The Metaphors of Sound : from Semantics to Acoustics. A Study of Brightness, Warmth, Roundness, and Roughness

Victor Rosi

Doctoral thesis from Sorbonne Université

Ecole Doctorale ED3C

The Metaphors of Sound: from Semantics to Acoustics

A study of Brightness, Warmth, Roundness, and Roughness

Author

Victor Rosi

Supervisor

Patrick Susini – STMS Lab, IRCAM-CNRS-SU, Paris

Co-supervisors

Olivier Houix – STMS Lab, IRCAM-CNRS-SU, Paris

Nicolas Misdariis – STMS Lab, IRCAM-CNRS-SU, Paris

Reviewers

Charalampos Saitis – C4DM, Queen Mary University, London

Sølvi Ystad – PRISM Lab, CNRS-Aix-Marseille University, Marseille

Examiners

Ophelia Deroy – CVBE, Ludwig Maximilian University, Munich

Bruno L. Giordano – La Timone Neuroscience Institute, CNRS, Marseille

Guest member

Mikhail Malt – STMS Lab, IRCAM-CNRS-SU, Paris

July 8, 2022

"Say it loud and there's music playing, say it soft and it's almost like praying."

— Stephen Sondheim, *West Side Story*

Abstract

The mysteries of the auditory world prompt us all to describe, as best as we can, what we hear. In some professional environments, the ability to verbally convey one's perception of sound qualities is crucial, whether you are a sound engineer, a musician, a sound designer, or a composer. Sometimes, talking about a sensation of any kind leads us to use metaphorical vocabulary. Thus, communication in the world of sound and music often depends on terms extracted from other sensory modalities like vision or touch. This is the case of four well-known attributes at the heart of this study, brightness, warmth, roundness, and roughness. But do we all have the same auditory sensation associated with such "extrasonic" concepts? To what extent are we able to faithfully describe a sensation expressed by these metaphors?

Brightness, warmth, roundness, roughness. The meaning of these terms used as sound attributes has been studied within the general framework of the semantic dimensions of sounds. However, the specific origins of such metaphorical terms and their mutual connections remain to be discovered. The aim of this study is to explore and expose the connection between these attributes and their projection in the sound domain. In other words, we aim to align their semantic definitions with mental representations expressed by their acoustic portraits. For each of the four attributes, we have reported on different layers of semantic descriptions that can be acoustic, metaphorical, or source-related. Through interviews and an online survey, we were able to develop definitions for each of the attributes based on the most relevant information from a population of sound professionals. However, the four terms depended on a lot of metaphorical elements that were still difficult to elucidate. To disambiguate these metaphorical descriptions, we asked three different expert populations (sound engineers, conductors and non-experts) to evaluate brightness, warmth, roundness and roughness in a corpus of orchestral sounds. We chose to use the new method of Best-Worst Scaling to fulfill that goal. This method allowed us to show that while some concepts transcend sound expertise, others can be specific to it. Gathering the data from the sound professionals brought forth a musical composition called

Quadrangulation – by Bertrand Plé – whose objective was to illustrate and transmit the meaning of the four concepts.

Through this interdisciplinary approach, we shed light on connections between our ability to understand a sound attribute's meaning and the mental representation associated with them. In addition, we uncovered potential incongruities between the perceptual projection of a metaphorical sound concept and the clarity of its definition. Finally, based on our results, we proposed a semantic explanation of the relations between the four concepts, thus inviting a better understanding of their use in professional conversations.

Résumé

Les mystères du monde sonore nous incitent tous à décrire, du mieux que nous pouvons, ce que nous entendons. Dans certains environnements professionnels, la capacité à transmettre verbalement sa perception des qualités sonores est cruciale, que l'on soit ingénieur du son, musicien, sound designer ou compositeur. Parfois, parler d'une sensation quelle qu'elle soit, nous amène à utiliser un vocabulaire métaphorique. Ainsi, la communication dans le monde du son et de la musique dépend souvent de termes extraits d'autres modalités sensorielles comme la vision ou le toucher. C'est le cas des quatre attributs au cœur de cette étude, la brillance, la chaleur, la rondeur et la rugosité. Mais avons-nous tous la même sensation auditive associée à de tels concepts "extrasonores" ? Dans quelle mesure sommes-nous capables de décrire fidèlement une sensation exprimée par ces métaphores ?

Brillance, chaleur, rondeur, rugosité. La signification de ces mots a été étudiée dans le cadre général des dimensions sémantiques du son. Cependant, la nature de leurs liens et leurs origines restent à découvrir. L'objectif de cette étude est d'explorer et d'exposer le lien entre ces métaphores et leur projection dans le domaine sonore. En d'autres termes, nous cherchons à aligner les définitions verbales de ces attributs avec les représentations mentales exprimées par leurs portraits acoustiques. Pour chacun des quatre attributs, nous avons fait état de différentes stratégies de descriptions sémantiques qui peuvent être acoustiques, métaphoriques ou liées à la source. Grâce à des entretiens

et à une expérience en ligne, nous avons pu formuler des définitions pour chacun des attributs à partir des informations les plus pertinentes provenant d'une population de professionnels du son. Cependant, les quatre termes dépendaient d'un grand nombre d'éléments tout aussi métaphoriques encore difficiles à élucider. Pour lever le voile sur ces descriptions métaphoriques, nous avons demandé à trois populations différentes d'experts (ingénieurs du son, chefs d'orchestre et non-experts) d'évaluer la brillance, la chaleur, la rondeur et la rugosité dans un corpus de sons orchestraux. Pour cela, nous avons choisi d'utiliser la nouvelle méthode de Best-Worst Scaling. Cette méthode nous a permis de montrer que si certains concepts transcendent l'expertise sonore, d'autres peuvent lui être spécifiques. La collecte des données auprès des professionnels du son a donné naissance à une composition musicale appelée *Quadrangulation* – du compositeur Bertrand Plé – dont l'objectif est d'illustrer et transmettre le sens des quatre attributs.

Grâce à cette approche interdisciplinaire, nous mettons en lumière les liens entre notre capacité à comprendre la signification d'un attribut sonore et la représentation mentale qui leur est associée. De plus, nous mettons en évidence les incongruités potentielles entre la réalité perceptive d'un concept sonore métaphorique et la clarté de sa définition. Enfin, sur la base de nos résultats, nous proposons une explication sémantique des relations entre les quatre concepts, invitant ainsi à une meilleure compréhension de leur utilisation dans les conversations professionnelles.

Acknowledgments

First of all, I wish to thank my supervision team. I am extremely grateful for their trust in offering me this remarkable research opportunity that allowed me to navigate between rich and diverse research topics. I thank Patrick Susini for having instilled a beautiful scientific rigor in me through always challenging questioning. I thank Olivier Houix for his support, his advice and his general help during the thesis. I thank Nicolas Misdariis for his supervision, his great management of the sound perception and design team, and his trust in the artistic project presented in this manuscript. Finally, I am deeply greatful to Pablo Arias, who, although he was not officially part of my supervision, has been a real mentor and source of many scientific insights that have allowed me to go much further than I would have imagined in this research.

I would like to thank *Fonds K pour la musique* which funded this thesis and allowed me to lead my research in great conditions at Ircam.

I wish to thank Charalampos Saitis and Sølvi Ystad for accepting the role of reviewer of this manuscript. I thank the members of the jury, Ophelia Deroy and Bruno Giordano who accepted to be examiners for my thesis defense. I am also grateful for the guidance of the members of my thesis committee, Geoffroy Peeters and Caroline Traube. Finally, I thank some inspiring scientists around the world who offered me their time and knowledge: Charalampos Saitis, Geoff Hollis, Svetlana Kiritchenko, Lindsey Reymore, and Eduardo Coutinho.

Ircam has been a wonderful institution for leading doctoral research. I would like to thank many people who helped me throughout my research with a lot of patience: Sylvie, Meryem, Alexandra, Patricia, Eric, Murielle, Cyrielle, Cristina, Anne-Marie, and Viktoriya. I also thank the great family of doctoral and post-doctoral students who helped me in every way during

my stay at Ircam and whom I now count among my friends: Axel, Constance, Clément, Léane, Lenny, Loïc, Théis, Vincent, Yann.

I am extremely grateful to the whole sound perception and design team for this great working environment and the strong camaraderie that resided there. More particularly, I thank my friends and brilliant doctoral students fellows for their support, their advice and their trust: Baptiste, Claire, Nadia, Valérian. I also thank Aliette, whom I had the chance to supervise for her master internship, and Emmanuel for his essential advice on on life after the thesis. I also thank all the great people who have undergone my experiments and interviews with patience and sympathy.

During my last year of thesis, I had the incredible opportunity to record a musical work based on my research. That is why, I would like to thank first of all the composer Bertrand Plé for his trust, the exciting work sessions, and of course the piece he composed. But this work would not have been possible without Ircam's production department. I would like to thank Jérémie, Clément, Aline, Cyril, Eric, and Quentin for their involvement in this project. I also thank all the musicians present for this project, Diane, Hélène, Simon, Luce, Iris, Solène, and Quentin.

Finally, I would like to thank a thousand times my parents Alexandre and Marie-agnès who have been of an unequalled support and kindness before and during this thesis as well as my dear sister Valentine who has created visual graphics for my thesis defense.

My final thanks go to my dear friends, Aurore, Théo, Timothée, Fanny, Ludo, Wilhem, Julien, Hélène, Mehdi, Maxime, Manon, Eve, Antonin, Tarik (and others that I may have forgotten) for their legendary patience during this period where I made them suffer a lot of complaints and anxiety episodes. Fortunately, thanks to them (and those I did not mention), I was able to go through with it.

Contents

1. Introduction

Foreword

In this thesis, we studied the semantics, acoustics, and mental representation of a metaphorical vocabulary commonly used by musicians, sound engineers or sound designers. We focused on pillars of the sound-specific terminology: **brightness**, **warmth**, **roundness**, and **roughness**. In this first chapter, I will introduce the variety of research that allowed us to build the theoretical framework to study the metaphorical vocabulary of sound. Specifically, I will present works that have explored functions of this vocabulary, contexts of use, its origins, and how it is expressed perceptually and acoustically.

First, I will describe the building blocks of sound perception and psychoacoustics (section 1.1). I will introduce a definition of timbre, its links to different listening modes and to the physics of sounds. Then, I will report on the perceptual dimensions of timbre.

Second, in section 1.2, I will present the diversity of sound semantics as studied from different experimental perspectives. Prior to that, I will explain some relevant notions of semantics and categorization for the study of this vocabulary. Then, I will introduce the salient semantic categories manipulated by sound experts and the most shared sound descriptions. Finally, I will report on the presence of metaphorical descriptors of timbre.

Third, I will clarify the notion of metaphorical sound description in section 1.3. Specifically, I will focus on descriptions that call upon other sensory

concepts and discuss the nature of their mental representations, their developments and origins. Importantly, this connection with other senses such as vision and touch resonates with the notion of crossmodal correspondences, i.e., the conceptual links between two sensory modalities. Thus, I will present theories on its levels of representation and the processes at the origin of this vocabulary.

Fourth, in the section 1.4, I will review studies that bridge the gap between the vocabulary, timbre perception and acoustic properties. We will start by evoking the methods commonly used to evaluate the perceptive qualities of timbre. Then, after having exposed some attempts to model timbre semantics, along with an acoustic characterization, I will present a state of the art of research exploring the meaning of the four concepts we wish to study, namely, brightness, warmth, roundness and roughness.

Finally, I will speculate on the benefits of a sound lexicon and a strategy to define metaphorical sound attributes. Importantly, I will depict a first attempt of a sound lexicon including metaphorical descriptors that is at the origin of the present study (Carron et al., 2017).

At the end of this journey, I will present the objectives of the thesis, which will be intimately linked to the purpose of defining the four sound concepts.

1.1 Sound perception

1.1.1 Basic knowledge on psychoacoustics

A sound is classically expressed along four perceptual dimensions, namely duration, loudness, pitch, and timbre. Through the lens of psychoacoustics – a branch of psychophysics that studies the relation between human auditory perception and sound properties – researchers have thoroughly been investigating these perceptual attributes of sounds. On the one hand, some attributes have been studied in depth and are well understood. Thus, loudness, pitch, and duration are psychoacoustical attributes for which definitions, computational models, and experimental methods are easily accessible (Moore, 2012; Zwicker and Fastl, 2013). For instance, loudness models have been standardized (Zwicker and Scharf, 1965) and revised (Moore and Glasberg, 1996; Moore et al., 1997). On the other hand, because they cannot be easily and unequivocally specified to a listener, timbral qualities requires indirect and multidimensional methods to be fully explored.

1.1.2 Three listening modes

Humans listening abilities have been theorized in three principal strategies: *causal listening*, *reduced listening*, and *semantic listening*.

The *causal listening* (Chion, 2019), or everyday listening (Gaver, 1993), is used for the identification of sound sources. Indeed, it is what allows listeners to recognize the voice of their loved ones, the sound of a clarinet among other orchestral sounds, or the classic sound of a Harley-Davidson motorbike engine (at least for the motorcycle-savvy).

Reduced listening is based solely on the intrinsic properties of the sound. This listening mode is independent of the source of the sound, its production mode, or its meaning in a specific context. Schaeffer formalized the idea of reduced listening to any kind of sounds, and proposed to decline it into a group of typo-morphological criteria mainly based on *facture* (i.e., the overall envelope of the sound) and mass (i.e., the spectral content) (Schaeffer, 1966).

This definition of reduced listening can be interpreted as expressing a sound through spectral and temporal qualities.

Finally, the *semantic listening* (Chion, 2019) corresponds to the listeners' interpretation of a sound as a code such as the the horn of a car that can translate a driver's impatience, or the vocal attitude of a speaker that can express friendliness or dominance.

The three listening modes are thouroughly detailed in Carron et al. (2017). The present work focuses on the concept of timbre, that is involved in the first two modes of listening.

1.1.3 The definitions of timbre

The notion of timbre originates from western musical tradition. Its definition has been widely discussed by musicians and auditory psychologists for years. Yet, there is still no single widely-accepted definition of it (Hajda et al., 1997). The most general way to understand timbre might correspond to this formulation by Risset and Wessel (1999):

> Timbre is a quality of sound. It is the perceptual attribute that enables us to distinguish among orchestral instruments that are playing the same pitch, and are equally loud. (Risset and Wessel, 1999) p. 113

This definition implies two uses of timbre which are essential to *causal listening*: the identification of a sound source, and the distinction between two sound sources. According to the ANSI definition of timbre, summarized in Krumhansl (1989), timbre is "the way in which musical sounds differ once they have been equated for pitch, loudness and duration". Once sounds have been "equalized" on pitch, loudness and duration, they can still be perceived as different due to their timbre. But coherently with the *reduced listening*, timbre can be evaluated on the basis of its intrinsic qualities, independently of the recognition of the sound. I will now report on how this understanding of timbre led to reveal its multidimensional perceptual and acoustic representations.

1.1.4 Perceptual and acoustic dimensions of timbre

In the last few decades, several works have sought to reveal the most salient perceptual and acoustic dimensions of timbre by means of the analysis of timbre spaces (Plomp, 1970; Wessel, 1979). A timbre space is classically obtained thanks to a Multidimensional Scaling (MDS) analysis (Shepard, 1962) of judgments of dissimilarity between pairs of sound. The MDS analysis generates a spatial configuration of sounds whose pairwise distances approximate the original perceptual dissimilarity (Winsberg and De Soete, 1993).

The three-dimensional timbre space provided by Grey (1977) paved the way for many studies seeking to uncover the perceptual dimensions of timbre and find their acoustic interpretations. The experiment consisted of asking participants to rate the similarity between 16 synthetic musical instrument sounds presented in pairs. Stimuli were created with additive sound synthesis of musical instruments (e.g., oboe, saxophone) and equalized in duration, pitch and loudness. These judgements were then summarized in a dissimilarity matrix. Then, by using INDSCAL[1] (Carroll and Chang, 1970) – an MDS analysis technique – the dissimilarity judgements were modeled as distances in a Euclidean space with dimensions expected to be the principal perceptual dimensions shared by the sounds. In this perceptual space, a large dissimilarity is represented by a large distance. Eventually, Grey identified three salient dimensions shared by the sound corpus.

In addition, Grey (1977) proposed an acoustic interpretation of each perceptual dimension on the basis of the spectral analysis of the sound corpus. The first dimension is correlated to the distribution of the energy spectrum, the second one to the synchronicity of the harmonics at the moment of attack and the third one to the presence of energy in high frequencies. At the end, both temporal and spectral (or spectrotemporal) features explained the perceptual results of this experiment on timbre.

[1] INDSCAL: INdividual Differences SCAling

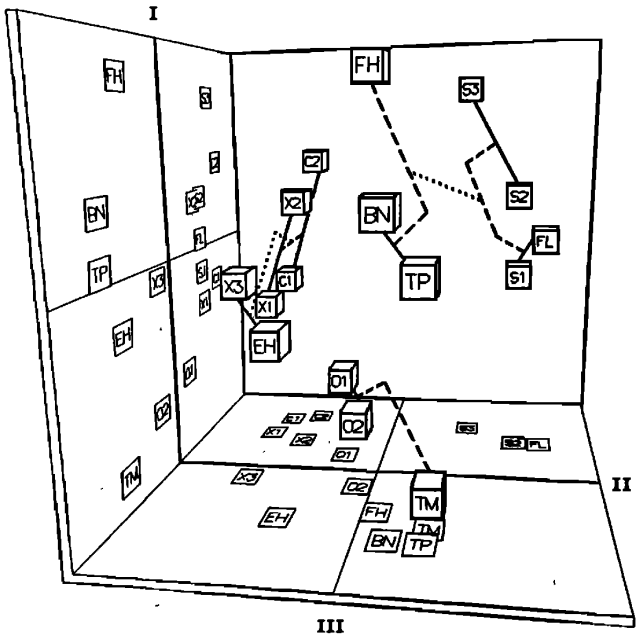


Figure 1.1: Three-dimensional timbre space obtained in Grey (1977) (Figure 1). BN: Bassoon; C1,C2,C3: Clarinet; EH: English horn; FL: Flute; O1,02: Oboes; TM: Trombone; TP: Trumpet; X1,X2,X3: Saxophones.

Later, several works have taken advantage of signal processing and mathematical improvements to address new issues related to the perceptual information provided by timbre spaces. Classically, these studies took up and extended the experimental protocol as Grey's (Eerola et al., 2012; Lakatos, 2000; McAdams et al., 1995; Susini et al., 2004). For instance, McAdams et al. (1995) wanted to verify that a hybrid sound created from a pair of instruments (e.g., Trumpar: trumpet/guitar) was located between the two instruments concerned in the perceptual space. In addition, they wanted to account for differences in group expertise using CLASCAL analysis (Winsberg and De Soete, 1993). Alongside these objectives, McAdams et al. (1995) revealed the acoustic correlates of the three dimensions of the timbre space they obtained as being related to the attack time, the spectral centroid (SC) (Krimphoff et al., 1994), and the spectral flux (i.e., temporal variation rate).

Subsequent studies (Caclin et al., 2005; Lakatos, 2000) also evaluated that the two main dimensions of timbre depended on the SC and the attack

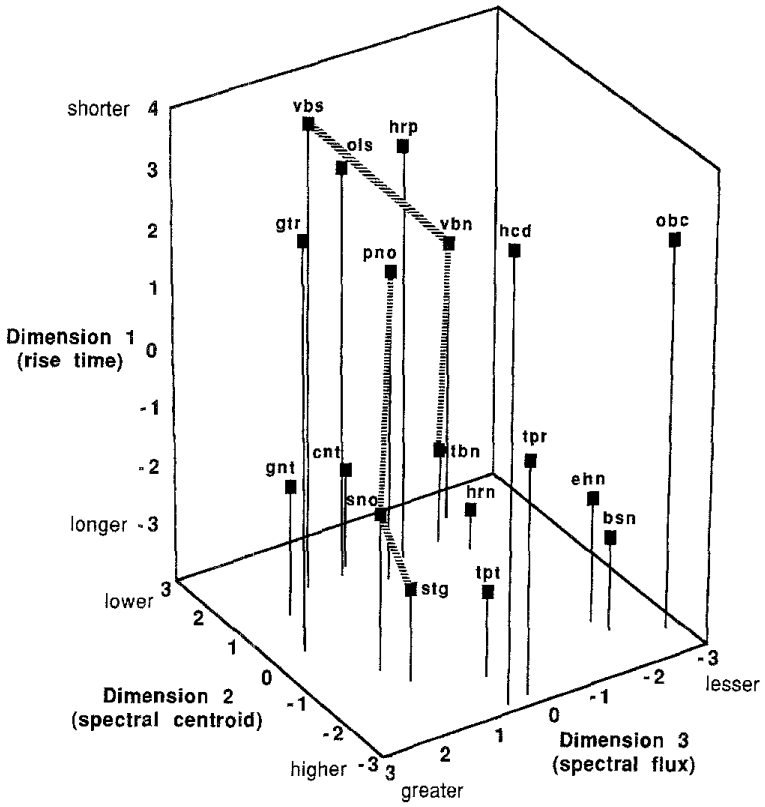


Figure 1.2: Three-dimensional timbre space obtained in McAdams et al. (1995) (Figure 1). bsn: Bassoon; cnt: Clarinet; ehn: English horn; gnt: *Guitarnet* (guitar/clarinet); gtr: Guitar; hcd: Harpsichord; hnr: French horn; hrp: Harpa; obc: *Obochord* (oboe/harpsichord); ols: *Obolest* (oboe/celesta); pno: Piano; sno: *Striano* (strings/piano); stg: Strings; tbn: Trombone; tpr: *Trumpar* (trumpet/guitar); tpt: Trumpet; vbn: *Vibrone* (vibraphone/trombone); vbs: Vibraphone.

time. Moreover, in a meta-analysis of timbre through different typologies of sounds, Misdariis et al. (2010) revealed that SC was a preponderant acoustic feature for sound description. From a similar perspective, a study reported on the similarities and specificities of the acoustic and perceptual results of 17 studies of MDS-generated timbre spaces (Thoret et al., 2021). A main conclusion is that while there are generic acoustic correlates of timbre that allow connections between timbre perception studies, a part of them remain specific to the results of one experiment. In the end, some timbral dimensions of sound seem to be sound specific and contextual.

It is important to note that timbre is not completely impervious to other perceptual attributes. For example, several studies have shown an interaction of timbre with pitch (Allen and Oxenham, 2014; Marozeau et al., 2003).

It seems clear that timbre is a multi-dimensional characteristic of sound. Its main perceptual dimensions are consistently related to the spectral centroid and the attack time of a sound. However, it is likely that these two dimensions are not sufficient to summarize the diversity of known sounds, especially in the musical world. In fact, several studies have reported the richness of sound descriptions, indicating a great diversity of sound perceptual qualities. There is therefore a challenge in bringing together this vocabulary with the perceptual and acoustic dimensions of sounds.

1.1.5 Section summary

Sound perception

In this section I presented three listening modes, namely, *causal listening*, *reduced listening* and *semantic listening*. Importantly, causal and reduced listening are intimately dependent on a perceptual sound quality called **timbre**.

Timbre is a multidimensional characteristic of sounds which has often been evaluated in the past based on three perceptual dimensions. In addition, several studies have subsequently attempted to explain these dimensions with acoustic quantities such as the spectral centroid and the attack time of a sound.

In the next section we will dive into the richness of the vocabulary related to these modes of listening and to timbre. I will report on different methods of approaching this vocabulary by first specifying the potential cognitive mechanisms that are responsible for it. Then, I will explore the variety of descriptions of sounds that we will study further in this thesis.

1.2 Communication and verbal sound descriptions

As human beings, we are used to verbally characterizing what we perceive through our five senses, and hearing is no exception. One can say that an alarm is too loud, or that a baby's crying is unpleasant. Naturally, sound professionals such as composers or sound engineers use a rich and technical vocabulary to describe sound properties and communicate in working contexts.

I have mentioned different modes of listening in section 1.1.2. While the type of sound description from causal listening seems quite intuitive – identification of the source and/or identification of a mode of excitation – reduced listening implies description of sound attributes or qualities that fall within the field of timbre semantics. In the next section, I will present the framework of cognitive semantics and the classical techniques of semantic studies.

1.2.1 Notions of cognitive semantics

To understand the linguistic mechanisms involved in assigning meaning to sound description, we need to introduce the link between the notions of semantics with the cognitive mechanisms of categorization, and perception.

Semantics and cognition

Semantics is a linguistics notion that is concerned with meaning of words, phrases, grammatical forms and sentences (Löbner, 2013). Cognitive science applied to semantics is concerned with how the human mind works, how it receives information from the environment via the senses and processes this information, recognizing what is perceived, comparing it to former data, classifying it and storing it in the memory. Language plays a central role in these theories. On the one hand, speech perception and production and the underlying mental structures are major objects of investigation. On the other hand, the way in which we use language to express what we 'have in mind'

can tell much about how the human mind stores information. In the sound domain, it can translate elegantly into a quote from Chion (2019): "We hear as we speak".

Figure 1.3 depicts a cognitive version of the semiotic triangle between an expression (i.e., a label), a concept, and a referent (or referents), i.e., a category. In the process, the expression calls upon the meaning of a concept. The concept determines the referents/situation relative to what is expressed, i.e., all characteristics or set of items pointed by a concept and that an expression refers to. In other words, it refers to the definition of its categorization, the cognitive mechanism at the origin of the modeling of a concept.

