Sound and Music in Narrative Multimedia

A macroscopic discussion of audiovisual relations and auditory narrative functions in film, television, and video games

Master's Thesis by Are Valen Lund

Institute of Musicology
Faculty of Humanities
University of Oslo

October 2012

Preface

Audiovisual storytelling has always fascinated me. As a musician and life-long consumer of films, television shows and video games, I simply could not pass on the opportunity to combine academic endeavours with personal interests. The initial motivation for this thesis was the desire to better understand that which I have spent so much of my time on over the years. All the while passively enjoying narrative multimedia, I have grown increasingly curious as to their nature as of late, and this project is an attempt to satisfy that curiosity.

While working on the thesis I repeatedly found myself facing problems that I was not equipped to deal with on my own. In this regard, I wish to thank my supervisor, prof. Rolf Inge Godøy, for all his guidance and support. Being able to draw on his knowledge and experience was an immense help when it came to distilling the myriad questions that presented themselves into manageable problems. Also, thanks to friends and family for support and for putting up with my occasional whining.

Oslo, October 2008
Are Valen Lund

Contents

Contents

Chapter 1: Introduction

This thesis presents a macroscopic, comprehensive discussion on audiovisual relations and contextual auditory meaning in narrative multimedia; here defined as films, television and videogames. While primarily a literature study, an assortment of illustrational case studies are also presented. My overall objective is twofold: from a learning perspective I pursue an overview – and consequently a deeper understanding – of the complex relationship between visual and auditory stimuli that composes the narrative experience. In a more practical sense I wish to compile a descriptive referential framework from which to identify and decipher the connotations and narrative implications of audiovisual stimuli within the setting of narrative multimedia.

Naturally, this approach poses a particular challenge with regard to structure and literature, in that there is a seemingly infinite number of potential implications that could each be the sole topic of a thesis such as this one. Consequently, this thesis favours the bigger picture over detail. When all is said and done, my reasoning for going with this large-scale approach within the limited format of a Master's thesis mainly comes down to a strong personal preference – during my studies, I have consistently found it the most stimulating to discover points and areas where initially separate lines of inquiry start to converge and make sense in a wider scope, and I wish for that to be reflected here.

Before getting started I should state my intent somewhat more precisely. At a basic level, the primary questions I concern myself with throughout the thesis are:

1) What is the nature of the relationship between what we see and what we hear when experiencing an audiovisual narrative?
2) What are the implications of this relationship in terms of how a narrative is presented and interpreted?
3) What are the specific functions of sound and music in this regard?

We can obviously break these problems down further, which I will do as we go along. Additional perspectives will also be presented throughout this introductory chapter.

1.1: Initial considerations

Theory and structure

In the attempt to achieve the above goals I have consulted theory from a variety of interrelated and

overlapping disciplines. This includes the broad fields of cognition, psychology, musicology, and semantics, along with some more localized areas such as psychoacoustics and narratology. As the subject matter is quite diverse, appropriate compartmentalization is necessary. The current chapter introduces some general perspectives, along with an initial terminology that will serve as point of departure. There is also a very brief historical runthrough that will serve as backdrop and reference point for some of the problems I deal with over the course of the thesis. Chapters 2 and 3 build the theoretical foundation that I apply to the case studies presented in chapter 4. The general argument that ties the various chapters and sections together goes like this: in order to systematically examine the functions of sound and music in narrative multimedia, we must first identify the underlying conditions for the audiovisual narrative experience. Since the one common denominator for all audiences is the human body and brain, I find that this is a logical place to start.

Thus, In chapter 2 I adopt the viewpoint of cognitive psychology and account for basic cognitive operations from an evolutionary perspective, focusing on the relationship between vision and audition. I then procede to discuss various implications of audiovisual perception in the context of narrative multimedia by presenting a selection of related studies. I also discuss the role of emotion in audiovisual perception. The fact that I am writing this thesis within the context of musicology means that there will be an added emphasis on certain musical aspects. The referenced theory and literature should, to an extent, reflect this.

Chapter 3 is dedicated to the task of establishing a terminology for the relationship between sound and image in narrative multimedia. For this purpose, the work of Michel Chion – particularly his book *Audio-Vision. Sound On Screen* (1994) – has been my main influence. In terms of empirical values, Chion's work typically represents an introspective approach that I hope to balance out somewhat by means of the more empirically grounded research presented in chapter 2. Still, it is in my mind one of the most thoughtful texts available on the topic, and thus an obvious source of inspiration. Chion considers the audiovisual relationship on the basis of a perceived (though illusionary) imbalance between vision and audition, with our perception favouring the former. His approach is therefore focused on sound and the ways in which it influences our perception of images. I adopt a similar approach as well as several of his terms, a few of which I shall introduce in this current chapter.

In chapter 4 I present a selection of case studies that illustrate some of the finer points made in earlier chapters that I personally find to be of special interest. For the purpose of visual representation I use a system of successive screenshots that emulate storyboards, with important visual and auditory events annotated.

Definitions of meaning

One theoretic aspect that is implicitly pervasive throughout the thesis, but that I have chosen not to explicitly define, pertains to semantics. As a field, the study of meaning is both vast and heterogeneous. Thus, in a macroscopic thesis such as this one, it makes little sense to invest too heavily in this area. There will be a somewhat more detailed discussion of various types of meaning in relation to auditory phenomena and their narrative functions in chapter 3. When I invoke semantics elsewhere, however it is typically in its broadest sense – I do no necessarily specify whether its implications pertain to semiotics, linguistics, etc., nor do I give detailed accounts of the relationships between signifiers and denotata.

Meanwhile, I *do* find a useful basic distinction between textual and emotional meaning. The reason for this is simple: in terms of structure and presentation, narrative multimedia are in many aspects perfectly analogous to literature, in that dialogue and other forms of speech are all representations of text. Emotional meaning mainly comes into play when dealing with auditory aspects of the narrative *other* than speech (i.e. text), especially music. As we shall see in later chapters, it is quite possible to perceive emotional meaning without actually experiencing emotion; another potentially useful distinction.

When it comes to textual meaning, we can further distinguish between what one might perhaps call internal and external forms. The former describes meaning that is specific to the current text or narrative, or founded in basic, general human experience. The latter describes meaning that pertains more or less entirely to something *else*, i.e. that requires specific familiarity. Here, we touch upon what is often referred to as *intertextuality*. Originally coined in 1966[1] by Julia Kristeva (1980), this initially poststructuralist term has since been adapted for a wide variety of uses. In very general terms it denotes cases where the meaning of a text is shaped in some way by another text. Referencing, allusion, and influence are commonly understood as intertextual mechanisms, although not everyone agrees with this expanding definition. Intertextuality is not restricted to literature, or even to similar types of texts. In the case of an audiovisual narrative, intertextual relations can extend to most any other type of media.

A famous and frequently analyzed example of intertextuality is the «Dawn of Man»-sequence from the opening of *2001: A Space Odyssey* (1968). The scene portrays a pre-historic group of ape-men at a defining moment in their evolution. As one member of the group discovers and starts to play around with the skeletal remains of a long-dead animal, the now unmistakable fanfare to Richard Strauss' *Also Sprach Zarathustra* is heard as nondiegetic music. This work was

1 The original text that introduced the term was in French. An English translation was included in the book *Desire in Language: A Semiotic Approach to Literature and Art* (1980).

famously inspired by the similarly titled novel by Friedrich Nietzsche, in which a recurring philosophical theme is that man in his current form is but a stepping stone on an evolutionary ladder between the ape and the so-called Übermensch, i.e. a superior individual that has reached its full potential. In the film, the music begins quietly and builds towards its triumphant climax as the ape-man picks up a large bone and swings it over his head, thus turning it into a weapon. The scene basically symbolizes evolution – by aquiring this knowledge our primitive ancestor has taken a vital step towards becoming something more than his current form. The intertextual implications of the music not only mirrors, but largely creates this symbolism.

Definitions of sound and music

Over the course of these pages, the word *sound* is used in multiple ways. First, it refers to the general phenomenon that is the sound *wave*, which may be defined as an oscillation of pressure transmitted through a substance. Second, in the tradition established by Pierre Schaeffer it refers to *a* sound, i.e. a specific auditory 'object' that is distinguishable from other sounds on account of its general or unique features, the details of which I will get back to in chapter 2. Third, I occasionally use it broadly to describe auditory phenomena that are not intuitively musical (as in the thesis' title).

What exactly constitutes music as opposed to other types of sounds is not a primary concern of mine, as we all have some intuitive notion of what music is and what it is not, even though we admittedly often find that verbalizing our thoughts on the subject can prove somewhat of a challenge. Also, as most of us are no doubt well aware, the parameters employed in distinguishing music from sound in general can differ greatly between individuals. But why, then, speak about sound and music as if they were separate phenomena? Briefly stated, the reasoning behind this decision is that music in a narrative context – what one might call narrative music – prompts a question that other types of sounds do not: why is it there, and what does it do?

We cannot close our ears the way we can our eyes. Consequently, we are surrounded by sound at all times. Given that virtually everything that goes on in our environment produces sound, and that speech is our primary means of communication, we do not ask ourselves why such sounds would be part of the audiovisual narrative experience, because the answer is obvious: they are there because we would expect them to be there. This is not the case when it comes to narrative music – there is no soundtrack accompanying our daily lives. So, in what way does the presentation of a narrative benefit from music? This is a recurring question that I will explore from various perspectives as this thesis progresses.

Music as sociocultural phenomenon and narrative device

As the title suggests, music has a special place within my project. Much has been said and written about the prospect of music being the 'universal language'; a phenomenon known in some form or other to virtually every human being on the planet. Despite the fact that local musical practises across the world can seem very different from one another, the common view among researchers seems to be that all forms of musical expression are recognizable as such because they invariably share at least *some* of the same, basic parameters – the ones most often referred to being a sense of rhytm and/or pitch.

The function of music within different cultures and societies throughout human history is also a frequent subject of scientific discourse, the takeaway so far suggesting that the phenomenon of music has traditionally situated itself somewhere in the intersection of utility and pleasure, and continues to do so to this day. In any case, music undeniably impacts the life of most people in some way or other, ranging from the subtle to the potentially life-changing – if nothing else then simply by its mere ubiquity. Technological advancements has seen the availability of music skyrocket during the past decade. Facilitated by portable devices (i.e. mp3-players, smart phones, tablet computers) and instantaneous internet distribution, our everyday music consumption may now take place virtually anywhere, at any time. And this is not even considering the aspect of live music, which recently has seen an upsurge due to the general decline in record sales; a development claimed by some to be mainly the result of internet piracy.

Digressions aside, the point is simply that music permeates modern society to such an extent that it has become second nature; the very idea of music being so ingrained in popular culture and the general public consciousness that we rarely find it pertinent to reflect on its presence in a given setting, nor the fact that things were not always this way. It seems a reasonable assumption that music's ubiquity is in large part due to the fact that music listening – both active and passive – only pertains directly to *one* of our senses, namely hearing. Thus, it does not necessarily demand our full attention whenever it is present, but rather affords us a certain degree of freedom in deciding our level of involvement, from letting ourselves be absorbed to almost completely ignoring it.

Although there are certainly settings in which music refuses to be ignored and quite literally takes center stage, we frequently find ourselves listening to music while at the same time engaging in some other activity. This occurs in any number of possible locations. Whether doing chores, working out, riding the subway, shopping, going out with friends or simply relaxing at home, music is a likely companion – courtesy of either the home stereo, a portable device, or the public address systems typically found in stores, coffee shops and clubs. Naturally, a distinction could easily be

made between music listening as a conscious decision, and being subjected to random music by our surroundings. However, from the fact that a great deal of everyday music listening occurs entirely by our own choice we can conclude that for most of us music represents an unequivocally positive aspect of life, and that it is generally desirable. What exactly constitutes this desirable quality is, of course, a very complex question that I will not be able to address satisfactorily here. When I bring it up, it is mainly in order to focus on the one aspects that is sure to come up repeatedly if pressing random subjects for an answer: that music seems to be able to 'speak' to our feelings, as well as our imagination.

Over the course of our lives, most of us will have experienced countless instances where music alone made us feel or 'see' something, conjuring up emotions and/or mental imagery. In films, television and video games, the emotional and visual aspects of music routinely come to the forefront; the fusion of music and real images having proven a very potent combination in terms of our emotional experience of an audiovisual narrative. For a vast majority of people, certain types of sensory input invariably evokes emotional responses, regardless of whether what is perceived is real or not. Humans thrive on emotion, and we would much rather experience the thrill of feeling *something* than feeling nothing. Even if the external cause of our emotional response is artificial and we are fully aware of this fact – as will be the case when we are experiencing a work of fiction – the emotional response itself is not. For many, film is the prime example of a medium where this phenomenon routinely occurs, and is just as routinely partially attributable to music.

1.2: Brief historical perspectives on narrative multimedia

Film and television

The story of the modern audiovisual narrative starts with film. Between the initial establishment phase of silent film in the late 1890's and the complete takeover of sound film by 1930, the medium saw an explosive growth in terms of production values, commercial success and artistic ambition that has largely continued to this day. As is often pointed out, the silent era was not really silent at all – films were accompanied by music from a very early stage, although it is somewhat unclear whether or not it was present during the very first public showings (Larsen, 2007, pp. 13–15). As the silent era progressed it was not uncommon for films to utilize both speech and various sound effects, courtesy of actors, narrators, and so-called 'noisemen' – i.e. personnel that operated various sound-producing contraptions – who would be either behind or in front of the screen, synchronizing their performance with the picture. Of course, these were all *external* sounds, i.e. they were not recorded and mechanically synchronized with the images. Hence, film sound had a performative

aspect that made each showing unique. Needless to say, perhaps, this was not always for the better – as the proficiency of these performers would have a big impact on the audience's experience, especially when lacking.

Initially, films were short in length and simple in terms of narrative. They would typically play as part of a mixed program in already established venues such as variety theaters, music halls, etc.; the notion of moving pictures being considered a spectacle in itself. As films became longer and narratively more complex, the medium eventually gained independence from other forms of entertainment. Dedicated cinemas were needed to accommodate its growing popularity. The transition into sound film that took place during the 1920's played an important part in the formation of modern narrative conventions, as it forced debate on certain fundamental issues that had previously been settled by default due to technical limitations. At the time, cinema was considered a visual medium and several prominent filmmakers and critics of the era regarded the prospect of synchronous sound as redundant or even detrimentary to its artistic integrity. While some fully opposed sound film, others (notably Sergei Eisenstein) were mainly reluctant towards dialogue while embracing other aspects. The experimental phase that resulted from the ongoing debate gave birth to the notion of audiovisual counterpoint, i.e. that sound and images should be asynchronous and avoid communicating the same meaning. Although this view would not establish itself as convention, it has continued to draw the interest of both filmmakers and historians, and its influences can be seen in the works of many later directors.

The eventual breaktrough and widespread acceptance of synchronous dialogue and naturalistic sound should be seen in conjunction with the so-called studio system that dominated Hollywood between 1920 and 1960. The five largest conglomerates[2] at the time each excerted full internal controll over production and distribution through direct ownership of both the studio facilities and theaters. As commercial enterprises, they naturally sougth to maximize profit by increasing efficiency in all stages of production. In terms of the construction of narrative, dialogue made films easier to make, as well as understand.

During the silent era music had been viewed as a necessary but ultimately inferior companion to the image. As naturalistic film sound became the norm, its functional role was by some thougth largely eclipsed by the possibility of realistic and perfectly synchronous sound. Hence, the question was raised as to what role music would now play, and whether it should still have a place in film at all. Hence, composers within the studio system experimented with styles and

2 Commonly referred to as 'the Big Five', these studios were Fox Film Corporation, Loew's Incorporated, Paramount Pictures, RKO Radio Pictures and Warner Bros..

different ways of implementing music. As some were met with critical acclaim, practices started to converge. Eventually, film music proved its value and settled into convention; however it has continued to evolve over the decades.

When it comes to the television medium, the most interesting aspect in terms of this thesis is the one type of narrative that is specific to it, namely the serialized drama. The prospect of telling a larger story in the format of weekly episodes and annual seasons has many advantages over film, and this previously untapped narrative potential is increasingly being recognized and realised. Within the last two decades the medium has seen its status change from being viewed as a type of poor man's cinema to being hailed as the future platform of choice by several established film directors.

However, for a TV-series to prosper it must typically seize its audience right out of the gate, or face unceremonious cancellation by its network. This is especially true in the United States, where executives have gained a reputation for being notoriously impatient in this regard. A trademark feature is traditionally a good way to distinguish oneself, and in terms of narrative many successful TV-series appear to have established their own subtle conventions. I look at two such cases that pertain to the auditory domain in chapter 4.

Video games

Similar to films, video games also had music before they could speak; at least as in 'over any extended period of time'. This was, of course, largely due to issues of computing power and storage space. The sound recordings required for speech would result in datafiles too huge to fit on any reasonable amount of storage units available at the time. Music on the other hand was much easier to accommodate, because no sound file was required – early arcade machines from the 1970's, as well as the first home systems of the 80's, had dedicated DA-converters able to generate analogue sound waves from simple playback information in the form of computer code that could easily be included on the game disc or cartridge. During the first decades of video game history, audio and visuals were at rudimentary levels and narrative was practically nonexistent. The use of music was initially very limited. Typically, music would occur only at specific points such as the title screen or upon starting/finishing a stage, or the same song or musical pattern would repeat for the entirety of the play session without there being any kind of actively developing relationship between the music and the visual action. Also, there was no polyphony, only short sequences of single notes. Even with the addition of multiple audio channels, which meant that more notes could play at the same time, a

term such as ‘narrative music’ hardly seems applicable to these early muscial practices[3].

As technology became more advanced, music would become more prevalent and increasingly involved with the visual aspects and gameplay. Different stages (or levels) would be given their own distinct musical track, sometimes accentuating a particular visual theme – with the possibility of more detailed visuals, designers would frequently draw inspiration from real-world cultures and architecture, past and present (e.g. roman, egyptian, chinese, etc.), and the music could then be composed in a complementary fashion, based on more or less stereotypical notions. Also, the player’s in-game actions would sometimes determine which music track would play or whether music would play at all. Thus, a more tangible relationship between audio and visual action was established.

The so-called graphic[4] adventure games that first began to appear during the early 1980’s and became increasingly popular throughout the decade pioneered audiovisual storytelling within the medium. While these were not the first games to tell stories, they were the first to combine a graphic interface and actual storytelling[5]. In many such games music is a near constant, if generic, companion to the visuals. We can safely say that the type of music heard in games like *Maniac Mansion* (1987) does not really participate in the narrative. For one thing, the argument could be made (at least in retrospect) that it does not really suit the mood of the game, even in a general sense. But most importantly, it is not at all affected by the action but plays continuously regardless of what is going on. It appears to be relevant merely because it is the only type of sound that the game is capable of producing.

Because the video game medium has to deal with the added element of player interaction, it becomes a much bigger challenge to implement music narratively when compared to film. Although game developers can easily control most parameters of the player experience, such as where you can go and what you can do, it is typically up to the player to decide *when* to perform the actions required to proceed. For instance, if tasked with solving a logic puzzle, the amount of time players use will inevitably vary. The *Interactive MUsic Streaming Engine*, commonly known as iMUSE

3 Compositions heard in many early games are typically public domain, or obviously derived from other well-known pieces. Because programmers were not necessarily musically trained, many early developers saw the inclusion of music as somewhat of an afterthought. While this obviously raises the question of its suitability, there are many instances where ‘generic’ music has become so strongly associated with a particular game that it may be regarded as an identifying feature. One such case is Alexey Pajitnov’s *Tetris*, first released in 1984. This classic game has become more or less synonomous with a particular piece of music called «Korobeiniki», a Russian folk tune that featured in the 1989 Nintendo Game Boy-edition. In a sense, this primitive use of music appears to have some aspects in common with that seen in the first silent films.

4 This prefix indicated player interaction based on images, so as to separate these games from their exclusively text-based predecessors. Computer graphics during the

5 The type of ‘narrative’ that could sometimes be found in more action-oriented titles would typically consist of just a basic premise or backdrop that was rarely developed or expanded upon.

(1990), represents a breakthrough for narrative music in video games. It was conceived by the company behind *Maniac Mansion*, LucasArts, and was the first comprehensive system capable of intelligently adapting music to the moment-to-moment narrative action in video games. It could generate and rearrange music on the fly using a combination of preprogrammed cues and algorithms. Comparing *Maniac Mansion* with later games that use the iMUSE-technology, such as *Monkey Island 2: LeChuck's Revenge* (1991) or *Indiana Jones and the Fate of Atlantis* (1992), the difference is rather striking.

It was also during the early to mid 1990's that speech and other sound effects were introduced to video games in a more extensive manner. This was made possible by the CD-ROM format, which had been available since 1985 and could store much larger amounts of data. *Indiana Jones and the Fate of Atlantis* was consequently reissued the year after its first release with the addition of full voiceovers and digitized sound effects. The 1990's also saw the revolution of 3D-graphics, which has since come to dominate video games. As we move closer to the present, games become increasingly complex, both technically and narratively, and the ambition of developers appears to have grown proportionally with the technology. Modern big-bugdet games have production values that rival and exceed those of blockbuster films, and it is becoming more and more common for creative talent such as directors, actors and composers to work with both mediums. It is for this reason that I have chosen to 'complicate' this discussion by including it.

1.3: Audiovisual media and narrative structure

Temporal elasticity

In most types of narrative, time is rarely continuous. A well-told story typically structures its temporal flow around events that have an actual function in terms of the overall narrative instead of wasting time on every inconsequential aspect of the diegetic timeline. Modern feature-length films typically have a runtime of somewhere between 90 and 180 minutes in which to tell a story that covers a significantly longer timespan. An episode of a TV-show typically last from 20 to 60 minutes, and even though the serialized format is becoming more and more prevalent, it is still the common conception that there should – on some level – be a contained story arc within each individual episode. In any case, it is clear that diegetic time is a highly flexible entity that may be stretched, compressed, reversed, etc. Chronology is not a given either, as storylines are often presented in nonlinear fashion. Hence, manipulation of the diegetic timeline is not just a matter of practicality, but an essential aspect of the aestethic and dramaturgic experience of the modern audiovisual narrative. We will therefore have to examine the role of audio in this regard. I have

chosen to distribute the discussion on auditory temporal implications between all chapters, starting here.

The main events around which the narrative flow is structured typically occur in real-time, and uses both visual and auditory cues to establish temporal continuity. That is, even as the point of view changes, certain characters or objects may remain visible, while music and dialogue continues uninterrupted between shots. This is further discussed in chapters 3 and 4. Depending on the nature of the action, time may also slow down beyond its normal speed (slow-motion). There is an example of this in chapter 4.

