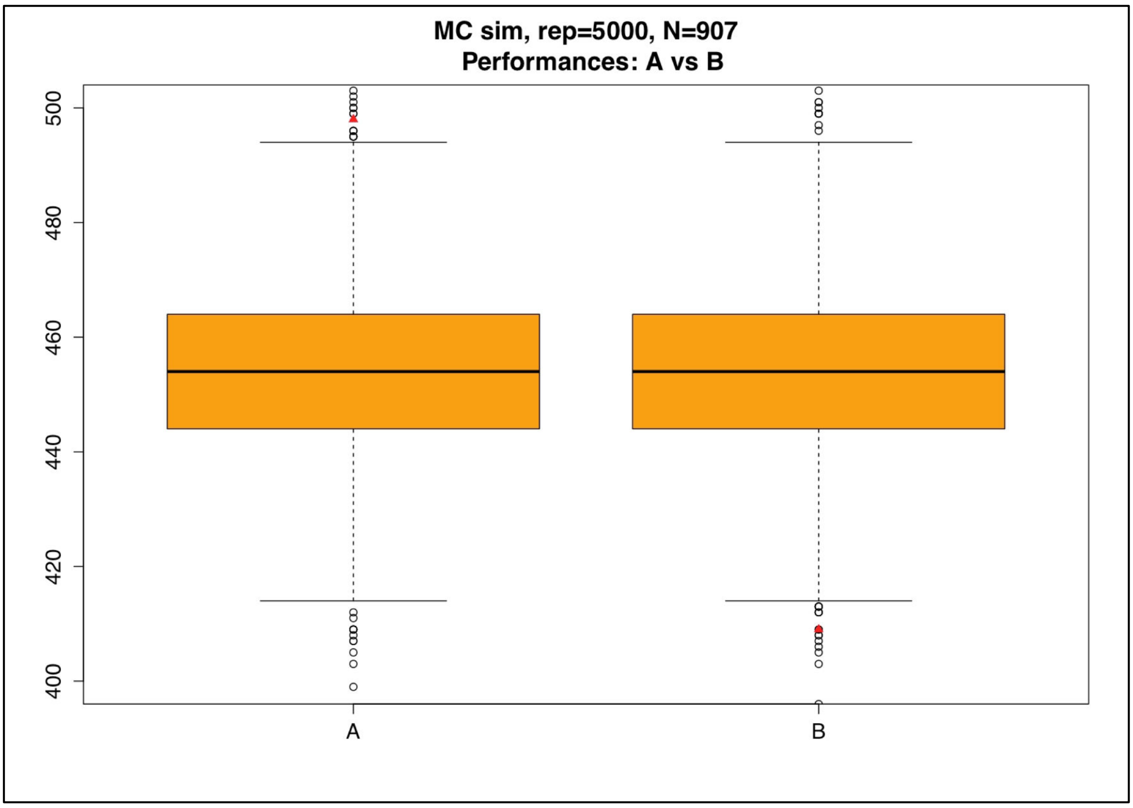


Fig. 23 Performance

MC simulation: comparison between the numbers of segmentations in the two performances

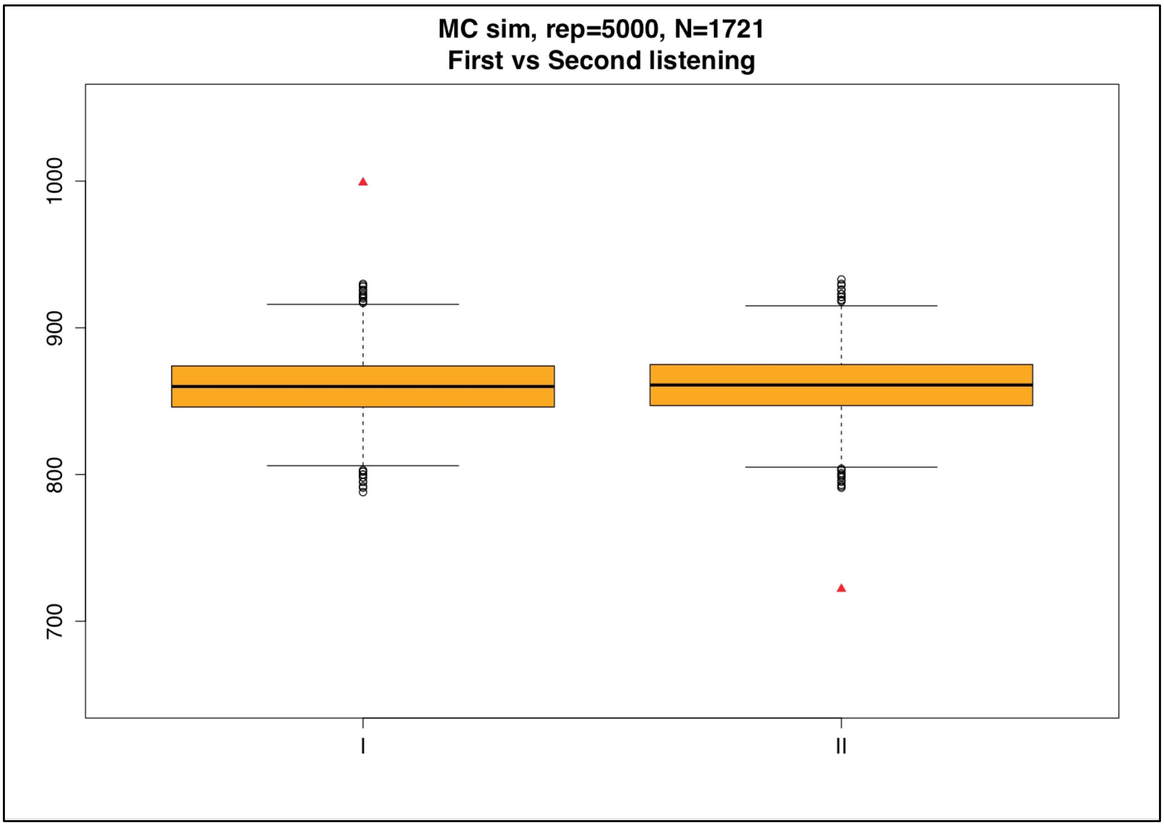


Fig. 24 First and second listening

MC simulation: comparison between the numbers of segmentations during the first and the second listening

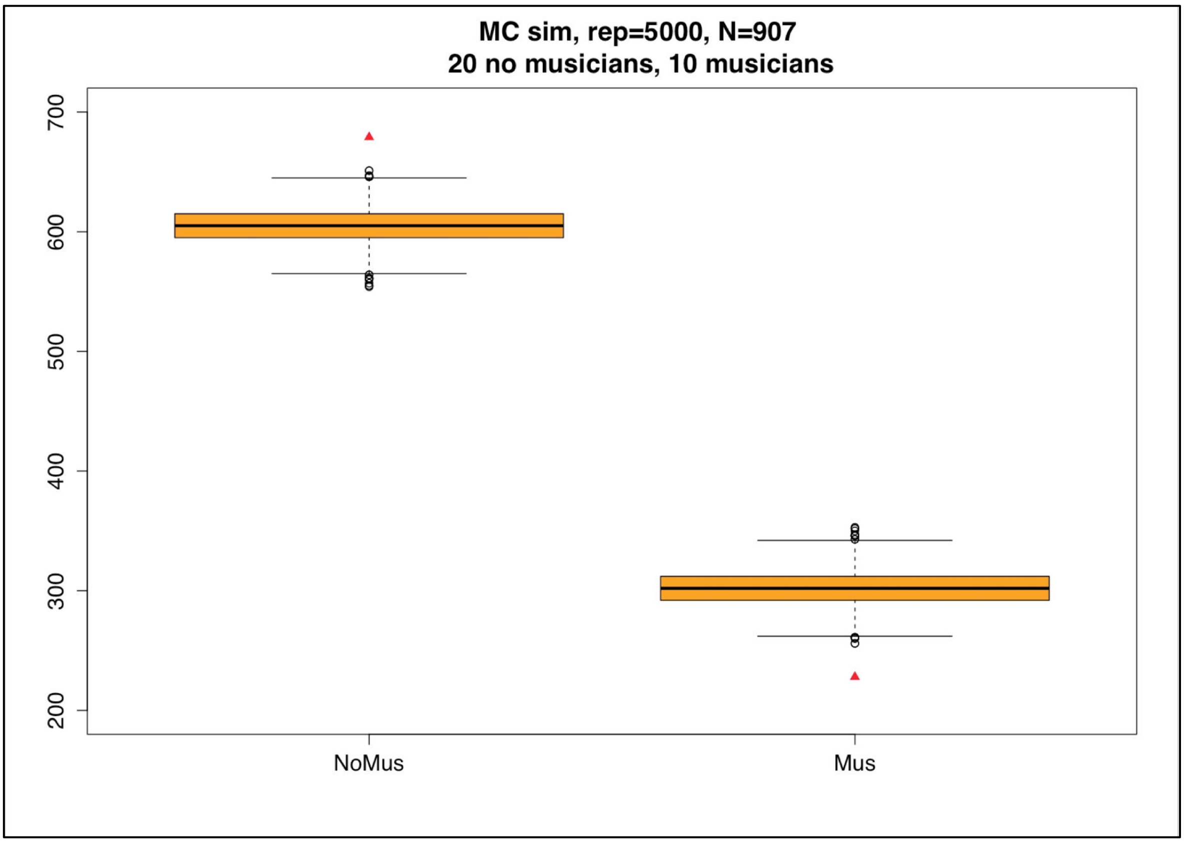


Fig. 25 Expertise

MC simulation: comparison between the numbers of segmentations of musicians and non musicians

Differently from the “Sequenza VI”, the segmentation of “Sequenza III” seems not to be influenced by the gender of the subjects. Taking into account the results of Olivetti (1996), we realized two further MC simulations to control whether some difference exists within the groups of musicians and non-musicians. The results shows an opposite trend for the two groups; while for non-musicians women tend to indicate less segmentations than men (f=11, m=9; N_F=351, N_M=330, p<.1), the opposite happens for musicians (f=4, m=6; N_F=121, N_M=105; p<.001).

Performances comparison

Using the same method of exp.1, each version of the Berio’s Sequenza III was divided in 24 classes of equal width (performance A= 22.29sec, B= 17.54sec).