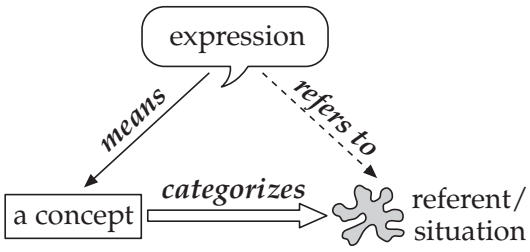


Figure 1.3: The cognitive understanding of semantics (Löbner, 2013).

Categorization

As defined in Smith (1989), a concept is “a mental representation of a class or individual and deals with what is being represented and how that information is typically used during the categorization” (p.509). Thus, through a categorization process, one expresses a concept/mental representation which may be carried out semantically (cf. previous section) and be based on attributes grounded in the perceptual domain Goldstone et al. (2013).

According to its Aristotelian definition, categorization is the act of classifying ideas and forming categories underlying one’s representation of a concept expressed or not by a linguistic utterance. It is a question here of defining the members of a category in accordance with necessary and sufficient conditions. A classic example would be to consider that the category of the utterance "string instrument" contains the classes "strings", "bridge", and "fingerboard".

In this case, the boundaries of a category are very clear and highlight a process based on the comparison of members of different categories.

The **prototype theory** (Lakoff, 2007; Rosch, 1975) argues that a category is defined by prototypes, i.e., a member item that comes directly to mind when talking about a category. Without a necessary condition, category membership is then a question of similarity with the prototype. Categories can be thought like probabilistic distributions that are maximal for prototypes and minimal for items that are not part of the category. Unlike its other definition, categorization by prototype bias does not have clear boundaries. For example, "violin" is a better prototype than "piano" – which is not considered a member of the string instruments organologically speaking – for the category "string instrument", and "trumpet" is too different from a violin to be considered in that category.

Another strategy of categorization is based on the exemplar theory. When presenting an item never seen before, one evaluates its similarity to all the members of a category rather than to the most representative item (i.e., a prototype). This corresponds to answering the question "is item X more similar to members of category A or to members of category B?".

Categorization represents the interface between a concept, its label as expressed linguistically, the perception, and the world. It involves the identification of representative exemplars or attributes that gives access to a concept. Depending on the ease of categorizing an item like a sound, its semantics representation will be more or less clear.

In the following section, I will introduce sound descriptors (or expressions) denoting concepts expressed by sound attributes that can be of abstract origin. Further on, I will show that the categorization strategies linked to sound description can vary (see section 1.2.3), coherently to listening modes. To introduce them, I will first depict studies that investigated the diversity of sound-related vocabulary. More precisely, I will highlight different approaches to reveal the concepts emerging from the discourse related to the description of sounds.

1.2.2 Exploring the terminology of sound

Many previous research dealing with sound semantic sought to reveal structural semantic categories, or prototypical and consensual descriptors of sound through a qualitative analysis of verbal data. In a qualitative analysis, some believe that the researcher should analyze all verbal or textual data without any assumptions, while others think the researcher should enter the field with their hypotheses in mind (Strauss and Corbin, 1994). Depending on the amount of data available and the research objectives, it is possible to refer to observations already made in the literature (prior knowledge), or to make semantic categories emerge only from the text data.

In this section I will show different approaches on how to understand the scope of sound description. First, I will provide an overview of sound semantics research highlighting the diversity of the vocabulary. Second, I will present the methodological focus of psycholinguistics, which aims to group this vocabulary into semantic categories conveying specific concepts used for different interactions with sounds (e.g., playing a musical instrument, soundscape). Third, I will name some interesting perspectives for revealing semantic categories of sound description.

Free verbalization

Free verbalizations are often the simplest way to have access to the most consistent descriptions of the timbre of musical instruments such as the piano (Bellemare and Traube, 2005; Cheminée, 2009), the organ, the guitar (Paté et al., 2015; Traube, 2004) or the violin (Saitis et al., 2017; Stepánek, 2006). In the same vein, an experiment conducted by Faure (2000) asked musicians and non-musicians to compare the timbre of the sounds used by McAdams et al. (1995) to unravel the semantic aspects of sound description. She compiled a very comprehensive list of descriptions from the most basic descriptions (e.g., «loud», «high-pitched») to extrasonic descriptions (e.g., «bright»). More recently, a study has provided a model for musicians' shared representation of Western musical instruments, based

on consensual and relevant verbal descriptions of musical timbre *qualia*[2] (Reymore and Huron, 2020). Researchers' interpretations led to a final 20-dimensional timbre *qualia* model including intensity-related dimensions (e.g., "soft", "smooth", "singing"), frequency-related dimensions (e.g., "rumbling", "deep", "shimmering", "bright"), harmonicity-related dimensions (e.g., "pure", "clear","airy/breathy"), and other acoustical features of sounds (e.g., "hollow","muted/veiled", "open"). However, research investigating sound descriptions through free verbalizations is not limited to the world of musical sounds. Free speech analysis has also been used for sonar sounds (Solomon, 1958), soundscapes (Dubois et al., 2006; Guastavino, 2003) and helicopters (Namba et al., 1991). Overall, many of the most important verbal descriptors are linked to extrasonic qualities on the semantic level like *round*, *bright*, *warm* or *velvety*.

Psycholinguistic approaches

The psycholinguistic approach (Dubois, 2000) has revealed itself to be a robust method to assess the semantic categorizations done by musicians when evaluating their instruments (Cheminée, 2009; Lavoie, 2014; Paté et al., 2015; Saitis et al., 2017). According to a definition from (Dubois et al., 2006) psycholinguistic techniques consist in :

> [...] discourse analysis techniques [that] aim at deriving relevant inferences about how people process and conceptualize sensory experiences [...]. The psycholinguistic analysis mediates between individual experiences and collective representations shared in language and elaborated as knowledge. (Dubois et al., 2006) p. 866

By getting as close as possible to the usual verbalization context of the participants, these methods aim to harness ecological conditions of sound descriptions. It is a rigorous method that seeks to understand the links between linguistic expressions based on grammatical attributes and linguistic rules

[2] the "phenomenal character" of a given sensory stimuli

involved (e.g., association, opposition, elaboration). Eventually, researchers reveal complex concepts that are supposed to be the building blocks of the relevant information contained in the data. A category is usually named after the most salient description it includes. For instance, Paté et al. (2015) led a psychological investigation on the influence of a guitar fingerboard wood on guitarists' perception. Through a free verbalization task with guitarists playing the guitars and a linguistic analysis of the verbalizations, they reported on the psychological descriptors that are relevant for the discrimination of the wood of the fingerboard, i.e., *precision* (referring to how each note stands out from others), *attack* (referring to the guitar's response to musician's gesture) and *balance* (referring to the frequency content). Interestingly, in many of these studies, the semantic definition of a category interweaves descriptions of various semantic nature. For example, one of the salient semantic categories used by violin player describing violin properties was *richness* (Saitis et al., 2017). Among other descriptions, the category of *richness* brought together metaphorical (e.g., "deep"), acoustic (e.g., "many harmonics") and affect-related elements (e.g., "expressive"). Overall, we observe that the description of sounds depends on different types of strategies.

Methodological prospects

Many qualitative analysis of verbal data presented above mainly relied on qualitative analyses. However, sound semantics research sometimes incorporated Natural Language Processing (NLP) techniques in its workflow that enable to build semantic categories out of the most occurring concepts from a text or free verbalizations (Faure, 2000; Wallmark, 2019b). Basic NLP steps consist of cutting the text into words (tokenization), sometimes grouping them by ethymological root (lemmatization) and then evaluating them on the basis of their frequency in the data (e.g., Type-token ratio). Researchers aiming to create semantic categories can then rely on their expertise, or evaluate the agreement (e.g., Cohen's kappa, Fleiss' kappa) between participants taking part in a classification task or sorting task of the verbal data.

To date, however, a number of NLP technologies based on automatic methods have been developed to evaluate salient semantic categories extracted

from text or speech. For instance, one can use automatic topic modeling techniques, such as latent semantic analysis (LSA) (Landauer et al., 1998) or Latent Dirichlet allocation (LDA) (Blei et al., 2003). Today's NLP approaches are based on the generation of semantic spaces enabled by models such as *BERT* (Devlin et al., 2018), *GloVe* (Pennington et al., 2014) or *word2vec* (Mikolov et al., 2013). Classically, *word2vec* is an algorithm that generates a semantic space by making use of the information on co-occurences of words in a text. In these multidimensional spaces, words are represented as vectors and the distance between two words is a semantic distance. For example, the vector operation $(king - man + woman)$ logically leads us to the neighbourhood of the vector $queen$. Although very popular, these technologies are not suitable for all issues as they need a large corpus of text to generate a semantic space. Indeed, to our knowledge, no timbre semantics study has employed such tools for the analysis of sound description and they rely on more classical qualitative analyses. This methodological reflection is a possible perspective of analyses for timbre semantic studies. We tried to set up such framework for the study 2.1 but it was unsuccessful due to the size of our dataset and the polysemy of several sound descriptions.

1.2.3 Description strategies of sounds

I introduced the fact that descriptions of sounds can take on many different semantic natures (see section 1.2.2). A line of research on timbre semantics investigated the strategies underlying sound description to uncover the habits of use of timbre and sound description. Here, we are specifically interested in the manner of describing sounds, i.e., the types of sound description are used predominantly in a certain context, rather than revealing the concept or sound attribute that is signified.

Based on free verbalizations of musicians and non-musicians, Faure (2000) proposed classes of sound descriptions of different types. In the experiment, musician and non-musician participants were asked to compare sounds from the timbre space of McAdams et al. (1995) in pairs. For each pair only descrip-

tions of the form "Sound (A) is more X than sound (B)" were retained. Overall, she reported fairly basic descriptions of sound (e.g., "high-pitch", "loud"), as well as more specific extrasonic descriptions (e.g., "natural", "bright"). Thus, she proposed eight categories of sound descriptions:

- *temporal* (e.g., "resonant", "disappearing", "continuous")
- *sensation-perception* (e.g., "high-pitched", "sonorous", "warm", "hollow", "bright")
- *shape* (e.g., "closed", "round", "large")
- *emotion-value* (e.g., "pleasant", "aggressive", "musical")
- *source-related qualities* (e.g., "metallic", "brassy", "synthetic")
- *intensity-energy* (e.g., "loud", "distant", "vivid")
- *sound name & imitation* (e.g.,"noise", "scream", "friction")
- *general* (e.g., "pure", "rich", "clean")

The *general* class corresponds to verbalizations that could not be placed elsewhere. Surprisingly, some of the proposed categories mixed descriptions of different nature. For instance, the *sensation-perception* category included descriptions related to sensory modalities, thus not distinguishing between psychoacoustic descriptions (e.g., "high-pitched") and descriptions related to other senses (e.g., "bright").

Porcello (2004) investigated the sound description strategies on recordings of conversations between sound engineers during recording and mixing sessions. One of these sessions involved a music producer and a recording engineer debating on drums sounds. Both speakers developed contextual verbal references in order to communicate on the nature of drum sounds. Porcello's study revealed the diversity of strategies available for communicating sound properties:

- *singing/vocables* (e.g., "hm", "pts", "dz")
- *lexical onomatopoesis* (e.g., "hollow", "ring", "muffling")
- *'pure' metaphors* (e.g., "tight", "deep")
- *association* (i.e., stylistic descriptions)

- *evaluation* (i.e., used to establish a mutual sense of solidarity between the two interlocutors, to mark a territory of shared musical aesthetics)

The vocabulary specific to orchestration for ensemble music is a good indicator of the diversity and descriptions of the sound and timbre of musical instruments. In a study on the semantics of timbre, Wallmark (2019a) analyzed the descriptions of the timbre of orchestral instruments through 11 orchestration treatises, some of which are still studied today, such as treatises by Hector Berlioz, Charles Koechlin, and Samuel Adler. As an index of lexical diversity, i.e., the relative level of redundancy or novelty of word choice, they used a type-token ratio (TTR), or ratio of unique words to the total word count. Then, through inter-rater agreement procedure consisting in the categorization of the 50 most recurrent adjectives, the author proposed seven categories for timbre description:

- *affect* (e.g., "rich", "expressive", "powerful")
- *matter* (e.g., "full", "round", "deep")
- *CMC*[3] (e.g., "bright", "soft", "sweet")
- *mimesis* (e.g., "voice-like", "breathy", "nasal")
- *action* (e.g., "penetrating", "piercing", "open")
- *acoustics* (e.g., "sonorous", "resonant", "shrill")
- *onomatopeia* (e.g., "buzzing", "rattling", "clicking")

Ultimately, by means of principal components analysis, Wallmark reduced the seven categories to three dimensions of musical timbre conceptualization: *material* (loaded positively onto onomatopoeia and matter), *sensory* (crossmodal and acoustics), and *activity* (action and mimesis).

Finally, a study investigated the description of the sound of products by consumers, such as hair dryers or toothbrushes (Özcan and Egmond van, 2012; Özcan and van Egmond, 2005). Although the participants were not sound professionals, we nonetheless identify many similarities with previous categorizations. This time, 11 categories were proposed:

[3] Crossmodal correspondence

- *actions* (e.g., "blowing", "moving", "cleaning")
- *emotions* (e.g., "annoying", "warm", "acceptable")
- *location* (e.g., "bathroom", "outside", "house")
- *material* (e.g., "plastic", "wooden", "metal")
- *onomatopeia* (e.g., "brr", "kling", "buzzing")
- *psychoacoustics* (e.g., "soft", "amplified", "high-pitched")
- *sound type* (e.g., "metallic", "synthetic", "electronic")
- *source* (e.g., "beep", "hair dryer", "door")
- *source properties* (e.g., "heavy", "old", "cold")
- *temporal* (e.g., "short", "long", "continuous")
- *meaning* (e.g., "time is up", "it is ready" for the sound of a ringing bell)

Interestingly, they added a category of *meaning* grouping descriptions of semantic information from a sound, which is exactly the manifestation of *semantic listening*.

The description strategies proposed by these four studies show both similarities and differences. For instance, all studies reported on descriptions that correspond to an imitation of the sound heard. Moreover, three of the four studies reported categories associated with emotional judgments on sounds. However, some descriptions were not integrated in the same way into categories. For example, the description "pure" is evaluated as a metaphor according to Porcello, as an evaluation of the value of a sound according to Wallmark, and unclassifiable according to Faure. Precisely such metaphorical descriptions were present across the categories of all studies.

Some description strategies echo the listening modes presented in 1.1.2. We have for example, descriptions of the sources which correspond to the *causal listening*. In addition, there were descriptions related to the intrinsic and acoustic qualities of the sounds, resonating this time with *reduced listening*. We note that this vocabulary is eminently preferred by expert populations of the sound (Lemaitre et al., 2010). But the question then arises as to what many of the metaphorical descriptions of sounds mean.

Classifying or define the metaphorical vocabulary of sound seems unclear. We found for example that the description "warm" belongs to categories related to other sensory modalities in Faure (2000) and Wallmark (2019a), but it was referenced as an emotional evaluation of sound in Özcan and van Egmond (2005). We thus have an expression of the polysemia of a metaphorical description that Porcello characterized as vague:

> "In making available to others through language what our ears hear, one might be left with the impression that there is very little alternative other than to use similarly vague metaphorical descriptions (for example, 'warm', 'bright', 'boomy'). But 'vague metaphorical descriptions' would of course prove insufficient as linguistic tools in a workplace defined by sound-creating and -manipulating technologies, and where the goal of work is to control and craft sounds with great precision." (Porcello, 2004) p. 734

Clearly, such descriptor has a meaning that can vary from one application to another, or from one expertise to another. However, despite the fact that Porcello criticizes its merits, it is clear that its use is necessary given its ubiquity in expert vocabulary.

Table 1.1 depicts description strategies that are common for at least two of the four studies with some examples. In the end, we can see that some categories including metaphorical descriptions (e.g., *psychoacoustic*, CMC, "pure" metaphor) intersect and are not classified in a consistent way, hence reinforcing the fuzziness of its meaning.

Table 1.1: Shared description strategies between sound semantic studies. The *Comparison* category refers, to the timbre similarity of a sound with sound sources different from the real one.

Category description	Study	Examples
Vocal imitation of a sound	Porcello (2004)	"buzzing", "hm"
	Wallmark (2019a)	"clicking"
	Özcan and van Egmond (2005)	"kling"
Affect/Value judgment	Faure (2000)	"pleasant"
	Wallmark (2019a)	"rich"
	Özcan and van Egmond (2005)	"annoying"
Sensory metaphor	Faure (2000)	"bright" (vision)
	Wallmark (2019a)	"warm" (touch)
Source description	Faure (2000)	"brassy"
	Özcan and van Egmond (2005)	"door", "wooden"
Comparison	Faure (2000)	"scream"
	Wallmark (2019a)	"voice-like"
Temporal description	Faure (2000)	"resonant"
	Özcan and van Egmond (2005)	"short/long"

1.2.4 Section summary

Communication & verbal sound descriptions

In this section, I presented different approaches to account for the richness of the technical vocabulary related to sound. In order to have an intuition on the generation and description of this vocabulary, I proposed a formalization of semantics from the perspective of cognitive science. In this framework, the meaning of sound descriptions is based on multiple categorization processes that report on the targeted concept through sound prototypes, and sound attributes.

Based on verbal sound description (free verbalizations), several studies intended to decipher the terminology of sound and proposed semantic models. These studies employ several methods included in a general process of qualitative analyses of verbal or text data which can call upon more or less sophisticated NLP tools. From a psycholinguistic point of view, the use of a sound vocabulary may be organized in semantic categories that are specific to the application and type of studied sounds.

More broadly, the vocabulary of sound consistently spans several description strategies whatever the context of expression. Generally, there are descriptions of the mode of production of a sound close to a prototypical understanding of the categories of sounds and causal listening, but also descriptions focused more essentially on the properties of the sound, on acoustic, psychoacoustic, emotional and metaphorical aspects.

Clearly, metaphorical sound descriptions are an essential tool for technical communication, in spite of its semantic ambiguity and polysemy. We are therefore curious to know its origin and the channels that would ensure that two experts in a conversation evoke the same type of concept when metaphorically describing a sound. In the next section, I will specify the nature of these metaphorical descriptions, the type of conceptual processes associated with them, and explore potential origins of their development.

1.3 Origins and shared aspects of metaphorical sound attributes

1.3.1 Why metaphorical?

According to its linguistic definition, a **metaphor** includes any concept, notion, model or image from one domain, the source domain, borrowed to describe things belonging to another domain, the target domain (Löbner, 2013). Moreover, the conceptual metaphors theory (Lakoff and Turner, 2009) indicates that a metaphor is central to thought and therefore to language. Here are some basic principles derived from this theory:

- metaphors structure thoughts
- metaphors structure knowledge
- a metaphor is at the heart of abstract language
- a metaphor comes from physical experience
- a metaphor is ideological

Understanding what metaphorical descriptions denote in a target domain corresponds to identifying the object and the type of categorization that occurs when using it (see 1.2.1). While categorization is quite simple when it deals with tangible and concrete concepts with obvious attributes or exemplars such as animals or musical instruments, it becomes more complex when it is about retrieving the meaning of metaphors that naturally enclose polysemia.

It is not clear how humans use and understand verbal attributes denoting concepts from other sensory modalities to describe sounds. Depending on the categorization, there can be some uncertainty regarding the origin of a concept, and the bi-directional relations between concepts, language, and perception (Goldstone et al., 2013). Let's take the example of brightness. Much research has associated the presence of high-frequency energy with the brightness of a sound (Allen and Oxenham, 2014; Faure, 2000; Saitis and Siedenburg, 2020). Therefore, brightness is explicitly linked to acoustic data and is a technical sound attribute. The same goes for roughness and

sharpness in their roles of psychoacoustic attributes (Ilkowska and Miśkiewicz, 2006; Pressnitzer and McAdams, 1999). However, it should be noted that throughout the literature, brightness has also been associated with other acoustic quantities such as fundamental frequency or zero-crossing rate (Alluri and Toiviainen, 2010; Spence and Deroy, 2012). Furthermore, it is important to remember that brightness, in its use for the characterization of sound – the target domain – is metaphorical because of its multiple applications in everyday life. Therefore, brightness is primarily used to describe visual sensations, e.g., "a bright light", "a bright filter applied on a picture"; but it can also be metaphorically characterizing the remarkable intelligence of a human being: "this child is bright for its age".

1.3.2 Mental representations and conceptual processing of metaphors

The multiple levels of treatment of brightness lead us to wonder about the origin of this metaphorical vocabulary of sound. A popular hypothesis would be to consider a metaphor as the lexical manifestation of a concept whose semantic attributes are linked to both a thought process and a sensory projection. According to its philosophical definition, a **concept** is the abstract mental representation of an object to which a verbal support is usually associated[4]. It is the basic element of thought that is crucial for psychological processes such as categorization, memory, or learning[5] (Goldstone et al., 2013). Unlike a percept, a concept does not depend on a specific sensory modality, and is made of the information enclosed in percepts.

Consequently, a **metaphorical sound concept** will have semantic attributes that are related to a specific use for describing a sound. However, these attributes depend on a broader lexical concept used for multiple cognitive processing like associating different sensations with it. For that reason, the mental representation of a sound concept like brightness or roughness

[4]https://www.cnrtl.fr/definition/concept
[5]https://plato.stanford.edu/entries/concepts/

may correspond to the thought process behind a symbolic projection of a sensory stimulation.

1.3.3 Development of shared sound concepts

From a cognitive perspective, accurate communication requires that individuals share a common mental representation of sound descriptions. Such mental representations may develop from explicit pedagogical learning, cross-modal associations, or from exposition to word-sound examples in professional contexts (Amodio, 2019). As a result, different populations may develop different mental representations (Jack et al., 2012). In other words, when two individuals with different professional backgrounds interact, they may be talking about different concepts, despite using exactly the same word.

In *Sense & Sensibilia*, Austin and Warnock (1962) highlight the phenomenon of philosophers (i.e., experts) bending the meaning of a word like "real" in order to use it in a professional settings:

> 'Real' is an absolutely normal word, with nothing new-fangled or technical or highly specialized about it. It is, that is to say, already firmly established in, and very frequently used in, the ordinary language we all use every day. Thus in this sense it is a word which has a fixed meaning, and so can't, any more than can any other word which is firmly established, be fooled around with ad lib. Philosophers often seem to think that they can just 'assign' any meaning whatever to any word; and so no doubt, in an absolutely trivial sense, they can [...]. (Austin and Warnock, 1962) p. 63-64

In the world of sound and music, previous studies have largely investigated the influence of sound expertise on sound perception tasks (Allen and Oxenham, 2014; Eitan and Rothschild, 2011; Lemaitre et al., 2010; McAdams et al., 1995; Pratt and Doak, 1976). Unfortunately, it remains largely unknown whether the mental representations related to metaphorical sound descriptions are influenced by expertise and shared between expert populations.

1.3.4 Sensory metaphors & crossmodal correspondences

In some professional areas, such as perfumery or oenology (Croijmans et al., 2020; Parr et al., 2002), experts often use metaphorical terms denoting concepts from other modalities to describe and categorize sensory experiences (Deroy et al., 2013; Suárez Toste, 2007). As demonstrated in the previous section, sound and timbre descriptions also relies heavily on such metaphorical terms. For instance, we observed descriptions that come from the sensory domains of vision, e.g., "bright", "round", "mat"; or touch, e.g., "rough", "soft", "harsh", "warm".

A **crossmodal correspondence** is the cognitive phenomenon of mapping sensory experiences in different modalities. Crucially, sound descriptors like "brightness" that comes from vision, or "warmth" that comes from touch, are potential linguistic evidences of crossmodal correspondences for sound perception (Saitis et al., 2020). In the past few years, both philosophical and perceptual research have investigated crossmodal correspondences between features of the five senses like pitch, geometry shapes, color, and taste (Deroy et al., 2013; Deroy and Spence, 2013; Klapetek et al., 2012; Wallmark et al., 2021). For instance, studies observed a pairing between the height observed by sight or hearing and frequency values of a sound (Jamal et al., 2017; Parise et al., 2014). However, as stated by Saitis et al. (2020), there is still much to be done in order to establish such connections concerning the multidimensional aspects of sounds.

To date, there are still uncertainties about the mechanisms underlying the coupling between sensory modalities. Nonetheless, this approach assumes that some metaphorical sound descriptions come from multisensory, amodal or supramodal phenomenons, rather than being mediated solely by linguistic channels. In a study on the evocative power of sound, participants were asked to pair words with acousmatic sounds (Schön et al., 2010). Based on their results, the authors proposed a tentative model of the semiotics of sounds that support the interaction of an amodal actor interfering with the description of a sound (Fig. 1.4).