Conversely, the flow of time is usually sped up or condensed whenever there is nothing going on in the narrative universe that requires moment-to-moment attentativeness (e.g. an uneventful journey, etc.). Thus, a story may leap between points on the diegetic timeline that are set hours, days or even years apart. Visually, such leaps typically occur as either 1) a direct cut, or 2) via some form of transitional image(s), e.g. a crossfade, intertitle, montage[6], or time-lapse[7]). It appears that more often than not, sound plays an important part in making the audience perceive these transitions as intended by the director, something that I examine further in chapters 3 and 4.

Nonlinearity

Many films have instances of isolated flashbacks/flashforwards, and some even use frequent shifting of the timeframe in a conceptual manner, as an element that comes to define their narrative structure. In such cases, it becomes very obvious that text is the primary structuring element if we exclude the sound. Since textual meaning is primarily conveyed as dialogue, this type of storytelling becomes very difficult to follow. The following examples illustrate this:

In *The Ususal Suspects* (1995), the narrative revolves around the interrogation of supposedly small-time criminal Roger 'Verbal' Kint (played by Kevin Spacey). As Verbal in the present recounts the events that brought him the attention of his interrogator, U.S. Customs-agent Dave Kujan (Chazz Palminteri), the narrative periodically launches into flashback-sequences. The twist is that Verbal is not always telling the truth – whenever he is caught lying his story changes, and we get to see an alternate, 'true' version of roughly the same events. Adding an interactive element, something similar was attempted (though perhaps not fully developed) in *Post Mortem* (2002), a film-noir inspired videogame set in 1920's Paris where you play as a detective investigating a

6 An editing technique wherein a series of brief shots are composed to form a short sequence that effectively condenses information relating to events that take place over a longer timespan.

7 A photography technique wherein images are recorded at a slower rate than the normal playback-speed, creating the impression that time is moving faster. Can be viewed as the opposite of slow motion.

double murder. Upon interrogating one particular suspect, the player will temporarily assume control of this character's past self as he explains his involvement with the case. Interestingly, the decisions made by the player during these flashbacks will to some extent determine the options that are later available when reassuming control of the main character. In *Memento* (2000) protagonist Leonard Shelby (Guy Pearce) suffers from a rare memory disorder that sees his short term memory reset every few minutes. This becomes a key narrative device, as the story is basically told in reverse: starting with the end, the film works its way backwards through each preceding memory fragment, gradually uncovering what lead to the events of the first and final scene.

Two of the case studies in chapter 4 continue this discussion to a an extent.

Interactivity

Video games separate themselves from other media by making its audience an active party. Physical interaction is required in order to advance the narrative, and, increasingly in modern games, the nature of the players' interactions can determine how the narrative develops. Games with a focus on story typically contain both interactive and noninteractive segments. The latter, which are often referred to as cutscenes, most closely resembles film and television. Indeed, it should be well known to anyone with a passing interest in the medium that emulating cinematic storytelling has been an expressed desire in the games industry for decades. Thus, there is an obvious narrative parallel to films and television, especially in terms of cutscenes. Video games typically convey significant plot developments via these noninteractive segments, and they are by far the easiest to implement into this thesis. I regrettably cannot satisfactorily account for the implications of the interactive nature of the medium in this context.

1.4: Terminology

Narrative and narrativity

While *narrative* refers to a story itself, *narrativity* denotes the various processes by which a story is presented and interpreted, both in terms of the intent of authors and the response of audiences. According to Abbott (2011), the term has taken on several connotations in modern usage:

> Though it has become a contested term, "narrativity" is still commonly used in two senses: in a fixed sense as the "narrativeness" of narrative and in a scalar sense as the "narrativeness" of a narrative, the one applied generally to the concept of narrative, the other applied comparatively to particular narratives. As such, it can be aligned with any number of modal pairings: e.g. the lyricism of the lyric/a lyric; the descriptiveness of description/a description. Depending on the context, these two uses of the term "narrativity" can serve their purposes effectively. But increasingly over the last three decades, the term has filled a growing and sometimes conflicting diversity of conceptual roles. In the process, other terms

> have, in varying ways, been drawn into the task of understanding narrativity, including "narrativeness" (used colloquially above), "narrativehood," "narratibility," "tellability," "eventfulness," "emplotment," and "narrative" itself (paragraph 2).

Because the focus of this thesis is what one might call audiovisual narrativity – i.e. the formation of a narrative in the intersection of sounds and images – I do not find it necessary to concern myself too much with the finer points of definition. I should, however, note that I use *narrativity* not in a scalar sense, but as a fixed, general concept – it is not in any way my intention to try and measure comparatively the *degree* of narrativity within that which I am analyzing. My point of departure is simply that a narrative is defined as such because it possesses narrativity to *some* (unspecified) extent. As for the question of its mechanics, the general perspective of this thesis is to view narrativity through the lens of basic perceptual mechanisms, which I believe can account for many of its implications, ranging from simple to complex.

In trying to define concepts such as narrative and narrativity we find that these are, really, qualitative descriptors of a perceived type of meaning. As per my earlier statement with regard to semantics, I will not pursue its intricacies here.

Diegesis

In all forms of storytelling the audience is experiencing both the narrative and their own reality at the same time. This brings us to the concept of the *diegesis*. In modern usage[8] this word refers to the internal reality of a narrative; i.e. the spatiotemporal world in which the characters live and the events of the story take place. Thus, the term *diegetic* describes any object or event that appear as if it is physically present in this world and as such can be seen and/or heard by both the characters in the action and the audience as external spectators. *Nondiegetic*, on the other hand, refers to any element in the presentation of a narrative that may be seen or heard exclusively by the audience. In the case of audio, this typically means narrative music and voiceover commentaries. On the visual side, the most common elements are title screens, opening/closing credits and subtitles. Additionally, nondiegetic can refer to the technical aspects of contructing an audiovisual narrative, such as writing, photography, editing, acting, etc.; elements that are only indirectly observable in the finished product.

There are many factors involved in determining whether or not an element has a physical presence in the narrative, but these I will discuss at a later time. I must note, however, that the distinction is certainly not always straightforward. The constant interplay between diegetic and

8 Dating back to ancient Greece, *diegesis* originally refers to the world of a story specifically being *told* or recounted by means of narration. It is contrasted with *mimesis* where, rather, a story is *shown* or demonstrated through enacting.

nondiegetic elements is a pillar of the narrative multimedia experience; one that authors frequently play around with.

The Audiovisual Contract and Added Value

Chion (1994) described the relationship between sounds and images as an 'audiovisual contract', claiming that «[t]he audiovisual relationship is not natural but rather a sort of symbolic pact to which the audio-spectator agrees when she or he considers the elements of sound and image to be participating in one and the same entity or world» (p. 222). In other words, perceived correlations between what we see and hear are more or less taken for granted when, in fact, there is every reason to raise questions regarding this relationship. Furthermore, in the context of cinema we are typically not explicitly aware of the fact that sound often conveys information that is not present in the image, and vice versa. However, this becomes evident if we separate them from each other, i.e. watching with the sound level turned all the way down, or listening with our eyes shut. Suddenly it becomes clear that both sounds and images respectively have vital narrative functions that the other does not, and often cannot, provide.

The term *added value* describes «[...] the expressive and informative value with which a sound enriches a given image so as to create the definite impression, in the immediate or remembered experience one has of it, that this information or expression "naturally" comes from what is seen, and is already contained in the image itself» (Chion, 1994, p. 5). Citing the opening scene of Ingmar Bergmans *Persona*, Chion demonstrated how integral sound is to the audiences' experience of the moment-to-moment narrative action, simply by asking us to remove it. With no auditory clues available, the images no longer tell the same story, because the sound carried vital information. Likewise, the sound alone is insufficient. In his own words; «Sound shows us the image differently than what the image shows alone, and the image likewise makes us hear sound differently than if sound were ringing out in the dark» (Chion, 1994, p. 21).

Although certain types of informative value are specific to either the visual or the auditory domain, there are also many instances where sound and image may be considered largely interchangable in terms of conveying a certain message or meaning. Chion (1994) called this *reciprocity of added value*, and used as example sequences by Aldrich, Cavani, Franju, and Tarkovsky, where sound and image combine to disturbing effect (Chion, 1994, pp. 21–24). Two of the aforementioned sequences portray acts of torture, but without explicitly showing the audience what is done to the victims. By means of camera placement the deeds are visually hidden from the viewer, and still the horrific nature of these acts is made almost completely unambiguous by the

sounds (screaming and gargling respectively) that the victims make as they are tortured. This is possible because sound and image reciprocally project their individual meaning onto each other, which in turn is attributable to the multimodal and pattern-oriented tendencies of perception, which I will discuss in chapter 2. Directors frequently treat violent content in similar fashion, both so as to avoid traumatizing the audience (particularly when targeting younger people) as well as for aesthetic and/or dramatic purposes. Making the audience 'fill in the blanks' using their imagination has repeatedly proven to be a very effective way of communicating certain aspects of a story.

Synchronization and Synchresis

An interesting aspect of screen-based audiovisual media is that images and sounds in their recorded state are entirely separate entities in terms of causality. Because images and soundtrack exist independently from one another and use different channels of presentation, events observed on screen are not the *actual* sources of the sounds we hear, although it certainly appears that way. This is largely due to synchronization – when a visual event coincides with an auditory event, it creates the illusion of a causal relationship. Conceptualizing this phenomenon, (Chion, 1994) devised the term *synchresis*:

> A *point of synchronization*, or synch point, is a salient moment of an audiovisual sequence during which a sound event and a visual event meet in synchrony. It is a point where the effect of synchresis (see below) is particularly prominent, rather like an accented chord in music (p. 58).
>
> Synchresis (a word I have forged by combining synchronism and synthesis) is the spontaneous and irresistible weld produced between a particular auditory phenomenom and visual phenomenon when they occur at the same time. This join results independently of any rational logic. [...] Synchresis is what makes dubbing, postsynchronization, and sound-effects mixing possible, and enables such a wide array of choices in these processes. For a single body and a single face on the screen, thanks to synchresis there are dozens of allowable voices – just as, for the shot of a hammer, any one of a hundred sounds will do (p. 63).

The exclusion of «rational logic» means that synchronization may produce an «irresistible weld» between sound and image even if the combination is completely unrealistic and we are fully aware of this fact. In animation and live-action slapstick comedy, inappropriate or exaggerated sound effects do not seem out of place because they somehow align with the exaggerated nature and intentional silliness of the visual action. But also 'realistic' films make frequent use of image-sound combinations that would simply not occur in real life.

One such combination in particular stands out; namely the punch. Chion argues that the typically exagerrated sounds we have come to expect when someone throws a punch in an action film is necessary to sell the illusion that contact was made and damage was inflicted. He refers to this as an *emblematic synch point* (1994, p. 60). Such sounds are often at their most exaggerated in

the martial arts genre, which typically favours elaborate fight choreography over storytelling. If we remove the sound, it is usually easy to see why: punches and kicks tend to lose their sense of impact, as if there were no weight behind them. Sometimes, particularly in older films, it even becomes blatantly obvious that the punch did not connect. This, of course, highlights the fact that the 'combatants' are really actors and/or stuntmen who are actively trying *not* to hurt each other, which will quickly disrupt our suspension of disbelief.

The synchresis-phenomenon is not without its limits and it is entirely possible to stretch these too far. One such instance in particular comes to mind: the so-called 'cork-screw' jump featured in *The Man with the Golden Gun* (1974) rank among the most famous car stunts ever portrayed on film. It is also notorious because of the completely out-of-place, flute-like sound effect that accompanies the slow-motion flight of the car. Few would argue that the result is anything less than ridiculous, and although it is very likely intended as tongue-in-cheek humour, it comes across as quite absurd in the context of the film, which does not really classify as a comedy.

The limits of our suspension of disbelief in relation to synchresis are also subject to change over time according to the culturally evolving conventions of narrative multimedia – just consider how modern audiences typically react with mild amusement and a headshake towards punch-sounds in films as recent as the late 1990's, while at the same time accepting the just slightly less unrealistic standard set by modern action films. Although the typical sound of a punch has been toned down considerably the past decade, it still cannot claim realism in any true sense of the word.

As the case of dubbing demonstrates, synchresis does not necessarily imply perfect synchronism. I discuss this somewhat further in the next chapter.

Chapter 2: Perception of sounds and images

The narrative experience of audiovisual media is comprised of two entirely separate entities that show a conspicuous tendency to act as though they were one and the same. At the core of Chion's audiovisual contract (see chapter 1) is the notion that the image is perceived as the dominant party, with sound essentially being relegated to a supporting role. In a truly audiovisual style of narrative this is obviously an illusion, because essential meaning is lost if one component is removed. It is hardly surprising that we tend to perceive sounds as less important than images in an entertainment setting, when this is also largely the case in real life. Just consider language – the words by which we describe auditory features largely seems borrowed from the visual domain. Why, for instance, does it make sense to describe a sound rich in 'high' frequencies as 'bright'?

It seems very likely that the everyday use of linguistically similar descriptors in the visual and auditory domains is no coincidence, but rather symptomatic of how we as humans perceive the world. Thus, in order to gain a better understanding of audiovisual relations, we need to first understand the nature of perception. Our tendency towards unifying auditory and visual information should come as no surprise given what science tells us about our sensory apparatus and how we perceive the world, nor should it surprise us that such a tendency can be exploited. This will be the topic of this chapter.

Perception is here defined as the general process of understanding our environment by organizing and interpreting sensory information. This chapter examines how we accomplish this and discusses various implications of the transition from a natural environment to an entertainment setting, with the ultimate intent of showing how certain inherent perceptual tendencies manifest themselves in the experience of narrative multimedia. I have chosen not to draw up any definite line between *perception* and *cognition*, as it appears to me that they share a significant degree of interdependency and that their individual definitions tend to vary somewhat. While it could be argued that the former mainly pertains to mechanisms involved in *acquiring* sensory information, whereas the latter pertains to the subsequent *processing* of such information, I do not find this distinction particularly useful in the context of this thesis. Thus, my usage of the term perception does not necessarily exclude higher-level mental processes traditionally labelled as cognitive. For the sake of clarity and consistency, however, cognition is preferred where it would generally be considered the appropriate term.

2.1: Cognitive perspectives on vision and audition

The survival agenda and perceptual biases

One fundamental aspect of perception is that it is subject to a multitude of mechanisms developed over the course of evolution, as well as our individual lifetimes, towards a very specific purpose which may be conveniently summed up as the ability of the individual to survive and propagate its species. This is the principal directive informing both conscious decision making and the basic, unconscious, cognitive processes that precede it. As we look at some of the under-the-hood processes of human perception, we should keep firmly in mind that the hard-wired survial instinct present in humans, as well as in most living organisms, impacts our moment-to-moment perception in some rather profound ways.

Generally speaking, the perceptual process starts with *something* existing or occurring in the physical world. This object or event may be referred to as the *distal stimulus*. With our sensory organs we can detect various forms of energy input (light, sound, etc.) from the environment, and thus collect information about the distal stimulus. This energy input is then transduced into neural activity, called the *proximal stimulus*. By processing the data provided by the proximal stimulus, the brain constructs a mental representation of the distal stimulus. This mental recreation of the original physical object or event is called a *percept*. However, since the process of recreating the distal stimulus uses only the information present in the proximal stimulus, the accuracy of our percepts are by no means absolute.

Studies have demonstrated the importance of taking into account the inherent subjectivity of human perception when dealing with auditory (e.g. Rocchesso & Fontana, 2003) and visual (e.g. Mack & Rock, 1998) phenomena. For instance, two separate individuals may sometimes perceive the same event in radically different ways; each focusing on a specific aspect of the distal stimulus that the other failed to notice. Occasionally we may even become convinced that we have seen or heard things that were simply not there. In short, all humans may under certain conditions experience differing, or downright false, interpretations of sensory information. This has to do with how the brain generally operates – because its capacity is limited, it must prioritize efficiently in order for us to comprehend, and survive in, the world. At any given moment the amount of information arriving at our sensory receptors is far beyond what any regular mind can handle. This means that the brain processes information in a highly selective manner, subconciously extracting useful information while discarding that which is deemed irrelevant. Thus, we only get to conciously evalutate a small fraction of the overall collected data. A direct consequence of the

capacity limitation, and one of the main reasons why the aforementioned errors occur, is that human perception is strongly biased towards logical connections and patterns, as opposed to randomness or chaos. Bluntly put, we have a natural preference for that which makes sense. Our natural biases and aquired knowledge of how the world works tells us that the likelihood of there being some kind of pattern is greater than the likelihood of there being none. Even when there is no real pattern to be found we typically assume there is anyway, and often procede to make errors. In general science terminology, a false positive – i.e. a perceived connection that is not real – is commonly referred to as a Type I error. Conversely, a false negative – i.e. failure to notice a connection that *is* real – constitutes a Type II error.

Furthermore, in our habitual behaviour we continuously project our experiences onto that which we are currently perceiving. When faced with a situation similar to one we have previously experienced, this causes expectations as to how events will unfold and their final outcome, which may in turn greatly influence our moment-to-moment interpretation of said events. In other words, we create *the best possible hypothesis* (Berthoz, 2000) based on the information we have and continue to search for evidence to support it – often disregarding more or less obvious clues to the contrary in the process. Lastly, because no two individuals lead identical lives, it seems likely that the criteria by which the brain organizes and interprets information are not entirely universal, but are also determined to some extent by our individual experiences.

Multimodality – sensory integration

We perceive the world through several sensory systems, popularly referred to as sight, hearing, smell, taste, and touch. Each sensory system is unique, in that it consists of a separate set of receptors (e.g. the retina) and neural pathways, and is triggered only by specific types of stimuli. Thus, each type of stimulus represents a separate sensory modality. The various sensory systems are coordinated by the brain, where the information they provide is interpreted and relied upon for decision making. Although these systems largely function independently from one another in a physiological sense, the interpretative process is one of integration. That is, the different types of stimulus information associated with each modality are not perceived as belonging to separate domains, but rather combine into a single perceptual entity. Thus, perception is *multi*modal – using our previous experiences as points of reference (more on this shortly), the brain continuously integrates the input from each modality and tries to form, as quickly as possible, a coherent sensory impression upon which we may decide the proper course of action.

In light of self-preservation this makes perfect sense, as speed is typically a deciding factor

when facing a potentially dangerous situation. In nature, a quick reaction may be the difference between life and death, and since the brain is ultimately responsible for triggering physical action it must be able to rapidly process huge amounts of information. At the same time, it is this sensory integration that causes us to forget that the visual and auditory modalities ultimately function on very different premises. When compared, it becomes clear that they both have inherent strengths and weaknesses (which I discuss later in this chapter). This means that they can and will influence each other in significant ways, some of which are essential to our experience and understanding of narrative multimedia. This appears to be the essence of the audiovisual contract (see chapter 1).

Grouping

When considering the sheer amount of information arriving at our sensory receptors at any given time, the fact that we are able to make any sense of it whatsoever is rather remarkable. Yet, we are usually capable of extracting and accurately analyzing relevant information even from very complex situations. This is partly accomplished through mechanisms of reduction where irrelevant or redundant sensory information is ignored and only a fraction of the data collected makes it into conscious thought. But how do we even distinguish one event from another, let alone decide what information is ultimately relevant? In the case of auditory stimuli this poses a particular challenge:

> The problem of scene analysis is this: Although we need to build separate mental descriptions of the different sound-producing events in our environments, the pattern of acoustic energy that is received by our ears is a mixture of the effects of the different events. It appears that our auditory systems solve the problem in two ways, by the use of primitive processes of auditory grouping and by governing the listening process by schemas that incorporate our knowledge of familiar sounds (Bregman, 1995, p. 641).

Grouping is «[...] the tendency for individual elements in perception to seem related and to bond together into units; the result of such a process» (Snyder, 2000, p. 259). The implication here is that elements which exhibit such properties must somehow be significant, while those that do not must be of secondary importance or irrelevant. According to the holistic view of gestalt psychology, the perceptual grouping of sensory information happens in accordance with certain laws, or principles. Shepard (1999) summarized the following principles:

> *Proximity.* Things that are located close together are likely to be grouped as being part of the same object. [...]
> *Similarity.* When objects are equally spaced, the ones that appear similar tend to be grouped as being related. If objects are similar in shape they are most probably related. [...]
> *Symmetry.* Because random unrelated objects in the world are not expected to exhibit symmetry, it would be most improbable for unrelated objects to exhibit symmetric relationships. [...]
> *Good continuation.* If objects are collinear, or arranged in such a way that it appears likely that they continue each other, they tend to be grouped perceptually. [...]
> The principle of *common fate* [...] dictates that objects that move together are likely to be connected. In the world, it is extremely improbable that two things move in a perfectly correlated way unless they are in some way connected. (Shepard, 1999, pp. 32–33)

While they are typically considered on the premise of visual stimuli, the principles of grouping apply to sound as well, as demonstrated by Bregman (1995, pp. 18–29).

The synchresis phenomenon is a clear indication that the grouping of sensory information is not 'locked' into separate modalities, but is indeed a multimodal process. Empirical evidence supports this claim. In one frequently referenced study, McGurk & Macdonald (1976) showed that when presented with images of a mouth speaking the 'word' «aga» while listening to the sound «aba», subjects would report hearing the word «ada». This phenomenon is commonly known as the McGurk effect, and demonstrates the reciprocal projections of visual and auditory meaning upon each other that tend to occur when we encounter their respective stimuli simultaneously. In the below excerpt, Chion (1994) clearly acknowledges the semantic and gestaltist implications of synchresis:

> [Synchresis] is not totally automatic. It is also a function of meaning, and it is organized according to gestaltist laws and contextual determinations. Play a stream of random stream of visual and auditory events and you will find that certain ones come together through synchresis and other combinations will not. The sequence takes on its phrasing all of its own, getting caught up in patterns of mutual reinforcement and phenomena of "good form" that do not operate by any simple rules. Sometimes this logic is obvious. When there is a sound that is louder than the others, it coagulates with the image it is heard with more strongly than previous or subsequent images and sounds. Meaning and rhythm can also play important roles in securing the synchresis effect (p. 64).

In another study, Lipscomb (1995) asked participants to evaluate the degree of audiovisual synchronization between perceived 'accent points' (which, incidentally, sounds very similar to Chion's synch points) in the music and images as audiovisual stimuli gradually increased in complexity. The results appear to indicate significant cross-modal influences when it came to the task of determining the salient features and their degree of synchronization. Also, precision dropped rapidly as the stimuli went from composites of basic music and simple animated sequences to incorporate actual film footage and sonically complex music. This appears to account for the 'elasticity' of synchresis, i.e. that which affords dubbing and other less precise forms of synchronism.

Categories

A category is basically an advanced form of grouping. While the gestalt laws of grouping are generally thought of as describing universal cognitive processes present in all humans from birth, categories are formed from our indiviual experiences as we grow and learn within the collective of our species. This is true for humans, as well as for other sentient lifeforms. Snyder (2000) writes:

> If they are to survive, all organisms must reduce the huge amount of information that comes in from the outside world, deciding which information is relevant to their survival. One of the primary mechanisms through which this is accomplished is *categorization*. [...] It is here defined as the ability (1) to group

> features together and thereby differentiate objects, events, or qualities; and (2) to see some of these as equivalent, and associate and remember them together in a *category* [...] Categories form the connection between perception and thougth, creating a concise form in which experience can be coded and retained (p. 81).