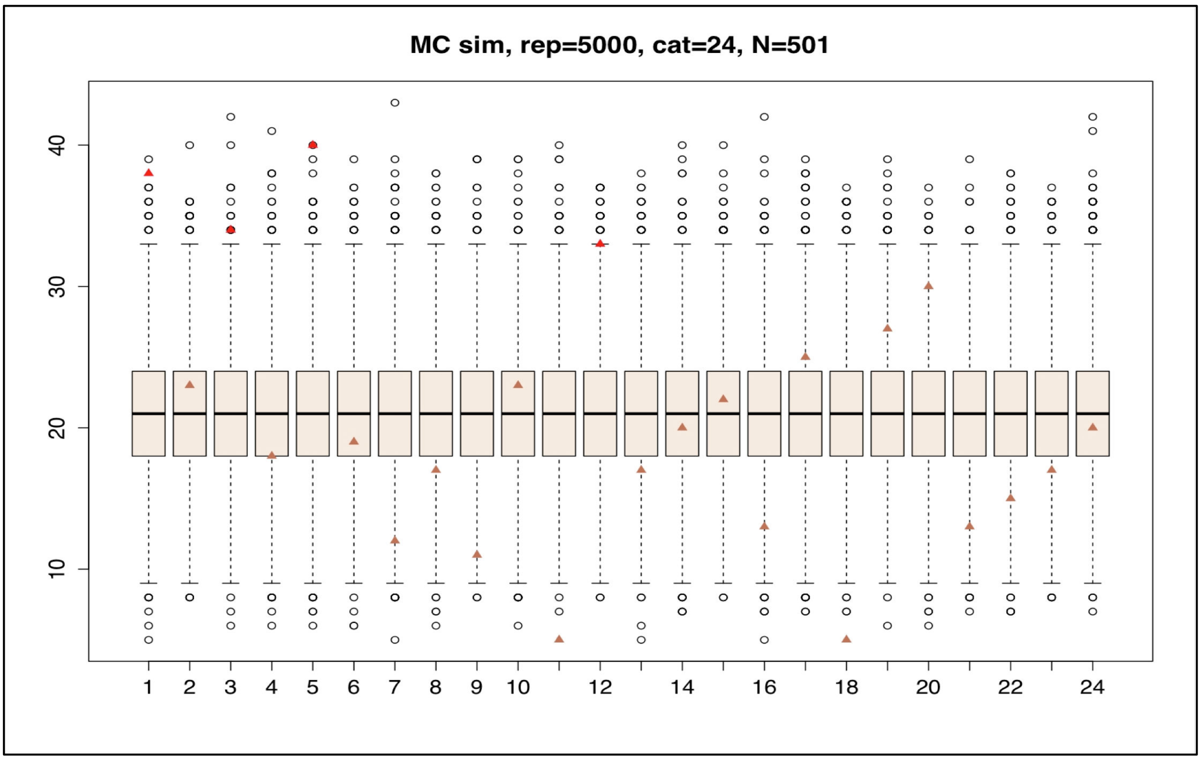


Fig. 26 Segmentation during the listening: Performance A

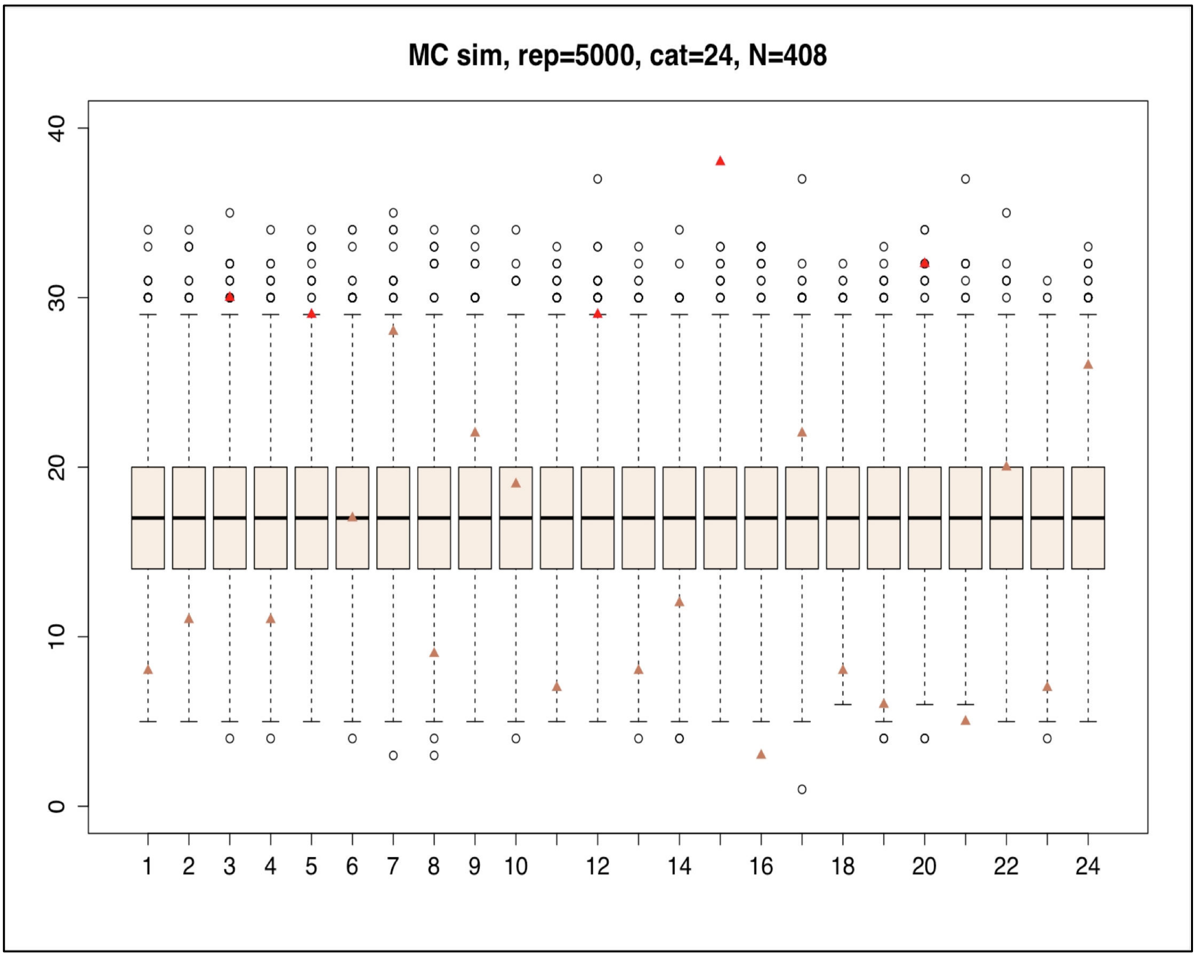


Fig. 27 Segmentation during the listening: Performance B

Then we used a new MC simulation to study the behavior of a randomized distribution of 499 (performance A, fig. 24) and 408 (performance B, fig. 25) answers in these classes. Again, the right tail of the distribution was considered (p<.1). Data show 4 MSA for the performance A (classes 1, 3, 5 and 12) and 5 for the performance B (classes 3, 5, 12, 15, 20), with three common areas. After a temporal mapping of the scores, it was possible to identify the common segments in the score in order to compare the peaks of answers in the two performances. We isolated 4 peaks in performance A and 5 in performance B, with three common pivots (67%) (Appendix B).

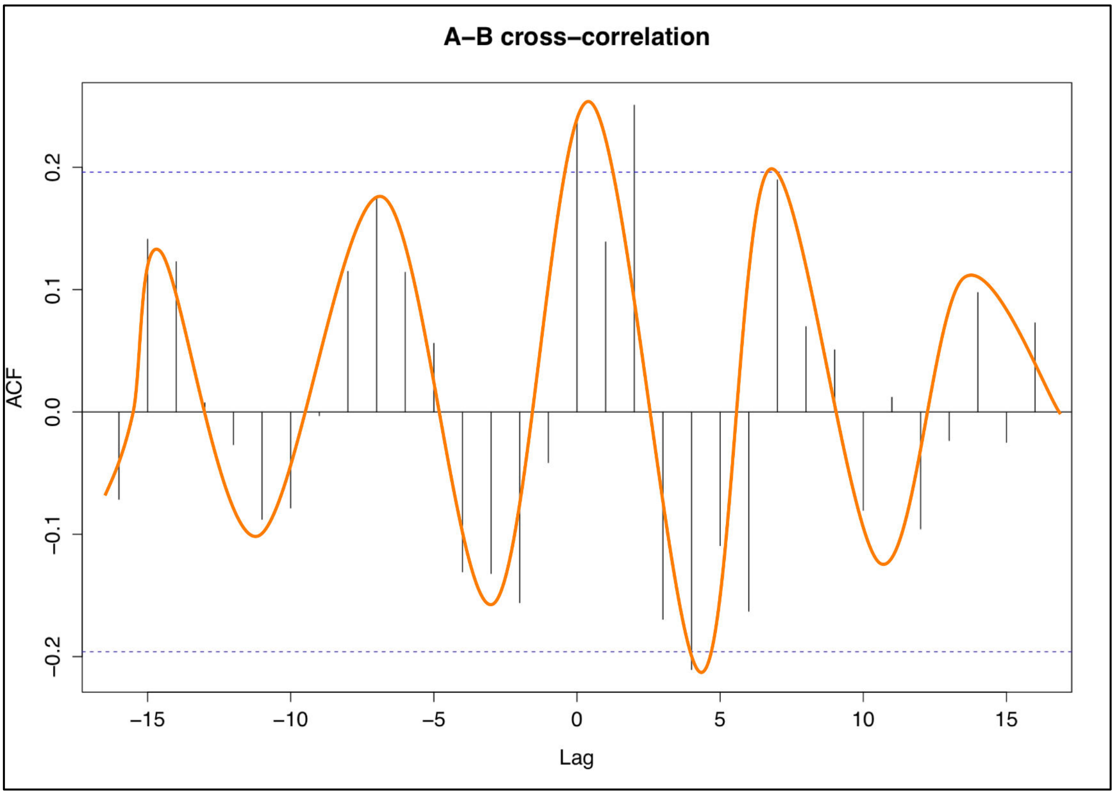


Fig. 28 Cross correlation analysis: comparison between the performances.

Correlational analyses

A cross correlation analysis displays significant results for LAG 0 and 2 (r=.242 in LAG 0, two tails, p<.05; fig. 26); This greater stability around

LAG 0, with respect to exp. 1, is thought to be due to the coincidence of the performer in the two versions that allows a similar interpretation of the piece even if changing in duration. Nevertheless, the imperfect overlapping between the corresponding clusters of the two performances can explain the correlation in LAG 2.

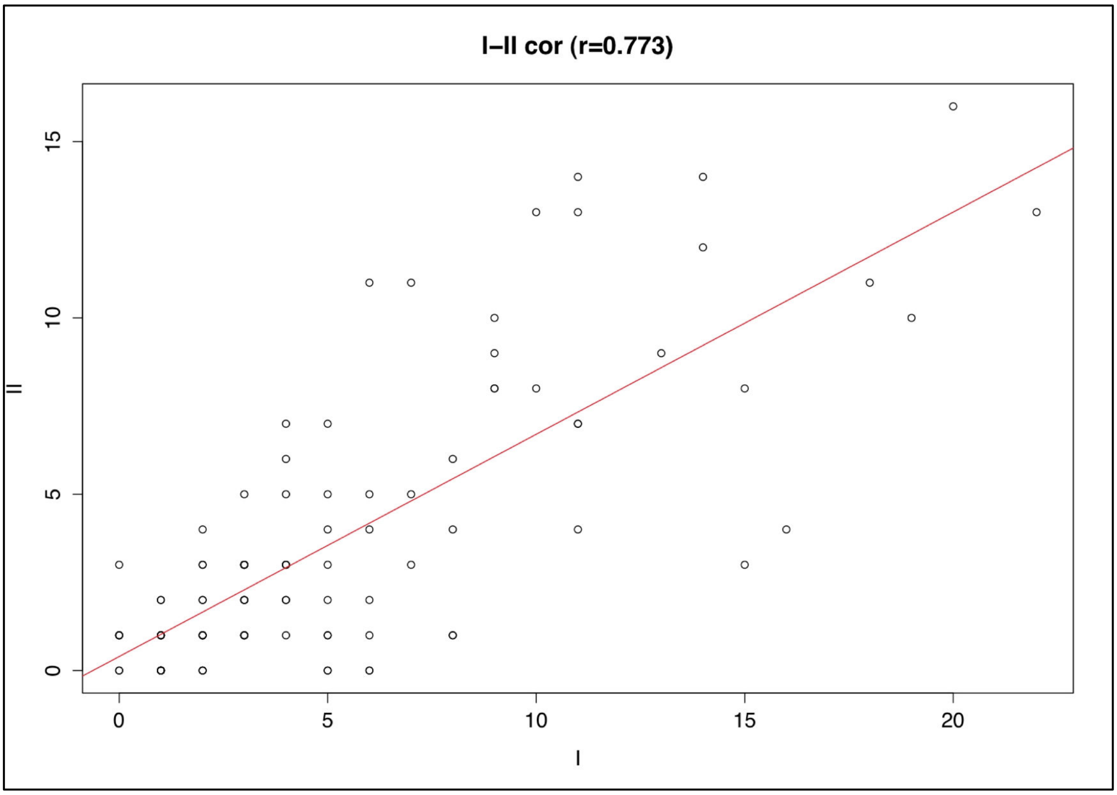


Fig. 29 Correlation analysis: first and second listening

As in Exp. 1, we found a very strong correlation between first and second listening (r=.773 in zero-shift point, two tails, p<.001, fig. 27, 28), with the only difference between the two represented by the number of answers.