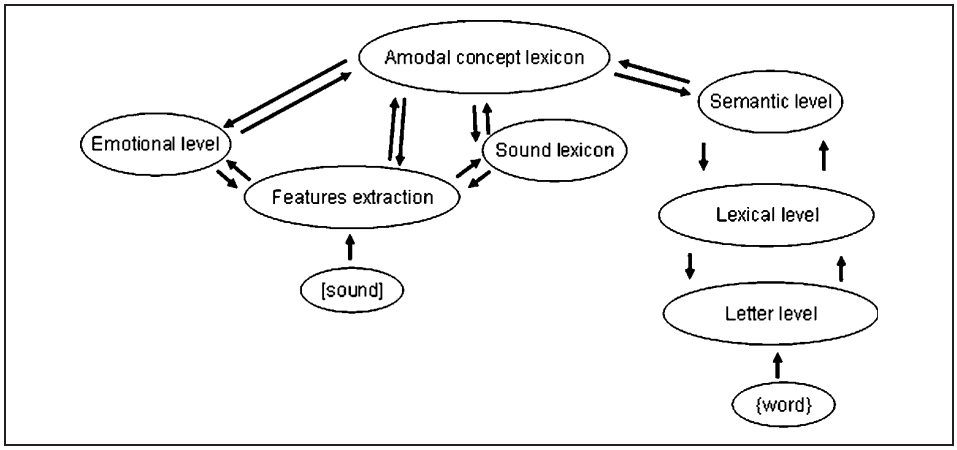


Figure 1.4: Tentative model describing how sounds can evoke concepts (Schön et al., 2010).

This model provides an explanation of the mechanisms of sound information processing that they deduced from their experiments with acousmatic sounds. First, the listener processes and obtains a representation of acoustic features. These features may correspond to an element of an internal sound lexicon of the source identity. If not, they may rely on an amodal representation of sound that may or may not be channeled through an emotional connotation elicited by sound features. The amodal representation of sounds can then either stimulate elements of the sound lexicon or be linked to concepts evoked by words. Finally, it spreads to the semantic level where it is associated with an additional auditory reference.

Interestingly, Schön's model (Fig. 1.4) does not evoke a direct link with another sensory modality and implicitly bases this link on a linguistic channel. Conversely, some studies question the existence of such a semantic pathway and argue that this cross-sensory link is independent of any linguistic apparatus. For instance, Walker et al. (2010) observed such pairing with "Preverbal infants". In the end, it seems that the construction of these concepts depends on situations and contexts, and can sometimes involve both perceptual pairings and semantic anchoring.

1.3.5 Section summary

Origins and shared aspects of metaphorical sound attributes

A metaphorical sound attribute is the application of a concept generally used in a source domain, transposed to the sound domain. I presented the technical use of metaphorical descriptions of senses that denote concepts generally originating from other sensory modalities. To date, it is difficult to generalize on the nature and the origin of this metaphorical vocabulary of sound. However, possible origins of this vocabulary are cultural, semantic, conceptual, or perceptual. Yet, there is no certainty that such sensory metaphors result from similar cognitive pathways.

Thus, along this manuscript, we will attempt to stimulate further reflection on the development of the metaphorical vocabulary of sound by considering it both as perceptual attributes (see chapter 2) and as concepts derived from a mental representations of sound (see chapter 4).

In the next section, I will introduce research in the perceptual and acoustical fields that is dedicated to the depiction of such metaphorical sound descriptions.

1.4 Perceptual and acoustic portraits of metaphorical sound attributes

To understand the use of the metaphorical vocabulary of sounds, previous works on sound semantics aimed to reveal its perceptual qualities alongside corresponding acoustic correlates. For that purpose, there have been propositions of different experimental approaches and innovations in the past decades. These are usually based on listening tests, where participants have to make a judgment on one or more sound stimuli at a time. To study a specific sound attribute like emotional attributes (e.g., valence, preference, arousal), or metaphorical attributes (e.g., brightness, roughness, roundness), several basic tasks are available for use by researchers, which can involve the use of rating scales, sound comparison, or sorting tasks. Crucially, the comparison of these methods and the experimental conditions of listening are also the subject of very interesting research (Dal Palù et al., 2017; Parizet et al., 2005). In this section, I will mainly focus on methods that have accounted for perceptual aspects related to timbre and the metaphorical vocabulary of sound.

1.4.1 Methods for assessing perceptual sound qualities

Rating scales

The most frequently used method to study the perceptual qualities of timbre is the rating scale. It generally takes two forms, semantic differential (SD) or Verbal Attribute Magnitude Estimation (VAME). The purpose of their use is often to assess perceptual and acoustic aspects related to descriptions of sound, either generated in preliminary studies of free speech (Faure, 2000; Reymore and Huron, 2020; Solomon, 1958), taken from the literature (Kendall and Carterette, 1993; Von Bismarck, 1974), or based on the expertise of the experimenter (Darke, 2005; Pratt and Doak, 1976). The choice of descriptions for the scales generally determines the results obtained. Interestingly, the descriptions chosen for these scales may not emerge much in free verbalizations, while retaining relevance when used as a perceptual criterion (Faure, 2000).

This idea then suggests a verbal use of these descriptions that differs from its perceptual relevance.

The SD is a response format that has been consistently used for many years in experimental psychology, particularly since the seminal study of Osgood et al. (1957). It consists of presenting a discrete bipolar scale that can take several points like a Likert scale. This scale constitutes a graduation between two contrasting semantic entities. For example, Solomon (1958) used seven-point SDs like "full-empty", "solid-hollow", or "rich-thin" to evaluate the use of 50 adjectives for the description of sonar sounds like. Following the impulse of Von Bismarck (1974), SDs have been widely used for the evaluation of musical timbre (Alluri and Toiviainen, 2010; Eerola et al., 2012; Pratt and Doak, 1976). It has also been used for the assessment of sound quality and annoyance on environmental sounds (Bjork, 1985; Jeon et al., 2007; Zeitler and Hellbrück, 2001). SDs are relatively ergonomic and maximizes the information obtained thanks to the presence of two descriptors at the ends of the scale. However, this position was criticized in Kendall and Carterette (1993) on the grounds that it implied a potentially invalid prior knowledge about the relation between the two descriptors. Thus, some will use the scale "bright-dark" (Alluri and Toiviainen, 2010), when other use "bright-dull" (Pratt and Doak, 1976).

To overcome the above-mentioned disadvantage of SDs, Kendall and Carterette (1993) introduced VAMEs. VAMEs are unipolar scales (e.g., "bright-not bright"). Following the suggestion, many timbre studies preferred using VAMEs than semantic differentials (Disley et al., 2006; Nykänen et al., 2009; Stepánek, 2002; Zacharakis et al., 2014). Crucially, such methods are supposed to give a better differentiation of sounds.

In a typical use of rating scales – SDs or VAMEs – participants are asked to rate stimuli with the scale. Then, the score of a sound corresponding to the studied dimension is calculated by averaging the ratings of all participants for this sound. Importantly, when the set of stimuli is presented beforehand, participants may have a relative use of the rating scales and adjust the

ends of the scale to the range of stimuli depending on the dimension being studied (Poulton, 1979). However, when used in many-item designs, rating scales may fail to differentiate between stimuli with a similar value along an underlying dimension evaluation and may show multiple consistency biases (Baumgartner and Steenkamp, 2001; Schuman and Presser, 1996) including:

- *Inconsistencies in annotations by different annotators*: one participant might assign a score of 7 to the word "good" on a 1-to-9 sentiment scale, while another participant can assign a score of 8 to the same word.
- *Inconsistencies in judgments* by the same participant: a participant might assign different scores to the same item when the annotations are spread over time.
- *Scale region bias*: participants often have a bias towards a part of the scale, for example, preference for the middle of the scale.
- *Fixed granularity*: in some cases, participants might feel too restricted with a given rating scale and may want to place an item in-between the two points on the scale. On the other hand, a fine-grained scale may overwhelm the respondents and lead to even more inconsistencies in judgments.

Sorting tasks and categorization

I presented earlier the cognitive process of categorization, which bridges the gap between a concept, its perceptual attributes and its semantic representation (see 1.2.1). Thus we can consider sorting tasks, as an attempt to simulate this process. Sorting tasks are very commonly used in cognitive psychology to address the questions of identification and categorization of sound sources (Susini et al., 2011). Classically, during a sorting task, or clustering task, participants must freely group stimuli into classes whose labels are disclosed or not. This is very close to the idea of categorization by exemplars or by prototypes. Participants' judgments can take the form of co-occurrence matrices which are used for cluster analysis through hierarchical clustering or additive trees representations to reveal the main categories (Houix, 2003).

Such a method was used for the categorization of sounds from everyday life (Guyot et al., 1997; Houix et al., 2012) or environmental sounds (Guastavino, 2007). Therefore, like description strategies, categories can take on different semantic levels in the context of free-sorting tasks. For example Lemaitre et al. (2010) showed that, in a free-sorting experiment of soundscapes, participants grouped together the sound samples into eight global categories that were verbally associated either to sound sources, human activities, room effect, type of space and metaphorical personal judgments.

Pairwise comparisons

As presented earlier, studies investigating timbre perception have employed dissimilarity ratings and MDS analysis to assess perceptual dimensions of timbre (Grey, 1977; McAdams et al., 1995). Dissimilarity ratings were based on a relative judgment format called pairwise comparison. In McAdams et al. (1995) participants had to indicate on a slider the dissimilarity between two sounds. Dissimilarity or similarity ratings are suitable to assess subtle differences between stimuli, as each stimulus serves as the standard in a series of relative judgments with the other stimuli. Thus, dissimilarity ratings were also used to evaluate the difference in brightness between stimuli (Saitis and Siedenburg, 2020). Crucially, these pairwise comparison methods impose a small corpus of sounds as the number of trials ($N(N-1)/2$) increases rapidly with the number of stimuli (N). This is an important characteristic of the method that may have negative consequences on participants' fatigue or motivation.

In addition, the popularization of the reverse correlation method (Ahumada Jr and Lovell, 1971), which aims to reveal mental representations associated with sounds, also relies on a pairwise comparison format that has been used for loudness temporal profiles (Ponsot et al., 2013) social attributes of speech prosody (Ponsot et al., 2018). When used with sound stimuli, reverse correlation is based on the temporal manipulation of a small number (one or two) of stimuli features like the fundamental frequency or the loudness. It necessitates a large number of trials that increases greatly with the number of manipulated features.

MUltiple Stimuli with Hidden Reference and Anchor

Along the way of studies dealing with perceptual aspects related to timbre semantics, other methods based on a comparison between sounds to be evaluated and reference sounds, such as the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) method have been used. Typically, the MUSHRA procedure consists in presenting participants with groups of sounds with the addition of two "anchor" sounds already observed. At each trial, participants are asked to to rank and rate the stimuli. Among other applications, the method has been used for a psychoacoustical study of wind buffeting noise (Lemaitre et al., 2015), for modeling timbral brightness (Saitis and Siedenburg, 2020), and for modeling hardness on the *freesound* dataset (Pearce et al., 2019). Although popular, this method can sometimes present some incongruities in the results, potentially attributed to the stimulus spacing and range equalizing biases (Zielinski et al., 2007).

1.4.2 Semantic dimensions of timbre and relations between concepts

Faced with the profusion of metaphorical descriptions of timbre, many studies, including some of those mentioned above, have sought to reduce and understand the relations between descriptors. As a result, using factor analysis or MDS, some works have proposed semantic representations or models of timbre. Table 1.2 reports on the different proposals for semantic dimensions made in the literature on timbre. The majority of studies have observed a main contribution of brightness under different names such as luminance (Zacharakis et al., 2014) or brilliance (Pratt and Doak, 1976).

In addition to providing an overview of the language of timbre, these studies report on the relations between different semantic descriptors. Naturally, semantic research attempts to characterize acoustically the semantic dimensions of timbre, the individual descriptors or the relations between them, following perceptual results.

Table 1.2: Non-exhaustive list of studies that investigated the main semantic dimensions of timbre. For each study, the semantic dimensions are organized in order of importance.

Studies	Semantic dimensions
Von Bismarck (1974)	Dull/Sharp Compact/Scattered Full/Empty Colourful/Colourless
Pratt and Doak (1976)	Dull/Brilliant Pure/Rich Cold/Warm
Kendall and Carterette (1993)	Dull/Brilliant Pure/Rich Cold/Warm
Alluri and Toiviainen (2010)	Activity Brightness Fullness
Zacharakis et al. (2014)	Luminance Texture Mass

1.4.3 Acoustic correlates

We have seen in section 1.1.4 that MDS-generated timbre spaces are often described as being dependent on acoustic features like the attack time and spectral centroid. Consistently, the researchers also wanted to depict acoustically the semantic aspects of musical timbre (Alluri and Toiviainen, 2010; Disley and Howard, 2004; McAdams et al., 2017; Zacharakis et al., 2014). For instance, in a cross-language study, Zacharakis et al. (2014) correlated the semantic dimension of *texture* with the energy distribution of harmonic partials, *thickness* (a term related to either *mass* or *luminance*) and *brightness*

with inharmonicity and variation of the spectral centroid, and F0 with *mass* or *luminance* depending on the language group. Typically, such results are based on linear models (PLSR and MLR[6] included) or factor analysis methods applied to relatively small corpora of sounds, not offering subtle differences between sounds.

Nowadays, we can make the most of machine learning and feature extraction tools to characterize these metaphorical terms acoustically (Bogdanov et al., 2013; McFee et al., 2015; Peeters et al., 2011). There have already been some attempts to model semantic dimensions or concepts using machine learning tools (Jiang et al., 2020; McAdams et al., 2017; Pearce et al., 2019). For example (McAdams et al., 2017), looking at the importance of fundamental frequency on the evaluation of subjective timbre qualities, used a neural network which was shown to perform better than a PLSR on accuracy. The challenge for such paradigms is to be able to collect enough judgement on sounds to allow a convergence of the learning model.

[6]PLSR: Partial Least Square Regression; MLR: Multi-linear Regression

1.4.4 Section summary

Perceptual and acoustic portraits of metaphorical sound attributes

In this section, I presented methods commonly used to bring out the salient perceptual aspects of sound, and more particularly of timbre. Several studies involving perceptual tasks have identified the semantic dimensions of timbre which include the metaphorical vocabulary of sounds presented in the previous sections. In the end, it was possible, as for timbre, to give the acoustic correlates of these dimensions. Taking a step further, I will propose in chapters 3 and 4 a more data-driven approach allowing a fine evaluation of the perceptual and acoustic aspects of the semantics of sounds.

In the next section, I will explore ways to define the sound's metaphorical attributes, based on the various types of data that we may gather for each one, i.e., semantic, perceptual, and acoustic.

1.5 Semantic and acoustic definitions of four metaphorical sound attributes

Whether it is to facilitate communication between sound professionals with different expertise or to enhance learning in a pedagogical setting, the prospect of being able to define the metaphorical vocabulary of sound is appealing. In this section, I will first describe a proposal of a sound lexicon created with the aim to serve as an interface between people with different expertise. Second, I will do a literature review concerning perceptual and acoustic aspects of four important metaphorical attributes present in the sound lexicon, brightness, warmth, roundness and roughness. These four attributes will be the main focus of this study. Finally, I will discuss the issues and pitfalls to be avoided when proposing a definition of sensory attributes such as those studied.

1.5.1 Creating a sound lexicon

Few years ago, a study on the sound identity of a brand proposed a sound lexicon that aimed to facilitate the communication between sound designers and non experts clients Carron et al. (2017). The content of this lexicon is based on common sound descriptors encountered in the domains of sound and music. To do so, the authors did a thorough analysis of the literature dealing with sound and timbre verbal characterizations. Through this process, they extracted a variety of the most frequently observed sound descriptions from studies in the field of timbre perception and sound semantics. Professionals such as sound engineers, composers, and sound designers were then approached through a questionnaire and interviews to narrow down the list of descriptions to the most frequently used and relevant. At the same time, it also helped to give meaning to each of the descriptions. In the final version, the definitions of 35 sound semantic descriptors were accompanied by contrasting sound examples for different types of source (e.g., voice, musical instrument, synthesis) that were either selected or created according to the definitions obtained through the description selection process. Today, the

sound lexicon is available under the name of *SpeaK*, in Max[7] and for web browsers[8]. Figure 1.5 shows the Max/MSP interface of *SpeaK*.

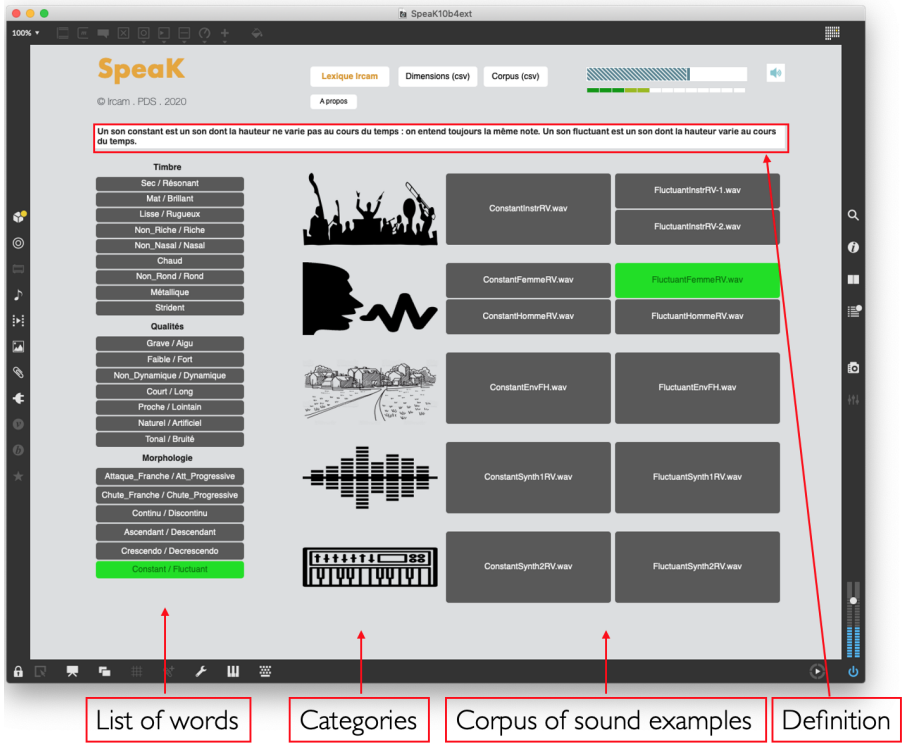


Figure 1.5: The sound lexicon of Carron et al. (2017) in Max, developed by Frédéric Voisin.

Carron's lexicon consisted of three categories of verbal descriptors, each dealing with an aspect of sound: general, timbre, and temporal. The timbre category thus consists of nine types of descriptions, some of which are in the form of word pairs:

[7] https://cycling74.com/products/max
[8] https://speak.ircam.fr/

- Bright - *Brillant*
- Rough - *Rugeux*
- Resonant - *Résonnant*
- Nasal - *Nasal*
- Rich - *Riche*
- Round - *Rond*
- Warm - *Chaud*
- Strident - *Strident*
- Metallic - *Métallique*

In this category, there are metaphorical descriptions denoting other sensory modalities that were observed in the above studies such as 'bright', 'warm', 'round' and 'rough'. In his treatise on musical objects, Pierre Schaeffer also accompanied his proposed listening formalism with prototypical sound examples varying in mass and *facture* (see section 1.1.2). However, he chose to present only 'acousmatic' sounds to focus the attention of the listener on intrinsic sound qualities. In a way, by presenting examples from multiple sound sources, the sound lexicon similarly invites listeners to focus only on extra-semantic (i.e., no source qualities) sound features.

1.5.2 Brightness, Warmth, Roundness, and Roughness

From the beginning of this thesis, our purpose has been to explore the meaning of timbre attributes present in Carron et al. (2017)'s lexicon. Hence, we wanted to accurately evaluate their semantic, acoustic descriptions, underpinning their associated mental representations. In addition, we wished to differentiate between warmth and roundness that seemed to be acoustically very close. Because of our desire to reveal subtle differences and similarities at different levels of representation, we chose to restrict this study to four metaphorical key terms referring to other sensory modalities, namely Brightness, Warmth, Roundness, and Roughness.

Warmth and roundness are two sound descriptors that seem to share many similarities in the description of sound as observed in previous studies

(Bernays and Traube, 2014; Carron, 2016; Paté et al., 2015; Zacharakis et al., 2014). On a spectral level, several studies have associated a low spectral centroid with warmth (Alluri and Toiviainen, 2010; Disley and Howard, 2004) and roundness (Paté et al., 2015; Zacharakis et al., 2014). Some implicated a link of roundness with the attack (Bernays and Traube, 2014). Both were associated with a third descriptor, namely "softness" (Alluri and Toiviainen, 2010; Eitan and Rothschild, 2011; Zacharakis et al., 2014). Moreover, some studies reported the difficulty of distinguishing the two terms in professional conversations between sound designers and industrial partners (Carron et al., 2015; Misdariis et al., 2021). Although not thoroughly documented, this issue is persistent in sound design workshops based on verbal descriptions of sound characteristics. Since the use of the two terms is also recurrent in other sound domains, we want to identify the similarities and possible differences between these two attributes.

Brightness is certainly one of the most studied descriptor in the literature since Helmholtz (1954). It is conventionally correlated to a high spectral content (Disley and Howard, 2004; Faure, 2000; Schubert and Wolfe, 2006; Wallmark, 2019b), and represents one of the main semantic dimensions of timbre in multiple studies (Alluri and Toiviainen, 2010; Kendall and Carterette, 1993; Pratt and Doak, 1976; Zacharakis et al., 2014). Therefore, it makes it an excellent reference for our study. Furthermore, some works suggest that brightness might not only be based on high spectral energy but also on other features like the attack time, pitch, and zero-crossing rate (Allen and Oxenham, 2014; Alluri and Toiviainen, 2010; Ilkowska and Miśkiewicz, 2006; Saitis et al., 2020).

Roughness is an attribute that has largely emerged from the aforementioned research dealing with sound design applications. It is defined psychoacoustically as the proximity of frequencies in critical bands (Helmholtz, 1954; Pressnitzer and McAdams, 1999; Terhardt, 1974) producing the sensation of sound modulation. Even with an explicit scientific meaning, it is not certain

that professionals such as composers or sound engineers use this definition to refer to the concept of roughness.

1.5.3 Defining a metaphorical sound attribute

As seen in section 1.2.3, the semantic categorization of the metaphorical vocabulary of sound is not straightforward and may depend on the type of sound or the domain of application (whether scientific, technical or artistic), within which the vocabulary is observed. For example, brightness and roughness can be considered simply linguistically as abstract metaphors in their own right, as the transposition of a sensory stimulation, or associated with psychoacoustic notions endogeneous of the research field of auditory perception. But even in a chosen field, a concept denoted by a metaphorical description can be ambiguous and polysemous. For example, despite its conventional spectral definition, brightness is not always obviously related to the value of spectral centroid (see section 1.3.1). As a result, the definition of these terms may include several points of view, be based on free verbalizations, or on sufficiently consistent perceptual and acoustic models, which makes the exercise of writing a definition laborious and potentially subjective.

In a study investigating the formalism of definitions for sensory descriptors, Giboreau et al. (2007) proposed guidelines to follow considering the form and the meaning of a definition. For instance, based on basic principles from (Landau, 1984), they suggest to:

- avoid circularity: do not use the defined word nor derivations in the definition,
- use general terms or define every complex word used in a definition,
- define the entry word and not its context of use or any side information,
- begin the definition with the most important elements of meaning,
- write a definition which is substitutable (that could be substituted) to the defined word in any sentence,
- use a comparable grammatical function,
- avoid difficult wording and try to write simple phrasing,
- do not use ambiguous words,

As a consequence, the question arises for such concepts to be able to formulate non-ambiguous definitions of sound concepts. When creating semantic categories or even defining certain terms, researchers often have no choice but to define one metaphorical description with another metaphorical description. For example, Saitis et al. (2017) revealed that the concept of richness is associated with emotional judgments. Thus, it may be sometimes easier to take sound examples and not to go through a semantic definition as it is proposed by (Carron et al., 2017).

1.5.4 Section summary

Perceptual and acoustic portraits of metaphorical sound attributes

I have introduced here the perks of a sound lexicon in the French language on which the present thesis is based. I have explained the strategies to follow in order to define metaphorical descriptions denoting extrasonic concepts. In line with the idea of categorization which can be based on exemplars or prototypes as well as on attributes, Carron et al. (2017) proposed to define them through definitions and sound samples. However, it seems that the meaning of certain metaphorical terms with ambiguous or polysemous attributes like 'warmth' or 'roundness' remain quite unclear.

Throughout this thesis, I will explore ways of defining such terms with the help of verbal and acoustic descriptions, sound samples, the semantic relations they share, and the mental representations they denote.

1.6 Objectives and overview of the thesis

The objective of this thesis is to reveal the meaning and the acoustic portraits associated with four metaphorical attributes used for sound description that are frequently used by sound professionals, namely, brightness, warmth, roundness, and roughness.

With the perspective of bringing together semantic and acoustic characterization of these terms, we wish to (1) propose semantic definitions for each of the descriptors/attribute; (2) find an experimental methodology to study the perceptual and acoustic aspects of each of the descriptors in a subtle way; (3) measure with the help of this new method the mental, and acoustic representations of these descriptors; (4) question the extent to which these representations are shared between populations of different sound expertise; (5) propose a musical work informed by the characteristics of each of the descriptors in order to highlight their prototypical forms as well as their relations.