The obvious benefit of categories is that they allow for a common denominator approach to sorting information – for all the observable features of the world, we are able to focus our attention on those that appear to have something in common with each other and/or our previous experiences, and to classify them accordingly as an object or event. Interestingly, this implied preference for similarity extends beyond the physical aspect – as we shall soon see, even things that outwardly appear thoroughly dissimilar from one another may bond together by some abstract quality.

Organizing sensory information by categories also has the benefit of significantly reducing the workload of information processing. There is a trade-off, however, as our ability to recall details and other seemingly nonessential elements suffers because of the categorical divide between that which is perceived as crucial information and that which is not.

Memory

For information to be useful, we must be able to store and remember it. The formation and retaining of memory structures in the brain is often referred to as having three stages: encoding, storage and retrieval. The encoding process begins as information arrives at our sensory receptors and involves the initial transducing of energy input into neural activity as well as the basic information processing operations of grouping and categorization. Encoded information does not automatically become permanent memories, however. Before information can stabilize in long-term memory (LTM) it goes through several stages of processing. Based on the amount of time we are exposed to a stimulus, different mechanisms come into play. Within the first 200-500 milliseconds after perceiving something only briefly, a so-called sensory memory (SM) rapidly forms, degrades, and disappears (Sperling, 1963). There are different subcategories of SM for each modality – e.g. visual-, auditory- an tactile perception correspond to iconic-, echoic-, and haptic memory respectively. The capacity of SM is very limited, both in terms of duration and the amount of information that can be recalled. Thus, when only briefly exposed to a stimulus, we remember having perceived something but are usually unable to report its details. Sensory memories are thought to be automated responses that we have no top-down cognitive control over. This means that they cannot be consolidated (stabilized) into lasting memory structures. There appears to be no way of expanding the timeframe or capacity of SM.

Extended exposure to the stimulus activates the processes associated with short-term

memory (STM). Here, information can be recalled for anywhere between a few seconds and up to about one minute without repetition before it is lost, depending on the nature of the stimuli and the experience of the subject. Similarly to SM, the capacity of STM is very limited. However, there are ways of stretching these limits, as we shall soon see. It must be noted that STM is a general term that encompasses multiple processes. The inner workings of STM are often described in terms of a model known as working memory (Baddeley, 2000, 2003; Baddeley & Hitch, 1974); so named because the processes it refers to form the premise for everyday thought and reasoning. This model breaks down STM into four components: 1) the *central executive*, which basically coordinates the other three; 2) the *phonological loop*, which stores and continuously rehearses auditory information; 3) the *visuospatial sketchpad*, which stores and processes visual and spatial information; and 4) the *episodic buffer*, which integrates different types of data into multimodal memory units that encompass entire scenarios, e.g. a film scene.

With repetition, information temporarily stored in STM solidifies in LTM where it eventually becomes stable and can potentially be recalled indefinitely. This associative process is referred to as consolidation. Unlike SM and STM, LTM has no definite limitations when it comes to duration and capacity, although various mechanisms and disorders can disrupt the usually rapid process of recalling information from LTM – after all, we do occasionally fail to remember things for no apparent reason. For instance, stress can have a negative impact on memory. According to (Anderson, 1976), LTM can be divided into two main categories: 1) *declarative* (or explicit) memory refers to information that requires conscious recall. Subcomponents include a) *semantic* memory, which stores factual and abstract information; and b) *episodic* memory, which refers to personal experiences that are contextual in nature – the episodic buffer component of working memory seems to be the basis for this component of LTM. 2) *procedural* (or implicit) memory refers to information that is recalled and used unconciously, most notably motor skills. We automatically improve such skills with repetition without actually learning anything new.

Chunking

With regards to how information is stored, the brain works similarly to an archive or database, in that data is associated with other data and organized accordingly. Even when stored alongside many other similar elements, each individual piece of data can still be accessed. Through the processes of grouping and categorization, information is sorted into perceptual units that may consist of a single element or multiple ones. Such units may be referred to as *chunks*. Via association, an initially small information block may be expanded to accomodate much larger sets of data, thus increasing the

capacity of STM. The term *chunking* refers to this process, i.e. the ability to organize data into larger or smaller chunks, and navigate between these different 'levels'. Miller (1956) proposed that the number of chunks that an individual is able to recall in working memory immediately after being presented with a memory task is typically seven plus/minus two, although the validity of this number has since been questioned. It has been demonstrated that significant deviations from this rule will occur based on the conditions of the memory task – the nature of the data, mode of presentation and expertise of the subject all factor in. For instance, auditory information is more easily remembered than visual information. Currently, the average number of chunks is generally considered closer to four.

Remembering telephone numbers is an everyday example of chunking in practice. When presented with a rapid sequence of eight or more single digits, as in 9-1-8-2-7-3-6-4, many will struggle to keep them all in STM sufficiently long for the information to be stored in LTM. If, on the other hand, the digits are grouped, as in 91-82-73-64 or 918-273-64, the amount of units does not exceed any hypothesized limit. An even more common example is language – when learning an unfamiliar word, analysis of each individual letter or syllable is initially required, whereas with experience the whole word is immediately perceived as such. Although chunking is a feature of STM, it is facilitated by the associative nature of LTM: we find meaningful connections in the data in working memory through previously learned, schematic knowledge. Thus, chunking is a process of interaction between STM and LTM.

For the purposes of this thesis, the perhaps most interesting aspect of chunking is the (unsurprisingly) analogous relationship between this basic cognitive operation and language, particularly the linguistic units which we rely on for describing images and auditory features in narrative media. For instance, if we consider the construction of a film, terms such as *scene* and *shot* seems to denote larger and smaller chunks of information respectively. The same could easily be said for musical terms such as motif, phrase, theme, etc.

Schemas

Schemas (or schemata) are long-term memory structures that pertain to semantic memory. They are in a sense the logical extension of categories, i.e. useful generalizations that simplify information processing and thus reduce reaction time. Schemas play a key part in our daily routine and are in many ways essential to our general understanding of the world, as well as our interaction with it. They function as described below by Snyder (2000):

> In addition to categorizing objects and single events, we generalize and categorize entire physical

> situations and temporal sequences of events. When a number of different situations occuring at different times seem to have aspects in common, they are eventually averaged together into an abstract memory framework. [...] Schemas function as norms or sets of ideas about how things *usually* are, and allow us to move through situations without having to repeatedly consciously evaluate every detail and its meaning: they operate unconsciously to contextualize current experience (pp. 95–96)

As schemas continually develop for every real-life situation we encounter that in some way differs significantly from our previous experiences, we must assume that the same goes for the way in which we experience a fictional, technologically mediated narrative. First, consider that well over a century has passed since the earliest motion pictures and that the medium of cinema, along with the array of narrative techniques it typically employs, has since become staples of popular culture – both as an art form and as mainstream entertainment. Then consider how technological advances has brought the audiovisual language of the screen into the domestic sphere through television, computers and gaming consoles, and that modern children are typically exposed to its wonders from a very early age. Even with the potential of cultural variances it seems quite plausible that by now the act of experiencing an audiovisual narrative is a thoroughly, schematically ingrained one. This goes for both diegetic and nondiegetic aspects.

From the practical viewpoint of an author every form of narrative relies on schematic knowledge to some extent, because the vast majority of the information that is needed to understand what is going is not contained in the actual presentation of the story, but is an implicit part of the relationship between author and audience. In narratology, this aspect of schemas largely comes to define the term, as shown in this excerpt from Emmott & Alexander (2011):

> Schemata are cognitive structures representing generic knowledge, i.e. structures which do not contain information about particular entities, instances or events, but rather about their general form. Readers use schemata to make sense of events and descriptions by providing default background information for comprehension, as it is rare and often unnecessary for texts to contain all the detail required for them to be fully understood. Usually, many or even most of the details are omitted, and readers' schemata compensate for any gaps in the text. As schemata represent the knowledge base of individuals, they are often culturally and temporally specific, and are ordinarily discussed as collective stores of knowledge shared by prototypical members of a given or assumed community (paragraph 2).

On another note, one might point out the apparent connection between schemas and synchresis – when an event portrayed in the image temporally coincides with a distinctive sound, we immediately assume a causal connection because in our experience this is usually the case. And because schemas operate unconciously such assumptions largely go by uncontested. That is, as long as there are no significant deviations from the established schematic form.

Metaphor

As previously mentioned, the specific features shared by the various elements in a category, need

not be concrete. Categories and schemas streamline information processing by means of purposeful reduction, forming templates of generalized experience which we continuosly refer to when encountering new situations. The obvious implication is that the usefulness of such templates stems, at least in part, from their flexibility. This might seem strange, as lack of specificity is typically not regarded as a particularly useful quality, and it certainly seems to contradict the need for immediate classification dictated by the survival agenda. However, what this allows for is association; the triggering of schematic experience by something that is only partially or indirectly similar. This happens all the time, and often in less than obvious ways – i.e. the connection between the triggering event and the experience of which we are reminded may not be immediately clear to us. In the following excerpt, Snyder (2000) states that such seemingly indirect connections between current and previous experience can be explained in terms of *metaphor*.

> An interesting aspect of associative memory is that we can be reminded of some things by others in totally different areas of experience. This happens when aspects of what we are currently experiencing or thinking are somehow correlated to aspects of some other thing in long-term memory. [...] Although no two situations are absolutely identical, we can still make generalizations that function across different situations. Imagine, for example, learning as a child to use the word "close" both to describe closing your eyes and closing a door. The relation between these two uses of the word "close" requires making a connection between two different types of experience. This way of experiencing one thing in terms of another is referred to as "metaphor" (p. 107).

One might say that metaphors allow us to perceive the unfamiliar in terms of the familiar, which in turn facilitates learning and enables us to grasp phenomena at the conceptual level.

A common aspect of everyday conversation, the general understanding of the word 'metaphor' is that of one rooted in language. This is not entirely the case, however – although metaphors certainly feature prominently in language, they do not necessarily originate from it. In the case of certain metaphors there may not even be an adequate verbal expression. Snyder (2000):

> Metaphors not couched in natural language are referred to as "nonpropositional metaphors." That is, either or both the experience that cues the reminding, and the memory of which one is reminded, while they may or may not be describable in language, do not necessarily originally occur as statements in language. They may be connections between any kinds of experiences or memories, such as those of sounds, images, and smells, and some of the memory may be implicit [...] (p. 107).

This suggests that our frequent use of visually rooted linguistic descriptors regarding auditory features – musical or otherwise – is based on perceived connections that are metaphorical in nature. Obviously, the kind of metaphorical thinking that we commonly demonstrate in conversation is not restricted to the individual, but shared by most, if not all, individuals familiar with the language in question. If this was not the case metaphors would hold little to no referential value. Considering this, it would seem very likely for there to be some identifiable principles guiding the formation of metaphors. Snyder (2000) confirms this in the following excerpts:

> Recent theory has suggested that metaphorical mappings are not arbitrary, but are grounded in fundamental embodied cognitive structures generalized from recurring physical experiences, especially the experience of our own bodies. These cognitive structures have been referred to as "image schemas." [...] Image-schemas are thought to be derived from commonalities in different experiences that seem related; as such they are believed to form a basis for our conceptual systems, indeed to connect our perceptual experience and concepts (p. 108).

> Among the rudimentary human experiences generalized into image-schematic form are those of up and down, of spatial centeredness, of one event being linked to another through some form of causal connection, of moving along a path toward a goal, and of containment or "inside" and "outside" [...]. All of these image schemas require some sort of imaginary "space" in order to make sense, but they need not consist of specific (detailed) concrete images. So basic are they to our idea of how the world works that they are used not only literally but also metaphorically to represent many other, more abstract types of ideas (p. 109).

As this chapter progresses, I shall further discuss some implications of an embodied cognition in terms of the audiovisual relationship. For now, suffice it to say that as an aspect of perception, metaphor is in all likelihood closely linked with the integrated nature of information processing. Considering that many different types of sensory data are routinely and beneficially combined in the name of survial, it seems likely that metaphors serve as 'bridges' between physical and mental; between concrete and abstract. Because they add flexibility to otherwise rigid categories, metaphors are essential to our understanding of the world we inhabit. We may therefore suspect that they also represent a vital component in understanding perceived relations between the auditory and visual modalities.

2.2: Semantic and spatiotemporal perspectives on vision and audition

Vision versus audition

At the core of this project lies the seemingly intuitive notion that the visual and auditory modalities are in some way perceptually related, i.e. that certain features of a sound may somehow represent *the same* as certain features of an image. Considering the multimodal nature of perception (as detailed above), this is hardly surprising. However, when we take into account the obvious physical and operational differences between sensory systems, it becomes clear that the implications of such a notion warrants further explanation. Let us start by examining the following statement by Chion (1994):

> Visual and auditory perception are of much more disparate natures than one might think. The reason we are only dimly aware of this is that these two perceptions mutually influence each other in the audiovisual contract, lending each other their respective properties by contamination and projection (p. 9).

It appears to me that 'contamination and projection' refers to the multimodal process by which the brain actively – yet unconciously – shapes raw sensory data so that the various pieces fit together, thus foregoing the implied initial 'purity' of the sensory data in favour of a coherent, meaningful

picture. As for the disparity between vision and audition, there are several notable aspects. For one thing, vision is firmly rooted in the material world, because that which we can see is inevitably made up of some physical substance which reflects light. Even intangible substances consist of particles, molecules, atoms, and so forth, and thus, when we perceive visually, the distal stimulus is a concrete *thing*. Sound, on the other hand, is more elusive. In auditory perception, the distal stimulus is the sound wave itself, rather than a material object. A sound wave may be defined as an oscillation of pressure or vibration transmitted through a substance, and the sound we hear is the interpretation of such vibrations by our sensory system. Although there must be some kind of material present for sound to propagate, the sound wave itself is not only intangible, but immaterial altogether.

Granted, this could be said for the electromagnetic radiation that is light as well – although the reflective object will always consist of some form of matter, light *itself* does not. However, there is still an important distinction to be made here: unless we are looking directly at something which radiates light (e.g. the sun), visual phenomena are only indirectly perceived in terms of their associated stimulus modality. While we may be perfectly concious of the fact that light is a prerequisite for vision, the physical features that light makes visible to us are ultimately those of a material substance, and thus the sensory stimulus (light) itself becomes secondary in terms of awareness. Auditory features, on the other hand, are not observable in terms of any material substance – the perceived representational value of sound is merely connotative, i.e. a product of accumulated schematic knowledge within an individual or a culture (although there seems to be in effect some notion of an analogous, nonphysical object or entity, as we shall see). Instead, we perceive directly the features of the sensory stimulus; i.e. the sound signal.

This difference is perhaps best illustrated using the following scenario: imagine having no previous knowledge whatsoever about how objects, events, and sounds generally interrelate. A sound heard for the first time outside of any visual context cannot tell us anything about how its source (i.e. the object(s) involved in generating the motion we perceive as sound) looks, smells, tastes, or feels like. With no connotative points of reference, we would only be able to evaluate the features of the sound signal itself.

Moving on, Chion (1994) notes a second inherent and significant difference between vision and audition, related to the temporal dimension:

> [...] each kind of perception bears a fundamentally different relationship to motion and stasis, since sound, contrary to sight, presupposes movement from the outset. In a film image that contains movement many other things in the frame may remain fixed. But sound by its very nature necessarily implies a displacement or agitation, however minimal (Chion, 1994, pp. 9–10).

Motion is perceived in terms of a temporal change in the state of the observed object. Provided we are willing to submit to the notion of time as a constant and linear 'flow', every event will have a starting point and an ending point, between which something will have moved. In real life time cannot be paused or reversed, but the diegetic time of a recorded video can.[9] When pausing a recorded audiovisual sequence, the perceived motion of the image ceases but the current frame remains visible on screen. The sound, meanwhile, disappears altogether. While it it possible for any visual element in perception to remain perfectly static, sound is intrinsically linked with motion. In fact, sound *is* motion, and thus an inherently temporal phenomenon – as soon as time stops, sound ceases. While it is possible to create the impression that a sound has been 'frozen' by taking a small fragment and having it play back continuously in a loop, this sound event cannot exist outside this timeframe, however brief it may be. Here, we encounter another temporally related difference between vision and audition, namely that of perception *speed*:

> Sound perception and visual perception have their own average pace by their very nature; basically, the ear analyzes, processes, and synthesizes faster than the eye. Take a rapid visual movement – a hand gesture – and compare it to an abrupt sound trajectory of the same duration. The fast visual movement will not form a distinct figure, its trajectory will not enter the memory in a precise picture. In the same length of time the sound trajectory will succeed in outlining a clear and definite form, indiviuated, recognizable, distinguishable from others (Chion, 1994, p. 10).

Similarly, Snyder (2000) notes; «[...] hearing has the highest level of temporal acuity achieved by any of the senses; vision, for example, is considerably slower» (p. 25). These statements refer to the generally accepted notion in psychology that the various sensory memory stores have different temporal characteristics, with echoic (auditory) memory having a longer duration than ichonic memory. For recent empirical evidence in support of the above claim that a sound trajectory will linger in memory as a «[...] clear and definite form [...]» (Chion, 1994, p. 10), one might look to the so-called *sound-tracing* experiments conducted by Godøy et al. (2006) and later analyzed by Haga (2008), where participants demonstrated the ability to instantly render the perceived visual shape of several musical excerpts by drawing on a digital tablet, with largely convergent results. So what is the reason for this discrepancy in perception speed? According to Chion (1994), it is a question of systemic workload:

> The eye perceives more slowly because it has more to do all at once; it must explore in space as well as follow along in time. [...] So, overall, in a first contact with an audiovisual message, the eye is more spatially adept, and the ear more temporally adept. [...] In the course of audio-viewing a film, the spectator does not note these different speeds of cognition as such, because added value intervenes (p.

9 The notion of moving images is, of course, an illusion – in reality, a film consists of a series of static images that plays out too rapidly for us to recognize each frame as a separate event. Both visual and auditory perception have such thresholds. In technical terms these are typically referred to as *frame rate* and *sample rate* respectively. When such rates drop below their respective thresholds, we are suddenly able to identify the individual events: video ceases to be fluid (as is common in lower budget animated films), and audio is heard as nothing more than a series of 'clicking' noises.

11).

Of course, spatial exploration figures into audition as well, but the statement is still a valid one – vision *is* more accurate when it comes to locating objects in space, and we *do* process auditory information more quickly. In that regard, vision and audition complement each other. This is an important aspect of the audiovisual contract that I will return to in later chapters.

Point-of-view versus point-of-audition

As we have just seen, vision is the more adept modality when it comes to spatial exploration. It enables us to pinpoint precisely the position of objects in three-dimensional space, including ourselves. As we shall see, this is not the case with sound.

If we look at the term *point of view*, we find that it has multiple meanings in relation to narrative multimedia:

1) The nondiegetic, spatial position from which we as audience members are watching the screen, which implies an *exact* point in space.
2) The spatial position from which the camera captures the action in the frame, specifically in association with the first-person perspective, i.e. when it appears that the audience is viewing the action through the eyes of a particular character. When this is not the case I will use the term *camera position*. This interpretation also implies an exact point in space.
3) The figurative angle or perspective from which a story is being presented, typically representing the position, personality or values of a narrator, or character. This definition is commonly associated with narratology. I shall refrain from using it here, however.

It is easy to forget that the camera is an active and influential nondiegetic party in narrative multimedia, not only in terms of how we perceive images, but sounds as well. The reason for this is that audition is significantly less adept at spatial judgements, and as such vision easily trumphs it. Since the sound wave is invisible and tends to propagate omnidirectionally, it is typically impossible to pinpoint the location of its origin with the same precision that visual perception affords an object.

Granted, we can usually discern the general direction of a sound source due to the shorter amount of time it takes for the sound wave to arrive at whichever ear is situated nearest to it, as well as judge whether the source is close by or far away. However, the overall accuracy is nowhere near that of vision. Also, the spatial acuity of audition is easily influenced by factors such as the size and reflectiveness of the space in which sound propagates, while vision is much more stable in this

regard. Thus, while we can certainly speak about a precise, fixed point of audition in the sense of the nondiegetic, spatial position of the listener, there is no auditory analogue to camera positioning. This does not mean that we can in any way eliminate the notion of a point of audition, however – because vision is spatially dominant, 'point of auditon' becomes intrinsically linked with 'point of view'. Our perception of a more or less specific point (because of the inherent, spatial inexactness of sound perception, Chion equates it to a place, or zone) from which we appear to be hearing diegetic sound is almost entirely determined by the image.

While not an uncommon device, the first-person point of view is typically used sparingly in films. Meanwhile, many video games are played entirely from this perspective, and thus camera position, point of view and point of audition are constantly one and the same.

Modes of listening

In light of the above, it is hardly audacious to make the claim that what is ultimately heard largely depends on how one listens, in terms of both the implicit biases of human perception and concious analysis of the auditory scene. While we are naturally more receptive towards certain types of auditory stimuli – the sound of the human voice and its myriad nuances being one obvious example – we must also keep in mind that perception is at the same time influenced by our level of mental awareness and active participation. A distinction between *everyday*- and *musical* listening was proposed and explored by Gaver (1993). Employing purpose as distinguishing factor, Chion (1994) suggested three listening modes which he labelled *causal*-, *semantic*- and *reduced* listening respectively (which, as far as I can see, appear to have at least some aspects in common with Gaver's terms). Among these, causal listening is the 'default' mode and ties into the previously outlined survial agenda guiding human behaviour:

> Causal listening, the most common, consists of listening to a sound in order to gather information about its cause (or source). When the cause is visible, sound can provide supplementary information about it [...]. When we cannot see the sound's cause, sound can constitute our principal source of information about it. [...] We must take care not to overestimate the accuracy and potential of causal listening, its capacity to furnish sure, precise data solely on the basis of analyzing sound. In reality, causal listening is not only the most common but also the most easily influenced and deceptive mode of listening (Chion, 1994, pp. 25–26).

Semantic listening, on the other hand is mainly employed for the purpose of communication:

> I call semantic listening that which refers to a code or a language to interpret a message: spoken language, of course, as well as Morse and other such codes. This mode of listening, which functions in an extremely complex way, has been the ocject of lingustic research and has been the most widely studied. One crucial finding is that it is purely differential. A phoneme is listened to not strictly for its acoustical properties but as part of an entire system of oppositions and differences. Thus semantic listening often ignores considerable differences in pronunciation (hence in sound) if they are not *pertinent* differences in the language in question (Chion, 1994, p. 28).