The significant differences in the number of segmentations related to the expertise and the order of presentation carried us to extend the cross-correlational analysis to these variables. In both cases we found a strong correlation in LAG 0, with r=.834, p<.001 between musicians and not musicians and r=.827, p<.001 between AB and BA.

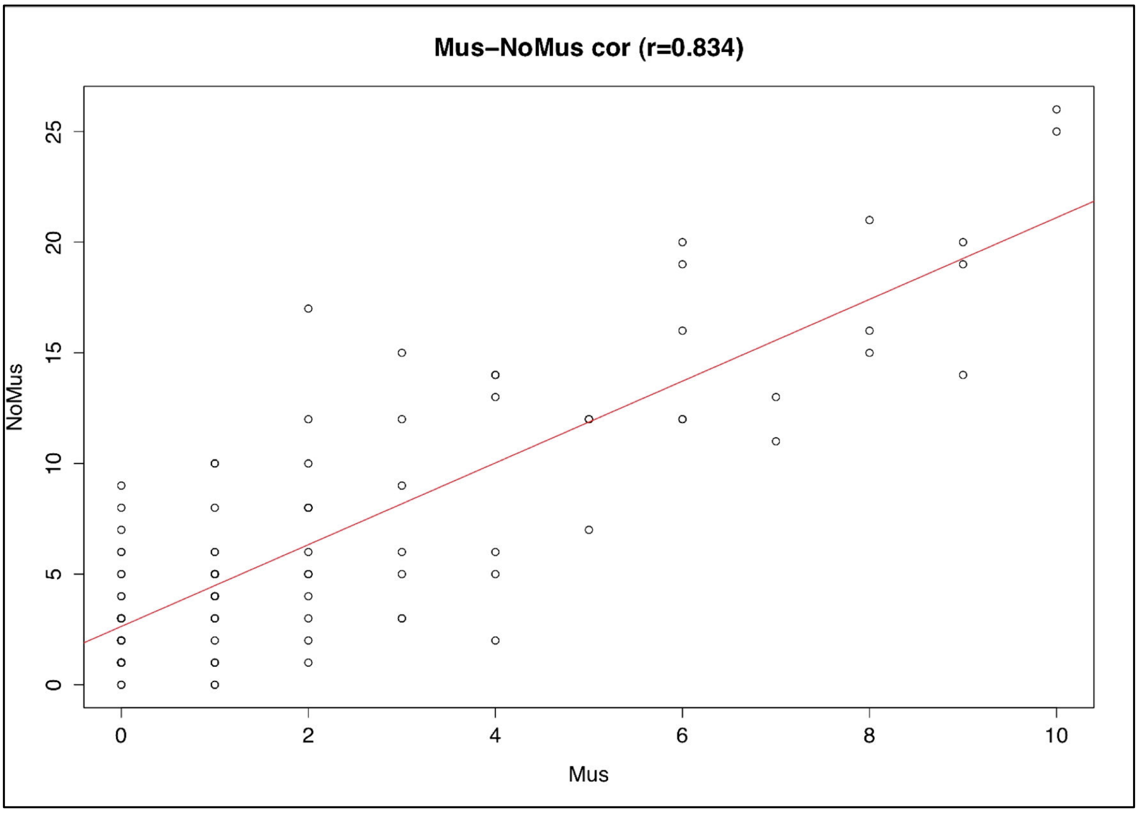


Fig. 30 Correlation analysis: Expertise.

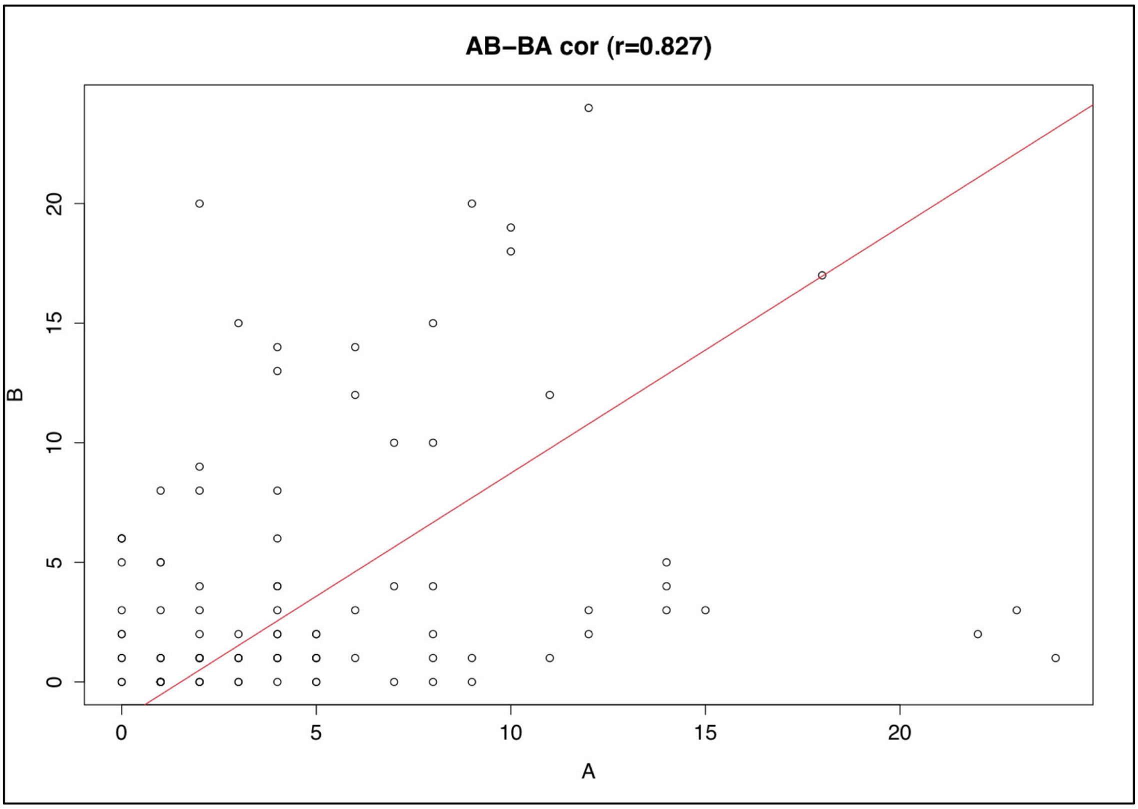
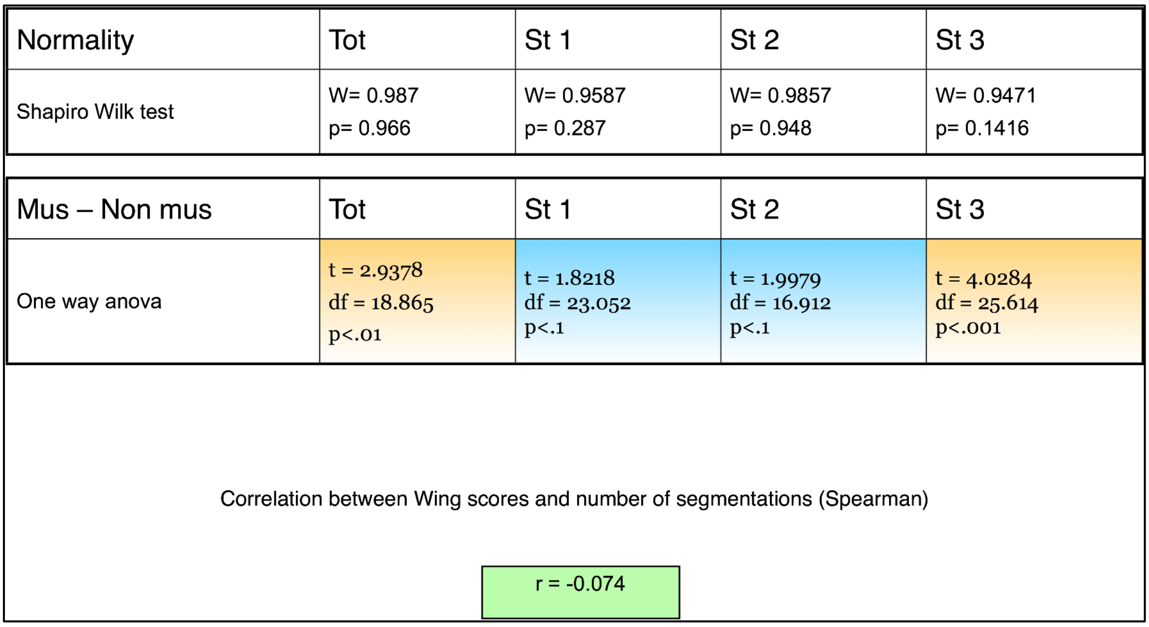


Fig. 31 Correlation analysis: Order of presentation

Musical aptitude

To analyze the effect of expertise on the Wing scores, we performed one-way Anova for the three sub-tests and for the total score between groups (musicians - not musicians). Results are significant for the total score (t = 2.9378, df = 18.865, p<.01) and for test 3 (t = 4.0284, df = 25.614, p<.001), and show a tendency toward a difference between the two groups in test 1 (t = 1.8218, df = 23.052, p<.1) and 2 (t = 1.9979, df = 16.912, p<.1) (tab.).

Table 2 Analyses of the Wing scores



Normality	Tot	St 1	St 2	St 3

Shapiro Wilk test	W= 0.987
	p= 0.966
		W= 0.9587
		p= 0.287
			W= 0.9857
			p= 0.948
				W= 0.9471
				p= 0.1416

Mus – Non mus	Tot	St 1	St 2	St 3

One way anova	t = 2.9378
	df = 18.865
	p<.01
		t = 1.8218
		df = 23.052
		p<.1
			t = 1.9979
			df = 16.912
			p<.1
				t = 4.0284
				df = 25.614
				p<.001

Correlation between Wing scores and number of segmentations (Spearman)

r = -0.074

A multivariate analysis of variance (tab.3) provided a measure of the influence of the three sub-tests on the differences in the total score produced by expertise (Test Pillai). Results are displayed in Tab.1. Test 2 and 3 seem to be influenced by musical training more than Test 1. Results for test 1, although not significant, show a p<.1.

Temporal distribution during the listening

A further MC analysis investigates the distribution of the segmentations during the listening of the piece. Each performance has been divided into four sections having the same duration. Then, each cluster represented in the figure contains the segmentations operated by the subjects in both the performances. Results clearly show that the number of answers of the listeners decreases during the listening.