In chapter 2, I will present a semantic study that aimed to reveal shared meanings of four well-used timbre attributes: bright, warm, round, and rough. We conducted two complementary studies with French sound and music experts (e.g., composers, sound engineers, sound designers, musicians, etc.). First, we led interviews to gather definitions and instrumental sound examples for the four attributes. Second, using an online survey, we tested the relevance and consensus on multiple descriptions most frequently evoked during the interviews. The analysis of the rich corpus of verbalizations from the interviews yielded the main description strategies used by the experts and definitions for the four metaphorical descriptors.

In order to obtain subtle perceptual and acoustic evaluations of the four descriptors, we need a method that allows us to consistently and ergonomically evaluate a large database of sounds. In chapter 3 we evaluate the viability of using Best-Worst Scaling (BWS). For this purpose, we asked a group of non-expert participants to evaluate the brightness of sounds according to a definition given to them. We show that the BWS is a good method for our

study, with a certain time advantage over the rating scale that allowed us to use it for the present study but also for investigating the perception of vocal attitudes.

In chapter 4, we measure the shared aspect of the mental representations related to the four metaphorical descriptors between three populations of different sound expertise, i.e., sound engineers, conductors and non-experts. For this purpose, the participants all rated brightness, warmth, roundness, and roughness within a database of orchestral instrument sounds with the BWS. With this study, we show the specificities of the sound expertise and the representation of each concept.

Chapter 5 will present the compositional process and structure of *Quadrangulation*, a creation by composer Bertrand Plé, which aims to introduce and illustrate the concepts of brightness, warmth, roughness, and roundness. This piece builds on the different facets of expert knowledge gathered in Chapters 2 and 4.

Finally, in chapter 6, I will introduce a deep-learning application perspective of our work. Then I will thoroughly compare the results obtained in chapters 2 and 4. In particular, I will compare the semantic and acoustic results. Finally, I will push further the reflection on the semantic links between the four concepts and the nature of the mechanism linked to the existence of this metaphorical vocabulary.

2. Uncovering the semantics of metaphorical sound attributes

Despite numerous insights on the general meaning of expressions like "bright, round, warm, and rough sounds" (see Chap. 1 sec. 1.5.2), it remains unclear whether the usage and verbal descriptions of metaphorical attributes are consensual or generalizable among sound experts with different profiles. In accordance with a definition of Guarino (1992), we consider such verbal attributes as the instantiation of a concept in the perceptual domain. Hence, through categorization, their meaning could be based on verbal explanations of its features or through prototypical members, e.g., sound samples (see section 1.2.1). Crucially, each metaphorical attribute can be denoted by the semantic traits and prototype sounds of the category it represents. Thus, the aim of this chapter is first, to formulate definitions on the basis of relevant description strategies related to the four metaphorical attributes. Second, to provide associated prototypical sound examples for each attribute and its opposite. Indeed, identifying a sound can be viewed as connecting auditory perception to concepts, and concepts to language, in a bidirectional relation (Goldstone et al., 2013). Therefore, this work is based on interviews (Study 2.1) and an online survey (Study 2.2) with sound experts from different fields. During the interviews, we asked participants to verbally define each attribute and to extend their definitions by selecting exemplary sound samples from a predefined sound library (Study 2.1). Then, the resulting descriptions were submitted via an online survey to an audience of experts to explicitly assess their consensus and relevance in relation to the definition of the four attributes (Study 2.2). Thus, it consists in determining to what extent the

descriptions obtained during the interviews for each attribute are relevant and shared among participants. In sum, by combining the two studies, this work provides a methodology for assessing the shared meaning of widely used metaphorical timbre attributes, and a consensual definition for each of them.

This chapter's contents (section 2.1 and 2.2) are published in the peer-reviewed journal **Music Perception** under the name: *Investigating the Shared Meaning of Metaphorical Sound Attributes: Bright, Warm, Round, and Rough* (Rosi et al., 2022a).

2.1 Study 1: Interviews with experts

Interviews were designed to address two goals: first, to obtain rich definitions with corresponding sound samples for the four attributes, and second, to reveal the sound description strategies used by experts to define the selected attributes. During the interviews, we also asked participants to illustrate their definitions with sound samples taken from a database of musical instruments.

2.1.1 Methods

Participants

Thirty-two French-fluent sound and music experts participated in the interviews (male: 23, females: 9, median age: 38.5, age range: 27–69). We selected a panel of experts that work in diverse audio fields to best represent the richness of description of the four attributes. The panel was mainly constituted of composers (10), sound engineers (7), classical musicians (7) and sound designers (6). See appendix A.1.2 for the full presentation of the professional profiles.

Sound corpus & Apparatus

To provide experts with sounds samples that could illustrate their definitions, the choice of a sound corpus was crucial. It had to be large, diverse, and easy to access. Therefore, we chose a corpus of musical instruments, showing multiple kinds of Western instruments, and playing techniques. The corpus of sounds was the result of the merger of the Studio-Online Library (SOL) (Ballet et al., 1999) and the Vienna Symphonic Library (VSL[1]). In addition to the usual instruments of strings, woodwinds, and brass, we added tonal keyboards (glockenspiel, vibraphone, xylophone, and marimba), an accordion, and a piano. For each instrument, we had a set of playing techniques, ranging from standard techniques (e.g., pizzicato, flatterzunge), to more contemporary ones (e.g., multiphonics, Bartók pizzicato). The instruments displayed variations in dynamics (from piano to forte) and pitch (octaves of C). Similarly to McAdams et al. (2017), and to avoid any potential bias created by intervals, we only presented octaves of C (except for multiphonics). Besides, some studies have observed an influence of pitch on the appreciation of timbre (Allen and Oxenham, 2014; Alluri and Toiviainen, 2010; Marozeau et al., 2003; McAdams et al., 2017; Siedenburg et al., 2021). For comfort reasons and to exclude the loudness as a main factor, we normalized the loudness of each sound sample (-23 LUFS) following the EBU norm on loudness (R-128). The loudness normalization was not noticed by the participants, except for one who felt the normalization denaturalized the sounds.

Interviews were led by the first author and lasted about two hours. They took place either in the IRCAM studios or at the participant's home or workplace. The setup was composed of a Max/MSP interface providing easy access to the sound corpus. Participants listened to sounds via open headphones Sennheiser HD 650. Each interview was recorded with a SHURE MV5 microphone.

[1]http://www.vsl.co.at

Interview procedure

During the interview, the four attributes were studied sequentially. The order of presentation followed a Klein four-group permutation (Klein, 1884) to avoid any order effect bias. Generally speaking, the design of an interview depends on the information to be extracted. Therefore, we designed a semi-directed interview in which some questions expect a certain type of response (e.g., selecting sound samples), while others leave more room for free verbalization (e.g., giving definitions), which is recommended for semantic study, as it has been done in formerly cited studies (Cheminée, 2009; Lavoie, 2014; Porcello, 2004; Reymore and Huron, 2020; Saitis et al., 2017).

The setup of an interview with experts often creates a hierarchy or asymmetrical interaction between both parties that could bring some sort of bias. This may come from the expert's assumption of lack of knowledge on the part of the interviewer, resulting in a lack of richness in the data collected. The status of the interviewer is thus defined as co-expert (Bogner et al., 2009; Van Audenhove, 2007). As a co-expert, the interviewer has similar knowledge of the technical terminology used by the expert, which allows for more depth in the conversation. To ensure clarity and relevance of answers, the interviewer must use a common vocabulary with all participants.

Before beginning the interview, both the corpus of sounds and the questions of the interview were introduced to the experts.

1. What is the context and frequency of use of the studied attribute?
2. How do you define the studied attribute?
3. Can you find at least three corresponding sound samples?
4. Can you find at least three sound samples in opposition to the studied attribute?
5. How do you define the opposite of the studied attribute?
6. Is there any affect related to the attribute under study and its opposite?

The first question was designed to obtain an overview of the context of use of each term as well as an indication of the frequency of use. In the end, it mainly tended to give a more ecological context to experts for formulating a definition.

For the second question, participants were asked to define the attribute. The interviewer helped the participants develop their responses by directing them to acoustic aspects of sound while trying to avoid definitions related to affect. The issue of affect was dealt with at the end of the questionnaire.

In the third question, participants were tasked to select three sound examples that corresponded to their definition of the attribute being studied. If necessary, the interviewer could help the expert find sounds, based on the sound descriptions given in the previous question (e.g., "low-pitched," "strings," "not too loud," etc.). The request for sound samples was not too restrictive as it was sometimes simpler for participants to select a playing technique or an instrument rather than a specific sound.

In a second part of the interview, we discussed the opposite concept to the studied attribute. The objective was to refine the answers given to the second and third questions. The fourth question had the same purpose as the third question but with sound samples in opposition with the definition of the studied attribute. Then, in the fifth question, participants tried to define the type of sounds opposite of the term studied.

Finally, the sixth question was the opportunity to question the presence of affect in the meaning of the studied attribute. It was also a way to remove any characterizations strongly related to affect from the second and fifth questions because of their lack of acoustic information. The answers to this question were used as complementary information for interpretation in the rest of this paper.

2.1.2 Analysis

After manually transcribing the interviews, we analyzed the descriptions given in the second and fifth questions. The verbalizations were filtered according to three basic steps of Natural Language Processing (NLP): (1) Tokenization of the text data with the nltk toolbox[2]. (2) Removal of stop words. (3) Lemmatization of the tokenized text, based on an adapted version of Sagot's lexicon (Sagot, 2010). In the end, we obtained the lemma/interviewee frequency

[2]https://www.nltk.org/_modules/nltk/toolbox.html

(i.e., the number of participants who cited a lemma for the definition of each attribute).

In an investigation of timbral attribute queries for sound effect libraries, Pearce et al. (2017) kept only relevant units of verbal description by following a few steps of manual filtering of their text data. We proposed a similar process that was run and reviewed by the four authors. Each ambiguous verbal unit was inspected according to its context in the sentence it is extracted from. One lemma was removed if its meaning was inconsistently identified more than 50% of the time. For instance, there was confusion about whether the term "aspect" was to be used to describe the metaphorical aspect of the sound or the fact that the sound had multiple aspects. Finally, if lemmas shared the same concept and root, they were grouped under the most frequent lemma out of the two. For instance, "bright" and "brightness" were grouped under the lemma "bright" rather than "brightness." Moreover, we did not consider the hapaxes for analysis.

Inspired by the literature that focused on the vocabulary employed by sound professionals Carron (2016); Faure (2000); Porcello (2004); Wallmark (2019a), we encoded the verbal data into categories of description strategies. The purpose was to better visualize the verbal data and to report the strategies most used to define each of the attributes.

As explained by Saitis et al. (2017) in a study on the evaluation of violin quality by professional violinists, there are two opposed perspectives regarding the qualitative analysis. Some believe that the researcher should analyze all data without any assumptions, while others think the researcher should enter the field with their hypotheses in mind (Strauss and Corbin, 1994). Here, we followed a hypothetico-deductive method and considered both prior knowledge from semantic timbre literature and information emerging from our corpus to create some of the categories of description strategies. Some of these categories emerged naturally from the transcriptions, such as the description of the source with musical instruments, and the playing technique, or spectral and temporal descriptions (all categories are reported in Table 1).

2.1.3 Results

We mainly observe descriptions that are either acoustic, source-related, or metaphoric. It is worth noting that in the French language, there can be confusion when classifying/lemmatizing descriptions of spectrum or pitch. For instance, *aigu*, and *haut*, will both describe high pitch or high frequencies. The same goes for *grave*, *bas* that designate either low frequencies or low pitch. Basse is more ambiguous as it can describe the bass clarinet, the bass guitar, or low frequencies. Considering source-related descriptions, the fact that experts mentioned instruments like the clarinet or the percussion is vastly influenced by the sound corpus. Finally, we counted numerous metaphorical descriptions such as pure (*pur*), full (*plein*), pleasant (*agréable*), and aggressive (*agressif*) that do not explicitly designate a physical aspect of the sound. Lists of the most occurring lemmas are reported for all attributes in appendix A.1.3.

Description strategies

The categories of description strategies, organized in three classes of acoustic, metaphoric, and source-related descriptions, are summarized in Table 2.1 along with examples. In total, we proposed 10 description strategies distributed over the three classes, to structure the verbalizations. For both acoustic and metaphorical categories, we have relied on a synthesis of the most recurrent semantic categories in different works on timbre (Carron, 2016; Faure, 2000; Porcello, 2004; Wallmark, 2019a). Source-related description were also inspired by research on environmental sound identification (Houix et al., 2012).

The first class gathers all the **acoustic** descriptions of sounds. There are **temporal** and **spectral** descriptions, but also **dynamic and intensity** aspects of sound, along with all of the lexical fields that are explicitly related to **sound**.

The second class collects the references to the **source**. It also corresponds to the causal listening evoked by Carron et al. (2017). There was information on the **source** mainly represented by naming the instruments present in the

Table 2.1: Description strategies (Left Column) Along with Samples From Most Occurring Lemmas (Right Column).

Acoustic	
Spectral	high (*aigu)*, harmonic (*harmonique*), low (*grave*)
Temporal	attack (*attaque*), sustained (*entretenu*), steady (*stable*)
Dynamic	*forte, piano, crescendo*
Sound specific	nasal (*nasal*), resonant (*résonnant*), noisy (*bruité*)
Source related	
Excitation mode	rub (*frotter*), *vibrato*, breathing (*souffler*)
Source	string (*corde*), voice (*voice*), clarinet (*clarinette*)
Metaphoric	
CMC	warm (*chaud*), harsh (*dur*), clear (*clair*)
Matter	round (*rond*), full (*plein*), organic (*organique*)
Effect	enveloping (*enveloppant*), itchy (*qui gratte*)
Affect	pleasant (*agréable*), aggressive (*agressif*)

sound corpus. There were also characterizations of the **excitation mode or the playing technique**.

The third class groups all of the **metaphoric** aspects of sound. The **cross-modal correspondence (CMC)** category that was extracted from previous studies contains descriptions related to other senses, such as sight, touch, and taste. A second metaphorical sub-category groups lemmas describing sound like **matter**, as specifically introduced by Wallmark (2019a). It shows descriptions of sound's shape, density, or material. The third metaphorical category groups all the descriptions of sound having an **effect** on the listener and its surroundings. The last category contains **affect, emotional value, and judgment** related to sounds. This category is present in all the studies cited above. The sixth question on the questionnaire was intended to prevent affect-related characteristics for the second question, but participants used this type of description anyway.

In order to test the validity of the description strategies, we performed an interrater agreement measure, as achieved by Wallmark (2019a). The four authors sorted the 50 top lemmas of both the second and the fifth question into the 10 categories. We noted incidental disagreement caused by the polysemy of some metaphorical items in the list, but we always considered the context

of the word and the definition from *Trésor de la langue française* database[3] to conclude on each classification. The measure of interrater agreement, Fleiss' κ, got a score of $\kappa = .69$, which reflects a substantial agreement (Landis and Koch, 1977). We then refined the categories and their definitions by collectively sorting the top 50 words one more time. Ultimately, we manually classified the lemmas cited by at least two experts in the categories.

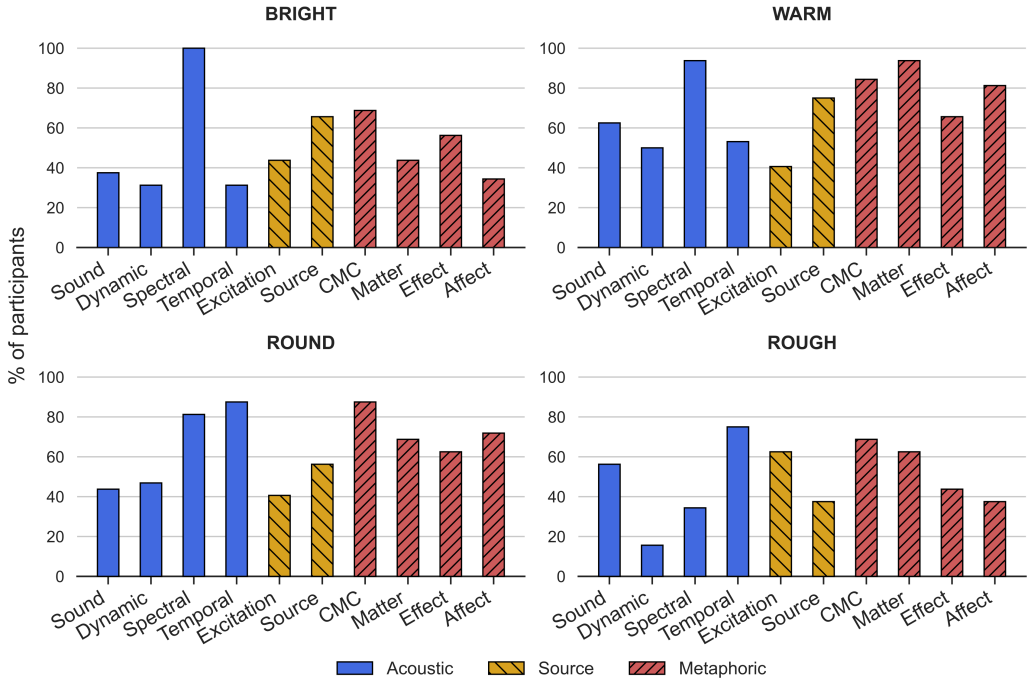


Figure 2.1: Percentage of participants using the different description strategies to define the four attributes in the second question.

Figure 2.1 presents the percentage of participants using the different description strategies. We noted that the acoustic aspects of bright were almost exclusively described through spectral features. To a lesser extent, the same is true for round and warm; but for round, there are also many descriptions of temporal characteristics of sounds. For both warm and round, there are many metaphorical descriptions. Finally, there are fewer descriptions

[3] http://atilf.atilf.fr/

Table 2.2: Descriptions cited by (N) experts for each of the four terms organized in the three classes of description strategies, along with the most frequently selected sound samples

	Acoustic		Source		Metaphoric		Sound samples
				BRIGHT			
(+)	Mainly high-frequency components (25) Medium/high pitch (10)	*Spectral*			Lexical field of light (18)	*CMC*	Glockenspiel - hard stick (29) Trumpet - *ordinario* (18)
	Sharp/strong/not soft attack (12)	*Temporal*					
(-)	Mainly low-frequency components (7) Low pitch (10)	*Spectral*	Muffled (11)	*Excitation*	Matt (11) Dull (*terne)* (7)	*Matter* *Affect*	Tuba - *ordinario* (15) Marimba - *soft stick* (12)
	Dull (*sourd)* (9)	*Sound*					
				WARM			
(+)	Mainly low/mid-low-frequency components (28) Low pitch (10) Harmonic richness (9)	*Spectral*	Voice (13) Strings (8) *Vibrato* (10) Breathed (9)	*Source* *Excitation*	Round (21) Pleasant (19)	*Matter* *Affect*	Bass clarinet - *ordinario* (22) Cello – *ordinario* (18) French horn – *ordinario* (11)
	Soft/little/not sharp/not harsh attack (12)	*Temporal*			Soft (13) Enveloping (9)	*CMC* *Effect*	
(-)	Mainly high-frequency components (19)	*Spectral*			Cold (11) Aggressive (11)	*CMC* *Affect*	Glockenspiel (16) Piccolo – *ordinario* (12) Accordion – *ordinario* (12)
				ROUND			
(+)	Soft/slow/little/not harsh/ not sharp attack (23) Resonant (14)	*Temporal*			Warm (20) Soft (13) Balanced / Homogeneous (10)	*CMC*	Double bass - *pizzicato* (19) Marimba - soft stick (18) Bass tuba - *ordinario* (17)
	Mainly low/mid-low-frequency components (17)	*Spectral*			Full (10) Pleasant (8)	*Matter* *Affect*	
(-)	Mainly high-frequency components (11)	*Spectral*	Brassy (8)	*Excitation*	Rough (13) Harsh (7) Aggressive (12)	*CMC* *Affect*	Brass - brassy (19) Strings - Bartok *pizzicato* (13)
				ROUGH			
(+)	Noisy (14) Unstable/irregular/ with variations (16)	*Sound* *Temporal*	Rubbing/ friction (13)	*Excitation*	Texture/asperities/ graininess (10)	*CMC*	Winds - *flatterzunge* (19) Bassoon - multiphonic (18) Strings - *Sul ponticello* (17)
(-)					Smooth (25) Pure (11) Round (10)	*CMC* *Matter*	Accordion - *ordinario* (18) Vibraphone (18)

for rough, which is associated more frequently with the mode of excitation than the source, unlike the other three attributes.

Verbal Descriptions and sound samples

Table 2.2 reports the descriptions most cited by the experts during the interviews when answering the second and fifth questions, along with the most frequently selected sound samples (third and fourth question). The descriptions are organized in the three classes of acoustic, source-related, and metaphoric descriptions (respectively in the first three columns). For each

description, we indicated the corresponding category from Table 1.1. The number of participants that cited a description either in the second or the fifth question are displayed in parenthesis in the table. We only presented descriptions evoked by at least 20% of the participants for the term (+), and the opposite (-). Some descriptions were grouped, e.g., homogeneous (*homogène*)/balanced (*équilibré*), if they were judged semantically closed according to the online dictionary of synonyms created by the Crosslanguage Research Centre on Meaning in Context (CRISCO[4]). We grouped descriptions that were expressed negatively in one question with corresponding descriptions expressed positively in the other question. For instance, the description "lots of high-frequency spectral content" for brightness was grouped with "few high-frequency components" for the opposite of brightness. The most frequently selected sound samples appearing in Table 2.2 were chosen according to the nature of the instrument and the playing technique. See supplementary materials[5] to listen to the sound samples presented in Table 2.2.

2.1.4 Discussion

The descriptions and sound samples cited by the experts allowed us to make multiple connections with results obtained in the literature on timbre semantics. Interestingly, although our study is in French, many of our results coincide with the literature on timbre semantics in English.

Coherently with research on timbral brightness, the great majority of experts evaluated brightness as being linked to a strong high-frequency spectral energy. As observed in previous studies (Alluri and Toiviainen, 2010; Marozeau et al., 2003; Schubert and Wolfe, 2006), the experts evoked the influence of high pitch on brightness. Several participants mentioned that a sound with a sharp attack is perceived as brighter, as was presumed but not proven in Saitis and Siedenburg (2020) in a study based on a pairwise comparison experiment. This may actually be due to a strong high-frequency spectral energy in the attack of the sound. Unlike the "bright-dark" semantic scales often used in the literature (Alluri and Toiviainen, 2010; Solomon,

[4]http://crisco.unicaen.fr/
[5]https://doi.org/10.5281/zenodo.6378886

1958; Von Bismarck, 1974), the opposition to brightness here is more consistently expressed by terms like muffled, muted or dull, like in Pratt and Doak (1976).

Most selected sound samples were high-pitched instruments played on their high register and rather loudly, in accordance with the spectral description of brightness. The choice of glockenspiel sounds played with hard sticks corroborates the potential relation between brightness and a fast attack time.

Warmth and roundness seem to be comparable attributes as they share many descriptions, but with some subtle differences. Participants evoked substantial low-frequency components for the definition of warmth and roundness coherently with studies involving the two attributes (Disley and Howard, 2004; Zacharakis et al., 2014). However, the number of overtones in a round sound was a point of disagreement among the participants, while some associate roundness with spectral richness, others imagine a sound poor in overtones. Concerning temporal aspects, the descriptions of the attack also appeared in both the definitions of warmth and roundness (no attack, little attack, soft attack, not a hard attack). However, roundness was more often described by the quality of the attack than warmth. Bernays and Traube (2014) also noted a relation between the nature of the attack and roundness in an experiment where pianists rated music recording rounder if the speed of the attack on the keys was slower.

Consistently with the acoustical descriptions from both the second and the fifth question, the round sound samples were quiet, low pitched, temporally stable, and with a soft attack. In addition, impact sounds such as the double bass playing *pizzicato* or the marimba enclosed a long resonance. Moreover, the opposition of the double bass *pizzicato* and the *Bartók pizzicato* in the selected sound samples confirms the importance of the attack for the roundness. As suggested by the source-related description of warmth, the selected cello sound displayed a strong *vibrato*. Importantly, in the evaluation of the "warm-cold" semantic scale with sounds, Eitan and Rothschild (2011) also correlated *vibrato* with the sensation of a warm sound. Finally, the breathy sound mentioned in the source-related category echoes the selected sound of

bass clarinet which, when played piano, lets us hear the air coming out of the mouthpiece.

Many of the descriptions for warmth and roundness were metaphorical. Several participants evaluated that a warm sound was also a soft sound, similarly to Eitan and Rothschild (2011) that measured a positive correlation between the semantic scales "warm-cold" and "soft-hard." In addition, we observed similar results to those of Zacharakis et al. (2014), as some participants contrasted round and rough, and others noted similarities between warm, round and soft.

Despite the design of our questionnaire, warm and round were often described through affect concepts in the second question. Hence, the experts characterized warm sounds, and to a lesser extent, round sounds, as pleasant and not aggressive. This result echoes findings in research treating valence and timbre. Valence has been depicted as being dependent on characteristics similar to observed descriptions of warm and round in this work, namely relatively long sounds with little energy and long transients (Eerola et al., 2012), and energy in the low spectrum (Wallmark et al., 2019). However, contrary to our results, McAdams et al. (2017) observed a correlation of the perceived valence on musical instrument sounds with a strong high-frequency spectral content. This opposition highlights the possible variability in affect judgments on sounds, as it has been observed in a preference-based sound quality assessment (Susini et al., 2004).