Both causal and semantic listening rely on experience, albeit in different ways. Whereas the former utilizes our general schematic knowledge of causal relations – thus enabling us to make informed (although not necessarily correct) assumptions as to what the source of a sound might be, as well as to whether or not this source could represent a potential threat – the latter requires familiarity with the particular code statement of the message, a far more specific skill. Also, both causal and semantic listening are completely common and (usually) effortless – we all practice them every day, and often in combination with each other.

As for reduced listening, it distinguishes itself by rejecting the characteristics of both former categories, along with any other criteria not *directly* related to the sound signal itself:

> Pierre Schaeffer gave the name *reduced listening* to the listening mode that focuses on the traits of the sound itself, independent of its cause and of its meaning. [...] Reduced listening takes the sound – verbal, played on an instrument, noises, or whatever – as itself the object to be observed instead of as a vehicle for something else. A session of reduced listening is quite an instructive experience. Participants quickly realize that in speaking about sounds they shuttle constantly between a sound's actual content, its source, and its meaning. They find out that it is no mean task to speak about sounds in themselves, if the listener is forced to describe them independently of any cause, meaning or effect (Chion, 1994, p. 29).

In contrast to causal- and semantic listening, reduced listening does not occur naturally. It is rather an activity we must conciously decide to indulge in, requiring full concentration. As Chion notes, the term was originally coined by Schaeffer (1966) and, as we shall see, the challenges posed by this exercise are in many ways representative of the issues this chapter deals with. As an experiment, reduced listening illustrates how difficult it is to observe and describe sound in terms of what we actually hear (the proximal stimulus) as opposed to semantic descriptors or physical properties of the perceived sound source event (both of which pertain directly to the distal stimulus). In other words, it highlights the perceptual ramifications of the survival agenda, which we discussed above. Consulting empirical research on sound perception, such as the studies conducted by the *Sounding Object*-project (Rocchesso & Fontana, 2003), we find that the results of several experiments (e.g. Giordano, 2003; Grassi & Burro, 2003) point towards a perceptual bias in favour of the distal stimulus – participants repeatedly displayed great precision in judging various physical properties of the sound source event, such as its dimensions, shape, and type of material, while the parameters of the sound signal itself generally proved to be much less reliable predictors of test results. These conclusions align with Chions stated bias towards causal listening.

Implications of perceived movement in sound

As we have previously seen, the nature of sound is one of motion and, therefore, it cannot exist outside the boundaries of time. In language, the relationship of sound and motion manifests itself

quite clearly, particularly when it comes to music. Whether in casual or formal contexts, the various auditory events that comprise a musical piece are usually identified and referred to by the listener in terms of their perceived movement along some imaginary axis – whether horizontal (time), vertical (pitch), some combination thereof, etc.. When we consider this, along with the gestalt principle of common fate (i.e. objects that move together are likely related and therefore relevant), it should be clear that movement is an important factor in understanding visual and auditory perception and their influences on one another.

The study of movement in relation to music has recently gained momentum and may provide a suitable frame of reference. With the book *Musical Gestures* (Godøy & Leman, 2010), the effort was made to map this growing field and suggest a direction for future studies. Here, Jensenius et al. (2010) argued that the word 'gesture' is preferable to 'movement':

> The main reason for [...] this is that the notion of gesture somehow blurs the distinction between movement and meaning. *Movement* denotes physical displacement of an object in space, whereas *meaning* denotes the mental activation of an experience. The notion of gesture somehow covers both aspects and therefore bypasses the Cartesian divide between matter and mind. In that sense, the notion of gesture provides a tool that allows a more straightforward crossing of the traditional boundary between the physical and the mental world (p. 13)

On the flipside, «gesture» is a rather vague term with several possible connotations and interpretations. Thus, Jensenius et al. (2010) proposed the following, threefold framework for definition:

> (1) *Communication* is involved when gestures work as vehicles of meaning in social interaction. This use of the word is common in linguistics, behavioral psychology, and social anthropology. (2) *Control* is involved when gestures work as elements of a system, such as in the control of computational and interactive systems. This is common in the fields of human-computer interaction (HCI), computer music, and similar areas. (3) *Metaphor* is involved when gestures work as concepts that project physical movement, sound, or other types of perception to cultural topics. This use of the term is common in cognitive science, psychology, musicology, and other fields (p. 14)

For the purposes of this thesis, these three definitions of gesture may all prove relevant to some degree. However, at this point we shall focus on the third definition – metaphor – since this chapter is mainly concerned with the fields of research mentioned thereunder, and because I previously argued that metaphor is vital to understanding the relationship between vision and audition.

According to Jensenius et al. (2010, pp. 17–19), a recurring theme among researchers seems to be that the perceived movements of sound, and specifically music, are mental constructs analoguos to movement occurring in a physical space. An implication of such correlations between mind and matter is that one may instigate the other – such as when music triggers the urge to dance. This is what Godøy (2010), employing a term from ecological psychology , calls gestural *affordance*; «[...] meaning that people, dependent on their individual background, expertise,

particular situation or mood at any moment, may focus on different features in any single phenomenon of the world [...]» (p. 103).

The notion of gestural affordance might not seem immediately relevant here, considering that traditional screen-based audiovisual media typically do not invite physical participation – either because of social conventions (i.e. sitting down, shutting up, and paying attention), a mere lack of space in which to move, or both[10] – but this shows precisely why 'gesture' is the preferred term. As previously mentioned, the nature of its meaning is not a uniformly physical nor mental one. Rather, it covers both these aspects and may therefore prove a useful concept. The following excerpt from Godøy (2010) gives us an idea of how this use of the word 'gesture' relates to the cognitive basics of perception which we have so far covered in this chapter:

> Studying gestural affordances of musical sound is [...] about how listeners extract movement-inducing cues from streams of musical sound. But it is also the other way around, i.e. about how listeners use images of sound-related movement in making sense of what they hear. Thus, there is a two-way process here where sound induces images of movement, and conversely, where previously learned images of sound-related movement are projected onto sound [...]. The constant shift between perceiving and acting, or between listening and making (or only imagining) gestures, means that music perception is *embodied* in the sense that it is closely linked with bodily experience [...], and that music perception is *multimodal* in the sense that we perceive music with the help of both visual/kinematic images and effort/dynamics sensations, in addition to the "pure" musical sound [...] (pp. 104–106).

Citing a famous scene from Chaplin's *The Great Dictator*, Godøy (2010) demonstrated how the audiovisual technique commonly known as *mickeymousing* – so named because it was popularized through animated features, such as those associated with the characters of Walt Disney – is a rather striking example of gestural affordance in the realm of audiovisual media. It involves the deliberate choreographing of on-screen movement to an existing soundtrack, or conversely, designing of the soundtrack so that its salient features appears to match the perceived accent points and developments of the visual action. The result is that the physical movement of the image appears to temporally correlate more or less perfectly with the perceived movement of the musical score or other nondiegetic sounds. When executed in an overt fashion this technique tends to evoke humour, and is most commonly employed for comic effect. The case of mickeymousing demonstrates the power of synchresis, as well as the embodied aspect of perception. The notion of a perception founded in bodily experience must, of course, be viewed in relation to schemas, as Godøy (2010) points out in the following quote:

> The idea of gestural schemas emerging on the basis of a combination of various hard-wired audio-motor coupling and learned sound – movement associations can be seen as an instance of what is now often called *embodied* cognition. Common to different variants of this concept is the idea that our perception of

10 This is, of course, excluding the recent developments of the home based interactive entertainment scene based around motion controls for video games; where, in contrast to the traditional keyboard/mouse-, joystick-, and gamepad configurations, direct body movement (registered by various sensors) translates into similar on-screen movement. High profile examples include the Nintendo Wii, Microsoft Kinect and Playstation Move.

> the world, and our mental activity in general such as reasoning, imagining, planning, etc., is a process of incessant mental simulation of various body movements, both those made by other people and those made by ourselves, as well as both those we can see and those we can only assume (p. 108).

Furthermore, a set of basic gestural categories was proposed; the initial distinction being that of *sound-producing* gestures on one hand, and *sound-accompanying* gestures on the other. Roughly speaking, the former category is restricted to body movements directly involved in producing and modifying sound. The latter, on the other hand, refers to all other types of movements that a given sound may *afford* as it is heard, but that are not directly involved in its creation. Such gestures (which typically imply some form of synchronism between sound and movement) include dancing, marching, and – interestingly – air instrument playing, which basically imitates sound-producing gestures. Both of these main categories may, according to Godøy (2010) be divided into three, basic sub-categories based on how gestures develop over time:

> *Iterative*, meaning rapid repetition of small movements such as to fuse these into a single gesture, e.g. as in [...] the rapid bouncing movements of a drum roll [...]
> *Impulsive*, meaning discontinuous effort such as in hitting, kicking, rapid stroking or bowing [...]
> *Sustained*, meaning continuous effort such as in continuous bowing or blowing [...] (p. 111).

When audiovisual relations are broken down into such basic action categories they become useful descriptors. I make further use of these terms in the next chapter.

While on the topic of gestures and air instrument playing, I find it hard not to draw a parallel to so-called ‘rhythm games’, particularly the *Guitar Hero* and *Rock Band* franchises. These games represent an interesting hybrid between playing an actual instrument and imitating one. Their main gameplay mechanic consists of trying to synchronize controller inputs with pre-recorded musical tracks, with the aid of on-screen visual cues. Though not the first game series to use this mechanic, *Guitar Hero* (2005) made its mark by introducing dedicated controller peripherals resembling actual (although somewhat downsized) guitars that were sold alongside the games. Instead of the frets and strings of a real instrument, these peripherals have five colour-coded buttons placed on the neck along with a ‘strum bar’ on the body for inputting note pitches and onsets, and the player has to match both left and right hand movements to the audiovisual prompts. The basic task is to press the indicated button while hitting the strum bar at the precise moment when the onset-prompt aligns with the target indicator. As long as the player is successful in synchronizing his inputs, the music will continue to play and his score will rise exponentially due to a multiplier that automatically increases over time. Whenever the player misses an input, however, the music is momentarily interrupted and this multiplier resets.

There are several notable aspects here. For one thing, it is not entirely clear (at least not in my mind) if these inputs should be considered sound-producing gestures. On one hand they *are*

required for the sound to play correctly – if the pitch or timing is off, various pre-determined noises will play instead, highlighting your mistake. On the other hand, the feeling that you are in control of the sound is otherwise mostly an illusion. I say mostly because the peripheral also includes a 'whammy bar' by which the player may, in fact, at any time 'bend' the pitch of the note that is playing down by a whole diatonic step in real time. Still, the overall relationship between the music and player interaction is entirely governed by how the developers chose to map controller inputs to each individual song. For instance, pitch is entirely relative, each fret-button being used to indicate several different notes over the course of the same track. Also, on lower difficulty levels, many audible notes do not show up as visual prompts, and the player is actually penalized if he tries to play them.

The notion of sound as visual shapes

We shall now return to an issue I briefly touched upon earlier; namely that of the challenges posed by reduced listening. Bearing in mind the perspective of embodied cognition previously outlined, we should be better equipped to adress its main implication; the question of *why* it is so difficult to consider sound solely on its own terms. As mentioned, the exercise of reduced listening was first suggested by Schaeffer (1966). The quote below by Godøy (2010) outlines what it entails:

> Schaeffer's point of departure was to encourage the listener to disregard the source and everyday significations of any sound fragment, of what is called the sonic object, and to focus on the various perceived features of the sonic object [...] the point of departure would always be the seemingly simple question of "what do we hear now?" The strategy involved thinking of all these features as shapes, shapes that reflect basic action categories such as sustained, impulsive, iterative, flat, curved, steep, etc. [...] (p. 114).

While determined to eliminate the need for secondary descriptors such as cause, meaning, and effect, Schaeffer seemingly embraced the notion of a visual terminology applied to the auditory domain. Coining the term *objet sonore*, or *sound object*, he proposed that sound can be viewed as analogous to a physical object, i.e. a finite entity that by its shape is perceptually distinguishable from others. In light of the general argument presented in this chapter, it is hardly surprising that such a concession (were one inclined to see it as such) would eventually have to be made. It is yet another indicator of sensory integration. As human beings we cannot escape the perceptual limits of our body and mind, however, we *can* be aware of them. Empirically, it appears that the task of rendering sound features as physical shapes – sound-tracing (discussed earlier in this chapter) – is performed intuitively, with a significant degree of consensus among subjects. It must be noted, however, that results tend to become increasingly divergent as sound becomes more complex (Haga, 2008). Subjects produce similar renderings as long as the sound in question exhibits just one

or a few salient features, but when multiple features are present and the task of rendering becomes more difficult, individuals select different features to focus on.

One aspect subjects rarely, if ever, deviate from is the basic schema of a two-dimensional plane, where the y-axis represents pitch (implying gravity), and the x-axis represents the flow of time. This bears an obvious resemblance to most systems of musical notation; the difference being that such systems demand familiarity with, and adherence to, particular sets of rules. This eliminates most of the variables that produce divergences in sound-tracing experiments. Furthermore, by means of computers we may produce a variety of precise, visual renderings of sound features, ranging from simple waveforms to more complex models such as spectrograms, etc. Calculated visual representations such as these are frequently relied upon in both the production and study of music, as advanced software allows for the extraction and/or calculation of auditory features on a microscopic level.

Common to all these visual representations of sound is that they are temporal renderings, i.e. that they visualize the *movement* of the sound using the horizontal and vertical axes. However, intuitive relationships between certain phonetic patterns and nontemporal visual figures have also been demonstrated. The initial discovery was made by Köhler (1929, 1947). In this study, subjects were presented with two geometric shapes, one rounded, the other angular (fig. 1.1), and were asked to identify each one as either 'Maluma' or 'Takete'. Reports overwhelmingly established the rounded figure as Maluma and the angular as Takete. This, of course, strongly suggests that the relationship between words and the objects that they describe is not entirely arbitrary, which seems perfectly reasonable in light of metaphor and the notion of an embodied cognition.

Fig. 1.1: 'Maluma' and 'Takete'

2.3: Emotional perspectives on audiovisual perception

As of today, it is becoming a more common view among researchers that the study of emotion and its role in perception has not been given sufficient attention up to this point. Consequently this field is somewhat plagued by a lack of consensus on certain key issues. For one thing, researchers have

yet to come to terms with the question of what emotion actually *is*. Juslin (2009, p. 131) listed six components by which emotion may be measured – cognitive appraisal, subjective feeling, physiological response, expression, action tendency, and regulation – affirming that while these characteristics of emotion are generally agreed upon, a precise definition is not. A result of this is that the question of how emotion should be studied is also being debated. For instance, it has been demonstrated that we may interpret from the sensory stimulus a particular emotional meaning without actually experiencing ourselves the emotion in question. In other words, a *perceived* emotion is not necessarily an *induced* one. However, as pointed out by (Sloboda & Juslin, 2010, pp. 82–83), empirical studies thus far has not consistently taken into account this distinction.

Furthermore, although the validity of studying emotion as an aspect of perception is no longer in question, there is a backstory of longstanding reluctance dating back to the 17th century separation of reason and emotion often attributed to Descartes. It used to be the dominant view that emotions should be largely ignored in favour of reason, because they represent irrational behaviour, and because they would simply not submit to scientific research because of their supposed subjectivity and elusiveness. Such notions were only recently challenged in a broader sense, and although emotions are still largely treated separately from cognitive processes, they are now increasingly being recognized as an integral part of decision making; one that demands study for its own sake (Peretz, 2010, p. 100). Many would no doubt have found such a change in attitude long overdue. After all, emotions indisputably impact human behaviour in significant ways. Based on the evolutionary perspective presented at the beginning of this chapter, we can safely assume that there is a reason for this.

Emotion, music and narrative multimedia

When it comes to music and its universal presence across cultures, it is a common conception among researchers that emotion is a key factor as to why our species has developed musical tendencies (Justus & Bharucha, 2002). As for how emotions manifest themselves in music, Juslin & Västfjäll (2008) and Juslin et al. (2010) proposed a model based on seven distinct mechanisms: Brain stem reflexes, Rhytmic entrainment, Evaluative conditioning, Contagion, Visual imagery, Episodic memory and Musical expectancy. Although I will not get into further detail here, it should be relatively clear that these mechanisms relate to the cognitive aspects of perception previously accounted for in this chapter.

When it comes to sound perception, music in particular is an area that may benefit from a better understanding of emotion. Furthermore, the emotional aspect may appear particularly

pertinent as music takes on the role of narrative device in film and its related media. As we shall see, it appears that music's perceived narrative value pertains to both its affective qualities and its structural aspects. Although audiences are generally more conciously aware of the effects of the former, the significance of the latter must not be underestimated. After all, having an emotional reaction to something comes from interpreting its meaning, which, in turn, is accomplished by recognizing patterns. In other words, emotional meaning is not a separate entity from the structural aspects of music.

As we have seen, a precise definition of emotion is lacking. And although certain characteristics are agreed upon, not all of these are necessarily present in music. Review studies on emotion in music (e.g. Gabrielsson & Juslin, 2003; Juslin & Laukka, 2003) suggest that some emotions are more prevalent than others. Largely on the basis that they appear to be universal across cultures, certain types of emotions are commonly referred to as 'basic' among researchers. The general contents of this term is explained by Peretz (2010) in the following quotation:

> 'Basic emotions' refer to emotions like happiness, sadness, anger, and fear. Such basic emotions are today the focus of neuropsychological studies, because these emotions are assumed to be innate, reflex-like circuits that cause a distinct and recognizable behavioural and physiological patterns [...]. Although basic emotions may differ from what most adults experience when listening to music [...], many researchers believe that music can induce happiness, sadness, and fear. These basic emotions are typically the target of film soundtracks, especially those intended for children. Moreover, these basic emotions are typically the easiest to recognize in music [...] (p. 101)

Although there are certainly differing views on the question of exactly which affective phenomena constitute these basic emotions, this descriptor is generally indicative of a view in which emotions are considered to be discrete values that are fundamentally different from one another, as opposed to there being just a single affective entity which may vary in intensity (Ekman, 1999). However, I do not intend to engage in any debate on this topic, nor shall I concern myself with the question of how specific musical features are connected with specific emotions, except for pointing out the obvious: when it comes to our emotional responses to a narrative and their perceived enhancement (or, for that matter, attenuation) on account of music, we are very likely dealing with a nonarbitrary multimodal relationship.

Emotion and the 'paradox' of nondiegetic music

Regardless of medium, believability is a critical element of successful storytelling. On some level, a story must be able to connect with its audience, which effectively means that it must facilitate emotional investment in the diegetic universe it portrays. From the practical viewpoint of the author, this is largely a matter of infusing said universe with a sufficient amount of familiar aspects

and qualities that people can recognize and appreciate in terms of their own life experiences. It does not matter if this universe is an 'augmented' one, in the sense that it contains impossible or fantastical elements, as long as an internal set of rules and sense of reality is established that the audience can understand and feel invested in.

When it comes to storytelling in audiovisual media, it is a common conception that the general purpose of music is to strengthen the emotional impact of the images, and thus to contribute to the audience's sense of immersion. However, by acknowledging that music has become an intrinsic part of the audiovisual language of film and related media, we are immediately confronted by the paradox that Cohen (2010) outlines in the following excerpt:

> Film theory commomly refers to the fictional, imagined narrative world of the film as the *diegesis*. In contrast, the *non-diegesis* refers to the objective world of the audience, the world of artefact, of film screens, projectors, proficiency of actors, and technical aspects of the film. In terms of physical reality, music as acoustic vibrations belongs to the non-diegesis. Logically – unless such sound were part of a scene portrayed in a film, as in a film about a musical instrument, or the life of a great composer – these sounds of music should *detract from* rather than *add to* the sense of reality of the film (p. 884).

This apparent contradiction is further emphasized when we consider that not only are audiences willing to accept the presence of nondiegetic music, they also tend towards negative responses when it is absent (Cohen, 2009, p. 442). Whether by convention or some other factor, it seems to be a widely accepted notion that certain narrative contexts simply *require* music. Taking into account all that which we have so far covered in this chapter, it is tempting to argue that film scoring works simply because we have learned to 'ignore' the physical separation between our reality and the diegetic universe; that the concept of nondiegetic music has become part of schematic experience.

This seems no less likely when we consider the historical evidence. According to Larsen (2007, pp. 83–85), the notion of nondiegetic film music was largely rejected by the industry upon the initial transition into sound film that took place during the late 1920's. In keeping with the newfound realism that sound afforded the narrative, film music during the early years of the sound film era was often strictly diegetic, and thus quite scarce (Larsen, 2007, pp. 83–85). At the same time, this period was also characterized by experimentation, and as successful composers increasingly ventured into nondiegetic territory, the initial position that music needed diegetic justification was gradually abandoned. Over the next decade most aspects of this new audiovisual style of narrative, including musical practices, eventually became more or less standardized. These standards largely survive to this day (Larsen, 2007, pp. 85–86).

Already prior to considering cultural and social factors, it seems evident that our fundamental perceptual limits and predispositions play a key part in explaining how such

conventions were formed. As we have previously seen, our brains process information based on what might be called a system of priorities, which in turn are informed by certain subconcious, hard-wired directives. As a result of this, we have a natural preference for that which makes sense, as well as for discarding that which does not. Occasionally, this causes us to misinterpret information or overlook it entirely. There are strong indications that the paradox presented by nondiegetic music is ultimately attributable to these inherent biases of perception. Take the following statement by Cohen (2010):

> [...] the audience selectively attends to only the part of the music that makes sense with the narrative. Selective attention is a common perceptual-cognitive operation. The phenomenon of 'inattentional blindness' is an example of it in the visual domain of film. Here it has been shown that people rarely noticed or were seldom distracted by impossible visual aspects represented in either a film or in their real-world experience (p. 884).

Based on her interpretation of this term, first coined by Mack & Rock (1998), Cohen proposed that a similar phenomenon affect audition, dubbing it *inattentional deafness*. As should be clear by now, selective attention is not random – when perception emphasizes certain elements at the expense of others, it does so because these particular elements are, for some reason, considered more important. So, by what criteria does selective attention operate in the case of film music? Taking into account that film music is experienced within a safe setting where we are under no threat from the environment, one could argue that the survival-based, evolutionary viewpoint established at the beginning of this chapter might not seem like a particularly well-suited approach to this question. A more pertinent phrasing might therefore be: in what way is musical information at all relevant to the visual action of a film? We have previously discussed the multimodal relationship between auditory and visual perception and their mutual influence on one another in the context of audiovisual media, which translates to what Chion (1994) called *added value*. It appears that the primary way in which music adds value to film is through emotion. According to Cohen (2010), «emotion characterizes the primary experience of both music and film» (p. 901). She further claims that emotion is a recurring factor in all but one among eight distinct functions of film music, which are as follows:

> First, music masks extraneous noises. Second, it provides continuity between shots [...].Third, [...] it directs attention to important features of the screen through structural or associationist congruence. Fourth, it induces emotion [...]. Fifth, it communicates meaning and furthers the narrative, especially in ambiguous situations [...]. Sixth, through association in memory, music becomes integrated with the film [...], and enables the symbolization of past and future events through the technique of leitmotif. [...] Seventh, music heightens absorption in film, perhaps by augmenting arousal, and increasing attention to the entire film context and inattention to everything else [...]. Finally, music as an art form adds to the aesthetic effect of the film (Cohen, 2010, p. 891).