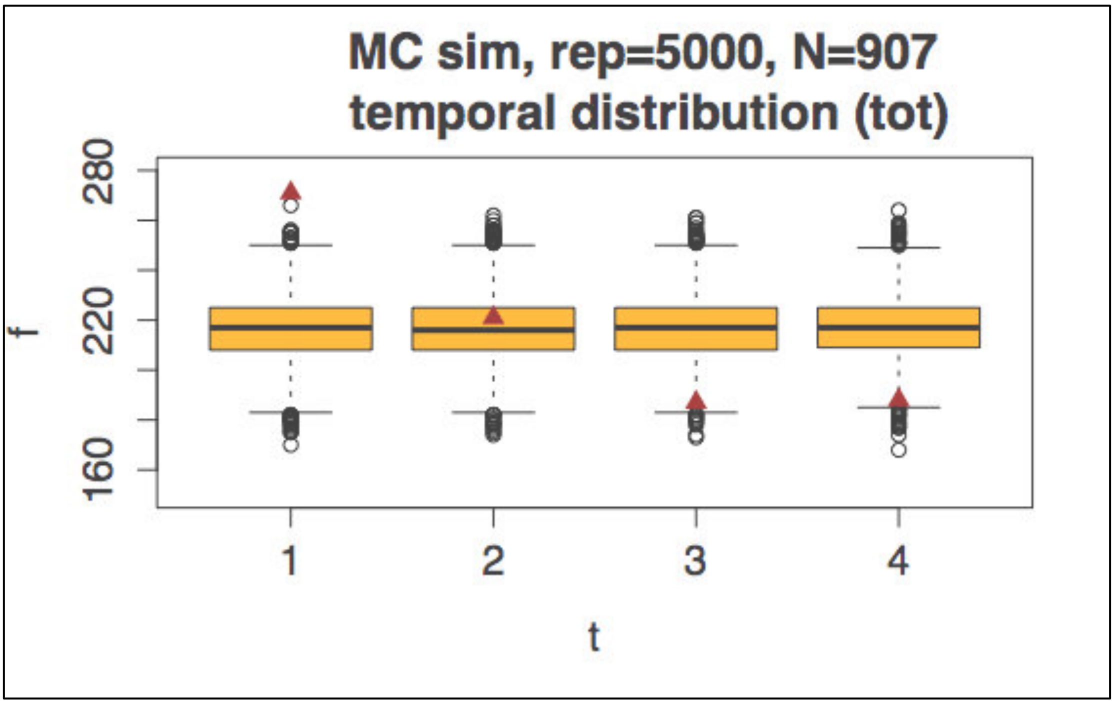


Fig. 32 Temporal distribution of segmentations during the listening

4.2.3 Discussion

As in the former experiment we found a strong correlation between the first and the second listening, with the only difference represented by the number of segmentations. Hence, subjects are able to identify a salient structure of the piece already from the first listening. False alarms are eliminated in the second listening. Interestingly, the longer-lasting performance produces more answers, even if with a decreasing from the first to the second listening that has not been found for the shorter version. Furthermore, performance B shows a great stability not only for the placement of the boundaries but also for the number of segmentations. These data suggest the possibility that duration is not the main cause of the found differences: the lacking of a decreasing trend in performance B can be explained if we state that this version is more simple to be perceived and/or represented. We can, in effect, acknowledge that what is lacking here is not as much a decreasing as the false alarms during the first listening. This reading also can explain the difference relative to the order of presentation: this is indeed only generated by the serial position of the performance A in the task.

As expected, we did not find differences relative to gender, but to expertise. Nevertheless, we found an opposite trend for musicians and non-musicians with respect to gender. Consistent with Olivetti (1996), naïve women mark less boundaries than naïve men. This difference is reversed for musicians. Even if this result goes over the Olivetti's data, we must be careful in claiming some new interpretations because of the small number of musicians we have. We can instead remark, in any cases, that expertise act on the difference between genders in segmenting a composition.

Subjects who have at least one year of formal musical training tend both to obtain higher scores in the Wing test and to divide the composition in larger groups. On the contrary, there is not correlation between Wing scores and segmentations number. While the expertise has a role both in the Wing scores and in the segmentation task, there is not any effect of the musical aptitude on the number of boundaries marked by the subjects. These results clearly suggest that formal musical training might produce some qualitative modifications in the way to analyze a composition. Subjects with a strongest musical aptitude, even if being more able to solve the musical problems posed by the Wing test, do not show different strategies in segmenting the piece with respect to less able listeners.

The segmentations analysis replicates the results of the former experiment, underlying a 67% overlapping of the peaks in the MSA. To conclude, texture seems to be the main feature involved in the formation of a plan of the composition. Then, the saliency of certain melodic or rhythmic patterns may reside in their-self structure more than in phenomenal accents produced by the performers.

4.3 General discussion and conclusions

The research presented in this paper had the main goal of inquiring the current knowledge on the auditory maps of salience with regard to the perception and elaboration of musical stimuli. The very broad and complex nature of this field made us introduce a consistent number of variables that could create some confusion about the linearity of our reasoning. To avoid it, the general discussion of the results will be anticipated by a brief summary concerning the main points highlighted by the former chapter and by a list of the main results of the two experimental sessions.

As exposed at the very beginning of this paper, this research arises from a doubt related to the elaboration, in the listener's mind, of a representation of a musical piece. During the recent years, this theme has been always more frequently related to the auditory maps of salience, computational models that try to replicate the human way of processing auditory stimuli. Since it, this work exactly starts with a brief review of the main models existing in literature. Our analysis puts in evidence that almost all of these models try to explain the processing of auditory stimuli exclusively with bottom-up processes. Even when including goal-directed attention, these models do not give to this level any possibility of influencing the map of salience. When building the representation of an auditory scene, thus, listeners should be inextricably tied to the sub-set of elements that have been already chosen through an automatic attentive process; they have neither possibilities of changing these elements nor of modifying their weight in the saliency map. Fortunately, this is not a universal point of view: Kalinly and Narayanan (2009), for example, admit the possibility that goal directed attention can enhance/inhibit the saliency scores of certain auditory features. Another possibility has been advanced by Coath (et al., 2005; 2009) with his 'banks of spectro-temporal response fields': the salience of an element is related to its context, not wholly to the element itself. Coath's bank of STRF, moreover, poses the accent on the importance, when segmenting an auditory stimulus, of the detection of similarity between chunks.

The models presented in the first part of Chapter I are not specific for the processing of musical stimuli, but provide indications that concern a more general auditory domain. Nevertheless, we can find conclusion similar to those above underlined (about goal-directed attention and the role of similarity) even in more specific theories on musical segmentation. Three considerations are at the basis of the choice of dedicating a chapter to the theories on segmentation:

1. Music is a temporal phenomenon: the saliency map elaborated during the listening of a piece is necessarily dynamic; the salience of each element cannot be explained without including it in a temporal context composed by other tones and silences.
2. If salience is tied to the element context, then in each context there are salient and not-salient elements.
3. If the detection of similarity is involved in the attribution of saliency scores within and between patterns of tones, then this mechanism cannot be independent from a continuous matching among elements dislocated in time.

The topic of the second part of Chapter II helps us to have a better comprehension of these points. Agreeing with the consideration of the point 2, the theorizations on the segmentation of music consider as "salient" the elements that arise from their context. In Dèliege's viewpoint, for example, these tones (or groups of tones) allow the listener to create a representation of the piece by using a sort of milestones. Each salient element is the label of a group of elements. Two groups starting with the same salient element can be represented as variations of the same cue. If it is true, then the study of segmentation allow us to inquire the issue of saliency maps from a perspective that is different from the computational one. All the theories analyzed in the second chapter show two common points:

1. Music is conceived as a hierarchic and temporal phenomenon.
2. Similarity is the main feature involved in segmentation

Unfortunately, it is clear that similarity cannot be the only variable implied in such kind of processes. An investigation on salience through a segmentation paradigm, hence, cannot be detached from the knowledge and the control of the other variables playing a role. The third chapter, thus, tries to identify the musical features that can influence the listener's segmentation of the piece. Furthermore, we also take into account that the listening of real music cannot exist without someone who plays the music and someone else that hears it. For this reason, the third chapter also focuses on the variables related to the performer as well as on those belonging to the specific listeners, such as expertise.

4.3.1 General discussion

Two experiments have been realized with the aim of inquiring the salience of musical stimuli through a segmentation paradigm. Subjects were asked to listen to atonal classical compositions, try to understand the plan of the pieces and mark the boundaries between different sections.

In the first experiment we used two versions of an instrumental piece, the Berio's Sequenza VI per viola solo, differing in dynamic and in duration. We can resume the main results of this experiment as follows:

1. Good number of coinciding segmentations in the two performances irrespective of the order of presentation.
    a) Temporal distribution of the segmentations marked by the subjects:
    b) Good correlation between the two versions (CCA); Cross-correlation is normal-shaped.
    c) High correlation between first and second listening.
2. Number of segmentations:
    a) The first listening produces more segmentations than the second one.
    b) Decreasing number of segmentations during the listening
    c) Women produce more segmentations than men.

The second experiment focuses on the effect of duration. To reduce the variability of dynamics among the performances we used two versions of a new atonal piece, the Berio's Sequenza III per voce solo, both recorded by the same singer. Differently from the composition used in the first experiment, this is a vocal piece. In this experiment we also take into account the expertise of the listeners and their musical intelligence as measured by the Wing test. The main results are:

1. Good number of coinciding segmentations in the two performances irrespective of the order of presentation.
2. Temporal distribution of the segmentations marked by the subjects:
    a) Good correlation between the two versions (CCA); Cross-

correlation shows is sinusoid-shaped.

    b) High correlation between first and second listening.
    c) High correlation between musicians and non-musicians.

Number of segmentations:

3. The first listening produces more segmentations than the second one.
    a) Decreasing number of segmentations during the listening.
    b) The longer version produces more segmentations than the second one.
    c) Women produce more segmentations than men.
    d) Non-musicians produce more segmentations than musicians.
4. More expert subjects have better scores in the Wing test, but there is not linear correlation between Wing scores and number of segmentations.
5. Subjects produce more segmentations if the longer version is played before the shorter one. Nonetheless, there is not difference in the temporal distribution of segmentations.

As we can see, in both experiments we can observe a consistent stability in the choice of the main boundaries between segments. This choice seems not to be influenced by other variables (e.g., performance, order of presentation, expertise). In other words, the simple sequence of tones that a musician find on the composition partiture seems to provide the listener with all the information necessary for the understanding of the piece structure. Furthermore, in both experiments we asked the subjects to represent the piece already during the first listening. In this condition the listeners do not know what is going to happen with the music; they can only use the processed information for elaborating an expectation of it. Nevertheless, the MC analyses clearly show a similar distribution of segmentations along the temporal line in the first and in the second listening. Thus, we can observe that the information contained in the simple ongoing of tones and rests is sufficient not only to make the listener understand the piece, but also to let him forecast how the piece will go on. For a better analysis of this overlapping, we extrapolated the areas of the pieces that seemed to elicit a greater number of answers by the subjects. Once again, it is possible to see that the listeners' segmentations fall in quite precise points, confirming our first impression. If we consider only this first analysis, it is maybe likely to think that the bottom-up models of saliency map are not so far from reality: if different people, listening a piece played by different performers on the basis of the same partiture, build a similar representation of the composition, then it means that there is a direct connection between the listener's representation and the common source of the performances: the partiture.

We have seen, in literature as in our experience, that a piece played by different musicians can sound very different; this is the case of the two performances used in the first experiment. Furthermore, experimental data demonstrate that performers cannot avoid of giving to a piece their own prompt, inextricably tied to their own motor patterns. Nevertheless, this difference seems not to influence the listener's comprehension of the piece structure. Then, how does performance influence the processing of music?