From the verbal results, roughness is related to noise, temporal patterns, or instability. Furthermore, the metaphorical descriptions represented by the lexical field of touch were a large part of the data. However, it is unclear to what extent the auditory definition of roughness relates to the *sul ponticello* sound, while it seems that multiphonics match the dissonance evoked by Helmholtz, and that both *flatterzunge* and multiphonics produce the typical envelope fluctuation of psychoacoustical roughness Pressnitzer and McAdams (1999).

In sum, the interviews offered a great diversity of verbalizations with representative sound samples for the four attributes. We observed quite different description strategies for the four attributes that allowed to establish consistent links with the literature on timbre semantics. While bright, warm and round seem to be spectrally and temporally related, this is not the case for rough whose spectral definition is almost non-existent. Warm and round retain many similarities both in their description strategies and in their acoustic definitions. Finally, the nature of the selected sounds highlighted certain aspects of sound over others (e.g. loud instruments denote high-frequency spectrum for bright, *flatterzunge* denotes temporal variation for rough).

Despite valuable insights on the descriptions strategies and meaning of the four terms, the results presented in Table 2.2 did not quite take advantage of the diversity of verbalizations as we had to summarize or group some concepts. Moreover, numerous metaphorical descriptions were difficult to interpret, and some participants had sometimes opposite points of view on them (e.g. richness for round). Therefore, in a similar fashion as some timbre semantics studies (Faure, 2000; Reymore and Huron, 2020), we sought to estimate which characteristics are the most important and relevant. To do so, in a second part that consists of an online survey, we focused on the level of consensus and relevance of descriptions given by the interview participants for the four attributes.

2.2 Study 2: Online survey

The goal of this survey is to find a way to select and rank the most relevant information contained in the verbalizations obtained during the interviews. For each of the four attributes, we built a corpus gathering the descriptions made up of the lemmas most frequently used for the second and fifth questions (see *Interview Procedure*). To investigate this question, we asked sound professionals to evaluate how one item of the corpus relates to the corresponding attribute, as part of an online survey. We also wish to evaluate presumably similar descriptions (e.g., "a sound with a soft attack," "a sound with a slow attack," "a sound without an attack") in order to derive the most relevant and consensual version.

2.2.1 Methods

Participants & Apparatus

Fifty-two sound experts participated in the survey. Similar to the interviews, all participants had one or multiple professional activities related to sound or music. Among the population of participants, 17 also participated in the interviews. They were mainly sound engineers (20), classical musicians (12), and sound designers (8). See appendix A.2.2 for the full presentation of the professional profiles.

We designed the survey with an online Javascript tool for psychology experiments called Lab.js (Henninger et al., 2021). The survey was deployed on JATOS (Lange et al., 2015) and available on all kinds of web browsers.

The phrase corpus

As we wanted to study the verbalizations obtained in the interviews, we extracted the descriptions from the responses to the second and fifth questions (cf. Study 2.1)—when participants were asked to define a sound attribute and its opposite. We selected phrases that included the most occurring lemmas for each of the two questions (i.e., a lemma was selected if it was evoked by

at least three persons). We discarded all the descriptions including names of instruments as participants of the survey could not listen to any sounds. We filtered the original corpus of descriptions to make the survey feasible in a reasonable amount of time and online following three rules:

- Discard metaphoric concepts evoked by only one person.
- Homogenize the description of the spectrum with quantifiers, e.g., the sentences "a sound loaded with high harmonics" and "a sound with a lot of high-frequency components" become "a sound with a lot of high harmonics/components."
- Eliminate a description from one of the questions that is opposed to another one from the other question, e.g., "a sound that is expressive" (second question) and "a sound that is not expressive" (fifth question).

By the end of the procedure, there were 45 phrases for bright, 67 for warm, 68 for round, and 34 for rough. Note that the corpus of descriptions is different for each attribute as it is based on the verbalizations proper to each attribute. However, some descriptions are common to more than one attribute (e.g., "rich," "smooth ," "low," etc.).

Questionnaire

The questionnaire invited participants to evaluate the adequacy and relevance of sound descriptions with the four attributes presented in a randomized order. When starting the experiment, they had to explain in which context they would use each of the four attributes. The idea behind this question is to get closer to the real context of use to enhance the reflection of the participant. The form was composed of two questions:

1. According to you, the description "X" is: *Accurate/Vague/Incomprehensible*
2. According to you, a [sound attribute] sound is X?
   2A. *Not relevant (Yes/No)*
   2B. *Strongly disagree-Strongly agree* (Likert scale)

An example of both question for bright is:

1. According to you, the description "a sound with a soft attack" is: ...

2. According to you, a bright sound is a sound with a soft attack?

The descriptions were not originally formulated by the participants of the online survey, so it could be difficult for them to relate it to a specific attribute. To address that issue, we first asked participants to express their degree of comprehension of the description (question 1). Then, participants proceeded to the second question only if they answered "accurate" or "vague" to the first question. The question 2 is two-fold. First, participants indicated if the description was not relevant to the attribute (question 2A). Second, if they felt the description was relevant, they indicated how well it matched or did not match the attributes on a 5-point Likert scale ranging from strongly disagree to strongly agree (question 2B). the additional information on relevance was motivated by Faure (2000), that evaluated the relevance of a group of descriptions with a collection of sounds. An example of the interface in French is reported in appendix A.2.1.

2.2.2 Analysis

In order to select descriptions that were familiar, relevant, and with a clear trend with respect to each attribute, we applied statistical tests to the answers of the three questions (i.e., 1, 2A, and 2B) sequentially. First, for question 1, to test whether a description's meaning is significantly familiar (i.e., "accurate"/"vague") or not (i.e., "incomprehensible") we used a chi-square test of homogeneity ($1, N = 51, p = .05$). Second, for question 2A, we used a similar chi-square test ($1, N = 51, p = .05$) to select the significantly relevant descriptions. Lastly, we applied a Wilcoxon signed-rank test to the Likert scale results of question 2B, to evaluate the central tendencies of a description whether it was matching the attribute, mismatching, or neutral. Only descriptions with significant tendencies from the Likert scale midpoint were retained. In other words, a description was not selected if it was not judged significantly familiar, and the Wilcoxon test was considered only if the description was judged significantly relevant.

Because tests were applied sequentially, the probability of type 1 errors is multiplicative and is generally low ($p < .053$). Thus, corrections for multiple

comparisons by the number of tested descriptions (e.g., the Bonferroni correction) would not affect the results and were considered unnecessary here. See appendix A.2.3 for more information on the statistical analysis of Study 2.2.

2.2.3 Results

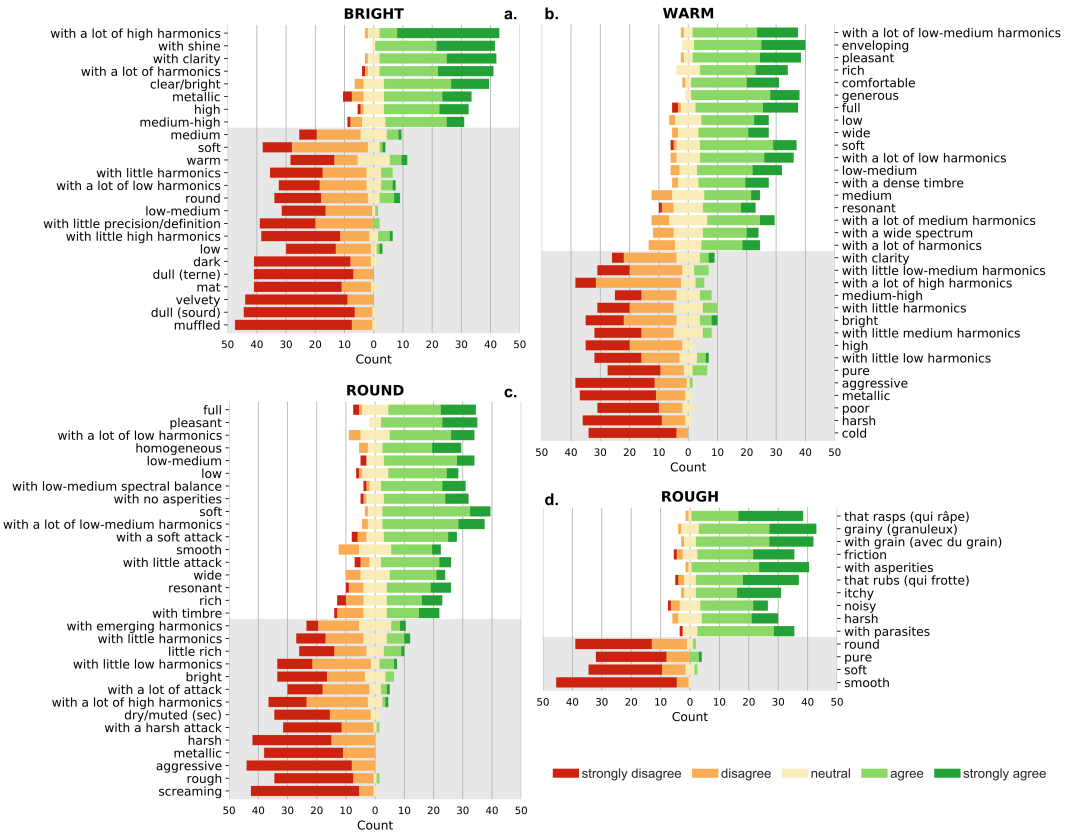


Figure 2.2: Relevant descriptions and distribution of answers on the Likert scales obtained through the online survey for (a) bright, (b) warm, (c) round, (d) rough. The grey area gathers the descriptions in mismatch with the attribute. Some ambiguous descriptions in English are followed with a French translation in parenthesis.

Figure 2.2 reports the most significant descriptions from the survey, hence giving a general meaning of the four attributes. Translations were formulated by the authors on the basis of the literature on sound semantics, and are therefore not perfectly accurate but rather an aid to understanding. See appendix A.2.4 for the original versions of the descriptions in French.

Table 2.3: Number of relevant and consensual descriptions compared to the total number of descriptions, along with the distribution of these phrases into the three classes of description strategies established in study 2.1.

Attribute	Total	Relevant & Consensual	Acoustic	Metaphoric	Source related
Bright	45	24	11	11	2
Warm	67	33	17	15	1
Round	68	31	15	15	1
Rough	34	14	2	9	3

Importantly, all the descriptions in agreement with the attribute under study are from the second question of the interviews and those in opposition from the fifth question. The statistical analysis revealed large consensus on the meaning of many attributes across participants. Table 3 reports the number of relevant phrases with a consensual meaning and their distribution into the three classes of description strategies, namely, acoustic, metaphoric, and source related. In sum, we observe strong shared meanings for the four attributes with still many metaphorical descriptions.

2.2.4 Discussion

The most important information emerging from the interviews is found in the expression of a strong consensus in the survey results, specifically on metaphorical descriptions. The shared meanings expressed through metaphorical descriptions may be due to their lexical relation with the studied attribute instead of an acoustic description. For instance, descriptions such as “comfortable,” “pleasant,” or “enveloping” might simply be the depiction of a pleasant warm feeling uncorrelated with acoustic features.

Interestingly, the absence of audio material in the online survey might have changed the results of the relevance of some of the descriptions. The most glaring example of this phenomenon is the temporal description strategy, widely used in the interviews, that almost disappeared in the final results of the survey. To a lesser extent, the same observation could be made considering the source-related descriptions. These observations lead us to reflect upon the optimal conditions for collecting sound descriptions. While some studies have

relied on listening support (Disley et al., 2006; Faure, 2000), others have done without it (Carron et al., 2017; Reymore and Huron, 2020). Findings in both cases reveal a consistent use of the descriptions employed. Our approach, which includes both types of verbalization context (with or without listening), allows us to gain insight into which strategies are dependent on the context of verbalization (e.g., temporal, source-related) and which are less dependent (e.g., spectral, affect).

In the end, the results for brightness are very similar to the ones obtained in the interviews with regard to its association with a high-frequency spectral content. Moreover, we observed a certain opposition of brightness with round and warm, based on their spectral descriptions. Overall, we noted a substantial consensus on the descriptions, with clear tendencies toward the meaning of brightness.

We noted many shared metaphorical descriptions for warmth and roundness. For instance, both were strongly associated with concepts like “full,” “pleasant,” or “soft” that were already emerging in the interviews results. While roundness is opposed to roughness, warmth has no significant trend with roughness. This absence of link between warm and rough is expressed by the fact that warm is opposed to the term “smooth” which is itself opposed to rough. Another distinction that may exist between the definitions of roundness and warmth is the relevance of the description ’rich’ that is more important and consensual for the description of warmth than for the description of roundness. Interestingly, the discrepancy between the relevance and tendency of “rich” and “with a lot of harmonics” in the results of warmth and roundness may suggest that richness does not depend essentially on spectral features as it was mentioned by some participants of the interview. These results are consistent with a study on the richness of violin timbre that have also evaluated a correspondence of timbral richness with nonspectral aspects such as warmth, vibrato, or the ability of a violinist to play a wide variety of different sounds (Saitis et al., 2019). Incidentally, the absence of a trend of “with a lot of harmonics” and roundness echoes the little consensus we noted on the relation between “spectral richness” and roundness during

the interviews. While the description of the attack for roundness was very prominent in the interviews, it appeared diminished in the survey results (e.g., “a sound with a soft attack,” “a sound with little attack”). However, it remains more relevant than for warmth. Surprisingly, recurring source-related descriptions for warmth from the interviews such as “breathy” or “vibrating” did not appear in the results.

Overall, the results regarding roughness were consistent with the interview findings. The dominant descriptions were related to the source and the lexical field of touch. Acoustic descriptions were limited to the association of roughness with ’noisy’ and the presence of ’parasites’ in the sound.

2.3 General discussion

With this study, we wanted to reveal the shared meaning and to clarify the definition of four well-used timbre attributes, bright, warm, round, and rough. To do this, we employed a methodology consisting of two complementary studies, interviews, and an online survey with a population of experts. The first is qualitative and allowed us to extract a rich vocabulary with various semantic characteristics, while the second statistically evaluated the consensus and relevance in a corpus of descriptions previously obtained. Consequently, we got three different outputs to understand the meaning of each attribute: free verbalizations structured in categories of description strategies (Figure 2.1, Table 1.1), sound samples (Table 2.2) and semantic portraits (Figure 2.2). We observed consistent descriptions across studies for warm, round, and bright that are in line with findings in the literature on timbre semantics. Furthermore, the overall results allowed us to highlight interactions between the four attributes, such as an opposition between bright on one side and warm and round on the other. Importantly, rough has no connection with bright but is opposed to round.

Due to the lack of richness in sound-exclusive terminology, sound or music experts borrow their vocabulary from other sensory domains or metaphors for sound description. The attributes of brightness and roundness are de-

rived from the sense of sight, while the attributes of warmth and roughness are derived from the sense of touch. Switching from one sensory modality to another with the same term necessarily implies multiple meanings for a term that seem to overlap in our results. The example of roundness is very illustrative to that matter. One can argue that the shape of a round object has a kind of perfection, homogeneity, and purity. These three words were found in the definitions of a round sound during the interviews. It is difficult to understand whether the person is referring to acoustic features, or visual characteristics, hence inferring a crossmodal representation of the attribute. This raises a question, "Is the perception of timbre linked to the visual perception of shapes?" Such phenomena of audiovisual crossmodal correspondences have been observed between pitch and shape (Marks, 1987), and between word morphology and shapes (i.e., the "bouba-kiki" effect). Moreover, a study revealed that judgments of roundedness and pointedness on pseudowords recordings are based on analogous sound and visual properties: smoother and more continuous, for roundedness, and disrupted, discontinuous, and strident, for pointedness (McCormick et al., 2015). Further research is needed to establish whether such sound-to-shape mappings are based on more general cognitive correspondences.

In the end, the results account for the meaning given to timbral attributes in two different situations, thus specifying the shared meaning of each attribute within an expert population. The novelty of our approach lies in the quantification of the consensus on the descriptions of the four attributes obtained in a conversational context. The results show that the conditions of evaluation influence the meaning of a timbral attribute. In particular, the context of the interviews favored temporal descriptions that were not retained in the online survey.

Nonetheless, despite these different evaluation conditions, the outcomes of the interviews and the online survey show many similarities. Among the three main types of description strategies, acoustic and metaphorical descriptions seemed to be the most suitable for defining the attributes. Interestingly, the

largest consensus involves descriptions semantically untied to sound (e.g., “full,” “rich,” “pleasant”). Thus, according to the survey results, a big part of the consensual descriptions for roughness are metaphorical, and the affect-related descriptions highly express a shared understanding of the terms round and warm.

Crucially, these results raise the question of our ability to formulate definitions for such perceptual attributes. In a study investigating the formalism of definitions for sensory descriptors, Giboreau et al. (2007) recommend avoiding ambiguous definition items. However, meeting this constraint would have been tedious given the polysemy and complexity of some of the most relevant and consensual descriptions according to the survey (e.g., “pleasant,” “full ,” “rich,” etc.). With the summaries, we have added terms that are either complementary or opposite to the attribute being studied. Nevertheless, inspired by the three types of results obtained, free verbalizations, sound examples (Table 2.1), and the results of the online survey (Figure 2), we derived definitions expressing the shared meaning of each attribute:

A *bright* sound has most of the spectral energy in the high frequencies. It is often a high-pitched sound, with clarity, definition, and similarities with a metallic sound. (*Non-bright*: Muffled, Dull, Velvety, Matte, Dark.)

A *warm* sound encloses substantial spectral energy in the low-mid frequencies. It is a rather low pitch sound. Temporally, it has a rather soft attack. A warm sound is pleasant, enveloping, and rich. (*Non-warm*: Cold, Harsh, Poor, Metallic, Aggressive.)

A *round* sound has a soft attack. It has a spectral balance localized in the low-mid range and is rather low-pitched. It is full, pleasant, and homogeneous. (*Non-round*: Screaming, Rough, Aggressive, Metallic, Harsh.)

A *rough* sound relates to a sound of friction. Listening to a rough sound feels raspy and itchy to the ear. It is a sound with grain, which has temporal

disturbances and can be noisy. (*Non-rough*: Smooth, Soft, Pure, Round.)

In sum, the meaning of timbral brightness is in line with previous research. The importance of the attack was briefly addressed during the interview but then was evaluated as not relevant in the survey. It is therefore difficult to conclude on the importance of the attack time.

Round and warm remain at the end of the study quite similar. They enclose common spectral and temporal definitions and are both positively weighted with affect or axiological adjectives, i.e., adjectives that enclose emotional reactions or value judgments (Kerbrat-Orecchioni, 2009).

In contrast to warmth, participants seemed to emphasize the temporal definition for roundness. Besides, the opposition of warmth and roughness with "pure" and the opposition of roundness with roughness may associate a stable and monotonic temporal envelope to round sounds while it is not a necessary condition for warmth. Thus, the definition of roundness could be considered as a more restrictive form of warmth from a temporal point of view.

For roughness, the survey results mainly display high consensus in metaphorical descriptions. It is relatively paradoxical with regard to the scientific definition of the term which seems clear and simple, and the choice of sound examples which seem to confirm it.

Additionally, we note that brightness and roughness seem to present two different dimensions from a semantic point of view. Brightness interacts a lot with roundness and warmth, notably from a spectral point of view, but it does not interact with roughness on an acoustic level. This echoes the results obtained by Zacharakis et al. (2014), who identified luminance and texture as two semantic axes of timbre in a interlanguage study involving English and Greek sound descriptions. To take the results further, we could imagine a cross-language study in the future (e.g., French - English), questioning the similarities and differences in the definition of these attributes.

2.4 Chapter conclusion

Uncovering the semantics of metaphorical sound attributes

In this chapter, we investigated the consensus on the definitions of well-known metaphorical sound attributes in sound and music circles. Hence, it aimed to understand the meaning of four sound attributes and the way experts describe them. We approached the question with interviews with experts and an online survey. Our results included rich descriptions and prototypical sound samples, an assessment of the consensus and relevance of these descriptions to the meaning of the four attributes, and definitions for each of them.
The characterization of each attribute is divided into three main categories of acoustic, metaphorical, and source-related descriptions. As a result, the semantics of each of the terms is robust and relies mainly on metaphorical and acoustic descriptions. Through our results, we were able to summarize the consensus regarding the meaning of the four attributes. But we note that the ambiguity of metaphorical descriptions makes the task of formulating definitions tedious.

In order to study the representation of these metaphorical attributes, we need to set up an experiment that can provide a perceptual and acoustic characterization of the four attributes. Thus, by bringing together the results of this chapter and those of a perceptual experiment (chapter 4), we wish to answer the question "Do verbal descriptions of sound attributes match acoustic descriptions obtained through analysis of a perceptual experiment?". In the next chapter I will introduce the chosen Best-Worst Scaling method and its comparison with the classical rating scale method. It will then applied to the four attributes in chapter 4.
A limitation to our approach was the difficulty to study the influence of the professional profile of the experts who participated. In chapter 4 we investigate the shared and specific aspects of each attribute between three populations of different expertise.

3. A method for the subjective evaluation of large sound corpora

As we have seen in chapter 1, several methods have been used for the subjective evaluation of timbre – from the classical ones such as the rating scale or the pairwise comparison, to methods with more specific formats such as the MUSHRA method (see section 1.4). In order to finely evaluate the representation associated with each attribute, we propose the use of the Best-Worst Scaling method (BWS) that does not suffer from the consistency bias observed for the rating scale method, while proposing sufficient ergonomics for the evaluation of large sound datasets.

In this chapter, we aim to evaluate the suitability of BWS for our purpose of labelling a large dataset of sounds with the four metaphorical sound concepts. First, I will give a definition of the method. Second, I will depict our empirical contribution that aimed to validate the BWS method by comparing it with the more classical rating scale method on the evaluation of perceptual sound qualities. Third, I will present an application of the BWS method on the perception of vocal attitudes, along with a reflection on its generalization.

3.1 Best-Worst Scaling (BWS)

3.1.1 Definition

Best-Worst Scaling (BWS) (Louviere et al., 2015) is a method used to gather subjective judgments. It was proven to be a valuable alternative to the rating

scale for consumer preference studies (Auger et al., 2007), semantic judgments (Kiritchenko and Mohammad, 2017), and face perception (Burton et al., 2019). A Best-Worst Scaling procedure consists of asking participants to select the best and the worst item in a subset of k items (e.g., $k = 4$) along the studied dimension at each trial. By counting the overall best and worst judgments, the BWS procedure allows to build a ranking of items from worst to best. Figure 3.1 gives an example of a single BWS trial for a study on face attractiveness.

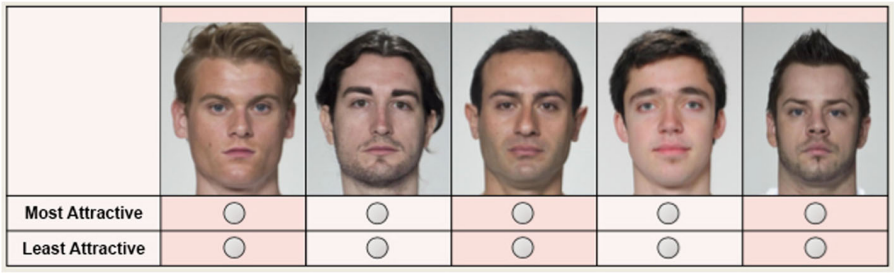


Figure 3.1: An example of a Best-Worst Scaling (BWS) trial. Participants are presented with a subset of the faces to be rated, and select the "best" (in this case, most attractive) and "worst" (in this case, least attractive) from the subset (Burton et al., 2019).

3.1.2 Many-items designs

Some studies adapted BWS to many-items contexts for semantic judgments tasks (Hollis, 2018; Kiritchenko and Mohammad, 2017). Specifically, Hollis (2018) proposed to consider each trial as a tournament paradigm where the choice of best and worst made by participants brings additional inducted information on the pairs of sounds within the subset of sounds. For instance, in a trial with 4 items [A, B, C, D], if a participant chooses A as best and D as worst, then, in addition to the deducted information that $A > D$, we also consider that $A > B$, $A > C$, $C > D$, and $B > D$. Crucially, this paradigm allows us to disseminate the information between different sequences of trials using a scoring algorithm, e.g., the Rescorla-Wagner model (Rescorla, 1972), to build the ranking of the dataset of items. To our knowledge, this method

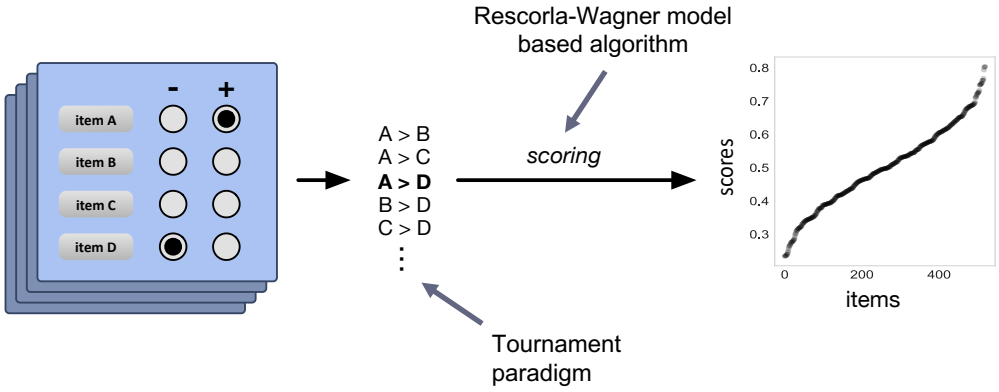


Figure 3.2: A diagram of the BWS operation according to Hollis' design.

has never been used to evaluate perceptual properties of sounds. Figure 3.2 is a representation of the BWS method when adapted with Hollis' design.