The first function listed is the sole exception mentioned above, while the third function must be seen in conjunction with an underlying cognitive framework called the *Congruence Associationist*

Model. When first proposed (Marshall & Cohen, 1988), it was geared towards animated films, and showed that audiences would attribute specific personality traits to different types of simple, geometric figures when the perceived structural and/or affective features of the accompanying music appeared to match the features of the visual shape. This was referred to as *cross-modal congruence*. Later iterations (Cohen, 2010) expanded the model for use with live action films as well.

There appears to be an obvious parallell between this model and the nonarbitrary pairings of certain phonetic patterns with geometric shapes (i.e. the maluma/takete-phenomenon). Furthermore, the idea of perceived audiovisual congruence based on shared structural and affective qualities seems consistent with the previously discussed research related to gesture, i.e. that certain sound features more or less intuitively translate into visual renderings of auditory movement. Lastly, we should also note how everyday language tends to apply the 'embodied' concept of pitch (which is commonly referred to as an image-schema) to the process of describing our emotions: words like 'up', 'high', or 'rising' usually denote positive aspects of ones emotional state, while 'down', 'low', or 'falling' refer to the negative end of this spectrum.

Although it is quite possible to make a logical argument that explains why nondiegetic film music has become so prevalent and accepted (as I have attempted to do throughout this chapter), we cannot completely ignore the historical fact that it used to be considered as inherently foreign to the narrative. This notion did not disappear, even as nondiegetic music established itself as a conventional feature of the film medium. Granted, the tables appear to have turned – the decision not to use this type of music is perhaps most often regarded as a stylistic choice, or a deliberate effort by an author to defy convention. Because audiences tend to take notice of it, the *lack* of music can take on a function of its own, e.g as a signature element. Nevertheless, it is still also strongly associated with the kind of storytelling that emphasizes a strong and immediate sense of realism.

In the TV mini-series *Generation Kill* (2008), the authors decided against using any form of nondiegetic music – even the title screen and closing credits seen in every episode are accompanied only by distorted radio chatter. Although fictionalized, the narrative is closely based on the personal account of a journalist who travelled with a squad of American soldiers during the invasion of Iraq in 2003, and is widely considered to be a very authentic and accurate portrayal of the U.S. military. It is certainly tempting to suggest that the decision to exclude nondiegetic music came from the assumption that it would somehow clash with the realism of the narrative. In light of this, I should specify that *realism*, as a perceived quality, need not necessarily equate *reality*. Whereas the former pertains to our judgements about the likelihood and/or accuracy of the events of a narrative when

compared to the real world, the latter could be regarded as an assessment of the internal consistency and coherence of the narrative universe, which ultimately is the more important factor when it comes to the audience's experience. We can safely assume that there is no directly proportional relationship between our level of immersion in a narrative and its perceived realism – generally speaking, we frequently find ourselves more absorbed by fiction than fact. Thus, one could perhaps argue that nondiegetic music might not be so paradoxical after all.

Chapter 3: Classification of sounds and images

In this chapter I examine audiovisual narrativity by presenting a system of sound categories based on three layers of definiton. These are:

1) *Spatial position*, i.e. the perceived point or zone inside or outside the narrative space where the sound source appears to be located, as well as the perspective from which we appear to be hearing the sound.
2) Sound type, i.e. which main category the sound belongs to based on its features, spatial position, and narrative context. I use intuitive descriptors of auditory phenomena such as speech and music as point of departure, and supplement these as needed.
3) *Narrative function*, i.e. the meaning or value that sound adds to the images. I previously distinguished textual and emotional meaning, both of which will be featured here.

This will be my starting point. Although the particular sequence in which the layers are introduced is not randomly chosen, I should note right away that they should not be viewed as hierarchical or territorial – on the contrary, they are by nature interconnected and overlapping. Also, though they may appear too static to properly reflect the dynamic nature of sound-image relations in the evolving narrative context, they should prove a useful foundation from which to define further descriptors that can be used to analyze sound at specific points in time within the audiovisual sequence. We shall proceed to look more closely at each of these layers in turn.

3.1: Spatial position

Diegetic or nondiegetic?

In the context of audiovisual narrative media, sound falls into one of two overall categories: diegetic and nondiegetic. The distinction between the two is a matter of whether or not there appears to be a meaningful, causal connection to the immediate spatiotemporal reality of the narrative. Previously, I used characters awareness to illustrate the difference, but this is clearly insufficient. After all, diegetic environments typically contain many sounds of which the characters never demonstrate awareness. As for direct synchronism with a visual event, this may seem like an immediate and compelling indicator, but it is no requirement, nor does it offer any guarantee that the sound pertains to the diegesis – the case of mickeymousing demonstrably illustrates this. It does not necessarily matter whether or not the perceived source of a given sound is visible in the image (or whether the

sound was recorded live or added in post-production[11]), as long as this source would generally be able to claim a 'natural' place within the particular environment of the scene in question. In any case, it is evident that there are several factors at play here which we will have to identify.

Explicit and implicit sources

In terms of listening we are causally oriented by default, i.e. we actively search for and 'locate' sound sources in three-dimensional space. Similar to real life, this natural conditioning applies to the imagined space of the diegesis as well. Any perceived connection between a sound and its source likely results from both direct observation and schematic knowledge. I discern four basic scenarios, the first two of which pertain to diegetic sound:

1) The source is *explicit*, i.e. an onscreen character or object. Chion (1994) prefers the term *visualized* sound. When there is movement, synchresis will typically be in effect.

2) In the narrative context, an *implicit* connection to an offscreen source is established in one of two ways: either a) we recognize the sound as that of a previously visible source that is now hidden from view, or b) based on experience we attribute to it a likely source that may be general or specific; the level of precision typically depending on the nature of the sound. Chion calls this *acousmatic* sound (I will return to this concept shortly).

Similarly, nondiegetic sounds may be determined as such on the basis that

3) there is an explicit (obvious) contradiction between sound and image in terms of the perceived sound source (e.g. symphonic music accompanying a mountaintop vista), or

4) there is no sufficiently clear indication of a diegetic source, which implicitly defines the sound as nondiegetic.

I should note an important difference between diegetic and nondiegetic sound: because the diegesis represents a concrete spatiotemporal world, the images inevitably impose upon diegetic sounds the sense of a (more or less) fixed spatial position (see chapter 2). This notion of an acoustic space does not apply in the same way to nondiegetic sounds. Being external to the narrative universe, such sounds tend to ignore spatial concerns; i.e. they do not appear to come from anywhere in particular. It is important to be aware that the information from which we distinguish the abstract nondiegetic 'plane' from the concrete diegetic space is not inherently contained in the sound itself – the illusion

11 As is today common knownledge, a rather large part of the sounds that for all intents and purposes appear to be part of a film's setting are routinely recorded separately by sound designers and synchronized with the images during post-production. Even 'real' background noise recorded on location is carefully mastered, so as to help define the space without interfering with dialogue, etc..

of sound situated in space is created by the images. On the basis of the sound signal alone, diegetic and nondiegetic sound cannot be told apart.

Foreground and background

It may or may not prove useful to think of the diegetic space in terms of a foreground and a background. The advantage of these terms are their intuitiveness. The drawback is that they are very unspecific. Still, I find it counterproductive to eliminate them entirely from the equation. Although it might seem contradictory given my previous statement about the diegesis being spatially abstract, I find the notion of a foreground and background potentially useful for nondiegetic sound as well, specifically in terms of audience awareness. When it comes to nondiegetic music, there is absolutely no doubt that our awareness level is dynamic; i.e. that the music becomes more or less pronounced according to the developing relationship between sound and image.

Internal and on-the-air sound

Under certain conditions, audiences can be made privy to the particular auditory perspective of a specific character, as if our point of audition becomes that of this character. In other words, it appears that we hear what the character hears, either in terms of 1) the physiological sounds generated by a) the environment, and b) his or her own body (e.g. breathing, heartbeats, etc.), or 2) an 'inner voice' that represents the character's thoughts. This is what Chion refers to as *internal* sound (1994, p. 76). In the first case the internal sound is *objective*, i.e. presumably audible to any other character in the action that is situated sufficiently close to the one whose point of audition we currently share. In the second case sound is used to represent an entirely inaudible, mental process and therefore *subjective*.

Just to be clear, the heard 'thoughts' of a character in the present action represents a different scenario from when a character narrates his past self from an external, nondiegetic position. It must also be noted that the impression of internal sound, whether objective or subjective, does not necessarily require that we also assume this character's point of view. Although the impression of internal sound is typically associated with a closeup shot of the character in question, it it also partly a result of the sound mix – e.g. when a sound that would normally not be heard by others (such as that of heartbeats) comes to the forefront.

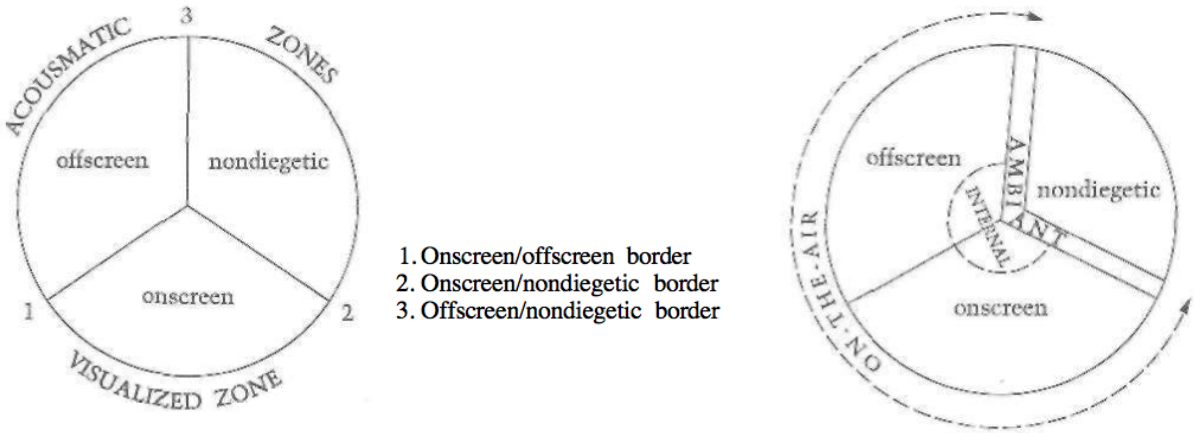
As we can see, identifying sources is oftentimes far from straightforward. Diegetic sound that is supposedly transmitted electronically, such as a radio transmission, phone call, or announcement made over a PA system further illustrates this – whether explicit or implicit, the

device that the sound is perceived as emanating from is not what initially produced it. Normally, not seeing a sound's originating cause would fit the description of an acousmatic situation. However, for cases where sound more or less clearly implies what we might call an indirect source, (Chion, 1994, pp. 76–78) devised the term *on-the-air* sound. We typically recognize it as such in films based on the subtle or obvious distortion of the sound signal typical of electronic amplification devices that do not emphasize fidelity.

What complicates matters is that on-the-air sound sometimes appear to propagate naturally within the diegetic environment, while at other times, it appears to ignore such physical laws, i.e. when the voice of a radio announcer is heard clearly as the camera pans slowly across a cityscape. Although Chion explicates many of these nuances, I shall not pursue these further here, except for one particular intance: when a character is using a sound receiver device that is situated very close to his or her ear (e.g. a phone, radio headset, earpiece, etc.) and we hear the received sound loud and clear, we have a case where sound can be interpreted as simultaneously internal and on-the-air – we assume the character's point of audition in order to listen in on the electronically transmitted sound that we would otherwise not be able to hear.

Chion (1994) illustrates the relationship between the zones with the tripartite figures seen below – the leftmost figure showing a basic version, while the rightmost includes the internal, ambient, and on-the-air zones:

Fig. 3.2: Auditory zones (Chion, 1994, pp. 74–78)



3.2: Sound type

Different sounds have different qualities. Certain sonic features affects perception in specific ways; instilling awareness, etc., and, to a greater or lesser extent, come pre-imbued with meaning. This

meaning is further influenced by the narrative context. It is therefore counterproductive, I find, to attempt to exclude semantic descriptors in favour of strictly sonic properties when determining *type*. Although the question of *function* will be discussed in more detail later, it will inevitably impose itself upon the current matter also. Furthermore, certain combinations of sound features are so intuitively recognizable that they seems to constitute categories by default, as is the case with speech and music. I see no reason to problematize these and will use them as starting point.

For the purpose of identifying useful categories for other types of sound, I shall first establish *distinct* and *indistinct* as a conceptual pair. I use the former to describe individual sounds that, by some feature or other, are immediately distinguishable from other sounds heard at the same time. Because an auditory environment is typically dynamic, distinctness is largely a contextual quality, although some sounds are almost universally recognizable. Whether by amplitude, frequency content, pitch contour or some other parameter, distinct sounds stand out to form separate and defined 'objects'. Conversely, indistinct sounds display a tendency to blend with other sounds and seem to lack definition, i.e. they possess no features that demand attention. I discussed some of these issues in more detail in the previous chapter.

In the case of most diegetic sounds, it is useful to think of them as having either an *active* or *passive* role.[12] Active sounds participate in the immediate action, courtesy of being produced by or acted upon by characters, or by raising questions as to their origin. Passive sounds, on the other hand, are not involved in the action but merely present as part of the setting. Active sounds naturally draw attention to themselves in a way passive sounds do not.

I use *action sounds* as a broad category for all distinct, non-speaking sounds with explicit or implicit sources (onscreen or offscreen), that somehow pertain to the narrative foreground. This includes the sounds that characters produce as they move around and interact with other characters and objects in the action (e.g. footsteps, clothing, moving doors, etc.). Here we can make a further distinction between 1) *active* action sounds; i.e. sounds that are produced with a clear communicative purpose or intention, such as when someone rings a doorbell or honks a car horn, and 2) *passive action* sounds; i.e. sounds that have no distinguishable purpose in and of themselves and thus can be regarded as mere byproducts of an action. I must note that when I speak of a sound not having a specific *purpose* from the viewpoint of the initiating character, this does not necessarily exclude it from having a narrative *function* from the viewpoint of an author or audience member.

Conversely, *ambient sound* denotes the indistinctive blend of passive sounds that typically

12 Chion uses these terms primarily for offscreen sounds (1994, pp. 85–86). I prefer to expand their use somewhat.

form the background of the auditory environment, and that is largely responsible for creating the sense of space. For explicit or implicit sounds that come across as perfectly distinct but have no active involvement in the action, I shall adopt Chion's term *elements of auditory setting*:

> I call elements of auditory setting (E.A.S) sounds with a more or less punctual source, which appear more or less intermittently an which help to create and define a film's space by means of specific, distinct small touches. Typical sounds of the auditory setting are the faraway barking of a dog, or the ringing of a phone in the office next door, or a police car siren (Chion, 1994, p. 55).

If the need should arise to further specify sounds by drawing directly from their features, we may draw upon the three basic action-categories discussed in chapter 2; i.e. continuous, impulsive, and iterative.

Acousmatics and acousmêtres

A heard sound with no visible source may also be referred to as an *acousmatic*[13]. Once its source is revealed we can speak of a *de-acousmatized* sound (Chion, 1994, pp. 71–72). Sometimes an 'unseen' sound is so ambiguous or unfamiliar that it holds little or no associative value. With no visual indicators as to its diegetic status (e.g. character awareness) we have no choice but to suspend definition. Usually, however, an acousmatic will have an implicit source courtesy of the sound signal, placing it either inside or outside the narrative. In the case of any vocal phenomenon we automatically identify the source as a person. We may not immediately know whether the voice is nondiegetic or merely offscreen for the time being, but this will usually become clear as the situation progresses. For instance, past tense speech indicates an external narrator, while present tense speech and character awareness indicates a physical presence in the action.

When acousmatic speech occurs systematically throughout a narrative, so as to conceal the visual identity of the character that the voice belongs to, we are dealing with a so-called *acousmêtre* (Chion, 1994, pp. 129–131). In such instances, the disembodied voice becomes a character trait that imbues said character with an almost supernatural presence. Its source may be hinted at, as if behind an object that the frame focuses on (e.g. a curtain), or even partially visible in certain shots (closeup of shoes, hands, etc.). However, with no face to it the voice represents a mystery that begs for resolution. Once this happens though, its supernatural power typically ends, since this power is intimately tied to the mystery.

From a spatial perspective the interesting aspect of acousmêtres is this: while the narrative establishes them as present in the action (character awareness) and capable of influencing it, their general lack of visibility grants these voices the special privilege of having no *exact* source position

13 Chion appears to use this word both as a noun and as an adjective. It was originally theorized by Schaeffer.

inside or outside the frame. Because of this they may or may not take on a status as all-seeing, omniscient, omnipotent and ubiquitous (Chion, 1994, pp. 129–130). From a narrative viewpoint acousmêtres represent the simplest of mechanisms: present a question, then suspend the answer in order to create tension. A famous example is the classic James Bond villain Ernst Stavro Blofeld who appears exclusively as an acousmêtre in both *From Russia with Love* (1963) and *Thunderball* (1965): his body is partially seen as he pets his trademark white cat, but his face is never revealed. A more recent example is the character of Professor James Moriarty in *Sherlock Holmes* (2009).

3.3: General narrative functions of sound

Temporalization

Chion (1994) describes three ways in which sound can «temporalize» the image:

> The first is temporal animation of the image. To varying degrees, sound renders the perception of time in the image as exact, detailed, immediate, concrete – or vague, fluctuating, broad. Second, sound endows shots with temporal linearization. In silent cinema, shots do not always indicate temporal succession, wherein what happens in shot B would necessarily follow what is shown in shot A. But synchronous sound does impose a sense of succession. Third, sound vectorizes or dramatizes shots, orienting them toward a future, a goal, and creation of a feeling of imminence and expectation. The shot is going somewhere and it is oriented in time (Chion, 1994, pp. 13–14).

Temporal animation largely translates as immediacy of attention. «[...] exact, detailed, immediate, concrete» refers to the here-and-now, implying a high degree of moment-to-moment attentativeness, whereas «[...] vague, fluctuating, broad» indicates a state wherein time becomes nonspecific and our perceptual focus is drawn away from the moment-to-moment visual action (or lack thereof); potentially in favour of some other narrative element (more on this shortly).

Linearization refers to instances where audio creates the impression that successive shots appear in chronological order, as opposed to being randomly organized or showing simultaneous events.

Vectorization refers to instances where audio effectively encodes a shot onto a one-directional timeline (usually visualized as left-to-right). Contrary to the vast majority of sounds, visual movements are for the most part ambiguous in terms of temporal orientation; i.e. they could just as easily be viewed in reverse without appearing much different. Only certain types of movements or events (e.g. walking, falling, a car accident, etc.) will automatically appear unnatural or downright impossible if their temporal orientation changes. Sounds, on the other hand, are (but for a few exceptions) naturally vectorized – they tend to change rather dramatically if we flip them around. Thus, when a shot has only ambiguous motion or no motion at all, audio will typically

create a temporal orientation by itself. Chion (1994) identifies the following two scenarios:

> First case: *the image has no temporal animation or vectorization in itself*. This is the case for a static shot, or one whose movement consists only of a general fluctuating, with no indication of possible resolution - for example, rippling water. In this instance, sound can bring the image into a temporality that it introduces entirely on its own.
>
> Second case: *the image itself has temporal animation* (movement of characters or objects, movement of smoke or light, mobile framing). Here, sound´s temporality *combines* with the temporality already present in the image. The two may move in concert or slightly at odds with each other, in the same manner as two instruments playing simultaneously (p. 14).

Additionally, the extent to which temporalization will occur depends on the type of sound. Briefly paraphrased, the most relevant conditions are: 1) *sustainment*, i.e. the manner in which a sound is maintained over time. An iterative sound will cause greater temporal animation (immediacy of attention) than a continuous one. 2) *predictability*, i.e. a sound's regularity, or lack thereof, as it progresses. An irregular pulse or rhythm generally causes greater temporal animation than a regular one. However, overly mechanical regularity also increases alertness, making us anticipate a sudden change. 3) *tempo*, i.e. the speed at which iterations of a sound occurs; particularly relevant when it comes to music. As the previous condition suggests, a rapid series of notes with regular intervals creates less temporal animation than irregular intervals at moderate speed. 4) *definition*, i.e. a sound's frequency content. High frequencies focuses attention more than low frequencies.

Unification

The one function that sound is most frequently called upon to fill in narrative multimedia is that of unifying the images by masking the inherently fragmented and discontinuous nature of a series of shots. As we have seen, the shot has both the advantage and disadvantage of representing a definite unit – regardless of the nature of the images contained within it, its borders are clearly defined. It may appear that the perceived visual continuity of a film establishes itself more or less automatically, when in reality this could not be farther from the truth. Because each shot represents an individual unit, great care must be taken to achieve a sense of coherence, and even then, the images alone may not be sufficient. Granted, techniques for establishing visual continuity are considered fundamental knowledge for every filmmaker and are by no means an exclusively auditory concern. However, in the auditory domain there is no all-purpose unit capable of 'containing' the sound signal the way the shot inevitably does for images. Thus, it is often far easier to accomplish with sound that which sometimes requires great effort with images.

The basic mechanism of unification is that of temporally overlapping visual cuts using either diegetic or nondiegetic sound. This could mean ambience, music, dialogue, or most any other type of sound, depending on the context. In the case of diegetic sounds there is also a spatial concern, in

that they define – to a greater or lesser extent – the narrative space outside of the frame. In such cases, the unifying effect ties as much into our perception of this surrounding environment created exclusively by sound. Typically, continuous sound (see chapter 2) display the strongest tendency for unifying images. However, impulsive and iterative sounds can also contribute towards visual unity, particularly when they occur as E.A.S.. Chion notes in his analysis of the prologue to *Persona* (1966) how the repeated offscreen sound of some dripping liquid (assumed to be water) creates continuity between shots of dead people in a morgue and a sleeping boy (in this case a 'false' continuity, as the boy is later shown to be alive and presumably located elsewhere).

The dripping sounds represent an interesting case in terms of the sound-producing event. The immediate association will likely be that of a faucet not fully closed, or a leaky pipe. In any case, our schematic knowledge of how liquid material behaves tells us that a drop – i.e. a small amount of liquid material that is held together by surface tension – is likely to be succeded by more similar events, because gravity is constantly in effect. Although each impact can be viewed as an isolated and impulsive sound-event, the initiating process by which water flows, accumulates into drops, and subsequently falls is a continuous one. Thus, in a sense it could also be viewed as a series of iterative events.