The answer is provided by the quantitative analysis of the segmentations between the first and second listening. Before all, we have to remember that in the paradigm we used the order of presentation of the performances is balanced across participants. Then, in the first as in the second experiment half of the subjects listened to the performances in a given order while the other half has heard in the reverse order. Regardless of the order of presentation, we can observe that all the listeners produce more segmentations during the first listening than during the second. An hypothesis on this behavior is that, despite of the difficult nature of the composition we used, the subjects can memorize the pieces, or at least the salient elements of the pieces. During the first listening the subjects should mark some boundaries between chunks on the basis of the portion of the piece they have already heard; going on with the listening, they can notice that the piece is not continuing in line with their expectation and, in some cases, the element that they thought to be salient loses its saliency in this new larger context. In other words, expectancies can produce some "false alarms" that can be recognized only when knowing if these expectancies are true or false. During the second listening the subjects are already aware of the false alarms, which can be avoided with the result of decreasing the number of total segmentations. Furthermore, we observe that the number of segmentations also decreases during the listening, regardless of other variables. The former hypothesis, hence, seems to be insufficient to explain this phenomenon. A better hypothesis requires to enlarge our perspective for including two more variables, not necessarily independent from each other: the hierarchic nature of music and the action of goal-directed attention in the attribution of saliency scores.

4.3.2 Hierarchic nature of music and the need for goal directed attention.

The hierarchic nature of music has been described and documented in a conspicuous number of works by several authors. It is mainly related to musical grouping and can be described as the possibility of perceive each group of tones in a melody as:

3. Formed by shorter groups of tones
4. Belonging to a larger group of tones

The decreasing number of segmentations during the listening of a piece, joined with the decreasing of segmentations from the first to the second listening, can be explained in relation to the hierarchies used by the subjects for representing the piece. When the listener hears for the first time a piece he has never heard before, he necessarily must analyze a certain number of elements on the basis of different parameters. An idea of these parameters can be obtained from the saliency map models described in Chapter 2. At this level, the subject tends to group the notes in short groups separated by pauses or by the arising of some other salient element (e.g., a variation in intensity). During the listening, the subject could observe that some element or group of elements repeats. For example, consider a piece with the following structure (each letter represents a group of notes): A-B-A-B. At the very beginning of this sequence the listener will notice a group A, followed by a group B, then another A… When the listener understands that there is a perfect repetition AB-AB, he has the possibility of changing the hierarchic level that he is using to understand the piece. Whatever was the salient element marking a boundary between the groups A and B in the lower hierarchic level, it loses a consistent part of its importance in the upper one. Musical compositions and, in particular, the compositions we used in the experiments, are quite more complex than this simple structure. Repetition, in effect, is never complete, but almost always concerns only a few aspects of a group (e.g., rhythmic structure, melodic contour…). Furthermore, a group played at the end of the piece can be similar to another one played at the beginning. However, the only possibility for the listener to use always-higher levels of representation is that of going on with the listening for comparing always-longer portions of the composition. The results of our experiments exactly show this behavior: during the first listening of a piece, the subjects build always-longer groups, which belong to always-higher hierarchic levels. During the listening, elements that were salient in a certain hierarchic level become progressively less salient in the higher levels. If the segmentation of the beginning of the piece refers to small groups, that of the final portion refers to very large groups.

The micro-analyses of the MSAs shows that the main mechanism involved in segmentation is the detection of similarities and differences between chunks (this point will be better argued in a further paragraph). If it is true, then we have to consider that the subjects, even the musically naïve listeners of the second experiment, are able to memorize a conspicuous amount of information and to use it for comparing also very long segments. The effect of this comparison is the loss of salience of some elements, while the salience of other elements becomes progressively stronger. On the other hand, looking for similarities means that a perceptual object is matched with an object represented in memory. Despite of the majority of the models of auditory salience, and in agreement with Kalinli viewpoint, stimulus-driven attention cannot be sufficient for explaining the building of an auditory saliency map. In other words, we can claim that:

The perceptual salience of an auditory element in a musical context can be enhanced/inhibited by goal-directed attention, where the goal is, in general, the detection of a temporal regularity and, specifically, the detection of complete/partial similarities between temporal chunks.

The current research, hence, indicates the need of re-considering the role of goal-directed attention in the elaboration of an auditory map of salience. Some of the current maps, indeed, are not sufficient to explain the data described in this paragraph, since they exclude a priori any top-down mechanisms (e.g., Kayser's map). Other models admit a late intervention of goal directed attention but only on the elements arising from an already built map of salience (e.g., De Coensel's map). Even if this work does not allow to exclude this hypothesis, this point deserves a consideration: if the subject has, as we demonstrate, the possibility of changing the saliency of the elements in the piece for obtaining a better representation of the same, it does not mean that he is necessarily aware of it. Models as such of De Coensel, thus, should need to hypothesize two different automatic, aware-less levels of elaboration before that the listener can have a conscious representation of the stimulus. On the opposite, a saliency map obtained trough a dynamic interaction of stimulus-driven and goal-directed attentive mechanisms seems to be more efficient even if more difficult to reproduce.

4.3.3 Expertise and other variables

The idea of shifting hierarchic levels during the listening is strengthened by the results concerning the variable expertise in the second experiment. We can observe, indeed, that musicians and non-musicians tend to choose the same points to mark boundaries between chunks, but with a difference in the number of segmentations. This behavior has been already described in former researches, which obtain the same results: musicians produce less segmentations than non-musicians. The main difference between the two categories is, exactly, the amount of musical training they had. The result, thus, can be explained with the hypothesis that musicians are faster in shifting to higher hierarchic levels, since their experience help them to quickly recognize similarities among groups of tones. In this way, they can avoid many false alarms already in the early stages of the first listening. Interestingly, while expertise influences the number of segmentations, this is not the case of musical intelligence. In our work we used atonal pieces with a weak rhythmic structure, and that could be the cause of this result. However, with regard to the pieces used, musical aptitude is not sufficient for a faster comprehension of the compositions, while musical training is.

Another interesting (and unexpected) result is the sinusoid-shaped cross correlation between the temporal distributions of the segmentations in the two performances of Sequenza III. As written in the description of the piece, Sequenza III can be conceived as an alternation of two components that we called “sung” and “spoken”. This alternation is quite regular, with every component having a similar duration before of changing in the other one. The cross correlation analysis, hence, seems to reflect this alternation, with the majority of the segmentations placed across the boundaries between sung and spoken or vice versa. Once again, this data suggest that the detection of similarities and differences is the main mechanism involved in music segmentation.

In the first experiment we observed that women produce more segmentations than men. We hypothesized that it was the effect of a bias due to the absence of professional musicians in the sample. We could not know, indeed, if there was a difference in the knowledge of music among male and female subjects. In effect, when introducing musicians in the sample (exp 2), this difference disappears, while appears the difference between musicians and non-musicians.

Finally, we obtained some significant result in the second experiment but not in the first concerning the two versions: The longer performance produces more segmentations than the shorter one. This data is interesting if considering that, while in the second experiment the performances differ almost exclusively in duration, in the first experiment they differ *also* in duration. Furthermore, the order of presentation of the two versions of Sequenza III influences the subjects' answering: they produce more segmentations if the longer version is played before the shorter one. It is unlikely to claim, thus, that this result is due to duration only. Something else may have a role in the different data of the two experiments. The most evident difference between the pieces is in their nature, respectively instrumental, the first, vocal, the second. In Sequenza VI there is not a real discourse, but only a few English words that can be not always recognizable by the listener. Nevertheless, it is possible that their presence in the piece can activate processes that are not usually related to the listening of music, due (mainly but not only) to the different semantic nature.

In conclusion, this work poses the basis for a careful re-consideration of the current models of the processing of auditory stimuli through the elaboration of auditory maps of salience. Many points deserve to be clarified with further experiments, both concerning the issue of goal directed attention in the detection of salient features and with regard to other interesting points arisen from this work such as the role of gender and speech in the segmentation of musical pieces.

5 References

Apel, W. (1972). Harvard dictionary of music (2nd ed.). Cambridge, MA: Belknap Press of Harvard University Press.

Bamberger, J. (1978). Intuitive and formal musical knowing: parables of cognitive dissonance. In S. S. Madeja (Ed.). *The arts, cognition, and basic skills.* New Brunswick, N. J.: Transactions Books. Cited in Povel, D. J., Essens, P. J. (1985). "Perception of Temporal Patterns". *Music Perception*, 2, 411–441.

Behreus, G. A., & Green, S. B. (1993). The ability to identify emotional content of solo improvisations performed vocally and on three different instruments. *Psychology of Music*, 21, 20-33.

Bengtsson, I., & Gahrielssun, A. (1983). *Analysis and synthesis of musical rhythm*. In J. Sundberg (Ed.), Publications issued by the Royal Swedish Academy of Music (No. 17): *Studies of music performance* (pp. 27-59). Stockholm: Royal Swedish Academy of Music.

Bigand, E., Pineau, M. (1996). Context effect on melody recognition: A dynamic interpretation. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition,* 15, 121-134.

Boltz, M. G. (1991). Some structural determinants of melody recall. *Melody and Cognition*, 19(3), 239-251.

Boltz, M. G. (1998). Tempo discrimination of musical patterns: Effects due to pitch and rhythmic structure. *Perception & Psychophysics*, 60, 1357-1373.

Bregman, A. S. (1990). *Auditory scene analysis*. MIT Press: Cambridge, MA.

Bregman, A. S. & Dannenbring, G. (1973). The effect of continuity on auditory stream segregation. *Perception and Psychophyics* , 13 , 308–312.

Butler, D. (1979). A further study of melodic channeling. *Perception & Psychophysics*, 25(4), 264–268.

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 (ICMC'2001)*. Havana (Cuba).

Cambouropoulos E. (2003) Musical Pattern Extraction for Melodic Segmentation. In *Proceedings of ESCOM5*, Hanover, Germany.

Canazza, S., Poli, G. De, & Rodà, A. (2001). Kinematics-energy space for expressive interaction in music performance. In *Proceedings of the MOSART Workshop on Current Research Directions on Computer Music*.

Chomsky, N.(1968) *Language and mind* (Third edit.). New York, New York, USA: Cambridge University Press (1972).

Clarke, E. (1982). Timing in the performance of Erik Satie 's "Vexations." *Acta Psychologica*, 50, I - I9.

Clarke, E., & Krumhansl, C. (1990). Perceiving Musical Time. *Music Perception*, *7*(3), 213–252.

Coath, M., Denham, S. L., Smith, L. M., Honing, H., Hazan, A., Holonowicz, P., & Purwins, H. (2009). Model cortical responses for the detection of perceptual onsets and beat tracking in singing. *Connection Science*, 21(2-3), 193–205.

Coath, M., & Denham, S. L. (2005). Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience. *Biological cybernetics*, 93(1), 22–30.

Coensel, B. De, Botteldooren, D., Berglund, B., & Nilsson, M. (2009). A computational model for auditory saliency of environmental sound. *Nature*, 60.

Coensel, B. De, & Botteldooren, D. (2010). A model of saliency-based auditory attention to environmental sound. *20th International Congress on Acoustics, ICA 2010* (pp. 1–8). Sydney, Australia.

Cooper, G., & Meyer, L. B. (1960). The rhythmic structure of music. Chicago: University of Chicago Press.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge, United Kingdom: Cambridge University Press.

Deliege, I. (1987). Grouping Conditions in Listening to Music: An Approach to Lerdahl & Jackendoff's Grouping Preference Rules. *Music Perception*, *4*(4), 325–360.

Deliège, I. (1989). A perceptual approach to contemporary musical forms. *Contemporary Music Review*, *4*(1), 213–230.

Deliège, I. (1996). Cue abstraction as a component of categorisation processes iin music listening. *Psychology of Music*, 24(2), 131-156.

Deliège, I. (2001). Introduction: Similarity perception -> categorization -> cue abstraction. *Music Perception*, 18(3), 233-243.

Deliege, I., & Ahnmadi, A. El. (1990). Mechanisms of Cue Extraction in Musical Groupings: A Study of Perception on Sequenza VI for Viola solo by Luciano Berio. *Psychology of Music*, *18*, 18–44.