3.2 Best-Worst Scaling, an alternative method to assess perceptual sound qualities

This work is based on a chapter of Aliette Ravillion's 5-month Master thesis internship (Centrale Nantes / Politecnico di Milano) which I co-supervised during the spring of 2021 with Olivier Houix and Patrick Susini. The associated empirical contribution (see section 3.2) of this section was published in the peer-reviewed journal **JASA Express Letters** under the name: *Best-worst scaling, an alternative method to assess perceptual sound qualities* (Rosi et al., 2022b).

Importantly, the choice of this method and its evaluation was only made possible thanks to personal advice from Geoff Hollis and Svetlana Kiritchenko, who are the authors of reference papers of this work (Hollis, 2018; Kiritchenko and Mohammad, 2017).

The present study aims to evaluate the suitability of the BWS method to assess the perceptual qualities of timbre. More specifically, we wish to

test whether BWS is a valuable alternative to the rating scale (RS) – in this case a VAME – for the evaluation of perceptual brightness – one of the main dimensions of timbre (Alluri and Toiviainen, 2010; Pratt and Doak, 1976; Zacharakis et al., 2014). To evaluate the performance of the two methods we considered different questions: (i) How valid are the two methods considering the explicit definition of brightness? (ii) How reliable are the participants? (iii) What are impressions participants have of the two methods? (iv) Is one of the two procedures faster than the other? To do so, participants were set to evaluate the brightness of a musical instrument sound corpus using both methods. In this study, we considered the brightness of a sound as essentially defined through the quantity of high-frequency components, as it was demonstrated in some studies (Faure, 2000; Saitis and Siedenburg, 2020; Schubert and Wolfe, 2006). Thus, in addition to a comparison of BWS – and RS-derived brightness scores on the stimuli, the validity of the two methods was also assessed through the correlation of their scores with spectral centroid values. However, as brightness does not solely depends on spectral centroid (Alluri and Toiviainen, 2010; Marozeau et al., 2003) participants were explained the definition of brightness before the experiment.

3.2.1 Methods

The experiment aimed to measure the performances of Best-Worst Scaling (BWS) and rating scale (RS) in evaluating the brightness of a corpus of musical instrument sounds.

Participants

20 volunteer participants (10 women and 10 men, mean age = 24.3, age range = 21-27) took part in the experiment. None of them reported having any hearing problems. They gave their informed written consent before the experiment and were compensated for their participation. Participants had no sound or music education and were not familiar with either of the two methods.

Setup

We presented sounds to listeners through a Beyerdynamic DT-770 PRO (80 Ω) headset at an average level of 65 dB SPL. We measured the sound level with the sound level meter type 2250-S by Brüel & Kjær. Participants were tested in a double-walled IAC sound-insulated booth. We coded the test interface with Max (v8) on a Mac Mini.

Stimuli

The corpus was built from N=100 musical instruments sounds (i.e., woodwind, brass and string). The sounds were selected from the Studio-Online Library (Ballet et al., 1999) and the Vienna symphonic Library[1]. Each sound is a recording of a musical instrument playing sustained notes for 5 seconds. All sounds were octaves of Cs ranging from C1 (32.70Hz) to C7(2093.00Hz). We selected 100 sounds in a corpus of 200 sounds along the constraint of a constant spacing between two consecutive sounds in terms of spectral centroid. Previous work measured the just-noticeable difference (JND) of spectral centroid between two sounds as being 5% bigger than the sound with the lowest spectral centroid (Allen and Oxenham, 2014). As a result, 98% of the pairs of stimuli presented in the four-sounds BWS trials have a difference in terms of spectral centroid equal to or greater than the JND. The spectral centroid of each sound was computed and averaged for each sound with the *Librosa* library (Klapuri and Davy, 2007). The loudness of each sound sample was equalized following the EBU norm on loudness (R-128) with the *ffmpeg* library [2].

Procedure

At the beginning of the experiment, the concept of brightness was introduced to the participants using the following definition: “Brightness is relative to the amount of perceived high-frequency components in the sound. A very bright sound has a large amount of high-frequency components. A less bright

[1] https://www.vsl.co.at
[2] https://pypi.org/project/ffmpeg-normalize/

sound has a small amount of high-frequency components, it can also be called muffled or dull". We illustrated this definition with four pairs of sounds of equal pitch and different nature (musical instrument, voice and synthetic sounds). Each pair of sounds shared the same sound source but differed in brightness, i.e., the brighter sound had more high-frequency components than the other sound. Participants would then proceed to use the two methods in a randomized order after a training session on the same interface (using eight sounds for each method). We added retest trials for both methods to assess intra-participant consistency. Finally, participants completed a questionnaire asking for their impressions of both methods at the end of the experiment. Specifically, they were asked to rate the pleasantness and difficulty of each method on a 7-point Likert scale, and finally to choose their preferred method for evaluating brightness.

During the rating scale (RS) procedure, single sounds (N=100) were presented to each participant in a random order. At the end of the sequence, 20 sounds were repeated as retest trials for the intra-participant consistency measure. This RS procedure was based on VAME procedures (see section 1.4.1). After listening to each sound, participants were asked to evaluate its brightness on a 9-point Likert scale, going from "not bright at all" to "very bright". The scale was presented on the computer's screen and responses were given by selecting a point of the scale with the mouse. The brightness scores were eventually obtained by averaging the ratings for each sound.

During the Best-Worst Scaling (BWS) procedure, sounds were presented to each participant in groups of four ($k = 4$). Thus, each participant could evaluate the entire set of sounds through 25 trials of four sounds, with the addition of five retest trials at the end of the procedure for the measure of intra-participant consistency. At each trial, participants had to select the brightest and the least bright sound in each group of four sounds using the mouse. As specified in Hollis' design and to maximize the information propagated, we generated the participants' trial sequences so that each pair of sounds in a trial was presented only once over all sequences. Brightness

scores were derived from the information provided for the pairs of sounds in groups of four, thanks to a scoring algorithm based on the Rescorla-Wagner model used by (Hollis, 2018). See Hollis' personal page[3] for details on the generation of trial sequences and the implementation of the scoring algorithm.

Figure 3.3 presents the overall experiment (**A**) along with the criteria considered for evaluating the performances of both methods (**B**).

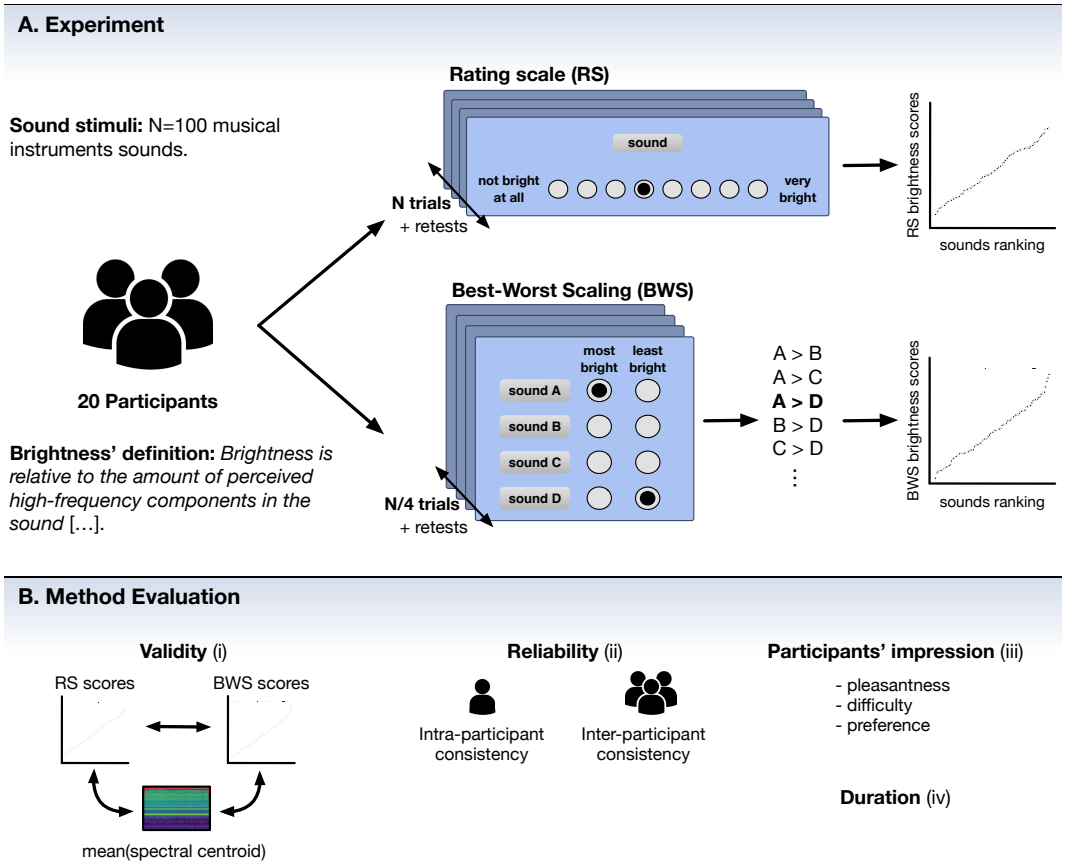


Figure 3.3: Schematic view of the methodology used to evaluate the Best-Worst Scaling (BWS) and the rating scale (RS) procedure. (A) Depiction of the two procedures. (B) Presentation of the evaluation criteria for the two methods.

[3]https://sites.ualberta.ca/ hollis/

3.2.2 Results

We used four main criteria to compare RS and BWS: validity, reliability – which can be broken down into inter-participant and intra-participant consistency – participants' impression of the methods, and duration (see Fig. 3.3-B). Validity is the extent to which a test measures what it claims to measure, based on a ground truth value. In this study, we evaluated validity by computing the correlation between the brightness scores obtained through the two methods with each other. Furthermore, since we explained to the participants that brightness is related to the amount of high-frequency components in the spectrum of a sound, we also assessed the validity of the methods through the correlations of the two sets of scores with the logarithm of spectral centroid.

The inter-participant consistency is a type of reliability measure that indicates the extent to which participants agree with each other. The intra-participant consistency measures how consistent a participant is with their previous answers when presented with retests trials. Finally, Participants' impression is assessed through a questionnaire asking them to rate the pleasantness and difficulty of the two methods, and to select the seemingly most adapted method for the task.

Validity

Figure 3.4 reports on the validity of the two methods, i.e., the correlation between the brightness scores obtained through both methods, and the correlations of these scores to the logarithm of the spectral centroids of the sounds. According to Figure 3.4-A, there is a strong correlation between the scores obtained for the two methods ($r(98) = .94, p < .001$). Moreover, for both methods, there are strong and comparable correlations with the sounds' spectral centroid ($r_{BWS}(98) = .89$, $p_{BWS} < .001$; $r_{RS}(98) = .88$, $p_{RS} < .001$) (Fig. 3.4-B). Steiger's test (Steiger, 1980) for the comparison of correlations from dependent samples did not reveal a significant difference between the two correlations with spectral centroid values.

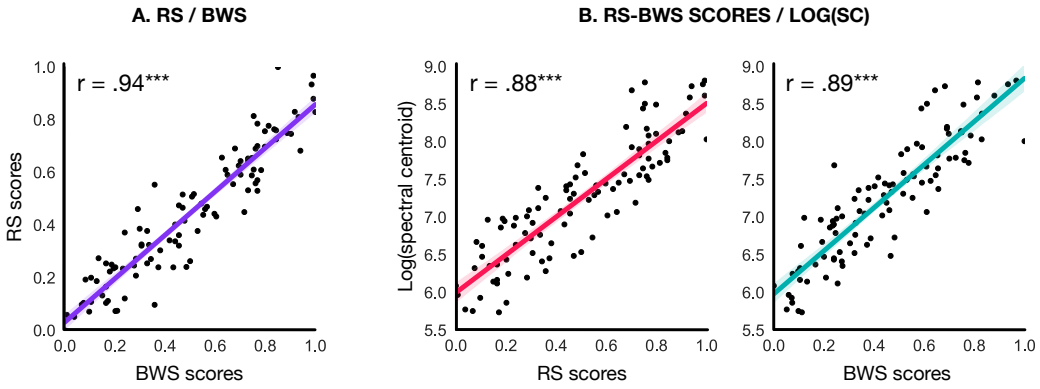


Figure 3.4: Validity evaluation of the BWS and RS procedures. (A.) Correlation between RS and BWS scores. (B.) Correlations between the logarithm of the sounds' mean spectral centroids (log(SC)) with the RS scores and BWS scores.

Reliability

Reliability was evaluated through the measure of inter-participant and intra-participant consistency.

Considering inter-participant consistency, it is worth noting that Hollis' design imposed a unique presentation for each sound pair. Hence, since the BWS procedure does not provide individual scores, well-known reliability metrics such as the Krippendorff's alpha or the Cronbach's alpha, were not suitable here. Therefore, we used the compliance to mean scores (C) as presented in Hollis (2018) to assess inter-participant consistency. Originally, Hollis introduced this metric to identify non-compliant participants in a BWS experiment. Each participant's compliance with the group was calculated as the proportion of matching duels between their own choices and the average score calculated for the group. In a trial, for example, if $A > B$, because a participant X chose A as the best and/or B as the worst, we then compared whether this inequality also holds with the brightness scores obtained by the BWS scoring. Thus, a participant giving random responses would receive a compliance score of 50% with the group scores. To adapt this measure to RS ratings – in order to provide a comparison point – we retrieved the sound pairs that appeared in the participants' BWS evaluation, and constructed inequalities based on their RS ratings (i.e., if sound A has a higher score than B, then A>B). In the case of equal RS ratings, a pair of sounds was not

considered either for the RS or the BWS measures of reliability. As for the BWS reliability evaluation, the obtained inequalities were then compared to the averaged RS scores for all participants, thus providing an average compliance value, hence a measure of inter-participant consistency. For both methods, compliance values were high ($C_{BWS} = C_{RS} = 83\%$) and did not differ significantly.

Intra-participant consistency is commonly measured by conducting tests and retests, i.e., presenting repeated trials in an experiment, and comparing the participants' responses. For BWS, we looked at the proportion of duels judged similarly in the test and in the retest. Based on five repeated groups of sounds, mean intra-participant consistency for BWS was 82% (random = 50%). For RS, we added 20 retest sounds. Here, intra-participant consistency was equal to 78% (random = 50%). Although we cannot compare the two measures due to the difference in the presentation format of sounds, we nonetheless can conclude that in average, participants were able to do both tasks without trouble.

Participants' impression on the two methods

Participants were asked to give their impression of each method by rating them on pleasantness and difficulty, and by choosing the most adapted method to evaluate brightness. Evaluations of pleasantness and difficulty between RS and BWS were not significantly different. However, BWS stood out as significantly more adapted for the evaluation of brightness (X^2(1, N=20) $= 5$, $p < .05$). Participants also elaborated on their impressions of each method in writing. On the one hand, one of them argued that they struggled to calibrate their use of the scale during the RS task. Others found that the rating scale was not very accurate, and that they were not able to extend their judgments to the extreme values of the scale. Additionally, some participants had the impression of contradicting their own previous judgments during the rating scale task. On the other hand for BWS, some participants found it difficult to choose between similar sounds and had to listen to them several times. In addition, some participants felt more concerned about making a mistake in

their judgment than they did with RS.

Duration

The BWS task lasted 9 minutes and 4 seconds on average. That is significantly faster than the RS task that lasted 9 minutes and 54 seconds on average ($T(19) = 5.22$, $p = .03$, $d = 0.36$).

3.2.3 Discussion

In this study, we evaluated the performance of two methods for assessing timbral brightness based on their validity, reliability, participants' impression, and duration. First, the scores of brightness obtained by both methods are highly correlated with each other and with spectral centroid values. This indicates that both BWS and RS methods provided accurate and similar evaluations of brightness, and are thus comparable alternatives for the study of the perceptual qualities of sound. Second, the results on reliability for the BWS and RS tasks were equivalent, supporting the idea that participants could perform both tasks consistently at group and individual levels. Third, we found that participants' impressions on the two methods were the same, with a significant preference for BWS in terms of suitability for the evaluation task. Thus, despite the similarities in performance, BWS may be a more satisfactory and comfortable method than RS for this type of task. Finally, The BWS procedure was faster than the RS procedure. Although the difference is only 50 seconds on average, it could tend to be greater in the context of annotating a larger corpus of sounds.

There is still some uncertainty about the extent to which brightness depends on spectral centroid. It could also interact with other features like fundamental frequency (F0) (Allen and Oxenham, 2014; Marozeau et al., 2003; Schubert and Wolfe, 2006). To avoid any possible confusion, we gave an essentially spectral definition of brightness to the participants at the beginning of the experiment. Therefore, we were curious to report both methods'

ability to identify brightness as being bounded to spectral centroid rather than F0. Indeed, brightness scores derived from both methods are strongly correlated with the F0 of the sounds ($r_{RS}(98) = .90$, $p_{RS} < .001$; $r_{BWS}(98) = .82$, $p_{BWS} < .001$). However, interestingly, Steiger's test applied on both correlations of scores with the F0 revealed that BWS scores are less correlated with the F0 than RS scores ($Z = 4.75, p < .001$). This suggests that in the BWS procedure, participants judgments were a little more faithful to the provided definition of brightness. It may be due to the fact that the sound presentation format of the BWS favors the comparison of the brightness of equal F0 sounds. Although the BWS procedure was comparable to the RS procedure for the measure of a perceptual property of sound, it showed specific disadvantages and advantages. On the researcher's side, unlike the RS procedure, the BWS procedure does not provide individual scores, which makes inter-participant analysis a challenge. In addition, the conditions of the BWS procedure (i.e., sequence generation, scoring algorithm) are more complex than those of the RS procedure. However, thanks to the contribution of Hollis (2018), implementing a BWS experiment (i.e., sequence generation) is fast and straightforward. On the participants' side, the BWS procedure was globally preferred by the participants, and took less time than the RS procedure. Thus, BWS seems to have crucial assets for the design of online experiments, where it is important to spare the attention and time of the participants.

Indeed, a growing number of sound perception studies rely on online crowdsourcing experiment designs to process larger quantities of sound stimuli (Cartwright et al., 2016). One motivation for this trend is the need to provide a more detailed analysis of perceptual sound qualities. In this context, and based on our results in terms of duration and performance, BWS could be an interesting experimental design choice. Moreover, according to other studies comparing RS and BWS, BWS can be conducted consistently on non-representative subsets of the entire sound dataset (Hollis and Westbury, 2018; Kiritchenko and Mohammad, 2017). Future works should therefore compare the two methods in a crowdsourcing context when applied to a bigger dataset. In addition, future method comparison experiments should also involve other perceptual property assessment methods, such as the MUSHRA protocol

which, based on a response format similar to that of BWS, was proved suitable when applied to timbral brightness (Saitis and Siedenburg, 2020).

3.3 Generalization and application of BWS to vocal attitudes

3.3.1 Application of BWS to vocal attitudes

In addition to our study, we had the opportunity to use BWS to evaluate shared mental representations of vocal attitudes. Specifically, we used the BWS experimental protocol previously introduced to perceptually validate a vocal attitude dataset.

Le Moine and Obin (2020) built Att-HACK, the first large database of acted speech showcasing four social attitudes: friendly, seductive, dominant, and distant. It comprises 20 speakers (male and female) portraying 100 utterances in 4 social attitudes. The immediate goal of this dataset is to enable voice attitude conversion using a deep learning model. A first attempt considering fundamental frequency (F0) and speech rythm as main parameters for attitude conveying has been proposed (Le Moine et al., 2021). Due to poorly sounding conversions, this first attempt revealed that those parameters were insufficient to account for actual speech attitude production. In this context, we showed that the production strategies of each attitude is specific by making connections between anatomic processes and traditional speech features. (Salais et al., 2022). In this line, we then directed our attention towards answering: "How are those attitudes perceived?". To tackle this issue, we applied our BWS experimental method to a subset of the Att-HACK dataset. Therefore, with an application of our experimental method (i.e., Best-Worst Scaling), we want to account for the alignment between the production and the mental representation of these attitudes in order to improve the conversion system performances.

In an experiment, we asked 100 participants to evaluate four distinct speech stimuli corpora representative of the four attitudes in the dataset. All

utterances in the corpora were evaluated according to the attitude initially produced. At the end of the experiment, we had four datasets annotated according to the corresponding vocal attitude. Specifically, these results brought out prototypes for each attitude in each corpus. Moreover, we can acoustically analyze the four datasets in order to account for the shared mental representation of each of the four attitudes. Like the experiment conducted in chapter 4, these results will also give fine information on the influence of groups of participants or speakers in the database. We could thus imagine evaluating the influence of parameters such as age, sex, etc.

3.3.2 Generalization to other research topics

Whether it is for the evaluation of metaphorical concepts related to timbre or vocal attitudes in the voice, Best-Worst Scaling has shown its flexibility to different research issues. We already discussed possible uses of Best-Worst Scaling in an online crowdsourcing context. But now we clearly expose its polivalency. Thus it is possible to generalize the method to other topics, whether cultural (i.e., the mental representation of non-binary voices), or for medical application (i.e., the discomfort of environmental sounds for hearing impaired participants). Each time it will be possible to access the mental representation shared between the participants, of a concept or a sound characteristic studied. Thanks to this, it will be possible to systematically compare groups of participants sharing a common characteristic related to their identity, from the most basic aspects such as age and sex, to the most cultural such as education or gender.

3.4 Chapter conclusion

A method for the subjective evaluation of large sound corpora

This work reported on the suitability of Best-Worst Scaling for an accurate measurement of timbral brightness perception. According to the criteria of performance (validity and reliability), participant preference, and overall duration, the coupled evaluation of a classic rating scale task and a Best-Worst Scaling task attests for the equivalence of both procedures, with a slight advantage in duration for Best-Worst Scaling. Therefore, Best-Worst Scaling, like the rating scale, stands as a viable relative judgment method for assessing perceptual qualities of sounds when processing many sound stimuli. Thanks to this, we were able to use this method to study the four attributes but also to apply it to the perception of vocal attitudes.

In the next chapter, I will describe the application of Best-Worst Scaling to annotate a larger dataset of musical instrument sounds (N=520) on brightness, warmth, roundness, and roughness. Specifically, this will help us reveal the acoustic portraits associated with the four concept according to three groups of participants with different expertise: sound engineers, conductors, and non-experts.

4. Shared mental representations underlie metaphorical sound attributes

In chapter 2, we saw that the semantic definitions of brightness, warmth, roundness and roughness can be of different semantic nature, with several descriptions constituted of causal (source-related), metaphorical and acoustic attributes. An objective of this study is therefore to pool these results with an acoustic depiction of brightness, warmth, roundness and roughness.

In the introductory chapter, we mentioned the difficulty of deciding on the origin of the four studied attributes (see section 1.3). Are they attributes of sounds rooted in their acoustical or perceptual definition? Or are they the projection of a mental representation, of an amodal or crossmodal concept onto sounds? In order to offer some diversity on this question, we explore the idea of mental representation associated with sound attributes.

The aim of this chapter is to investigate whether the mental representations of metaphorical sound attributes are shared across three populations, namely sound engineers, conductors, and non-experts, based on their acoustic portraits. We built this study upon the Best-Worst-Scaling (BWS) method introduced in chapter 3. We then compared the BWS ratings across populations and ran machine learning algorithms to unveil the contribution of a manifold of acoustic properties to each concept and population. Ultimately, we report acoustic portraits for each concept for the three populations.

4.1 Empirical contribution

In the present study, we investigate whether the mental representations of well-known sound concepts, i.e., brightness, warmth, roundness, and roughness, are shared between populations with different sound education backgrounds. For this purpose, three groups of participants, namely sound engineers, conductors, and non-experts evaluated a musical instrument sound corpus (N=520) on brightness, warmth, roundness, and roughness. We chose these concepts because they show both strong similarities (e.g., roundness vs. warmth) and specificities (e.g., brightness/warmth, roundness/roughness) (see chapter 2; Rosi et al. 2022a).

The three participant groups display intrinsic homogeneity in terms of expertise. The sound engineers have a rather technical knowledge of sound, whereas conductors have an intertwined knowledge of music and sound. Both populations, however, are accustomed to the use of sound concepts, unlike the non-expert group, who reported a basic metaphorical use of these concepts that is not influenced by sound or music education. As part of the experiment, participants rated each concept independently using Best-Worst Scaling (BWS), a method based on sound comparisons that has shown good performance in measuring perceptual sound properties (see Chapter 3). Subsequently, participants indicated how frequently they use said concept to talk about sounds in their professional life. Through the analysis of the judgments consistency and acoustic modelings of BWS scores, we show the influence of the groups' sound expertise on their shared understanding of those sound concepts. Figure 4.1 provides a schematic overview of the study conducted.