Anticipation

In real life, our understanding of current affairs is continuously influencing, as well as being influenced by, the prospect of what is to come. The nature of perception dictates that we never seize hypothesizing and trying to predict future developments. We construct such hypotheses from perceived sensory stimuli and schematic knowledge. This, of course, is also the case when it comes to our perception of an audiovisual narrative, where the individual auditory and visual components we encounter have the potential to trigger this natural inclination to a greater or lesser extent. Chion (1994):

> From a horizontal perspective sounds and images are not lined up like fenceboards in a row. They have tendencies, they indicate directions, they follow patterns of change and repetition that create in the spectator a sense of hope, expectation, and plenitude to be broken or emptiness to be filled. This effect is best known in connection with music. Musical form leads the listener to expect cadences; the listener´s anticipation of the cadence comes to subtend his/her perception. Likewise, a camera movement, a sound rhythm, or a change in an actor´s behaviour can put the spectator in a state of anticipation. What follows either confirms or surprises the expectations established-and thus an audiovisual sequence functions according to this dynamic of anticipation and outcome (1994, p. 55).

In the overall context of this thesis, the above quote represents an obvious throwback to chapter 2, where I basically argued that perception is a continuous process of meaning-making: the interpreting of past and present events enables us to anticipate and prepare for future events.

Punctuation

If we apply to a sequence of sounds and images the notion of an audiovisual syntax – i.e. that its construction and internal logic can be viewed as analogous to that of a linguistic sentence – nonvocal sounds can take on a punctuative role, standing in for the commas, periods, exclamation marks, and so on, that modulate or define the rhythm and meaning of a written or spoken text (Chion, 1994, pp. 48–49). Because film is a voco- and verbocentric medium (Chion, 1994, pp. 5–6) and most scenes feature speech in some form or other, we can usually speak about this punctuative function of nonvocal sound as an extension of the textual punctuation that actors apply to their spoken lines as per the instructions of the script and/or the director. Alternatively, or simultaneously, it can relate to the visual flow of the scene; e.g. the sequencing of shots, movements of the camera or elements in the frame, etc.. Deliberate placement of action sounds, E.A.S. or musical elements is a common way of emphasizing words or gestures, of spacing, and of 'bookending' shots or scenes by signalling their starting- and/or ending point.

The punctuative effect of nonvocal sounds can be subtle, such as when discreet background noises fill the auditory spaces created by pauses in the dialogue, or overt, such as when a gun-toting character speaks a threatening line and procedes to cock the hammer of his weapon. A common use of musical punctuation can be found in many martial arts films. Typically, during those tense moments where two soon-to-be combatants stare each other down, the silence is broken by a single instrument, percussive or otherwise. Mickeymousing and similar forms of music-image synchronism are other examples of punctuation.

As (Chion, 1994, pp. 48–49) notes, synchronous sound allowed for much more subtle ways of punctuating. Expanding the array of punctuative options beyond that which directly relates to the text has the potential benefit of making the narrative appear more fluent and natural, because it grants the author and production crew more flexibility in terms of the composition, shooting and editing of a scene. Also, actors are relieved of having to pause and articulate in unintuitive or exaggerated ways in order to make their character's intentions clear, something that might have otherwise hindered their performance.

Separation

From a syntactic viewpoint, one might argue that silence is the auditory equivalent of the spaces and gaps that separate the words, sections and chapters of a text. Although a somewhat clumsy analogy, I find that it is not entirely without merit – after all, pauses are as vital to the rhythm and structure of speech and music, as separation is to a text. In an audiovisual sequence, the ceasing of

sound can take on a similar function.

I refer to silence as an *auditory* device because its meaning comes from it being perceived as the negative of sound – if there is no sound to begin with, silence has no meaning. Also, the word *silence* should not be regarded as synonomous with a complete absence of sound, since this is rarely the case in narrative multimedia, nor elsewhere – in the real world everything that moves produces sound on some level. Consequently, no environment in which we regularly find ourselves is ever fully silent. In fact, complete silence would imply stasis, which is so unnatural to us that we tend to take instant notice. Because sound is largely responsible for creating our sense of a scene's physical space, some auditory presence is typically required in order to maintain the illusion of a real environment. When there is no dialogue, audiences will focus their attention on other auditory elements. Thus, managing the background noise or ambience of any given setting is an important aspect of modern sound design in films.

In the same way that the various sounds generated by an environment help us define it, the absence or sudden disappearance of certain types of sounds are interpreted to mean more or less specific things. The uncomfortable sensation that we may experience when an environment is 'too quiet' can likely be attributed to the hard-wired survival agenda that guides perception – in nature, the abrupt silencing of common animal sounds (i.e. birds, insects, etc.) could indicate the presence of a potential danger, whether in the form of a human with hostile intentions or some predatory animal. Incidentally, this type of 'menacing' silence is also commonly seen in films, a famous example being *Once Upon A Time In The West* (1968). Here, in an early scene, the characteristic and continuous sound of crickets in the background is conspicuously interrupted shortly before a family is gunned down by bandits hiding in the bushes around their home.

In the same scene, a rather loud, howling wind continues to be heard after the crickets fall quiet. This demonstrates how the impression of silence is typically a product of contrast, i.e. it is contextually determined by what came before (Chion, 1994, p. 57). It is also common to indicate silence by focusing on a diegetic action sound or E.A.S. (e.g. breathing, the ticking of a clock, the distant hum of traffic, etc.) that would normally not be heard in the typical auditory environment associated with a setting. When such sounds default to the foreground because there are no other auditory elements to compete with, it tends to evoke a feeling of emptiness (e.g. the difference between nighttime and daytime in a city).

Guiding and spotting

In chapter 2 we looked at how vision and audition differ from one another when it comes to

temporal and spatial acuity. In cinema, a consequence of vision's slower perception speed is that sound becomes all the more important as the visual action speeds up. Because auditory perception is significantly faster, filmmakers routinely and deliberately rely on audio cues in order to *guide* audience perception. The visually dense sequences characteristic of modern day action films are often comprised of many brief shots, with shifting camera angles, multiple characters and objects on screen, rapid movements, and so forth. Additionally, the audience do not get to influence the point of view from which they see the action. Instead, they must rely on the director to show them what they need to see. Naturally, adequate framing of each shot goes a long way, but sound often plays an equally important part in highlighting the essential elements of a shot.

As we shall see, the guiding function is commonly associated with video game sound, where its purpose typically is to indicate possible actions that the player may choose to perform. This function many types of sounds, including music. An example of this is found in *The Elder Scolls V: Skyrim* (2011). This is a role playing game set in an enormous, open, fantasy world in which the player may largely do as he wishes – a large number of possible activities are available, and they are entirely optional. One such activity is the collecting of plants and herbs, from which various beneficial items can be manufactured. One of the rarer plants, which goes by the name «nirnroot», has a distinctive chiming sound associated with it, that quite literally is intended to guide the player towards the spot where it grows: as the player approaches, the sound is first heard only faintly, but becomes louder the closer one gets.

The term spotting refers to instances where sound is used to trick the audience into believing they have seen something that did not actually happen. Chion (1994) specifically mentions this example:

> We find an eloquent example in the work of sound designer Ben Burtt on the *Star Wars* saga. Burtt had devised, as a sound effect for an automatic door opening (think of the hexagonal or diamond-shaped doors of sci-fi films), a dynamic and convincing pneumatic "shhh" sound. So convincing, in fact that, in making *The Empire Strikes Back*, when director Irving Kershner needed a door-closing effect he sometimes simply took a static shot of the closed door and followed it with a shot of the door open. As a result of sound editing, with Ben Burtt´s "psssht", spectators who have nothing before their eyes besides a straigth cut nevertheless think they see the door slide open (p. 12).

This shows how the discrepancy in perception speed can be exploited – the faster audition basically convinces the slower vision that it 'saw' the movement of the door.

3.4: Narrative functions of speech

Theatrical and textual speech

Chion defined three categories of cinematic speech, which appear to denote different ways of conveying semantic information. The first of these is called *theatrical speech* and is seemingly straightforward. It refers to the (usually) scripted dialogue spoken by characters in the action. This type of speech constitutes the primary method of conveying linguistic information relevant to the overall plot, and the narrative as a whole is largely structured around it:

> In theatrical speech, which is the most common, the dialogue heard has a dramatic, psychological, informative, and affective function. It is perceived as dialogue issuing from characters in the action [...] Theatrical speech conditions not just the soundtrack but the film's mise-en-scène in the broadest sense. From screenplay to editing, by way of setting, acting, lighting, camera movements and so on, everything is in fact conceived, almost unconciously, to make the characters' speech into the central action and at the same time to make us forget that this speech is what structures the film (Chion, 1994, p. 171).

The second category is called *textual speech*. So named because it is analogous to the way in which a story is presented in literature, this category denotes linguistic speech that occurs outside the narrative action of the diegesis, but relates to and structures it nonetheless:

> Textual speech – generally that of voiceover commentaries – inherits certain attributes of the intertitles of silent films, since unlike theatrical speech it acts upon the images. Textual speech has the power to make visible the images that it evokes through sound – that is, to change the setting, to call up a thing, moment, place, or characters at will. If textual speech can control a film's narration, of course, there no longer remains an autonomous audiovisual scene, no notion whatever of spatial and temporal continuity. The images and realistic sounds are at its mercy. [...] This great power is generally reserved for certain privileged characters and is only granted for a limited time. [...] The [voiceover narrator] can actually be a protagonist, or a secondary character as witness, [...] or, alternatively, an external, all-seeing novelistic narrator (Chion, 1994, pp. 172–173).

In films, textual speech is for the most part used sparingly. An exception, at least if we are to believe the cliché, is the the ever-so-popular sub-genre known as *film noir.* Here the protagonist-as-narrator has become something of a cinematic trope. Such films famously tend to feature 'hard-boiled' detective protagonists whose on-screen exploits are more or less frequently accompanied by their own past tense, non-diegetic commentary. These comments typically serve as exposition and are typically heavy on creative allegories and cynicisms. These trademark features have been the subject of countless parodies and comedic efforts over the years, an example of how conventions evolve and take on new meanings over time.

Emanation speech

Although Chion has chosen to reserve the term *textual* for one of the previously mentioned categories, it is evident that they both pertain to textual meaning. This is not necessarily the case

with the remaining category, which he refers to as *emanation speech*. It denotes instances where spoken dialogue is deliberately obscured in various ways so that, in terms of informative value, the properties of the sound itself become more important than the textual (i.e. linguistic) message. In other words, speech functions as a sonic device, instead of primarily serving as a vehicle for text. The following quote contains Chion's (1994) explanation of this term:

> Emanation speech is speech which is not necessarily heard and understood fully, and in any case is not intimately tied to the heart of what might be called the narrative action. The effect of emanation speech arises from two situations. First, dialogue spoken by characters [that] is not totally intelligible. Second, the director may direct the actors and use framing and editing in ways that run counter to the standard rules – avoiding emphasis on articulations of the text, the play of questions and answers, important hesitations and words. Speech then becomes a kind of emanation of the characters, an aspect of themselves, like their silhouette is – significant but not essential to the mise-en-scène and action (p. 177).

I take this to mean that a line of dialogue can still relay important information even as its textual meaning is de-emphasized. However, when the linguistic message is unclear meaning also becomes more ambiguous. Consequently, such information is rarely of vital importance to the narrative, but is relevant nonetheless, as it relates to the nature of the speaking character(s).

In order to fully understand what emanation speech entails, we must first understand why filmmakers would want to employ speech in such 'indirect' ways in the first place. It appears that the reasons have to do with both the nature of perception and the history of the film medium. Since, by nature, we are incredibly sensitive to both the sonic and semantic aspects associated with the human voice, speech automatically becomes a focal point of our attention whenever it is present. Consequently, in the context of cinema, language inevitably comes to dictate both what we see and what we *fail* to see in the images. Historically, this was considered somewhat of a problem, as many prominent filmmakers of the silent era had revelled in the unique qualities afforded by images not structurally bound by language and synchronous sound[14].

Emanation speech should thus be viewed as a product of the various efforts made throughout film history to circumvent this perceived problem by «relativizing» speech (Chion, 1994, p. 178). As filmmakers came to realize that omitting dialogue entirely would often make scenes appear awkward or unnatural, several techniques were attempted in order to remove or reduce the impact

14 It appears to me that a characteristic feature of silent films, when compared to sound cinema, is a less immediate sense of reality. Figuratively speaking, there appears to be a greater distance between the audience and the narrative action of the film. This may be largely attributed to the absence of realistic sound, and, by extension, text. For one thing, the fact that we do not hear the characters speak is a constant reminder that they are not 'real', and the general lack of textual meaning – save for some (usually) scarce intertitles – removes their attachement to any one particular language or culture. Additionally, because there are no sounds to help define the physical space of the narrative, its temporal as well as spatial aspects tend to appear less fixed. As dialogue and synchronous sound gradually came to dominate the film medium, formerly flexible entities such as time, place, and the overall sense of reality were rendered absolute. Those who continued to advocate the silent cinema likely regarded these particular qualities as an advantage rather than a detriment.

of language upon the images, while still retaining a vocal sound presence. Paraphrasing Chion (1994), such techniques include 1) limiting speech to sequences where it is absolutely required ('rarefication'); 2) creating a verbally dense sound environment where the meaning of individual words is effectively cancelled out ('proliferation'); 3) using one or more languages with which audiences are most likely unfamiliar; 4) having a voiceover narrator partially cover the sound of the dialogue; 5) having dialogue alternately submerge and reemerge from a 'sea' of diegetic noise, i.e. a crowd; 6) deliberately making dialogue unintelligible by technologically manipulating the recorded sound; and finally, 7) having the dialogue conflict with the images, as well as the nondiegetic, technical aspects of the scene (i.e. acting, camerawork, editing), thus encouraging audiences to ignore it ('decentering') (Chion, 1994, pp. 178–183).

If we interpret Chions description of emanation speech (see above) in a more literal sense, it could conceivably also denote an unclear speaking pattern as a character trait – such as when a character mumbles or mangles words for no apparent reason other than habit, so that the audience cannot fully make out what is being said. For instance, he may be insecure (afraid to raise his voice), quirky or eccentric (i.e. inadvertently thinking out loud in a seemingly nonsensical fashion), physically handicapped in a way that impedes speech, or mentally unstable (speaking in a fragmented fashion while living in his own mind, oblivious to the world around him). In such instances, the informative value pertains not to the words, but to how the speech sounds and the contextual meaning this creates in combination with the images – we learn something about the nature of the character in question.

A curious and fairly recent example of deliberately unintelligible (i.e. relativized) speech can be found in the popular claymated[15] children's television series *Pingu* (1986). The titular character is an anthropomorphic penguin that lives with his family at the South Pole and speaks in a gibberish-sounding, made-up language. Understanding the words is not necessary (nor is it, for the most part, possible to do so) in order to follow the narrative, but this 'nontextual' speech is still important in terms of conveying the moods and intentions of the characters, alongside body language and visual context. Much of its popularity is often attributed to the fact that no dubbing was needed upon export to different countries.

3.5: Narrative functions of music

Throughout film history, the notion that music should directly support the narrative has been

15 Abbreviated verbalization of 'clay animation', i.e. films starring figures made out of clay that are animated using the stop-motion technique.

increasingly championed and institutionalized. As the the silent era progressed through its early stages, the question of suitability apparently became an increasingly debated topic – there are documented complaints by critics about unsuitable music and that more concious effort should be put into selection and performance (Larsen, 2007, pp. 23–24). Since the nature of the musical accompaniment in silent cinema depended on the availability, reportoire and proficiency of musicians, the relationship between images and music in terms of narrative development was initially a very loose one, and little to no musical consistency could be expected between different showings of a film. Early film music was for all intents and purposes an on-the-spot phenomenon. This changed over time, as the growing demand for suitable music led to a more systematic approach where both selection and performance would follow specific guidelines. Between instruction manuals, anthologies (where pieces were classified according to their perceived mood) and detailed cue sheets that were distributed to theaters alongside each individual film, an increasingly formalized system started to materialize. Eventually, film scoring became an industry in itself, as original and 'suitable' compositions became the norm.

The question of suitability is really a question of function. So, again, what does music actually *do* in the context of an audiovisual narrative? It has often been argued that music is incapable of telling a story because a sound signal does not actually represent anything in the world other than itself. A common counterargument is that music is practically never experienced in isolation, but always in some kind of context. Even if music does not directly represent anything other than itself, it can still reference an external meaning (see chapter 1) or invoke historical, cultural, and social connotations. Larsen (2007) has this to say on the subject:

> Music [...] is not a representational art form, but under special circumstances it can be used to convey meaning. Melodies, types of music, genres, styles, etc. can in certain instances function as signs and refer to something other than themselves. There are a number of different possibilities, but most of them can be traced back to two basic sign-functions: music can represent non-musical phenomena by virtue of *structural resemblance*, and music can evoke *associations with other music* (p. 66)

We briefly looked at various manifestations of perceived structural resemblance between music and other phenomena in chapter 2. Wingstedt (2004, 2005, 2010; Wingstedt et al. 2010) proposed a classification framework for the functions of narrative music using the three social semiotic *metafunctions* of communication originally coined by Halliday (1978) as point of departure. The below summary is excerpted from Wingstedt et al. (2010):

> The *ideational* metafunction is the content function of communication. Kress et al. [...] describe it as representing what goes on in the world; material, verbal, mental and relational processes – 'who does what, with or to whom and where' [...].
> The *interpersonal* metafunction is the participatory function of communication, communication as doing something. It is the component through which the communicator expresses her own attitudes and judgments – and seeks to influence the attitudes and behaviour of others. It establishes, maintains and

> specifies relationships between members of societies or groups through expression of social relations, interrelations of power and knowledge [...].
> The *textual* metafunction is the component which provides the texture, the organizing of a text (in a broader sense) as a coherent message through textual resources of a mode (in relation to the environment). The textual component has an *enabling function* with respect to the other two. It is only in combination with textual meanings that ideational and interpersonal meanings are actualized [...]. A modal expression is an instantiation of all three functions interwoven (p. 4).

On this foundation, six categories that denote various functions of narrative music were suggested (Wingstedt et al., 2010, p. 3):

- The *Emotive* function refers to music's ability to communicate emotive qualities, either experienced by the audience (induced) or just cognitively identified [...]. The expressed emotions may be attributed to individual characters of the story or represent relationships or events – they can also describe overall emotive aspects of situations or forebode future implications of the plot.

- The *Informative* function comprises situations where music expresses or 'explains' phenomena or events by communicating information on a cognitive rather than on an emotional level. Music can for example evoke certain cultural settings or time periods, clarify ambiguous situations, indicate social status or simply represent a character or phenomenon, for instance by the use of a leitmotif [...].

- The *Descriptive* function is related to the informative function in certain aspects, but differs in that the music is actively (or programmatically) describing something rather than more passively representing certain values. It is usually a matter of describing the physical world, such as physical setting, appearance or movement.

- The *Guiding* function includes musical functions that, so to speak, turn directly to the audience aiming to 'direct' the eye, thought and mind. This could include indicative or imperative functions. The latter function is prominent in computer games or advertising, where the purpose is to bring the audience to perform specific actions.

- The *Temporal* function foregrounds the time dimension of music. Especially important is music's ability to provide continuity (immediate, longer or overall) as well as how music can contribute and define structure and form.

- The *Rhetorical* function refers to how music sometimes 'steps forward' to comment the narrative events or situation. This is often achieved by having the musical expression contrast the visuals or by referring to well-known musical material.

As far as I can tell, we have so far touched upon all of these functions in some form or other, either in general auditory terms or specifically related to music. Do note how these categories account for both structural and emotional aspects. If we apply this framework to random cases of narrative music, we typically find that multiple functions will be in effect at the same time. Fig. 3.3 shows how these categories are distributed between the three metafunctions (although I have created this table, it is arranged in accordance with Wingstedt's illustration).

Fig. 3.3: Narrative functions of music and their metafunctions (Wingstedt, 2008, p. 47)

Metafunction	**Ideational**	**Interpersonal**	**Textual**
Narrative function	Emotive (observed)	Emotive (experienced)	Temporal
	Informative	Guiding	Intermodal
	Descriptive	Rhetorical	

Note how the additional «intermodal» function refers to such factors as placement, timing, and, unsurprisingly, synchronicity (Wingstedt, 2008, p. 47). I could here mention that there is a subtle distinction between the terms synchronicity and synchronism, but as it appears to be of little consequence in this context, I shall refrain from further explication.

Empathetic and anempathetic music

We have previously discussed the relationship between nondiegetic music and the visual action in terms of perceived structural and/or emotional congruence, i.e. how structural and affective aspects of the music can be perceived as corresponding or aligning with their visual counterparts. Since our brains will be actively searching for patterns that unite visual and auditory information, we are very likely to find such patterns – even in random sound-image combinations there will usually be perceived points of synchronization, etc. Interestingly, music can add value also when its relationship with the images is perceived as contradictory. We may thus discern two helpful categories, as outlined in the below quote by Chion (1994):

> On one hand, music can directly express its participation in the feeling of the scene, by taking on the scene's rhythm, tone, and phrasing; obviously such music participates in cultural codes for things like sadness, happiness, and movement. In this case we can speak of empathetic music, from the word empathy, the ability to feel the feelings of others. On the other hand, music can also exhibit conspicuous indifference to the situation, by progressing in a steady, undaunted, and ineluctable manner: the scene takes place against this very backdrop of "indifference". This juxtaposition of scene with indifferent music has the effect not of freezing emotion but rather of intensifying it, by inscribing it on a cosmic background. I call this second kind of music anempathetic (with the privative a-) (p. 8).

As we can see, the term *empathetic* basically denotes cases of perceived music-image congruence, while *anempathetic* music could perhaps be said to represent an overt form of audiovisual counterpoint. The wording of the above excerpt also appear to be generally consistent with the emotional perspectives on audiovisual perception discussed in chapter 2. Notably, words that denote both structural (e.g. rhythm, phrasing, movement) and affective (e.g. sadness, happiness) components are used alongside each other to describe empathetic music. Given that we are dealing with quite basic metaphorical relationships that has been widely studied, I find it safe to assume that most readers will agree that determining whether or not a musical cue ‘participates’ in the narrative is a more or less intuitive task; one that is largely performed unconciously. Thus, the term empathetic should require no further explanation.

As for its counterpart, I shall briefly refer to the prologue of *Watchmen* (2009) which I analyze more extensively in the next chapter. Demonstrating what is probably the most common use of the anempathetic effect, this scene depicts the murder of a character called Edward Blake (played

by Jeffrey Dean Morgan) which serves as catalyst for the film's overall plot. Blake – an ageing, former vigilante crime-fighter with dubious morals and a murky past – is attacked in his high-rise apartment by an unknown assailant. An extended fight scene ensues, but Blake is ultimately defeated and falls to his death after being thrown out of the window. Accompanying this deadly encounter is the romantic jazz ballad *Unforgettable*, performed by Nat King Cole. Despite the obvious disparity between this song and the nature of the action, we can safely say that it enhances the emotional impact of the scene. It is as if the contrast underlines the hopelessness of Blakes desperate struggle versus a superior opponent. Despite him proving to be quite a formidable combatant himself, his eventual defeat seems inevitable – the indifferent music effectively dooms him from the outset.