Deliège, I., Mèlen, M., & Sloboda, J. (1997). Cue abstraction in the representation of musical form Perception and cognition of music. (pp. 387-412). Hove England: Psychology Press/Erlbaum (UK) Taylor & Francis.

Deutsch, D. (1972). Octave generalization and tune recognition. *Perception & Psychophysics*, *11*(6), 411–412.

Deutsch, D. (1975a). Two-channel listening to musical scales. *The Journal of the Acoustical Society of America*, *57*(5), 1156–60.

Deutsch, D. (1975b). Musical Illusions. *Scientific American*, *233*(4), 92–104.

Deutsch, D. (1978). Delayed pitch comparisons and the principle of proximity. *Perception & Psychophysics*, *23*(3), 227–230.

Deutsch, D. (1979). Binaural integration of melodic patterns. *Perception & Psychophysics*, *25*(5), 399–405.

Deutsch, D. (1982). Organizational processes in music. In M. Clynes (Ed.), *Music, Mind and Brain, M. Clynes, ed* (pp. 119–136). Plenum Publishing Corporation.

Deutsch, D. (1991). Pitch Proximity in the Grouping of Simultaneous Tones. *Music Perception*, *9*(2), 185–198.

Deutsch, D. (Ed.) (1999). *The Psychology of music* (second edition). San Diego, CA: Academic Press.

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology*, *5*(3), 322–337.

Dowling, W.J., & Harwood, D. (1986) *Music Cognition*. New York: Academic Press.

Dowling, W., & Hollombe, A. (1977). The perception of melodies distorted by splitting into several octaves: Effects of increasing proximity and melodic contour. *Perception & Psychophysics*, *21*(1), 60–64.

Drake, C. (1993). Perceptual and performed accents in musical sequences. Bull. Psychon. Soc. 31, 107–110.

Drake, C., & Botte, M. C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. Perception and Psychophysics, 54, 277–286.

Drake, C., Dowling, W., & Palmer, C. (1991). Accent Structures in the Reproduction of Simple Tunes by Children and Adult Pianists. *Music Perception*, *8*(3), 315–334.

Drake, C., & Palmer, C. (1993). Accent Structures in Music Performance. *Music Perception*, *10*(3), 343–378.

Duangudom, V., & Anderson, D. (2007). Using auditory saliency to understand complex auditory scenes. *Proceedings of the 15th European Signal Processing Conference EUSIPCO 2007*, (1), 1206–1210.

Duke, R. A. (1989). Effect of melodic rhythm on elementary students' and college undergraduates' perception of relative tempo. *Journal of Research in Music Education*, 37, 246-257.

Eichert, R., Schmidt, L., & Seifert, U. (1997). Logic, gestalt theory, and neural computation in research on auditory perceptual organization. *Music, Gestalt, and Computing Lecture Notes in Computer Science*, *1317*, 70–88.

Engel, K. C., Flanders, M., and Soecht- ing, J. F. (1997). Anticipatory and sequential motor control in piano playing. *Experimental Brain Research*, 113, 189–199.

Farnsworth, P., Block, H., Waterman, W. (1934). Absolute tempo. *Journal of General Psychology, 10*, 230-233.

Fraisse, P. (1956). *Les structures rythmiques*. Louvain: Publications Universitaires de Louvain.

Fraisse, P. (1963). *The psychology of time*. New York: Harper &Row.

Fraisse, P. (1982). Rhythm and Tempo. In D. Deutsch (Ed.) *The Psychology of Music*. New York: Academic Press.

Friberg, A., Bresin, R., Frydén, L., Sunberg, J. (1998) Musical Punctuation on the Microlevel: Automatic Indentification and Performance of Small Melodic Units. *Journal of New Music research* 27(3), 271-292

Gabrielsson, A., Juslin, P. N. (1996). Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. *Psychology of Music*, *24*(1), 68–91.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts.* (p. xii, 369 pp.). Oxford, England: Wiley.

Garner, W. R. (1974). *The processing of information and structure.* (pp. xi, 203). Oxford, England: Lawrence Erlbaum.

Garner, W. R., & Gottwald, R. L. (1968). The perception and learning of temporal patterns. *The Quarterly journal of experimental psychology*, *20*(2), 97–109.

Geringer, J. M., Madsen, C. K., MacLeod, R. B., & Droe, K. (2006). The Effect of Articulation Style on Perception of Modulated Tempo. *Journal of Research in Music Education*, *54*(4), 324–336.

Geringer, J. M., & Johnson, C. M. (2007). Effects of Excerpt Duration, Tempo, and Performance Level on Musicians' Ratings of Wind Band Performances. *Journal of Research in Music Education*, 55(4), 289–301.

Halpern, A. R (1988). Perceived and imagined tempos of familiar songs. *Music Perception*, 6, 193-202.

Handel, S. (1992) The Differentiation of Rhythmic Structure. *Perception & Psychophysic*s 52, 497–507.

Handel, S. (1993). The effect of tempo and tone duration on rhythm discrimination. *Perception & Psychophysics*, 54, 370-382.

Handel, S., & Oshinsky, J. S. (1981). The meter of syncopated auditory polyrhythms. *Perception & psychophysics*, *30*(1), 1–9.

Hang, B., & Hu, R. (2010). Spatial audio cues based surveillance audio attention model. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 289–292.

Imberty, M. (1987). L'occhio e l'orecchio: Sequenza III di L. Berio. In Stefani, G. & Marconi, L. *Il senso in musica.* Bologna, CLUEB, 136-163.

Imberty, M. (2000). Prospettive cognitiviste nella psicologia musicale odierna - Cognitivistic perspectives in modern musical psychology. *Rivista Italiana di Musicologia, 35*(1-2), 411-484.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506.

Iyer, V. (2002). Embodied Mind, Situated Cognition, and Expressive Microtiming in African-American Music. *Music Perception*, *19*(3), 387–414.

Jones, M. R. (1987). Dynamic pattern structure in music: recent theory and research. *Perception & psychophysics*, *41*(6), 621–34.

Jones, M. R., & Pfordresher, P. Q. (1997). Tracking musical patterns using joint accent structure. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *51*(4), 271–291.

Juslin, P. N. (1997). Perceived emotional expression in synthesized per- formances of a short melody: Capturing the listener's judgment policy. *Musicae Scientiae*, I, 225-256.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: relating performance to perception. *Journal of experimental psychology. Human perception and performance*, *26*(6), 1797–813.

Justin, P. N., & Madison, G. (1999). The role of timing patterns in recognition of emotional expression from musical performance. *Music Perception*, 17, 197-221.

Kalinli, O., & Narayanan, S. Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information. , 17 Ieee Transactions On Audio Speech And Language Processing 1009–1024 (2009).

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current biology : CB*, *15*(21), 1943–7.

Kristofferson, A.B. (1980) A quantal step function in dural discrimination. *Perception and Psychophisics,* 27, 300-306. Cited in: Clarke, E., & Krumhansl, C. (1990). Perceiving Musical Time. *Music Perception*, *7*(3), 213–252.

Krumhansl, C. (2000). Rhythm and pitch in music cognition. *Psychological bulletin*, 126, 159-179.

Kuhn, T. L. (1987). The effect of tempo, meter, and melodic complexity on the perception of tempo. In C. K. Madsen & C. A. Prickett (Eds.), *Applications of research in music behavior* (pp. 165-174). Tuscaloosa: University of Alabama Press.

Lapidaki, E., & Webster, P. (1991). Consistency of Tempo Judgements when Listening to Music of Different Styles. *Psychomusicology: Music, Mind & Brain*, *1*, 19–30.

Large, E. W., & Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science*, *26*(1), 1–37.

Lerdahl, F. (1989). Atonal prolongational structure. *Contemporary Music Review, 4*(1), 65-87.

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Luce, G.,G. (1972). *Body Time.* London: Temple Smith. Cited in: Clarke, E., & Krumhansl, C. (1990). Perceiving Musical Time. *Music Perception*, *7*(3), 213–252.

Lund, M. (1939). *An analysis of the "true beat" in music*. Unpublished doctoral dissertation, Stanford University. Cited in: Lapidaki, E., & Webster, P. (1991). Consistency of Tempo Judgements when Listening to Music of Different Styles. *Psychomusicology: Music, Mind & Brain*, *1*, 19–30.

MacKenzie, C. L., and Van Eerd, D. L. (1990). Rhythmic precision in the performance of piano scales: motor psychophysics and motor programming. *Atten. Perform.* 13, 375–408.

Marshburn, E., Jones, M. R. (1985). Rhythm recognition as a function of rate: Relative or absolute? Paper presented at the meeting of the Midwestern Psychological Association, Chicago (1985, May).

Mèlen, M., & Deliège, I. (1995). Extraction of cues or underlying harmonic structure: Which guides recognition of familiar melodies? *European Journal of Cognitive Psychology*, 7(1), 81-106.

Meyer, L. B. 1956. *Emotion and Meaning in Music*. Chicago: University of Chicago Press.

Meyer, L., & Cooper, G. (1960). *The rhythmic structure of music*. Chicago: Univ. Chicago Press.

Michon, J. A. (1972). Processing of temporal information and the cognitive theory of time experience. In J. T. Fraser, F. C. Haber and G.H. Müller (Eds.), *The study of time.* Heidelberg: Springer Verlag. Cited in: Clarke, E., & Krumhansl, C. (1990). Perceiving Musical Time. *Music Perception*, *7*(3), 213–252.

Monahan, C.B., and Hirsch, I.J. (1990) 'Studies in Auditory Timing: 2. Rhythm Patterns'. *Perception & Psychophysics* 47, 227–42.

Ockelford, A. (2004). On similarity, derivation and the cognition of musical structure. *Psychology of Music, 32*(1), 23-74

Ornstein, R. E. (1969). *On the experience of time*. Harmondsworth: Penguin Books.

Oshinky J. S., Handel, S. (1978). Syncopated auditory polyrhythms: Discontinuous reversals in meter interpretation. *Journal of the Acoustical Society of America*, 63, 936-939.

Pearce, M. T., Wiggins, G. A. (2006). Expectation in melody: the influence of context and learning. *Music Perception*, 23(5): 377-405.

Povel, D. (1977). Temporal structure of performed music. *Acta Psychologica*, 41, 309-320.

Povel, D.J. (1981). Internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 3-18.

Povel, D J. (1984). A theoretical framework for rhythm perception. *Psychological Research*, 45,315-337.

Povel, D. J., Essens, P. J. (1985). "Perception of Temporal Patterns". *Music Perception*, 2, 411–441.

Povel, D., & Okkerman, H. (1981). Accents in equitone sequences. *Perception & Psychophysics*, 7, 565–572.

Pfordresher, P. Q. (2003). The role of melodic and rhythmic accents in musical structure. *Music Perception: An Interdisciplinary Journal*, 20 (4): 431-464.

Repp B., H. (1997) The aestethic quality of a quantitatively average music performance: two preliminary experiments. *Music Perception*, 14, 419-444.

Repp, B. H. (1999). Control of expressive and metronomic timing in pianists. *Journal of Motor Behavior,* 31, 145–164.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review*, 77, 481-495

Restle, F. (1973). Serial pattern learning: higher order transitions. *Journal of Experimental Psychology*, 99(1), 61-69)

Reyna, V., & Brainerd, C. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*(1).

Rosch, E., "Principles of Categorization", pp. 27–48 in Rosch, E. & Lloyd, B.B. (eds), *Cognition and Categorization*, Lawrence Erlbaum Associates, Publishers, (Hillsdale), 1978.

Royer, F. L., & Garner, W. R., Response uncertainty and perceptual difficulty of auditory temporal patterns. *Perception & Psychophysics*, 1(2): 41-47.

Royer, F., & Garner, W. (1970). Perceptual organization of nine-element auditory temporal patterns. *Perception & Psychophysics*, *7*(2), 115–120.