4.1.1 Methods

Participants

24 volunteer participants (mean age = 33, age range = 25-65) took part in the experiment. They were organized in three groups of eight participants of different expertise: professional sound engineers (mean age = 31, age

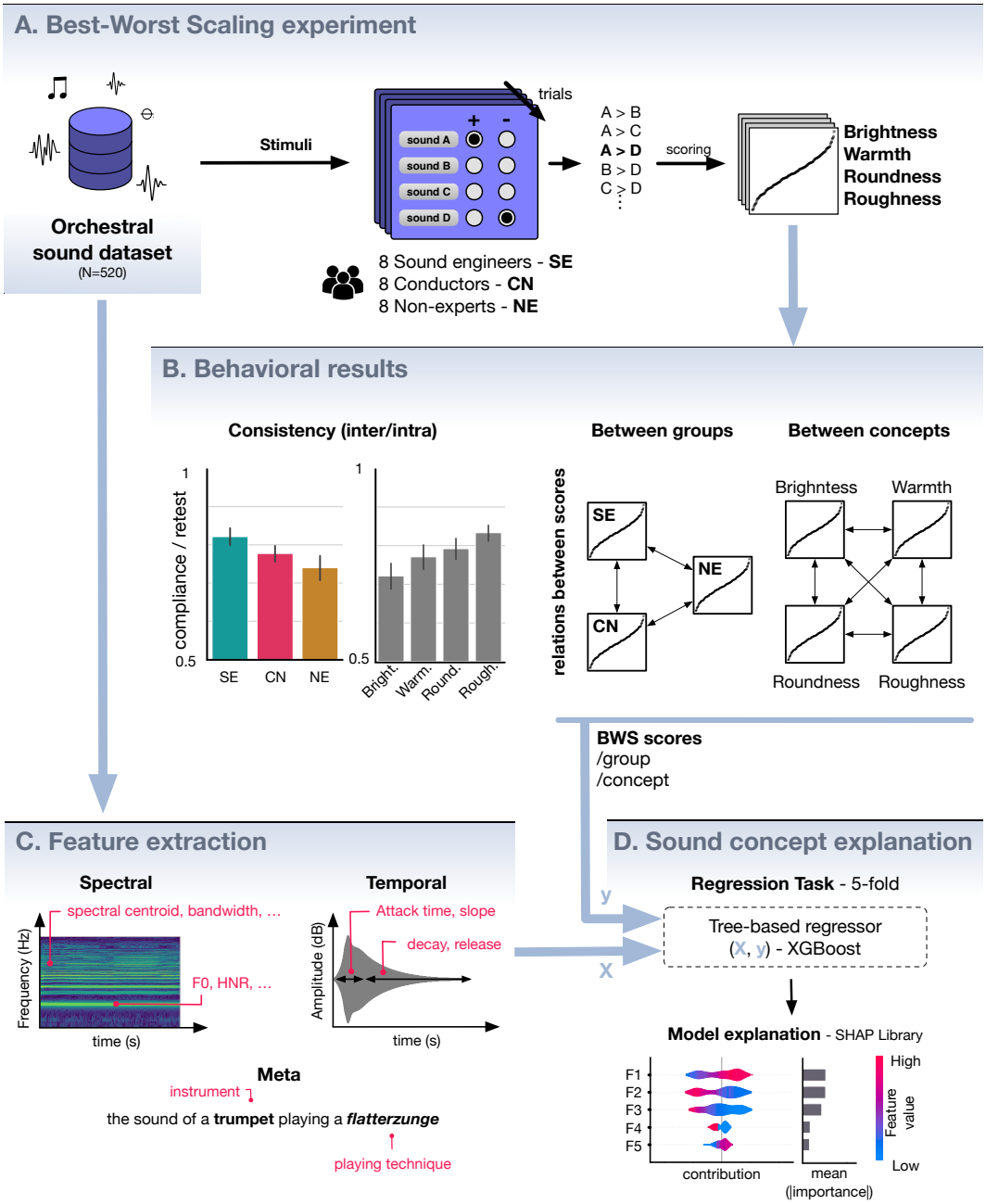


Figure 4.1: Schematic view of the methodology used to investigate the mental representations associated with specific sound concepts for different populations. (A) We collected ratings from three participant groups on an orchestral sound dataset, using four sound concepts and the Best-Worst Scaling method. (B) Using these ratings we computed consistency metrics and measured similarities and differences between concepts and between groups. (C) We extracted acoustic features (e.g., spectral centroid, attack slope) and meta features (i.e., instrument, playing technique) from the sound dataset. (D) We trained a tree-based model and assessed the most important features for the prediction of the BWS scores.

range = 25-33; seven men, one woman), professional conductors (mean age = 37, age range = 30-60; eight men) and non-experts (mean age = 33, age range = 25-65; four men, four women). The non-expert group only included participants who reported no amateur nor professional practice related to sound or music (less than 2 years of music practice). None of the participants reported having hearing problems. They gave their informed written consent before the experiment and were compensated for their participation.

Setup

Sounds were presented diotically to listeners through a Beyerdynamic DT-770 PRO headset (80 Ω) at an average level of 65 dB SPL. The sound level was measured with the sound level meter type 2250-S by Brüel & Kjær. Participants were tested in a double-walled IAC sound-insulated booth. The test interface was coded with Max (v8) on a Mac Mini.

Stimuli

The sound corpus consisted of 520 musical instrument sounds (i.e., strings, brass, woodwinds, and keyboards) from the Studio-Online library (Ballet et al., 1999) and VSL[1]. As in chapter 2, the sounds were selected arbitrarily on the basis of source, playing technique, variety of dynamics, and registers. Specifically, we retained 22 instruments with different playing techniques, e.g., sul ponticello, multiphonics, flatterzunge (see appendix B.1.2 for the presentation of all instruments and playing techniques). To ensure that the stimuli covered the full spectral range, while controlling for harmonic interactions, we selected instrumental samples playing over several octaves of Cs ranging from C1 (32.70Hz) to C8 (4186.01Hz) with different dynamics. The loudness of the sound samples was equalized following the EBU norm on loudness (R-128) with the *ffmpeg* Library[2].

[1]https://www.vsl.co.at
[2]https://pypi.org/project/ffmpeg-python/

Procedure

The BWS procedure introduced in this section is similar to the one introduced in chapter 3. However, we have kept both descriptions for the readers' convenience.

We used Best-Worst Scaling (BWS) (Louviere et al., 2015) to collect ratings on the sound corpus. BWS is a subjective annotation method based on a stimuli comparison format that showed great performance for the evaluation of perceptual sound qualities. In the context of sound evaluation, a BWS procedure consists of presenting k-tuples of sounds (e.g., k=4), and asking participants to choose the best and the worst sound depending on the studied concepts. Final scores for each sound are computed by counting the number of best and worst judgments. Recent works have adapted BWS for the annotation of a large corpus of items (Hollis, 2018; Kiritchenko and Mohammad, 2017). Specifically, by considering each trial as a tournament paradigm (Hollis, 2018), the information taken from a trial are the nature of duels between sounds, rather than just the information of the best and the worst sound. For instance when evaluating brightness, if a participant chooses A as the brightest sound and D as the least bright sound in a group of sounds [A, B, C, D], then, in addition to the deducted information that $A > D$, we also consider that $A > B$, $A > C$, $C > D$, and $B > D$. Crucially, this paradigm allows us to propagate the information between different sequences of trials using a scoring algorithm based on the Rescorla-Wagner model (Hollis, 2018; Rescorla, 1972), and hence, compute scores for all the sounds. To maximize the information propagated for the calculation of scores, a pair of sounds can only be presented once together. Finally, each participant evaluated the whole dataset with different trial configurations for each concept. We optimized the number of participants for each group based on the number of evaluations necessary to obtain consistent scores (Hollis, 2018). The generation of trials sequences and the scoring algorithm are reported in the supplementary materials. At the end of each sequence, participants used a 7-point Likert scale how often they used the concept to describe a sound in professional settings.

4.1.2 Analysis

Analysis of behavioral data

To measure whether the mental representations of specific sound concepts are shared between populations, we computed compliance scores — an individual measure of inter-participant consistency. Specifically, compliance is the proportion of matching duels of sounds between participant choices and means scores computed with the BWS scoring algorithm. For instance, if a participant from the sound engineer group answered that sound A > sound B, because he or she chose sound A as 'best' in the trial [A, B, C, D], then, that participant's compliance will increase if the BWS score of sound A is indeed greater than the one of sound B for the whole sound engineer group. In other words, a consistent group will have a higher average compliance score. Random responses from a participant in the experiment would result in a compliance score of 50%. We tested for the influence of the concept and the group of participants on compliance with two Kruskall-Wallis tests, because of the non-normality of the data distributions. We performed a non-paired test for the influence of the concept because the 'concept' variable did not have a clear paired nature due to its computation (i.e., compliance is calculated for each participant and depends on the mean score obtained for each group). As post hoc tests, we used Mann-Whitney U tests to measure the significance of eventual differences between concepts and between groups.

We measured intra-participant consistency by comparing test and retest trials. To do this, we calculated the proportion of duels of sounds with identical results both in test and retest trials. Because the retest scores were not normally distributed, we performed a Friedman test to test for the effect of concept and a Kruskal-Wallis test to test for the effect of group. Then, we performed post hoc Wilcoxon and Mann-Whitney U tests for the concept and the participant group.

We computed Pearson's correlation coefficients between sets of scores to compare results between concepts and populations. Additionally, we assessed

statistical differences between all correlations with the Steiger test. See Figure 4.2 for a presentation of the behavioral results.

Finally, we evaluated the differences in the frequency of use of the concepts by the expert participants (i.e., sound engineers, and conductors) with a one-way ANOVA and post hoc t-tests.

We applied a Bonferroni correction to all post hoc tests to correct for multiple comparisons. All statistical analyses were performed in Python 3.8 with the Pingouin library[3].

Feature analysis

In this section, we detail the analyses we led to explain the BWS scores associated with each concept and each population (Fig. 4.1-C & D). First, we trained a machine learning (ML) model on a regression task for predicting scores of brightness, roundness, warmth, and roughness based on static (i.e., collapsed over time) acoustic features and meta features. Second, we evaluated the contribution of all features to the BWS score of each sound with Explainable Artificial Intelligence (XAI) (Gunning et al., 2019) – a process that aims to make sense of the learning/predicting process of an ML model.

We extracted spectral and spectro-temporal features (median value and interquartile ratio) with the Librosa library (McFee et al., 2015), and temporal features with the Python version of the timbre toolbox (Peeters et al., 2011) (see Figure 4.1-C). We pruned the feature set and narrowed it down to 16 by performing a multicollinearity check and manually removing redundant features (see appendix B.1.1 for the list of selected features). We included meta features associated with the instrumental specificities of each sound, i.e., the type of instrument and the playing technique with the one-hot encoding approach (i.e., either one or zero depending on the presence/absence of the property). See the appendix B.1.2 for the full collection of acoustic and meta

[3]https://pingouin-stats.org/

features.

Next, we trained an ML model to predict the scores associated with a sound concept. For each concept and population, we performed a 5-fold regression task using a tree-based model in the *XGBoost* gradient boosting framework (Chen and Guestrin, 2016). The model would take the acoustic features and meta-features as input (X), and the BWS scores as output (y) for each concept and each population. We assessed the predictive accuracy of the model for each concept and population by computing the coefficient of determination (R^2) between the model's predictions on the test set and the actual score values (see Figure 4.1-D).

We measured the contributions of the retained features for each concepts by computing their SHAP values. Conveniently, the SHAP library is an XAI tool that provides a wrapper to explain any type of ML model and task (Lundberg and Lee, 2017). For a given sound, the SHAP value of a feature is based on the computation of Shapley values (Shapley, 1953), a game theory tool that evaluates the marginal contribution of a feature to the output prediction of an item. SHAP values can be positive or negative. Thus, the explanation of the model strategy for predicting scores lies in the assignment of a SHAP value to each sounds. We used the *treeExplainer* function to evaluate the contribution of features to our prediction of BWS scores. Such a tool allowed us to explain any dependence of the concepts studied on the acoustic features, whether linear or not.

4.1.3 Results

Consistency across populations

Figure 4.2-A reports on the compliance (left) and retest (right) results across participant groups (SE: sound engineers; CN: Conductors; NE: non-experts) and concepts. Our results show a main effect of concept on compliance scores ($H(3) = 27.6, p < .001$). Among all three groups, the most consensual concept was roughness (86%), which was significantly different from all others (

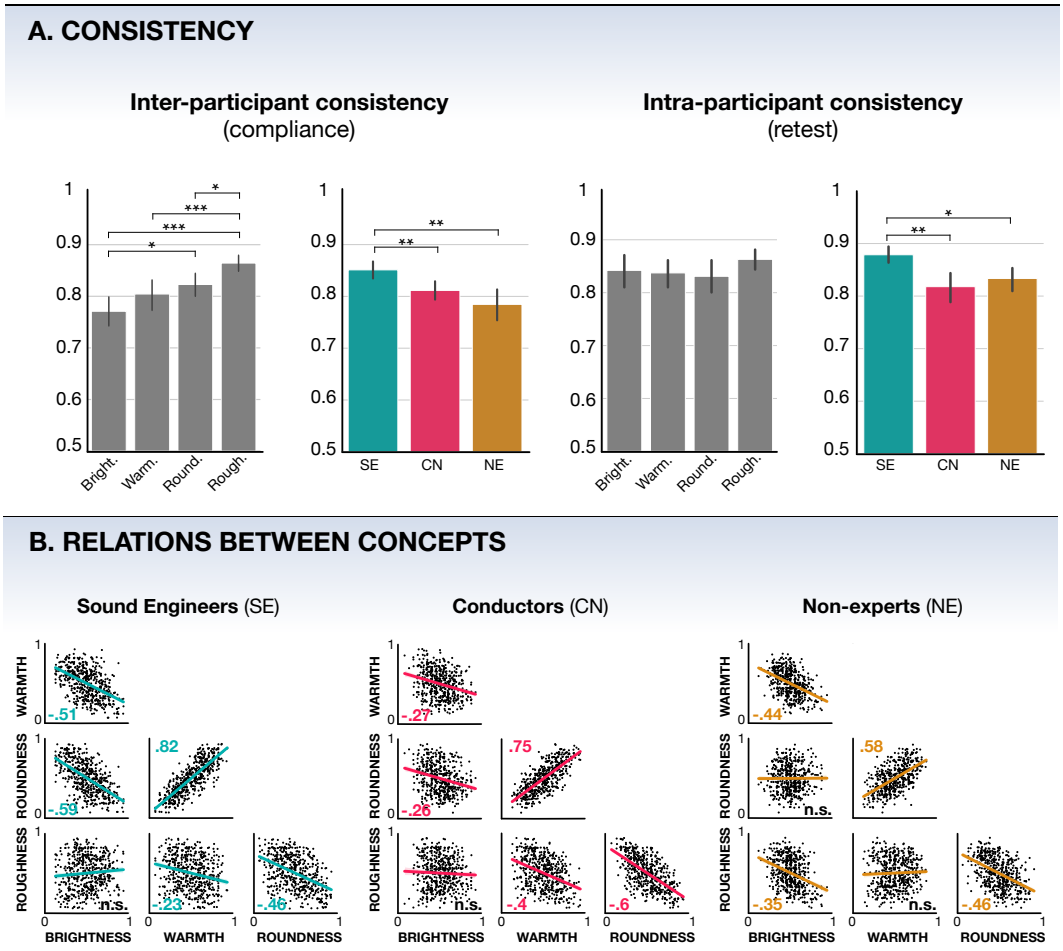


Figure 4.2: Behavioral results of the BWS experiment. (A) inter-participant (left) and intra-participant (right) consistency across concepts and populations. (B) Correlations between BWS scores of each scores for each group of participants. SE: Sound Engineers (teal), CN: Conductors (red), NE: Non-Experts (yellow).

$U_{rough./bright.} = 70.0, p < .001$; $U_{rough./warm.} = 105.0, p = .001$; $U_{rough./round.} = 149.5, p = .030$). The second most consensual concept was roundness (82%), where participants showed significantly more consistency than for brightness ($U_{round./bright.} = 159.0, p = .048$). The third most consensual concept was warmth (80%) and the least consensual was brightness (77%). We also observed a main effect of group on compliance score ($H(2) = 15.5, p < .001$). Specifically, sound engineers were significantly more consistent (85%) than the other groups of conductors (81%) and non-experts (78%) ($U_{SE/CN} = 719.0, p = .017$; $U_{SE/NE} = 798.0, p < .001$). Regarding intra-participant

consistency, we found a main effect of group on retests scores ($H(2) = 12.6, p = .002$). Once again, the sound engineer group showed a significantly higher intra-participant consistency (87%) compared to non-experts (83%) and conductors (82%) ($U_{SE/NE} = 751.5, p = .004$; $U_{SE/CN} = 720, p = .015$).

Relations between BWS scores

We then investigated the relations between concepts by correlating the BWS scores associated with each concept between them (Figure 4.2-B). For the three groups, brightness was negatively correlated to warmth ($r_{SE}(519) = -.51, p_{SE} < .001$; $r_{CN}(519) = -.27, p_{CN} < .001$; $r_{NE}(519) = -.44, p_{NE} < .001$), roughness was negatively correlated to roundness ($r_{SE}(519) = -.46, p_{SE} < .001$; $r_{CN}(519) = -.60, p_{CN} < .001$; $r_{NE}(519) = -.46, p_{NE} < .001$), and warmth was positively correlated to roundness ($r_{SE}(519) = .82, p_{SE} < .001$; $r_{CN}(519) = .75, p_{CN} < .001$; $r_{NE}(519) = .58, p_{NE} < .001$). Additionally, for the two experts population, roundness was negatively correlated to brightness ($r_{SE}(519) = -.59, p_{SE} < .001$; $r_{CN}(519) = -.26, p_{CN} < .001$), warmth was negatively correlated to roughness ($r_{SE}(519) = -.23, p_{SE} < .001$; $r_{CN}(519) = -.40, p_{CN} < .001$), and roughness and brightness were not significantly correlated. In contrast, for the non-expert group, brightness was negatively correlated to roughness ($r_{NE}(519) = .35, p_{NE} < .001$), and the pairs brightness-roundness and warmth-roughness were not significantly correlated.

Acoustic portraits of sound concepts

This section provides a description of the sound concepts for each group of participants, based on an ML-based analysis (see Methods 4.1.1). Figure 4.3 reports the five most important features, along with the nature of their contribution, for the modeling of each concept according to the BWS ratings of each population. The contribution of a feature is based on the averaged SHAP values computed on the test sets of the 5-fold regression task. The mean accuracy of the model across the 5-fold is reported with R-squared values in Figure 4.3.

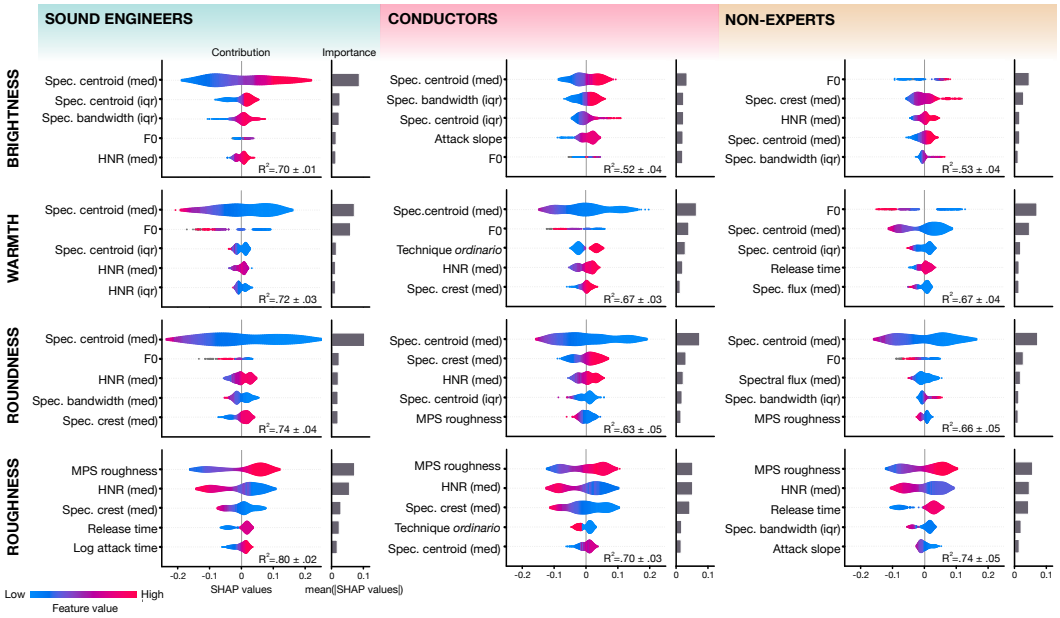


Figure 4.3: Top-5 features most explaining the regression model strategies for predicting the scores associated with sound concepts according to each group of participants. The figure represents both the nature of the contribution of each feature and its importance. Violin plots represent the contribution of each feature (SHAP Value on the x-axis) according to its value (hue color gradient). The thickness of the violin plot reflects the density of sounds for a feature value and contribution. The importance, i.e., the average of the absolute value of the contribution, is expressed in grey bar plot. med: median, iqr: interquartile range.

The use of non-acoustic features such as source and playing mode did not drastically change the model results ($\sim .02$ on average compared to the presented scores). However, we kept them in the pool of features because of their positive, albeit small, impact on the prediction of each concept.

Overall, we found that roughness was the concept with the highest accuracy scores ($R^2_{SE} = .80$; $R^2_{CN} = .70$; $R^2_{NE} = .74$), followed by roundness ($R^2_{SE} = .74$; $R^2_{CN} = .63$; $R^2_{NE} = .66$), warmth ($R^2_{SE} = .72$; $R^2_{CN} = .67$; $R^2_{NE} = .67$), and brightness ($R^2_{SE} = .70$; $R^2_{CN} = .52$; $R^2_{NE} = .53$). Moreover, sound engineers' ratings were predicted with more accuracy than the two other populations. Although some accuracy scores are low (e.g., $R^2_{CN} = .52$ and $R^2_{NE} = .53$ for Brightness), previous studies have shown that the interpretability offered by the SHAP library and a model created via XGBoost remains valid even for low predictive accuracy (Liu and Udell, 2020).

Here, we present the features underlying the shared representation of the concepts according to each group's ratings. While roughness and roundness have similar top contributing features across groups, warmth, and above all, brightness show discrepancies. First, for all groups, we found that roughness mainly grows with noise components like HNR, spectral crest, and most importantly a Modulation Power Spectrum (MPS) roughness – a metric corresponding to the average energy present in the 30Hz to 150Hz range on the time modulation axis of the MPS (Arnal et al., 2015)[4]. Second, roundness ratings relied heavily on low spectral centroids, and to a lesser extent, on low fundamental frequencies (F0). Moreover, roundness is negatively impacted by noise according to the high contributions of median harmonic-to-noise ratio (HNR), spectral crest, and MPS roughness. Third, the results show that for all populations, warmth is strongly dependent on low F0 values, more so than roundness. In addition, according to expert groups, a warm sound should also not be too noisy (e.g., HNR and spectral crest), which is less relevant for non-experts. Fourth, sound engineers mainly associated brightness with a high spectral centroid. The conductors also associated brightness mainly with spectral centroid, but its contribution is more shared with other features such as the spectral bandwidth, the attack slope, and the F0. Finally, according to non-experts' results, brightness relies heavily on F0 and noise components. In other words, according to the non-expert group, a bright sound is roughly a high-pitched sound with low noise. See supplementary materials for a more exhaustive presentation of the feature contributions for the prediction of each concept.

Frequency of use of sound concepts

Finally, with no significant distinction between sound engineers and conductors, expert participants evaluated that they use roughness significantly less than brightness ($t(15) = 5.4, p < .001$) and roundness ($t(15) = 5.2, p < .001$).

[4] see the appendix B.1.3 for a presentation of the MPS of sounds along with roughness scores and the computed metric

4.1.4 Discussion

In the present study, we investigated the mental representations of four sound concepts, namely, brightness, warmth, roundness, and roughness within groups of sound engineers, conductors, and non-experts participants. To do this, we used a dataset of orchestral sounds showcasing a great diversity of instrument timbres and playing techniques that participants rated on the four sound concepts using Best-Worst Scaling.

The results in terms of concept relations and acoustic portraits echo many findings of previous sound semantics research. First, we found that the spectral centroid is unanimously the principal feature of warmth and roundness (Eitan and Rothschild, 2011; Zacharakis et al., 2014) and that expert participants, also associated it with brightness (Allen and Oxenham, 2014; Disley et al., 2006; Pratt and Doak, 1976; Schubert and Wolfe, 2006; Zacharakis et al., 2014). Second, we found that roughness strongly depends on noisiness and time-varying features (Arnal et al., 2015; Eitan and Rothschild, 2011). Third, regarding relations between the sound concepts, most of our results (see Fig. 4.2-B) are congruent with findings observed in the literature, such as the proximity of the concepts of warmth and roundness and their relative opposition to brightness, the opposition of roundness and roughness as well as the absence of correlation between roughness and brightness (Zacharakis et al., 2014).