Both sides of the empathetic/anempathetic-dichotomy ultimately denote cases where music is actively commenting on the narrative, and as such they invoke several of the functions described by Wingstedt et al. (2010) . refers to as the rhetorical function of music. However, some instances of nondiegetic music do not seem to fit either of them. According to Chion (1994), «[...] there also exist cases of music that is neither empathetic nor anempathetic, which has either an abstract meaning, or a simple function of presence, a value as a signpost: at any rate, no precise emotional resonance» (Chion, 1994, p. 8). In terms of added value, this type of music is intentionally ambiguous, and therefore becomes highly subjective. Without directly participating in the scene in any immediately meaningful way, it may still influence the narrative in ways that can be difficult to define. I look more closely at one such case in the next chapter.

Lastly, Chion (1994, p. 9) also notes that the anempathetic effect can be produced by nonmusical sounds as well, specifically mentioning the sound of the running shower in the aftermath of the violent murder of Marion Crane (played by Janet Leigh) in Psycho (1960).

Chapter 4: Case studies

4.1: *Lost* – flashback-sound effect

The TV-series *Lost* (2004) made the flashback-structure into a trademark. Following a group of plane-crash survivors on a mysterious island somewhere in the pacific, the majority of episodes sees the timeline alternate between present and past events. The characters' present day struggle to survive in their new, unfamiliar environment is regularly interjected with protracted flashbacks that show their backstory and motivations. Because these sequences are such a frequent device, they become a defining element of the show's narrative identity. From an auditory perspective, the most interesting aspect is the way that these jumps in the timeline are introduced and concluded via a particular sound effect, which is quite consistent throughout the series.

Typically, the impending transition from present to past is signaled with a brief closeup of whichever character the narrative is currently focusing on while a characteristic swell of white noise forms a crescendo that ends abruptly when the image cuts to the first shot of the flashback. This sound effect usually signals the return to the present as well. The sound, when isolated, is reminiscent of the type of ambience typical of a seaside environment, i.e. the 'roar' of the wind and ocean waves. Thematically, this would be a natural fit with regards to the island setting, as the surrounding ocean represents both a physical and metaphorical barrier that separates the characters from their former lives, which is what the flashbacks portray. This case shows a sound assuming a textual function, while also cleverly establishing itself as a signature element.

4.2: *Under a Killing Moon* and *Beneath a Steel Sky* – monologue

These two PC adventure games, both of which are generally considered classics, share a number of superficial likenesses. In addition to their common genre heritage and thematically similar titles, both were published in 1994 and are set in distant, dystopian future cities ravaged by war and pollution, where the player is tasked with navigating various environments, gathering clues and items in order to solve logical puzzles and, ultimately, thwart the evil schemes of shady conspirators.

Here the similarities seem to end, however. Although the underlying mechanics are largely the same, the two games represent very different directions in terms of gameplay that ties directly into their vastly different technical aspects. Whereas *Beneath a Steel Sky* is a traditional 'point-and-click' game where you move an animated character around on static, hand-drawn 2D-backgrounds

using the mouse, *Under a Killing Moon* features full 3D-environments and is played from a first-person perspective with 360° freedom of both vision and movement. Also, the latter fuses recorded footage of real actors with computer generated environments, a technique considered revolutionary at the time of its release.

Most adventure games requires the player to frequently engage in conversation with non-player-characters (NPCs). These two are no exception, and both games feature fully voiced dialogue. As is a trademark of the genre, this applies to both the frequent conversations between characters and the just as frequent bits of dialogue that the main character issues whenever the player interacts with something in the environment. The interesting question here is, of course, who are they talking to? In the case of *Under a Killing Moon*, the voice of protagonist Tex Murphy has a noticeably different quality to it in the form of a reverb effect whenever it is heard outside of a conversation. Thus, it comes off as a case of subjective internal sound; i.e. the voice acts as a projection of his thoughts. In other words, Murphy's inner monologue does not necessarily address the player or acknowledge his/her presence. In *Beneath a Steel Sky*, on the other hand, main character Robert Foster speaks these lines the same way as he would during conversation, displaying his talking animation while doing so. Talking to oneself out loud in the presence of others is generally considered strange behaviour, but apparently this is not the case in this game. Granted, Foster spends the majority of the game accompanied by his robot sidekick Joey, who appears to be able to hear Foster at all times and even chimes in with the occasional observation or comment of his own. But still, it is hard to interpret this as anything other than Foster speaking directly to the player. In films, a similar situation would be referred to as 'breaking the fourth wall' [16]

4.3: *Twin Peaks* pilot episode: Cooper's introduction – music

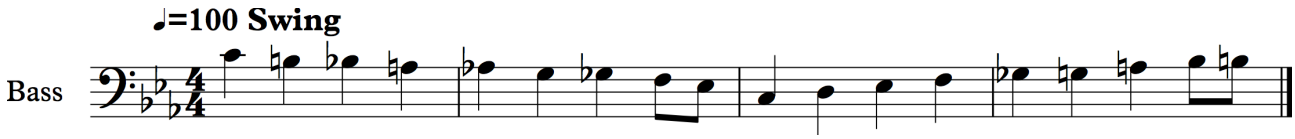
In groundbreaking television series *Twin Peaks,* (1990), I find the score by composer Angelo Badalamenti to frequently stray into the 'uncharted' territory between empathetic and anempathetic music. Revolving around a murder investigation within a small-town community of more or less eccentric characters, the narrative combines mundane everyday situations with elements of horror and the supernatural. The many absurd and humorous moments constitute a lighthearted counterpart to the darker investigation aspects of the story which deals with murder, abuse and madness. The frequent juxtaposition of seriousness and silliness became somewhat of a trademark of the series, as did its instantly recognizable score. In my opinion, the inherent 'strangeness' for which *Twin Peaks* became known owes a lot to its musical component. Largely built around a handful of recurring

16 This expression refers to the common theater setup of, for instance, a livingroom, where the open front of the stage could be understood as an imaginary wall through which the audience view the action.

themes and motifs, the score often appears conspicuously overt, almost as if it conciously rejects the label ‘underscoring’. In the series’ pilot episode we are introduced to main protagonist Dale Cooper (played by Kyle McLachlan), an FBI-agent assigned to investigate the murder of a young girl, as he approaches the town of Twin Peaks by car. While driving, he is making notes to himself using a tape recorder, his various comments serving as expository dialogue for the audience.

Accompanying this scene is a jazzy tune, the main component of which is a ‘walking’ bassline in the key of C minor. Following a few introductory bars with alternating, semiquaver octaves on the root, the bass instrument – apparently an upright played with fingers – starts to play a characteristic four-bar pattern which descends, then reascends within the aforementioned pitch-interval, using a sort of extended blues scale with several added chromatic steps along the way:

Fig. 4.1: Dance of the Dream Man – repeating bass pattern



Percussive elements include a drumset played using wisps, as well as what sounds like finger snaps on the second and fourth beat of each measure. A reverberated saxophone is eventually added to the mix, playing a melodic line that may or may not be improvised. It might not be entirely clear whether this music is in fact nondiegetic, as its source could conceivably be the car’s speaker system, but other than this circumstance there is nothing that suggests Cooper himself can hear it.

In isolation, this could be seen as a typical example of a traditional leitmotif; its first appearance coinciding with the introduction of a lead character. Over the course of the series, however, the theme appears repeatedly in different settings and in several variations (i.e. slow/fast; one instrument/more instruments; etc.), and in my opinion its usage does not clearly indicate a connection to Cooper, nor any other particular character, location or event. Although the official soundtrack has it titled *Dance Of The Dream Man*, the character whom this title refers to has yet to be introduced when the track is first heard, and his appearances throughout the series are extremely rare. It acts not so much as leitmotif in the regular sense of the word, but rather, I find, as a sort of symbolic signature of this particular diegetic universe in its entirety. It appears to me that its general function is to remind the audience that even though this universe resembles the real world, the ‘rules’ here are somewhat different. This is mere speculation, though. In any case, its emotional resonance – if any – does not pertain to any basic and easily recognizable type of emotion, nor does it come off as a demonstration of ‘cosmic indifference’. For all intents and purposes, it is simply

there. While we can safely assume that there are aesthetic considerations underlying the implementation of this theme (which would translate as a form of emotional meaning), the relationship between the music and the on-screen action is, at best, an indirect one. As such, it escapes definition in terms of empathetic/anempathetic.

4.4: *The Dark Knight:* Joker murders Gambol

The Dark Knight (2008) is an interesting film in terms of reciprocity of added value, as it repeatedly uses suggestive techniques to avoids showing directly the brutality committed by the characters, and instead relies on the audience's imagination to fill in the blanks. Despite its generally grim themes and subject matter, this film had quite low age restrictions, which undoubtedly was due to the fact that most of its violence is visually nonexplicit. One scene sees the main antagonist known as the Joker (played by Heath Ledger) and his crew overpower and subsequently murder rival crime boss Gambol (Michael Jai White), presumably by mutilating his face with a knife. At least, so it appears we are meant to believe – we do not actually see or hear any of this. Rather, this meaning is implied through substitute images and, perhaps somewhat surprisingly, musical sound.

The first shot of the scene shows Gambol at his pool table. The precise moment in which the film cuts to this shot is punctuated with an action sound, i.e. the 'clack' as the cue ball collides with another ball. The reverb from the last chord of a staccato string theme that played in the preceding scene can also be heard. The music does not subside entirely and continues into some low-frequency percussive sounds as Gambol is informed by one of his lieutenants that he has visitors. A group of men enters, carrying a body wrapped in plastic. The corpse is laid out on the pool table, and Gambol removes the wrapping around its face in a closeup (CU) shot, confirming that it is the body of Joker, whom he has placed a bounty on earlier in the film. However, when Gambol turns away, the not-quite-dead Joker suddenly rises from the table and stabs (presumably killing) two of his men in the process. Visually, this moment passes by extremely quickly. However, it is punctuated by an initial loud and dissonant orchestral 'stab'. A few more of these appear in succession, perhaps so as to underline the lethal consequences of the action that was just commited, but not followed up on by the image. Joker subsequently moves directly towards Gambol and takes him hostage. This is shown from an over-the-shoulder (OTS) view.

Fig. 4.2: The Dark Knight – Joker murders Gambol part 1

Joker's henchmen hold the lieutenants of his soon-to-be victim at gunpoint, while he himself performs one of his trademark monologues – a likely made-up story about how he received his distinct facial scars (he repeats this procedure later in the film, telling a different story) – his knife suggestively resting in the corner of Gambols mouth. Tension builds rapidly as the monologue draws to a close, fuelled by subtle but insistent musical underscoring, which appears to have two main elements: 1) a rhytmical, iterative bass drone played by an electronic-sounding instrument; likely a bass guitar or a synthesizer, and 2) a dissonant sounding instrument whose pitch slowly rises in a continuous and controlled glissando-movement. This is actually a sustained two-note interval (somewhat less than a semitone) played on a cello, gradually moving up the fretboard. This particular musical device is first introduced during the film's opening sequence and functions as a true leitmotif for the Joker-character. It possesses an eerie quality which the audience quickly learns to associate with his fearsome appearance and ruthless, unpredictable nature. Interestingly, while

17 =Audio. Indicates sound events

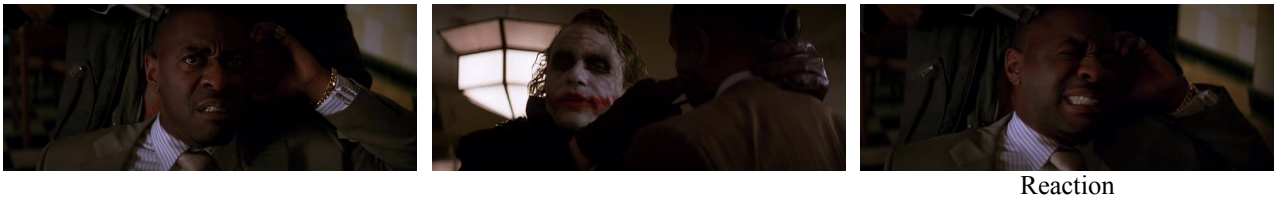
the first instrument retains a stable pitch, its amplitude fluctuates. Certain onsets, i.e. iterations of the sound-producing event, appear louder than others; perhaps suggesting a programmed delay or echo. The second instrument does quite the opposite, retaining a somewhat constant amplitude while its frequency content continually changes.

Suddenly The Joker looks directly at one of the subdued lieutenants, rhetorically asking the question "Why so serious?". Next, there is a closeup of the frightened lieutenant whose face suddenly makes a horrified grimace as a final loud and dissonant orchestral 'stab' is heard; an example of music being used punctuatively to represent an offscreen action. Our immediate assumption is that Gambol has met his demise, which is all but confirmed in the following shot. Here the Joker is briefly seen from behind, and as he casually lets go of his victims now limp body it falls to the floor. Again, nothing is actually seen or heard of the presumed murder; both images and sounds have been substituted. Still, because of reciprocal projection of meaning between modalities, the brutality and emotional impact of the off-screen murder is preserved.

Fig. 4.3: The Dark Knight – Joker murders Gambol part 2

A: Joker leitmotif
(Joker's monologue)

A: Joker leitmotif
(Joker's monologue)



Reaction

A: Joker leitmotif — Loud, dissonant orchestral stab
(Joker's monologue) — "Why so serious?"



V[18]: Limp body falls to floor

A: Resonance from orchestral stab

4.5: *Watchmen:* prologue

I previously referenced the prologue to *Watchmen* (2009) in relation to the anempathetic effect. I shall now attempt a more extensive audiovisual analysis, as this sequence has a number of interesting things going on. If we look at the film's very first moments (which lead up to the fight scene I described earlier), we find our view shifting between several different locations. After an initial, silent presentation of production company logos on a yellow backdrop (which are, of course, nondiegetic images), the film transitions seamlessly into its first diegetic shot via a clever visual effect: after the last logo disappears and leaves the screen all-yellow, a sustained, high-pitched sound is heard and we see some black shapes come into view from the edges of the frame. Because the source of this sound is currently unknown, we are here dealing with an acousmatic sound. Next, a similarly acousmatic male voice loudly speaks the words «wrong as usual». This voice also has the markings of on-the-air-sound due to noticeable compression of the sound signal, which gives it a slightly unnatural, almost nasal quality. At about the same time, we realize that we are actually watching the camera 'pull back' from an extreme closeup (ECU), revealing that the black shapes and the yellow backdrop form a 'smileyface'-icon that is printed on a button (this is a recurring visual motif in the film, as well as in the graphic novel it is based upon).

As the view continues to pull back, it turns out that this item sits on a bathrobe-collar worn by a grey-streaked, cigar-smoking man. We do not yet know his identity at this point, but the film later informs us that his name is Edward Blake. Directly in front of him is a kettle with steam coming out of it, which explains (de-acousmatizes) the high-pitched noise. Blake appears to be preparing a cup of tea. The noise suddenly stops, at the same time as the continuous opening shot cuts to an extreme closeup of Blake's face. Meanwhile, the voice from before continues speaking about escalating political tension between Soviet Russia and the United States. This is the first clue as to the setting of the film, which eventually turns out to be an alternate version of 1985 where the cold war is at its peak (although the exact year is not mentioned during the prologue).

18 =Video. Indicates visual events that occur over multiple shots

Fig. 4.4: Watchmen – prologue

Yellow backdrop | Black shapes enter frame

V: Pull-back shot

A: High-pitched acousmatic noise

Acousmatic voice ("Wrong as usual")

V: (Pull-back shot)

A: (High-pitched acousmatic noise)
(Acousmatic voice)

Kettle visible

V: (Pull-back shot)

A: (High-pitched acousmatic noise) Noise deacousmatized; i.e. kettle
(Acousmatic voice)

Kettle. Pours water. | Looks up (at TV)

V: Cut to ECU Blake.

A: Kettle noise ends.
(Acousmatic voice)

TV visible | Blake sits down to watch

V: Panning shot

A: Voice deacousmatized; i.e. TV

Based on its sonic properties, specifically the self-important tone and meticulous phrasing, we may

have already guessed that this voice is part of some sort of broadcast. In a subsequent shot there is indeed a television tuned in to what looks like a debate program. Thus, the voice is firmly established as on-the-air. As Blake sits down to watch, the voice – which we now understand as belonging to the program host – announces a statement by president Richard Nixon, who in this alternate reality is serving his fourth subsequent term in office.

The following shot suddenly transports us inside the very room from which this statement is being televised, presumably inside the White House. I should note that the program host actually announced the president's statement in the past tense, indicating that it is a previously recorded segment. Thus, the shots from the White House also transports us back in time, as we are watching an event that already happened in the timeline of the narrative. First we see the teleprompter that the president is reading from, then two shots of Nixon (played by Robert Wisden) on the podium from which he adresses the camera-crew. This is the first appearance of a real-world figure, of which there are four in total during this prologue. The sound of his voice bears no indication of being on-the-air, which further reinforces the notion that we are experiencing this speech as if present at the exact time and place of its delivery.

As soon as the statement ends, we find ourselves inside the studio where the debate program playing on Blake's TV is being filmed, as if it were a live broadcast (although we cannot know this for certain). Although not explicitly stated in the film, the program is clearly a fictionalized episode of real-world american talk show *The McLaughlin Group*, which debuted in 1982 and as of september 2012 is still running. The format usually sees moderator John McLaughlin accompanied by two regular cast members, conservative Pat Buchanan and liberal Eleanor Clift. Though actors stand in for the actual people featured in the real show, the host's characteristic mannerisms (particularly the trademark phrase «on a scale of zero to ten, zero meaning impossibility, ten meaning absolute metaphysical certitude») identify him as McLaughlin, while the other two are identified by name as the scene progresses. The reference is presumably obvious to american audiences, but the images alone do not establish this connection – the faces will be unfamiliar also to those with prior knowledge of the program. From a narrative perspective, the inclusion of clearly nonfictional elements at the very start of the film (in the graphic novel they come later) serves two obvious purposes. First, in a practical sense, it limits the need for extensive exposition. Because the audience is likely familiar with this recent period in history and its tense political climate, context is quickly and unobtrusively established. Second, it creates a foreboding atmosphere that is further strengthened by the knowledge that the overhanging nuclear threat was very real just a few decades ago.

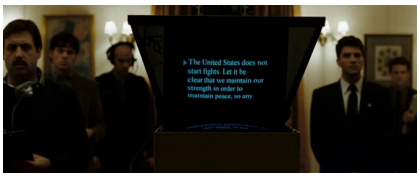
The ensuing dialogue between the three, which no longer displays any of the sonic properties associated with on-the-air sound, is interspersed with three more cutaway[19] (CA) shots that demand a closer look. The first of these occurs as McLaughlin states that «...the watchdog group of nuclear scientists moved the doomsday clock up to five minutes until midnight». We see some men in white laboratory coats in front of a model clock-face that symbolizes the estimated likelihood of imminent world-wide nuclear destruction (the clock is another recurring visual motif and symbol in both the film and graphic novel) There are apparently some offscreen photographers attending this event, as indicated by three quick, successive flashes of light synchronized with the distinctive sounds of cameras. Similar to the White House sequence, this shot also references a past event. However, is is apparently not meant to be viewed as featuring in the current broadcast, but rather intended exclusively for the film's audience, as a piece of exposition regarding the doomsday clock reference.

There are several indications of this. For one thing the host's voice continues to be heard throughout the shot, whereas it temporarily fell silent during president Nixons's earlier statement. Also, the shot is in slow motion and the camera sounds are heavily reverberated, as if they were propagating in a vast and highly reflective space. This has the effect of rendering our perception of time less immediate and exact, distancing the shot from what we perceive as the diegetic present – i.e. the TV studio and Blake's apartment. Additionally, it can be seen as adding gravitas to the situation by underlining the seriousness of the prospect of nuclear annihilation. At the same time, the host's voice continues as if in its original environment (i.e. the acoustically 'dry' TV studio), which keeps it firmly in the here-and-now. Thus, there are two separate auditory spaces operating in parallel during this shot.

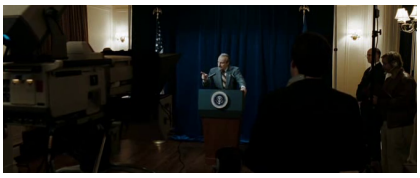
The second cutaway shot shows a silhouetted Blake from behind, still sitting in front of the TV. As one would expect, the host's voice regains its former on-the-air quality – a subtle, but important change that enforces the notion that the diegetic space follows established laws of physics. Third and finally, there is a shot from inside the production room where the ongoing broadcast is being edited in real time; another indicator that the program is live. From an over-the shoulder view, we see some monitors showing the studio cast. We hear the host introduce his female cast member, as the producer cues a closeup image of this woman with a hand gesture.

19 When a shot is momentarily interrupted with another shot that shows something different. Cutaways are brief and typically revert back to the preceding shot, or one similar to it. For instance, it is a common technique to insert a cutaway between a medium and closeup shot of the same character or object.

Fig. 4.5: Watchmen: prologue (cont.)



White House, teleprompter.



'Richard Nixon'

TV Studio

CA#1: Scientists, doomsday clock

'John McLaughlin'

A: Reverb ---------- No reverb

T[20]: Slow motion ---------- Normal speed

CA#2: Blake's appartment

'Pat Buchanan'

CA#3: Production room

Cues closeup of 'Eleanor Clift'

'Eleanor Clift

Up to this point, the prologue is basically a big chunk of exposition, i.e. background information regarding the general state of affairs in this alternate universe. Though nothing dramatic has happened so far, we have seen a rather complex series of shots that jump back and forth between different locations as well as points in time. Yet, it somehow makes perfect sense to us. How is this possible? Naturally, the shots themselves are carefully composed and the sequence deliberately structured. Having elements in one shot carry over to the next effectively creates visual continuity. However, the audio is really the glue here. The continuous sound of speech throughout the sequence softens the abrupt visual cuts, and the expository dialogue establishes a narrative context that ties the different locations together. Consequently, we have no trouble following along with the changing visual perspectives. We understand that they are merely different aspects of the same

20 =Time. Indicates temporal elasticity, i.e. slow motion, etc.

sequence, centered around a man watching television.