Schulze, H.- H. (1978). The detectability of local and global displacements in regular rhythmic patterns. *Psychological Research*, 40, 173-181.

Schwarzer, G. (1997). Analytic and holistic modes in the development of melody perception. *Psychology of music,* 25, 33-56.

Sloboda, J. A. (1983). The communication of musical metre in piano performance. *Quarterly Journal of Experimental Psychology*, 35, 377- 396.

Sloboda, J. (2000). Individual differences in music performance. *Trends in cognitive sciences*, *4*(10), 397–403.

Stainsby, T. & Cross, I. (2009). The perception of Pitch. In (S. Hallam, I. Cross, & M. Thaut, Eds.). *The Oxford Handbook of Music Psychology. Science (New York, N.Y.).* Oxford University Press. Oxford (UK).

Sternberg, S., Knoll, R. L., Zukofsky, P. (1982). Timing by skilled musicians: perception, production and imitation of time ratios. In D. Deutsch (Ed.) *The Psychology of Music*. New York: Academic Press.

Tenney, J., & Polansky, L. (1980). Temporal Gestalt perception in music. *Journal of Music Theory*, *24*(2), 205–241.

Tekman, H. G. (2002). Perceptual integration of timing and intensity variations in the perception of musical accents. *The Journal of general psychology*, *129*(2), 181–91.

Treisman, M. (1963). Temporal discrimination and the difference interval: Implications for a model of the internal clock. *Psychological Monographs*, 576. Cited in: Clarke, E., & Krumhansl, C. (1990). Perceiving Musical Time. *Music Perception*, *7*(3), 213–252.

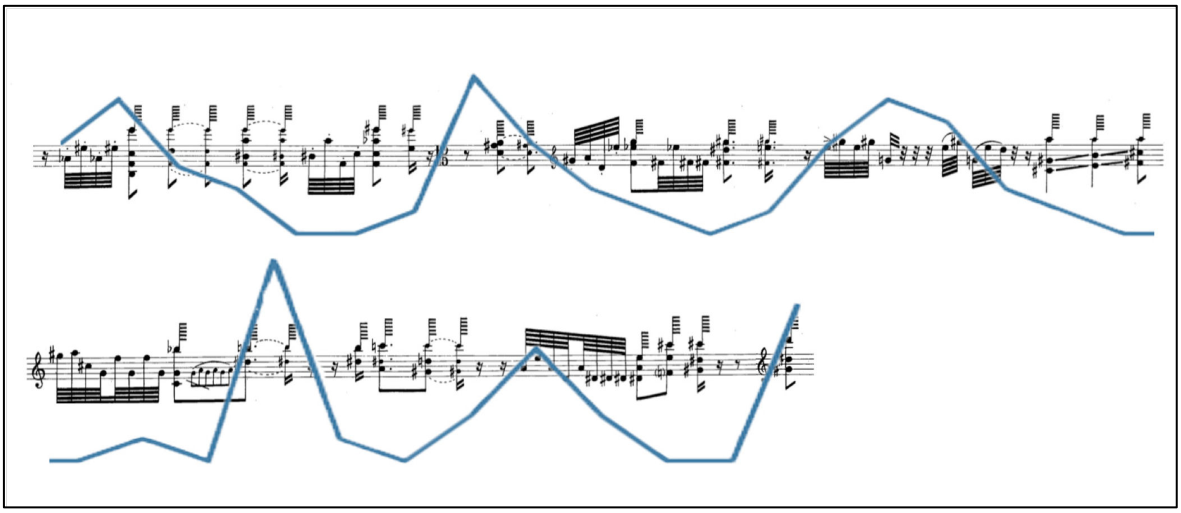
Van Vugt, F. T., Jabusch, H.-C., & Altenmüller, E. (2013). Individuality That is Unheard of: Systematic Temporal Deviations in Scale Playing Leave an Inaudible Pianistic Fingerprint. *Frontiers in Psychology*, *4*(March), 1–10.

Wang, C. C. (1983). Discrimination of modulated music tempo by music majors. *Journal of Research in Music Education*, 31, 49-55.

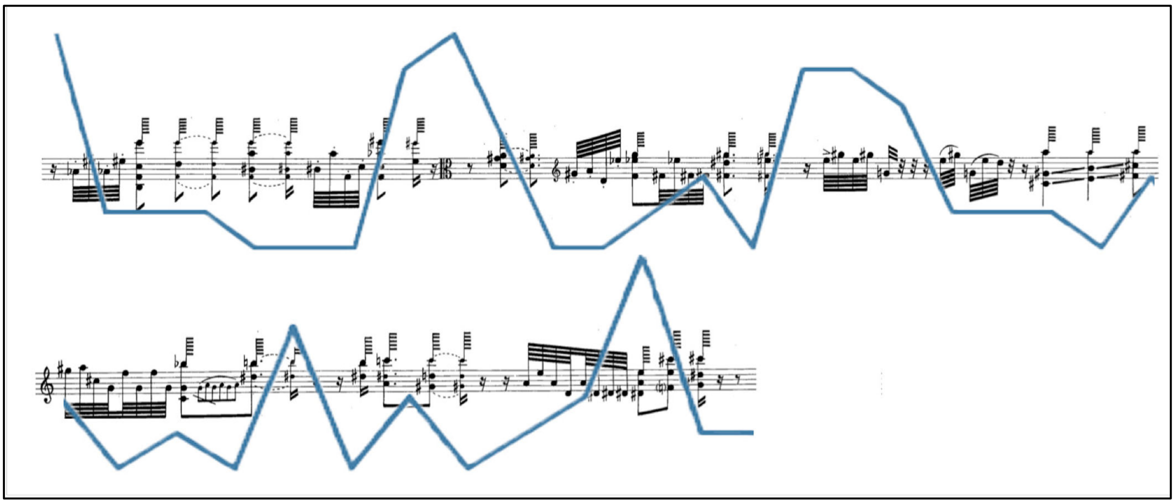
Wang, C. C., & Salzberg, R. S. (1984). Discrimination of modulated music tempo by string students. *Journal of Research in Music Education*, 32, 123-131.

Wang, K., & Shamma, S. (1995). Spectral shape analysis in the central auditory system. *Speech and Audio Processing, IEEE Transactions*, *3*(5), 382–395.

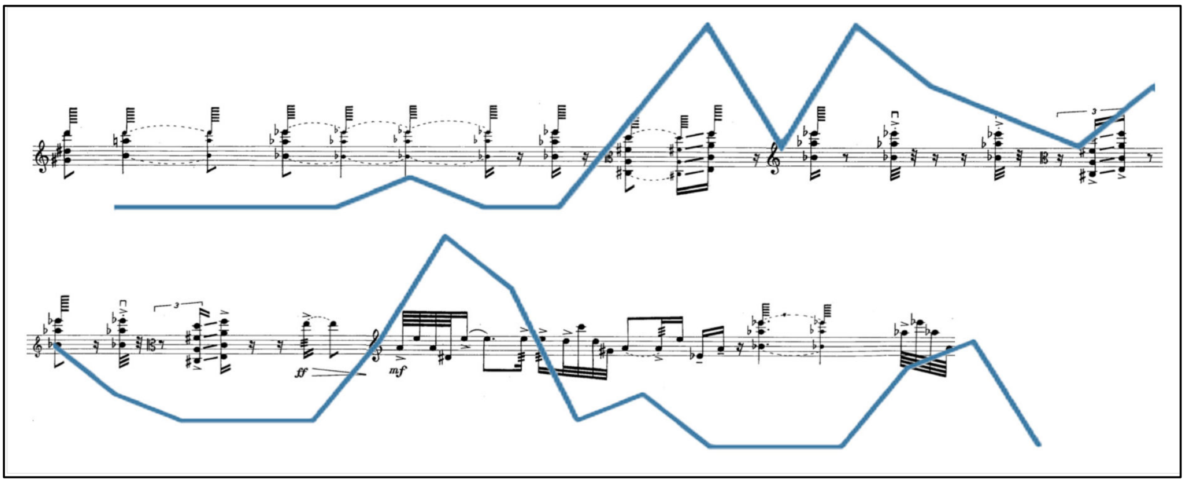
6 Appendix A



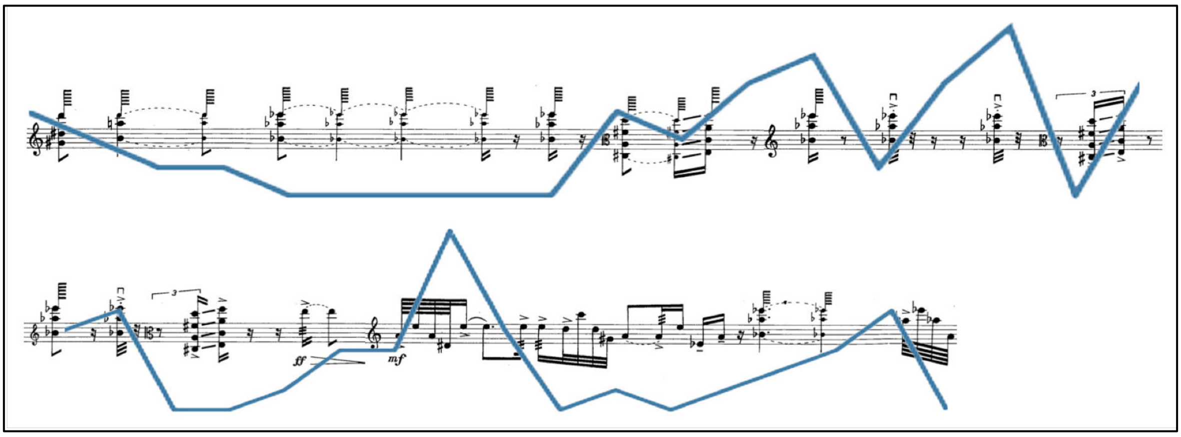
A 1 Performance A (IV)



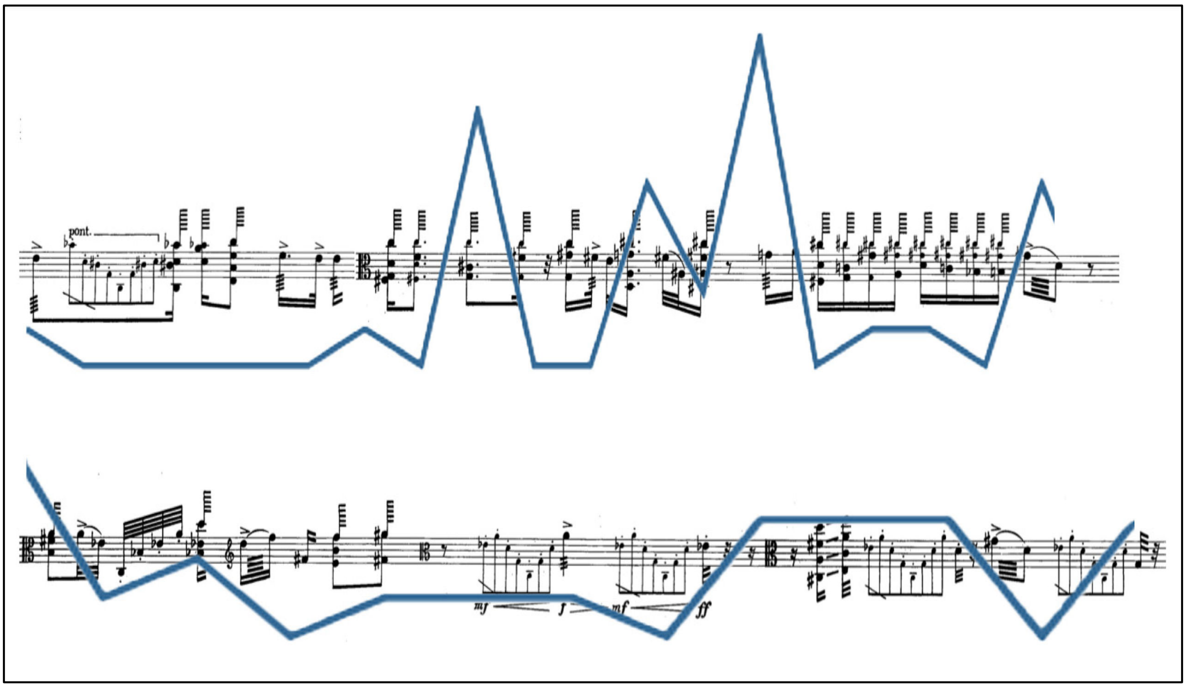
A 2 Performance B (IV)