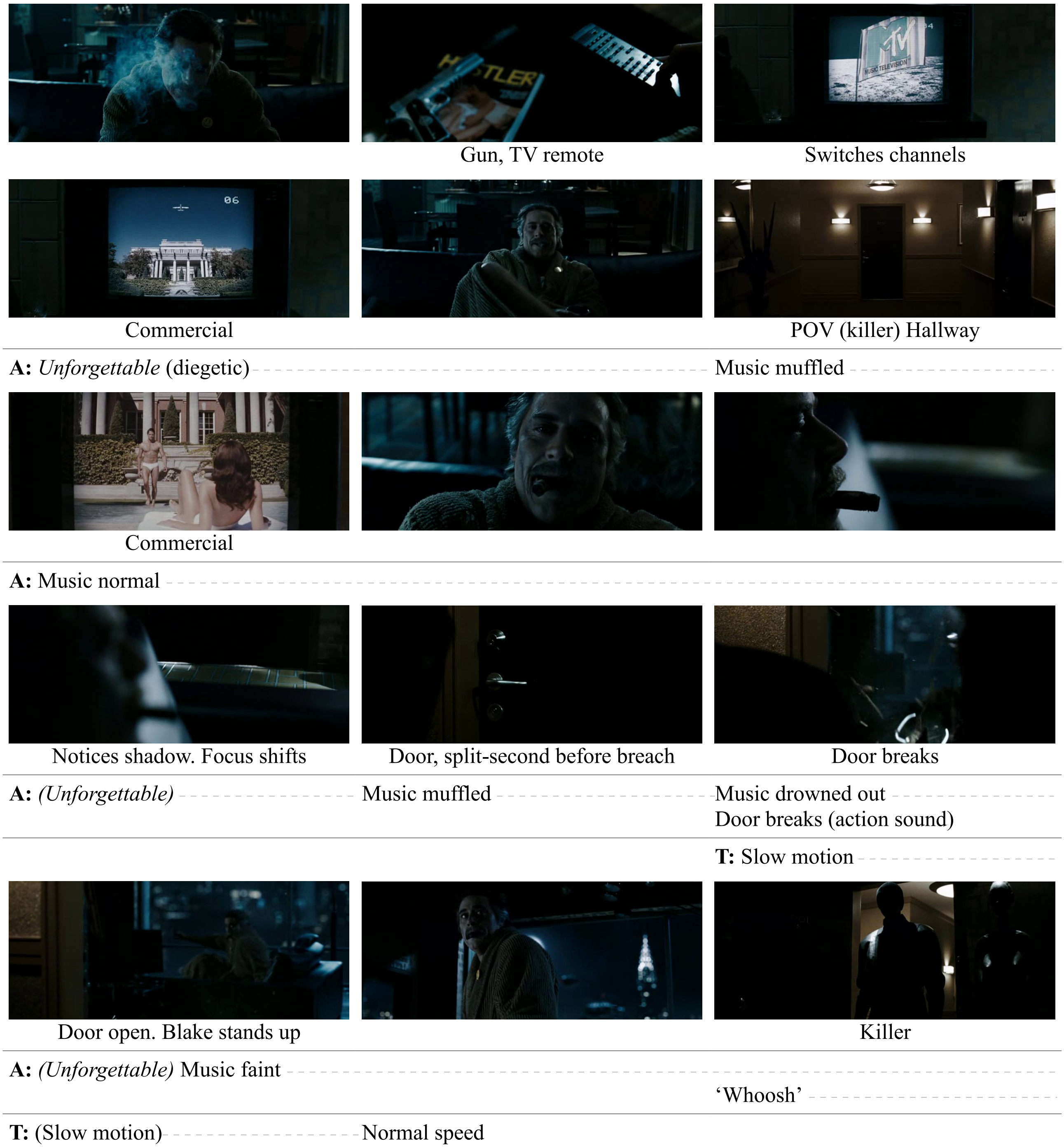
Crossing the diegetic divide

Interesting effects may arise when the line between the diegesis and nondiegesis is intentionally blurred. As Larsen (2007, pp. 58–59) notes in his analysis of *Metropolis*, it was not uncommon for the mostly nondiegetic music of silent films to temporarily take on a diegetic function, briefly appearing as if audible to the characters within the fiction. Instruments would sometimes be used to imitate environmental sounds, or the music would blend with the setting in such a way that it appeared to be *part* of the narrative instead of commenting on it from the outside. Since music would typically play continuously for long periods of time (often for the entire duration of the film) and there were no *actual* diegetic sounds, it made perfect sense for music to take on this function from time to time. In modern films, however, this technique should be considered on the premise that there is no longer any *need* for directors to use nondiegetic music in this fashion – when it occasionally happens, it is typically for stylistic reasons.

The second part of the *Watchmen* prologue exemplifies this. Initially, the music heard in this scene has an explicit on-screen source, as it is clearly shown as emanating from the TV. It starts just prior to the intruder's entry, when Blake changes channels (briefly stopping by MTV, another real-world reference) and arrives at the one where the song *Unforgettable* is playing (see fig. 5.4). As he leans forward and grabs the remote, there is a quick shot showing a silenced pistol embedded with the same smileyface-motif as earlier lying on the table. This serves as the first clue as to Blake's true nature as a character, and also indicates that he may be expecting trouble (why else have the gun lying out in the open?). In the narrative of the film, the program turns out to be a commercial advertising a perfume titled *Nostalgia*, made by a company owned by another major character (this is an example of foreshadowing). Logically enough, the music has the same on-the-air quality to it as the debate program.

Next, we are presented with a brief hallway shot of the apartment door that is about to be breached, as if seen from the intruder's point of view (POV). In accordance with this new visual perspective the sonic properties of the music instantly change as the loudness is reduced and high frequencies are cut. This results in a 'muffled' type of sound that effectively places the source of the music on the other side of the door.

Fig. 4.6: Watchmen: prologue (cont.)



In the following shot, once again showing the TV, we are back inside the apartment. As one would expect, the acoustic properties of the music return to their original state. Blake then notices a moving shadow under his door. The next shot is in slow motion and shows a brief closeup of the same door from the outside again as the intruder breaks it open with a kick, revealing a startled Blake as he stands up from the couch. The loud action sounds resulting from the breach momentarily drowns out the music. Like in the earlier slow motion shot involving the doomsday

clock, there is a reverb effect applied to these sounds. When the music returns in the next shot, which plays at normal speed, it sounds farther away. A presumably nondiegetic, 'whoosh' (not unlike the signature sound effect from *Lost*) sound effect is heard as the soon-to-be murderer appears in the doorway. The two men stare each other down for a few moments, until Blake takes a sudden initiative by throwing his cup at the intruder and making a leap for his pistol. Unfortunately, the incredibly fast intruder manages to grab and divert the gun before it fires, and the bullet from the ensuing shot accidentally hits the TV instead.

Fig. 4.7: Watchmen: prologue (cont.)



Blake speaks | Looks at gun

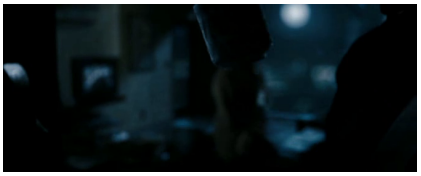
A: *(Unforgettable)* «Just a matter of time, I suppose» | Glove sqeaks (action sound) | Another 'whoosh'

T: Slow motion.



Empties cup

A: *(Unforgettable)* ('Whoosh')


V: Throws cup | Focus shifts | Cup smashes on door

A: *(Unforgettable)* Drowned out
Cup thrown (action sound) | Cups smashes (action sound)

T: Slow motion



Blake grabs gun | Jumps away from killer | Rolls on floor

T: Slow motion | Normal speed

Confused, aims in wrong direction

Killer's hand on gun. Fires, hits TV

Bullet exits chamber

A: *(Unforgettable)* Faint

Gunshot (action sound)

T: Slow motion

TV with bullet hole. *Nostalgia* advert

V: Screen goes black

A: (*Unforgettable)* Loud (nondiegetic)

T: Normal speed

At this point, something interesting happens with the music. From being just faintly audible, it suddenly comes rushing back in as its source is destroyed, the complete opposite of the behaviour one would expect. Its timing is impeccable as well, in the sense that a crescendoing phrase is just starting. By continuing to play after the TV goes dark it renders itself nondiegetic. This new status is confirmed when the scene cuts to an exterior shot, this time showing the two combatants from the outside, looking in through the apartment's panoramic windows. While the 'realistic' (i.e. they are, in fact, anything but) sounds of the struggle are predictably altered to account for this new perspective, the acoustic properties of the music remain unchanged – because it is now nondiegetic, it is no longer subject to the acoustic space of the narrative universe. By this point, it is becoming clear that Blake is unlikely to survive the encounter, and the otherwise serene *Unforgettable* sounds increasingly sinister due to the anempathetic effect. The music reaches its climax as he is picked up by the other man and tossed through the window.

Chapter 5: Closing

Summary and conclusions

As previously stated, the aim of this thesis was to present a macroscopic, comprehensive view on sound and music in narrative multimedia. I have attempted to do this in the form of an evolving discussion, that approaches complex implications of the audiovisual relationship from an initial platform consisting of quite basic factors. I have drawn on multiple theoretical sources, and while much of the theory presented here is extremely condensed, and therefore somewhat lacking in nuance, I believe this to be justifiable considering the wider scope of the thesis. In terms of structure, I have to the best of my ability attempted to organize the various theoretical aspects in a logical, sequential manner, so that they form a coherent argument.

While, on the one side, I have made an effort to continually add elements to the overall discussion, I have also made an equal effort to maintain a few core perspectives throughout. For the sake of summarizing these, let us briefly review the three problems I presented at the very beginning:

1) What is the nature of the relationship between what we see and what we hear when experiencing an audiovisual narrative?
2) What are the implications of this relationship in terms of how a narrative is presented and interpreted?
3) What are the specific functions of sound and music in this regard?

I started out by focusing on audiovisual perception from the viewpoint of cognitive psychology. This allowed me to base the further discussion on what I consider to be a solid and empirically grounded theoretical foundation. I initially assumed a bottom-up approach and gradually moved towards a more top-down cognitive perspective. The initial point I made was that perception is strongly biased on account of being the evolutionary product of the fundamental, primary directive instilled in us as living organisms; i.e. survial and proliferation. Having accounted for basic information processing tasks, sensory integration, and memory, I looked at differences between vision and audition in terms of temporal and spatial acuity. I assumed a view of cognition as embodied experiences, and discussed notions of an analogous relationship between sound and movement and sound as shapes, or objects. This segued into a discussion of the role of emotion in perception, specifically in relation to music in narrative multimedia, via the notion of audiovisual

congruence based on structural and affective qualities. Along the way I also introduced a few topics related to perception of a real environment versus a fictional narrative.

It seems evident from the research discussed in chapter 2 that semantics are ingrained in perception. The process of attributing meaning to audiovisual stimuli appears largely automated, while also, to an extent, subject to cognitive control. It permeates even the most basic levels of information processing and cannot be viewed independently of our fundamentally biased and multimodal perception. Any attempt to theorize the construction, presentation and interpretation of an audiovisual narrative should take into consideration this basic knowledge. In this regard, the different, yet integrated natures of the visual and auditory modalities are routinely on display in the audiovisual narrative experience. Since they appear to have complementary strengths and weaknesses, narrative meaning should be regarded as the product of mutual influences between sounds and images, although the spectator is typically not aware of this on a concious level. While audition is the more temporally attuned modality, vision is spatially dominant. Because of sensory integration, however, the removal of either visual or auditory components is typically required in order to highlight the full extent of the effects of one modality upon the other.

Having established this theoretical foundation, I proceeded to focus on the narrative experience itself, and the more detailed functions of auditory stimuli in this regard. The approach I attempted can in some aspects be viewed as a continuation of the previous chapter, in that I sougth to establish a spatiotemporal foundation on which to compile from literature and observation a descriptive framework for narrative functions of sound. There are some inherent problems with this approach, as I believe should be evident, but I still found it to be the most productive in terms of the overall discussion. Several auditory categories were suggested, which I then was able to use as points of reference in the task of determining the contextual narrative functions of specific types of sound, including music.

I applied this descriptive framework to a number of case studies. Those presented here are the ones I personally found to be of most interest in light of the core perspectives and general direction of the discussion. What I primarily have attempted to demonstrate is the value of having a set of descriptors that work on multiple levels and can be used in combination with one another for increased precision. Although far from complete, I believe this framework represents a largely successful initial effort.

Additional reflections and closing comments

As my entire academic background relates to musicology, I started work on this thesis possessing

very limited knowledge on many of the topics I would eventually have to deal with. Partly, this was due to a conscious decision to use this opportunity to aquaint myself with some unfamiliar areas that I found intriguing. However, it was also in large part due to the fact that I found my focus shifting more and more away from music as the primary object of study. I gradually found that my interest in music in the context of narrative multimedia was more related to its overall function as an auditory component *alongside* nonmusical sounds when accompanying images, than to specific muscial features. Thus, the process and resulting thesis turned out quite different than I had initally envisioned. I have made a point to myself throughout to keep the perspective of the learner in mind, and from this viewpoint, I find it only appropriate to acknowledge some of the inherent flaws in my project as I have come to understand them over the course of this process.

One valid objection in particular is the lack of an empirical study. Although the decision not to conduct a data collection an subsequent analyis was made early on, I repeatedly found myself raising this concern during the process. Recounting my initial arguments against doing such a study, they mainly boil down to that I found it difficult to devise an experiment that would be sufficiently coherent with the macroscopic view I had in mind – while it is entirely possible that conducting a localized study in the context of a wide discussion would have proven both relevant and illuminating, I eventually decided against doing so. Another factor in this decision was that studies already existed that covered the most intuitively relevant aspects. There were two possibilites in particular that I initially considered, which relates to a point that I believe to be quite clearly demonstrated in this thesis: that the question of functionality in the context of a narrative involves both structural and affective aspects. Thus, I initially considered synchronization and emotional responses to be the most relevant topics for an empirical study. As such, these are obvious candidates for potential future efforts on my part.

It also strikes me that by introducing the more elaborate case studies towards the very end of the thesis, seeking to establish theory prior to attempting any analysis, I may have committed an error of judgement. It might, in some ways, have proven a more fruitful approach to reverse the process, or at the very least integrate these cases more closely in the discussion. However, this is a matter of hindsight, and thus only worthy of a brief mention.

Upon reviewing it at this final stage, I admittedly find the discussion presented in this thesis to be somewhat lacking in depth. Many areas and elements are covered only superficially, and the majority of points made througout are of a very general and basic nature. As such, I cannot say that I am entirely satisfied with the end result. Still, when considering the process as a whole, I do no lament this, nor my decision to adopt a macroscopic perspective.

References

Literature

Abbott, H. P. (2011). Narrativity. In P. Hühn, J. Pier, W. Schmid, & J. Schönert (Eds.), *The Living Handbook of Narratology*. Hamburg: Hamburg University Press. Retrieved from hup.sub.uni-hamburg.de/lhn/index.php?title=Narrativity&oldid=1580

Anderson, J. R. (1976). *Language, memory and thought*. Hillsdale, NJ: Erlbaum [u.a.].

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, *4*(11), 417–423.

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature reviews. Neuroscience*, *4*(10), 829–839. doi:10.1038/nrn1201

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press.

Berthoz, A. (2000). *The brain's sense of movement*. Cambridge, Mass. [u.a.]: Harvard Univ. Press.

Bregman, A. S. (1995). *Auditory scene analysis : perceptual organization of sound*. Bradford Books.

Chion, M. (1994). *Audio-vision : sound on screen*. (C. Gorbman, Trans.). New York: Columbia University Press.

Cohen, A. J. (2009). Music in performance arts. *The Oxford Handbook of Music Psychology* (pp. 441–451). Oxford; New York: Oxford University Press.

Cohen, A. J. (2010). Music as a source of emotion in film. *Handbook of Music and Emotion* (pp. 879–908). Oxford; New York: Oxford University Press.

Ekman, P. (1999). Basic Emotions. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). Chichester, England; New York: Wiley.

Emmott, C., & Alexander, M. (2011). Schemata. In P. Hühn, J. Pier, W. Schmid, & J. Schönert (Eds.), *The Living Handbook of Narratology*. Hamburg: Hamburg University Press. Retrieved from http://hup.sub.uni-hamburg.de/lhn/index.php/Schemata

Gabrielsson, A., & Juslin, P. N. (2003). Emotional expression in music. *Handbook of affective*

sciences (pp. 503–534). Oxford; New York: Oxford University Press.

Gaver, W. W. (1993). What in the World Do We Hear? An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, *5*(1), 1–29.

Giordano, B. L. (2003). Material categorization and hardness scaling in real and synthetic impact sounds. In D. Rocchesso & F. Fontana (Eds.), *The Sounding Object* (pp. 73–94). Firenze: Mondo estremo.

Godøy, R. I. (2010). Gestural Affordances of Musical Sound. *Musical Gestures: sound, movement and meaning* (pp. 103–125). New York: Routledge.

Godøy, R. I., Haga, E., & Jensenius, A. R. (2006). Playing "Air Instruments": mimicry of sound-producing gestures by novices and experts. In S. Gibet, N. Courty, & J.-F. Kamp (Eds.), *GW 2005, LNAI 3881* (pp. 256–267). Berlin Heidelberg: Springer-Verlag.

Godøy, R. I., & Leman, M. (Eds.). (2010). *Musical gestures : sound, movement, and meaning*. New York: Routledge.

Grassi, M., & Burro, R. (2003). Impact sounds. In D. Rocchesso & F. Fontana (Eds.), *The Sounding Object* (pp. 47–72). Firenze: Mondo estremo.

Haga, E. (2008). *Correspondences between Music and Body Movement* (Ph.D. thesis). Univeristy of Oslo, Oslo.

Halliday, M. A. K. (1978). *Language as social semiotic : the social interpretation of language and meaning*. Baltimore: University Park Press.

Jensenius, A. R., Wanderley, M. M., Godøy, R. I., & Leman, M. (2010). Musical Gestures: Concepts and Methods in Research. *Musical Gestures: sound, movement and meaning*. New York: Routledge.

Juslin, P. N. (2009). Emotional responses to music. In S. Hallam, I. Cross, & M. H. Thaut (Eds.), *The Oxford Handbook of Music Psychology* (pp. 131–140). Oxford; New York: Oxford University Press.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. doi:10.1037/0033-2909.129.5.770

Juslin, P. N., Liljeström, S., Västfjäll, D., & Lundqvist, L.-O. (2010). How does music evoke

emotions? Exploring the underlying mechanisms. *Handbook of Music and Emotion* (pp. 605–642). Oxford; New York: Oxford University Press.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(05). doi:10.1017/S0140525X08005293

Justus, T. C., & Bharucha, J. J. (2002). Music Perception and Cognition. In H. Pashler & S. Yantis (Eds.), *Steven's Handbook of Experimental Psychology* (3rd ed., Vols. 1-4, Vol. 1, pp. 453–492). New York: Wiley.

Kristeva, J. (1980). *Desire in Language: A Semiotic Approach to Literature and Art*. New York: Columbia University Press.

Kubrick, S. (1968). *2001: A Space Odyssey*. Science Fiction, MGM.

Köhler, W. (1929). *Gestalt Psychology*. New York: Liveright.

Köhler, W. (1947). *Gestalt Psychology* (2nd ed.). New York: Liveright.

Larsen, P. (2007). *Film music*. (J. Irons, Trans.). London: Reaktion.

Lipscomb, S. D. (1995). *Cognition of Musical and Visual Accent Structure Alignment in Film and Animation* (Ph.D. thesis). University of California, Los Angeles.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, Mass.: MIT Press.

Marshall, S., & Cohen, A. (1988). Effects of musical soundtracks on attitudes toward animated geometric figures. *Music Perception* (Vol. 6, pp. 95–112).

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. doi:10.1038/264746a0

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81–97.

Peretz, I. (2010). Towards a neurobiology of musical emotion. *Handbook of Music and Emotion* (pp. 99–126). Oxford; New York: Oxford University Press.

Rocchesso, D., & Fontana, F. (Eds.). (2003). *The Sounding Object*. Firenze: Mondo estremo.

Schaeffer, P. (1966). *Traité des objets musicaux : essai interdisciplines*. Paris: Éd. du Seuil.

Shepard, R. (1999). Cognitive Psychology and Music. *Music, cognition and computerized sound*. Cambridge, Mass.: MIT Press.

Sloboda, J. A., & Juslin, P. N. (2010). At the interface between the inner and outer world. *Handbook*

of Music and Emotion (pp. 73–97). Oxford; New York: Oxford University Press.

Snyder, B. (2000). *Music and memory : an introduction*. Cambridge, Mass. [u.a.]: MIT Press.

Sperling, G. (1963). A model for visual memory tasks. *Human factors*, *5*, 19–31.

Wingstedt, J. (2004). Narrative Functions of Film Music in a Relational Perspective. Presented at the ISME - Sound Worlds to Discover.

Wingstedt, J. (2005). *Narrative Music - Towards an Understanding of Musical Narrative Functions in Multimedia* (Licentiate Thesis). Luleå University of Technology, Luleå. Retrieved from http://epubl.ltu.se/1402-1757/2005/59/LTU-LIC-0559-SE.pdf

Wingstedt, J. (2008). *Making Music Mean: On Functions of, and Knowledge about, Narrative Music in Multimedia* (Ph.D. thesis). Luleå University of Technology, Luleå. Retrieved from http://pure.ltu.se/portal/files/2172525/LTU-DT-0843-SE.pdf

Wingstedt, J. (2010). Narrative media music - functions and knowledge. *Music, education and innovation: Festschrift for Sture Brändström* (pp. 53–66). Luleå: University Press. Retrieved from http://pure.ltu.se/portal/files/4928208/Festskrift.pdf#page=48

Wingstedt, J., Brändström, S., & Berg, J. (2010). Narrative Music, Visuals and Meaning in Film. Retrieved from http://pure.ltu.se/portal/files/4930736/Narrativemusic_shortened_final.pdf

Film

2001: A Space Odyssey. (1968). Kubrick, S. MGM.

From Russia With Love. (1963). Young, T. Eon Productions.

Once Upon a Time in the West. (1968). Leone, S. Paramount Pictures.

Memento. (2000). Nolan, C. Newmarket Capital Group.

Persona. (1966). Bergman, I. Svensk Filmindustri.

Sherlock Holmes. (2009). Ritchie, G. Warner Bros. Pictures.

The Dark Knight. (2008). Nolan, C. Warner Bros. Pictures.

The Man with the Golden Gun (1974). Hamilton, G. Action, Eon Productions.

The Usual Suspects. (1995). Singer, B. Polygram Filmed Entertainment.

Thunderball. (1965). Young, T. Eon Productions.

Watchmen. (2009). Snyder, Z. Paramount Pictures.

Television

Generation Kill. (2008). Boom.

Lost. (2004). Bad Robot.

Pingu. (1986). HIT Entertainment.

Twin Peaks. (1990). Lynch/Frost Productions.

Video games

Guitar Hero. (2005). Harmonix.

Indiana Jones and the Fate of Atlantis. (1992). LucasArts.

Maniac Mansion. (1987). LucasArts.

Monkey Island 2: LeChuck's Revenge. (1991). LucasArts.

Post Mortem. (2002). Paris: Microïds.

The Elder Scrolls V: Skyrim. (2011). Bethesda.

Software

iMUSE. (1990). McConnell, P., & Land, M. LucasArts.