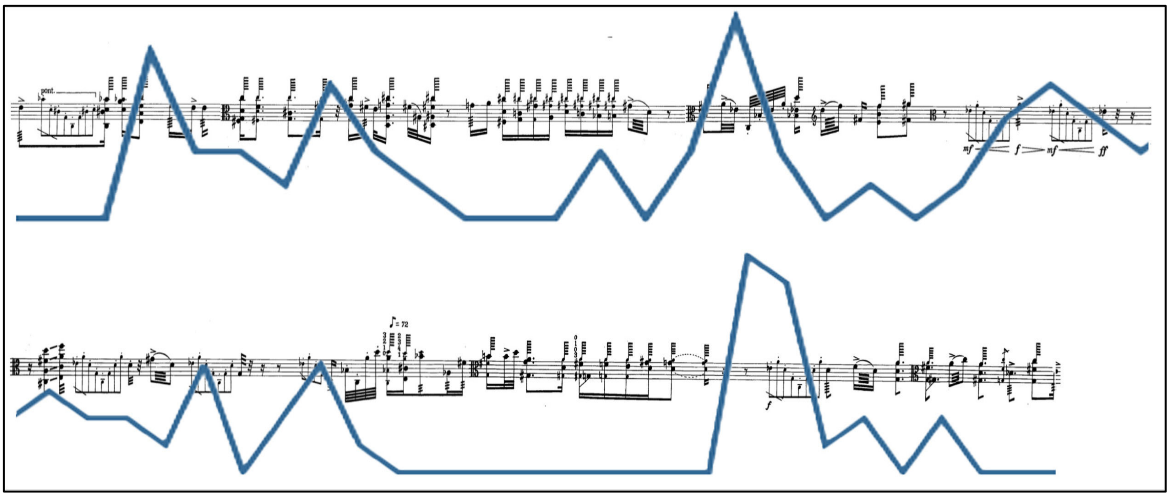
A 3 Performance A (V)



A 4 Performance B (V)

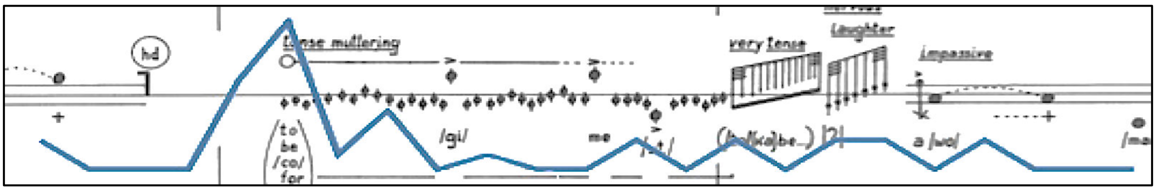


A 5 Performance A (VII)

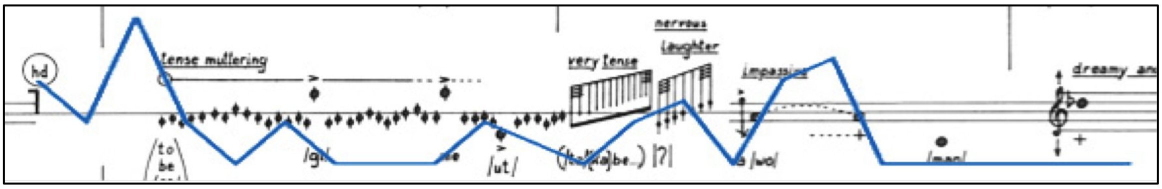


A 6 Performance B (VIII)

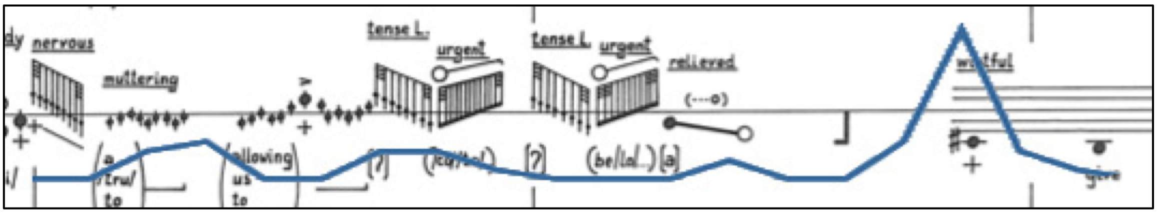
7 Appendix B



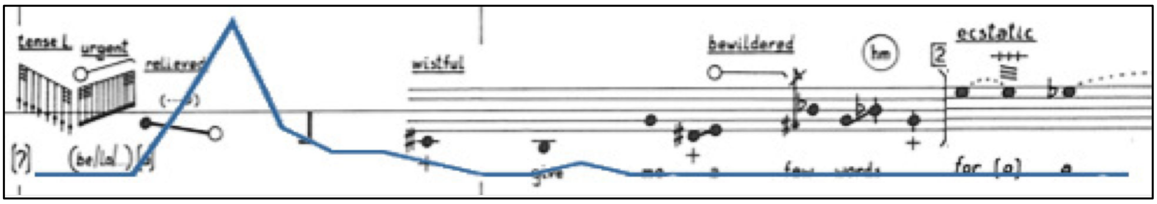
B 1 Performance A (III)



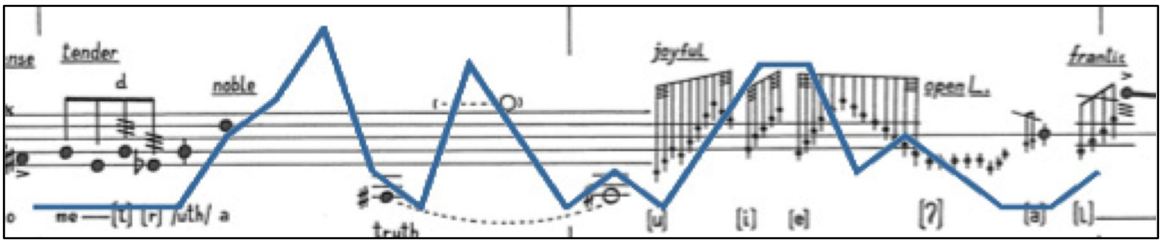
B 2 Performance B (III)



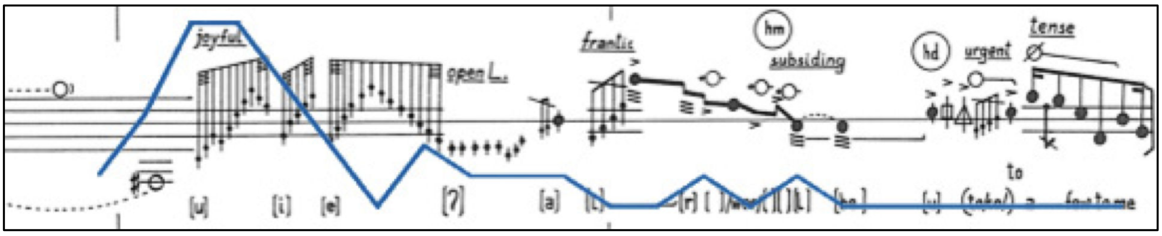
B 3 Performance A (V)



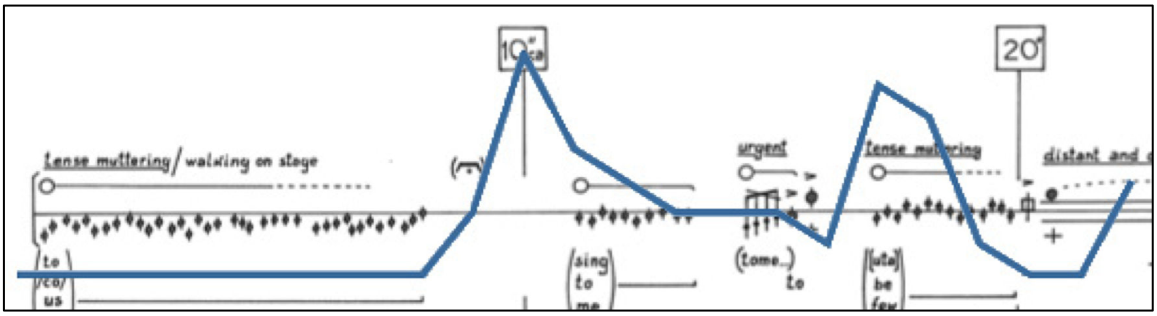
B 4 Performance B (V)



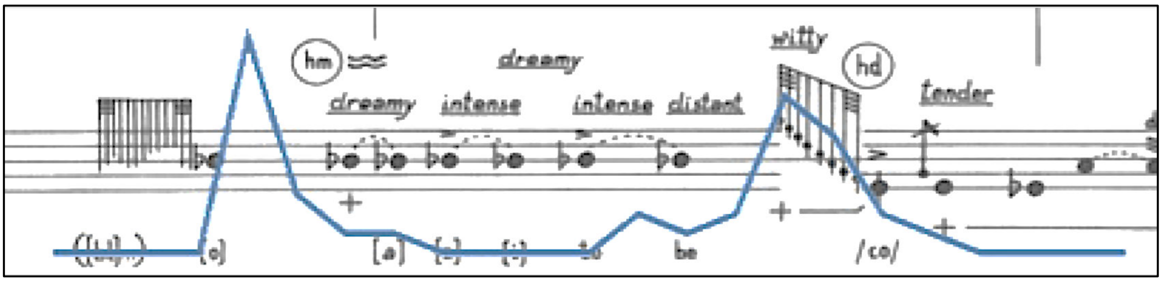
B 5 Performance A (XII)



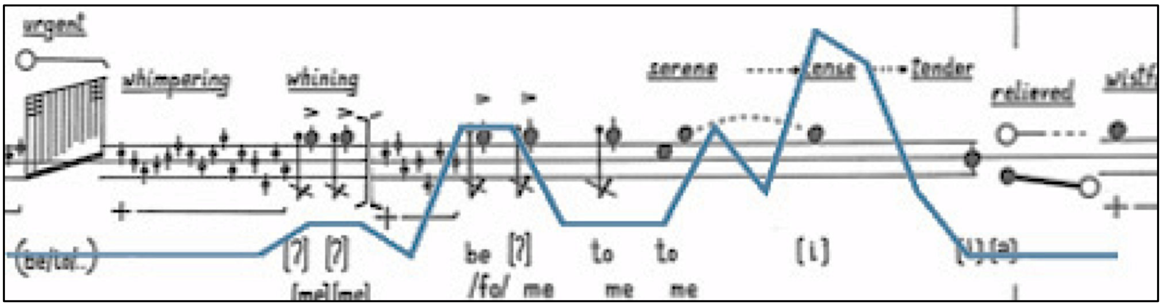
B 6 Performance B (XII)



B 7 Performance A (I)



B 8 Performance B (XV)



B 9 Performance B (XX)