Thanks to the fine-grained acoustic descriptions obtained we can unravel the specific representations of warmth and roundness. First, for all groups, the resemblance between roundness and warmth seem to be mostly explained by their dependency on low spectral centroid values. Second, one may notice that the two concepts differed in that a low pitch has more impact on warmth than roundness. Third, we note that non-experts and to a slighter extent, sound engineers evaluated that a low noise constraint is stronger for round sounds than for warm sounds. Interestingly, the contribution of median HNR to warmth scores is non-linear. Finally, these observations corroborate the fact that sound engineers and non-experts evaluated roughness – which strongly

depends on noise MPS metric proposed by Arnal et al. (2015) – as being more negatively correlated to roundness than warmth.

According to sound engineers and conductors, they frequently use brightness for sound description, while they rarely use roughness. In contrast, our results show that roughness is the most consensual concept across groups, unlike brightness. Brightness has been generally associated with strong high-frequency components (Allen and Oxenham, 2014; Rosi et al., 2022a; Schubert and Wolfe, 2006) and high fundamental frequencies (Klapetek et al., 2012; Rosi et al., 2022a). While being faithful to these findings, our acoustic results and conceptual relations account for discrepancies between groups in the mental representation of brightness. First, coherently with the aforementioned research, sound engineers mostly associated brightness with the median spectral centroid. This explains the nature of its relation with roundness and warmth which have an inverse dependence on spectral centroid. Second, the conductors also associate brightness with spectral centroid, but its importance is more distributed with other features like the spread of spectral bandwidth, the attack slope, and the F0. This specificity explains the significantly lower correlation of brightness with roundness and warmth ($Z_{bright./warm.}(519) = 4.39$, $p < .001$; $Z_{bright./round.}(519) = 5.92$, $p < .001$; Steiger's Z test) for the conductors compared to the sound engineers (see Fig. 4.2-B). Third, in contrast with the experts, non-experts mainly associated brightness with the F0 and the quantity of noise (i.e., HNR and spectral crest). In other words, for the non-expert group, a bright sound is a high-pitched sound with low noise. This explains why, according to this group, brightness is opposed to warmth, which is also strongly related to F0, and to roughness, which is strongly dependent on noise features. This is also expressed in the measured correlations between scores (see Figure 4.2). The negative correlation between warmth and brightness seems to be mainly based on opposite F0 dependencies, while the lack of correlation between roundness and brightness may stem from their common relation to the amount of noise which is compensated by their opposite F0 dependencies.

Previous research has provided evidence of the superiority of sound and music experts when evaluating the acoustic aspects of sounds (Allen and Oxenham, 2014; McAdams et al., 2017). Going further, we anchor the question in the vocabulary of sound professional communication and show, through inter-participant consistency and acoustic explanations, that individuals with different sound expertise working together – like a sound engineer and a conductor in a mixing session, or a marketing representative and a sound designer – do not necessarily have the same fine understanding of well-known sound concepts. Thus, concepts like roundness and roughness are the most consensual whereas brightness and, to a lesser extent, warmth express specific understandings across participant groups (see Fig. 4.2-A). Moreover, according to consistency results, sound engineers provide greater agreement than other groups for the understanding of sound concepts. Incidentally, we found a correlation ($r(11) = .89$, $p < .001$) between inter-participant consistency and the accuracy of the models (R^2) for each group and each concept. The performance of the model thus seems to depend strongly on the consistency within groups rather than on the nature of the acoustic features.

Current views on sound semantics aim to make sense of the mechanism involved in the pairing of a metaphorical sound concept with its source domain (e.g., touch for warmth or roughness) under the scope of crossmodal correspondences (Deroy et al., 2013; Klapetek et al., 2012; Saitis et al., 2020; Wallmark, 2019b). Our results do not give any indication of the actual sensory coupling that might underlie the mental representations of these sound concepts. Nonetheless, we wish to question the immutability of the four concepts' shared mental representation in expert communities that use them in professional settings. While roughness is the least used concept, it is the most consensual, and its acoustic representation is very stable across participant groups (see Fig. 4.2). This suggests that, despite any sound or music education, the common metaphorical use of roughness remains unchanged. In contrast, our findings regarding brightness – a key term in expert sound communication – seem to express a certain diversity in the shared mental representations for each group, both through consistency

scores and acoustic analyses. This result may indicate that the meaning of brightness got specified through its use in a professional context or through the sound education of expert participants. The specificity of brightness is such, that even between two groups of experts, the concept has different levels of complexity (see Fig. 4.3). Although the explanation for such a phenomenon remains to be thoroughly explored, our results suggest that brightness is reminiscent of the concept of dead metaphor (Pawelec, 2006). A dead metaphor is a figure of speech derived from the repeated verbal use of a metaphor in a specific community. Thus a term originally metaphorical (i.e., using a term coming from a source domain in a distinct target domain) becomes a term endogenous to the discourse attached to the domain of interest, here the sound domain. Thus, brightness, unlike roundness which is also widely used, but shared across populations, has seen its meaning evolve with the expertise of our participants. Sound and music professionals interact in partially independent discursive domains, which makes possible processes of individuation of linguistic uses such as the metaphorical description of sound.

4.2 Chapter conclusion

Shared mental representations underlie metaphorical sound attributes

In this chapter, we assessed the impact of the collective sound expertise of three groups of participants on their mental representations of metaphorical sound concepts. To do so, we acoustically explained brightness, roundness, warmth, and roughness according to the evaluation of a sound dataset by sound engineers, conductors, and non-experts of sound. Surprisingly, the term most used in the expert domains (brightness) is much less consensual than the least used term (roughness). Furthermore, we went deep into the acoustic descriptions of the concepts revealing the existing relations between concepts according to the ratings of each group of participants. For example, we studied and acoustically explained the subtle differences between roundness and warmth, which are otherwise spectrally very similar, for all participants, but also for each group. Beyond the understanding of these sound concepts, we proposed a methodology that provides a fine behavioural and acoustic understanding of a sound concept. It relied on the Best-Worst Scaling method and a Machine Learning-based analysis that can be applied in the future in crowdsourcing contexts, paving the way for studies of other complex sound concepts (e.g., richness, fullness) as perceived by other populations (e.g., brass instrument player vs string instrument player), but also on other issues (e.g., voice identity, sound dataset validation).

Results from chapter 2 and 4 gave a rich overview on the shared aspects of metaphorical sound descriptions according to a sound experts. Our results range from semantic to acoustic, from verbalizations to mental representations. In order to join the results of chapter 2 and 4, I will introduce in the next chapter the BWS results of the expert population composed of both sound engineers and conductors (see section 5.1.2). The pooling of these data will be the foundations of a musical composition in Chapter 5, followed by a discussion of their congruence in chapter 6.

5. *Quadrangulation*: A semantically informed composition

I chose this title, ***Quadrangulation****, simply for what it evokes: four elements that form four angles. In this piece, «Quadrangulation» should be understood as the pairwise interaction of four timbral concepts: brightness, warmth, roughness and roundness. The piece explores all the possibilities of coupling between these concepts. They are sometimes close, sometimes opposite and sometimes even unrelated to each other, and therefore offer very different possibilities of interaction. The piece is built on two ideas. First, to expose the concepts in their most salient characteristics. Second, to propose relevant compositional mechanisms to move from one concept to another. Interpolation processes from one concept to another allow listeners to clearly perceive and distinguish each concept. Then, the more the piece progresses, the more the interaction becomes complex, superimposed and mixed. The interest of hearing this piece lies in the perception, at each moment, of the concept or concepts being exposed.*

Bertrand Plé - Preface of *Quadrangulation* (2022)

In this thesis' introduction, I presented a sound lexicon that aimed to introduce a sound-specific terminology to non-experts (Carron et al., 2017). Conveying or teaching the meaning of timbral attributes is usually done in educational settings whether one is a musician, a sound engineer or a sound designer. From my experience, this is not done explicitly in instrument lessons, resulting in incidental misunderstandings when it comes to producing the right sound. With another take on the transmission of definitions for timbral concepts, we imagined an explanatory musical work showcasing our four studied concepts, with the aim of highlighting their acoustic and semantic aspects, and conveying their specificities and similarities.

In this chapter I will present a collaboration with the above-mentioned composer Bertrand Plé which resulted in the composition and production of *Quadrangulation*, a 6-minute piece of music for seven musicians.

The purpose of *Quadrangulation* was to introduce and explain the four sound concepts to listeners. We based the design of the piece on the results of chapters two and four. The challenges were to auditorily present the results in the best possible way, and to reflect on a piece structure that invites the listener to understand the concepts and their relations with each other.

I will first introduce the idea and the issues governing the construction of the project. Second, I will present the process of designing such a composition. Third, I will report on the conditions of the recording session. Finally, I will speculate on possible experimental uses of this musical material.

The piece was recorded in IRCAM's Studio 5 thanks to the precious support of the production department. It is available on Bertrand's website[1] and on IRCAM's media library website [2].

[1] https://www.bertrandple.fr/audios-vidÃl'os/quadrangulation/
[2] https://medias.ircam.fr/xb9db8a_quadrangulation-bertrand-ple

5.1 Introduction

5.1.1 Artistic explanations of timbre

Some well-known masterpieces of music history were composed with the purpose of inviting listeners to recognize the timbre of various musical instruments. Such compositions are generally intended for a rather young audience. Thus, composers needed to think of compositional mechanisms that serve the information they wished to convey. Two musical works stand out for having successfully carried out this compositional process. First, Benjamin Britten's *Young person's guide to the orchestra* (1946) introduced the different instrument families (i.e., Strings, Woodwinds, Brass, and Drums) by using a pre-existing theme from Henri Purcell's *Abdelazer* (1676). Hence, all the instrument families would play this theme by turns at the beginning of the musical piece. Second, *Peter and the Wolf* (1936) by Sergei Prokofiev introduced some key orchestral instruments by presenting them figuratively as characters (humans or animals) in a tale. For instance, the duck's theme is consistently played by an oboe and Peter's grandpa is played by the bassoon. Although our purpose is different, our piece seeks to give an account of the timbral concepts underlying the orchestration of a piece. Our characters were the four concepts and the existing relations between them compose the narrative of the piece.

5.1.2 Expert views on the four sound concepts

The piece's composition was based on the empirical contributions of the present thesis. We therefore took inspiration from the semantic definitions of the four concepts obtained in Chapter 2 and from the results of the experiment presented in Chapter 4 for the grouped populations of sound engineers and conductors (N=16). Specifically, we re-analyzed the BWS judgments in section 4.1.1 of chapter 4.

Figure 5.1 presents an overview of the data we used for *Quadrangulation*, i.e., the definitions, the acoustic portraits, the relations between BWS scores,

and prototypical sounds that we obtained from our panel of sound experts. Indeed, in this case, the BWS scores for each concept were calculated for the grouped expert population of conductors and sound engineers. We computed and analyzed using the same process as presented in section 4.1.1.

Definitions

Despite their observed limitations, we communicated the definitions obtained in the experiments conducted in chapter 2 (Figure 5.1-A) – considering that elaborating by means of other semantic concepts of timbre (e.g., "full" for round) would help the composer understand the meaning of the concepts.

Acoustic portraits

Our acoustic explanations of the metaphorical concepts result from the evaluation of the relative contribution of a manifold of acoustic features for a BWS score prediction model (see chapter 4.1.1). Due to the larger number of participants (16 instead of 8) the accuracy score of the model is globally higher than the ones obtained for the three populations in separate groups (see Figure 4.3). See appendix C.1 for more specific representation of the contribution of some of the most important features (i.e., spectral centroid, F0, HNR, spectral crest) for the prediction of each concept.

- According to the results in Figure 5.1-B, **brightness** is mainly defined by the spectral centroid (median, iqr). The median spectral bandwidth and the F0 also seem to play an important role. That result is coherent with the definition obtained in chapter 2 that says that "A bright sound has most of the spectral energy in the high frequencies [...] It is often a high-pitched sound,".
- **Warmth** is defined mainly by the spectral centroid and the F0. Hence, coherently with the definition of chapter 2, a warm sound "is a low-pitched sound" that "encloses substantial spectral energy in the low-mid frequencies".

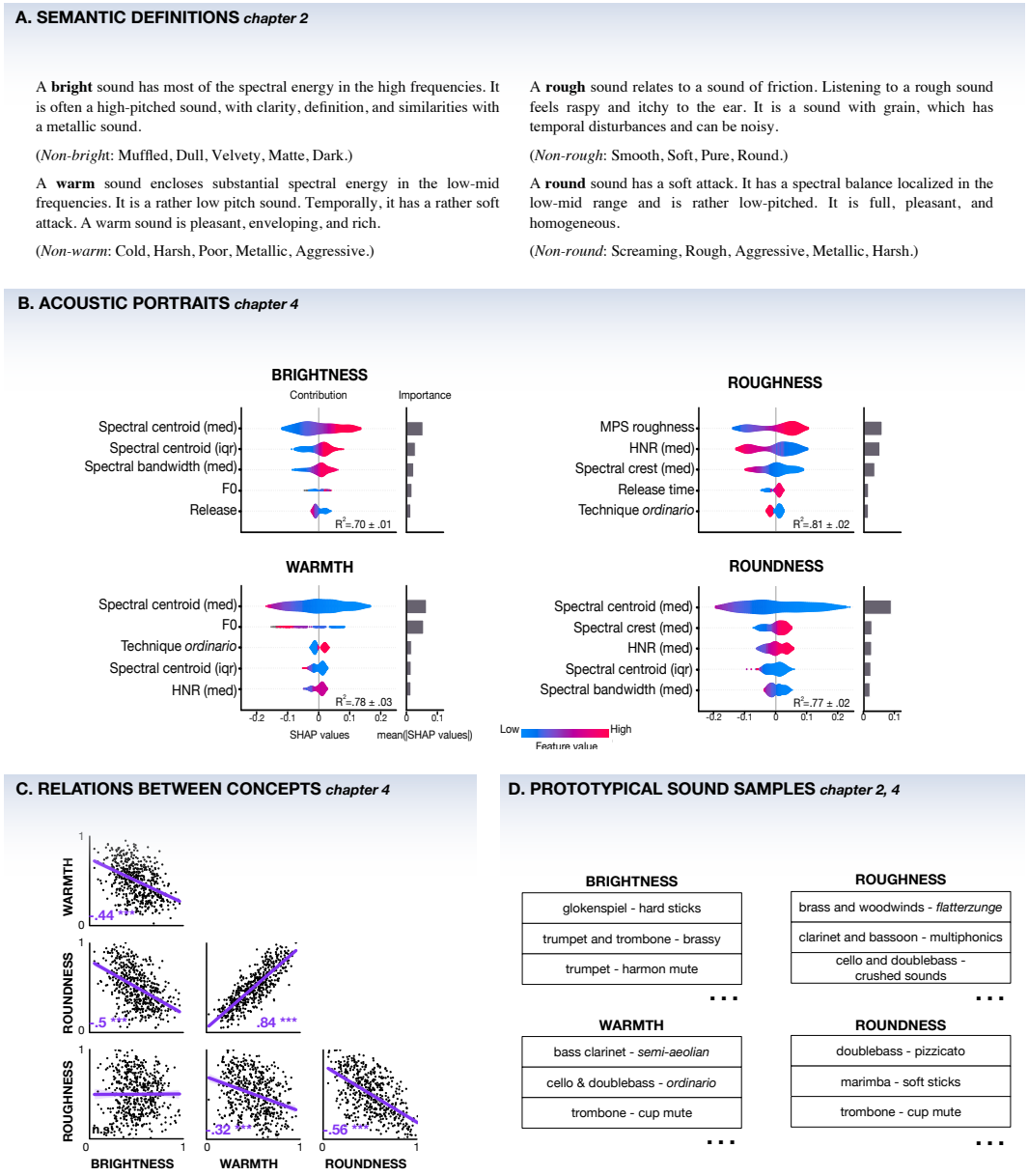


Figure 5.1: The full data communicated to the composer – it constituted a pool of expert knowledge about the four concepts. (A) The definitions of each concept. (B) The most important features for the prediction of each concept (in terms of SHAP values). (C) The existing relations between the BWS scores of the concepts. (D) The prototypical sound examples according to the sound examples of chapter 2 and tp the most extreme BWS scores (chapter 4).

- **Roughness** is mainly defined through three features related to noise, the MPS roughness, the HNR and the spectral crest.
- Finally, **roundness** is defined almost essentially by the median value of the spectral centroid. We also note a contribution of features related to noisiness, i.e., HNR, spectral crest. Thus, coherently with its definition, a round sound has a rather low spectral content and is opposed to roughness that is rather noisy.

Relations between concepts

From the correlations between the BWS scores for each concept (Figure 5.1-A), we find similar relations to those observed for the populations of conductor and sound engineers (see Figure 4.2). Thus, brightness was negatively correlated with warmth ($r(519) = -.44, p < .001$) and roundness ($r(519) = -.5, p < .001$). Roughness was negatively correlated with roundness ($r(519) = -.56, p < .001$) and warmth ($r(519) = -.32, p < .001$) Warmth was positively correlated with roundness ($r(519) = .84, p < .001$), and brightness was not correlated with roughness.

Prototypical sounds

Composers do not necessarily have the ability to decipher acoustic data. We therefore complemented the definitions with a list of the most prototypical sounds for each of the concepts, based on the analysis of the BWS experiment (chapter 1) for the expert population. Figure 5.1-D presents some examples of prototypical sounds, according to the BWS scores. It is important to note that in general, the choices of prototypical sounds selected during the interviews (see Table 2.2) are consistent with the scores they obtained via the BWS experiment. Table 5.1 lists some of the selected sounds with their overall BWS rank (out of 520) they obtained.

Thus, the composer could access the sounds with the highest scores for each concept. That was the easiest information to manipulate at first hand

Table 5.1: Prototypical sounds selected by sound professionals during the interviews for each concept along with their rank in the sound dataset according to their BWS scores.

	sounds	**rank (1-520)**
Bright	glockenspiel - hard sticks	1st-6th
	trumpet - *ordinario*/brassy	7th
Warm	bass clarinet - *ordinario*	4th
	cello - *ordinario*	7th
Round	double bass - *pizzicato*	1st
	marimba - soft sticks	4th
Rough	winds - flatterzunge	1st
	woodwinds - multiphonics	2nd

(see appendix C for a visualization of scores as a function of instruments and playing techniques). Interestingly, the fact that these sound examples were more eloquent to describe these concepts might reminds us of the prototype theory of categorization (Lakoff, 2007; Rosch, 1975). In other words, the composer would mentally refer to certain sounds (e.g., glockenspiel or brass trumpet for brilliance) to construct the orchestration that corresponds to each concept.

It is important to note that we did not want to give a harmonic or figurative meaning to these concepts. The inclusion of such parameters would go beyond the scope of our results.

5.1.3 A semantically informed composition process

> *"The more constraints one imposes, the more one frees one's self. And the arbitrariness of the constraint serves only to obtain precision of execution."*
>
> — Igor Stravinsky

We articulated the piece's conception in four steps. First, we introduced the empirical results of the expert population to the composer. We gave him definitions, instrument rankings, acoustic descriptions for the concepts, and acoustic explanations about the relations they have. Second, we established the instrumentation of the piece (see section 5.2.1). Third, we elaborated its structure to best illustrate the sound concepts and the relations that exist between them. With this information, the composer finally proceeded to compose the piece on his own. A major challenge for him was to stay in line with our empirical studies, whose results may differ from his personal representation of one of the concepts.

When starting the collaboration, Bertrand was invited to take part in the experiment presented in the chapter 4 as a starting point for the discussion of the four concepts' nature[3]. For example, to discriminate brightness, it appeared that the nature of the attack was a more discriminating criterion to him, than to the average participant. The point was to highlight the fact that this creative process should not depend on his taste in terms of instrumentation and orchestration and rather had to stand as an objective synthesis for the results of the experiments.

Providing the data on the four concepts (see section 5.1.2) (definitions, acoustic portraits, and prototypical instruments) then allowed us to identify the salient aspects of each concept that we wished to highlight. We thus reflect on the instrumentation and orchestration of *Quadrangulation* and put special care into imagining the transitions or temporal interpolation between

[3]His individual results were not taken into account for the analysis of the study in the second chapter

two concepts.

In the following section, I will detail the instrumentation, orchestration, and structure of *Quadrangulation*.

5.2 Design and analysis of *Quadrangulation*

5.2.1 Instrumentation & Orchestration

Instrumentation: set of instruments present in a musical work.
Orchestration: the art of combining the instruments' timbres, according to their registers and playing modes.

Instrumentation

Quadrangulation's instrumentation had to meet two constraints – related to our data, and to the recording conditions of the piece. Indeed, as the recording took place in IRCAM's studio 5, we had a maximal capacity of seven musicians.

We therefore proceeded to select the instruments that would be played by the seven musicians, based on acoustic data and prototypical sounds for each concept (see Table 5.1. We needed instruments that could both bring out characteristic aspects of each concept (e.g., glockenspiel for brightness) and be versatile enough to take part in the depiction of multiple concepts (e.g., the trombone with its various timbre-changing mutes). We avoided using instrument that would have redundant roles (e.g., clarinet and saxophone). Finally we also optimised by sometimes choosing pairs of instruments that could be played by one musician like keyboards and clarinets.

Figure 5.2 reports the pruning process for the final instrumentation that consisted in a clarinet, a bass clarinet, a bassoon, a trombone, a trumpet, a cello, a doublebass, a marimba, and a glockenspiel.

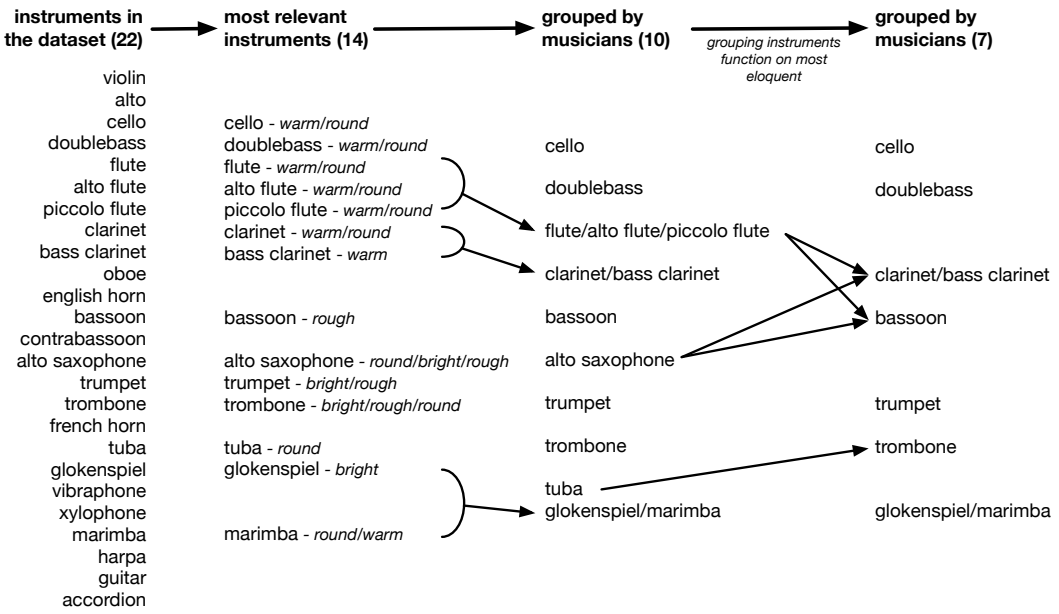


Figure 5.2: Selection process of the instruments heard in *Quadrangulation*.

Orchestration

The orchestration had to reflect the acoustic results but also to showcase the prototypical instrumental features that we obtained for each concept in chapter 2 and 4. For instance, expert participants labelled a doublebass *pizzicato* playing a low-register piano note as the roundest sound. Moreover, they, labelled a glockenspiel played with hard sticks as the brightest sound. Thanks to the composer's expertise, we could also imitate specific instruments that we did not selected in the instrumentation, such as a high-register, piano tuba sound that was played by a trombone with a mute. The composer also suggested playing modes that were not included in the labelled dataset but fitted the acoustic description well, such as crushed sounds for the double bass and cello.

5.2.2 Harmonic content

To focus the listeners' attention on timbral and orchestral characteristics, we made sure to neutralize harmonics that could interfere with the timbral concepts on emotional levels. In particular, in the first study of chapter 2,

we observed that warm and round sounds were considered pleasant, and that many of the participants in the study found that rough sounds tend to be aggressive (see Table 2.2 and Figure 2.2). We therefore had to avoid the obvious pitfall of using triads or complacent chords for roundness and warmth and dissonant chords for roughness. For this purpose, we organized the piece around the superposition of chords formed by three multiphonics from the roughest sounds in the database. These chords were reproduced in multiple copies to present each of the concepts in their most prototypical form, accounting for the pitch and register constraints that correspond to the acoustic descriptions (e.g., high register for brightness, low registers for warmth). These chords act as signals that a new concept is being presented to the listener. Figure 5.3 presents the harmonic arrangement of these three multiphonics to form the core of the harmonic content – and the variations they undergo for each concept.

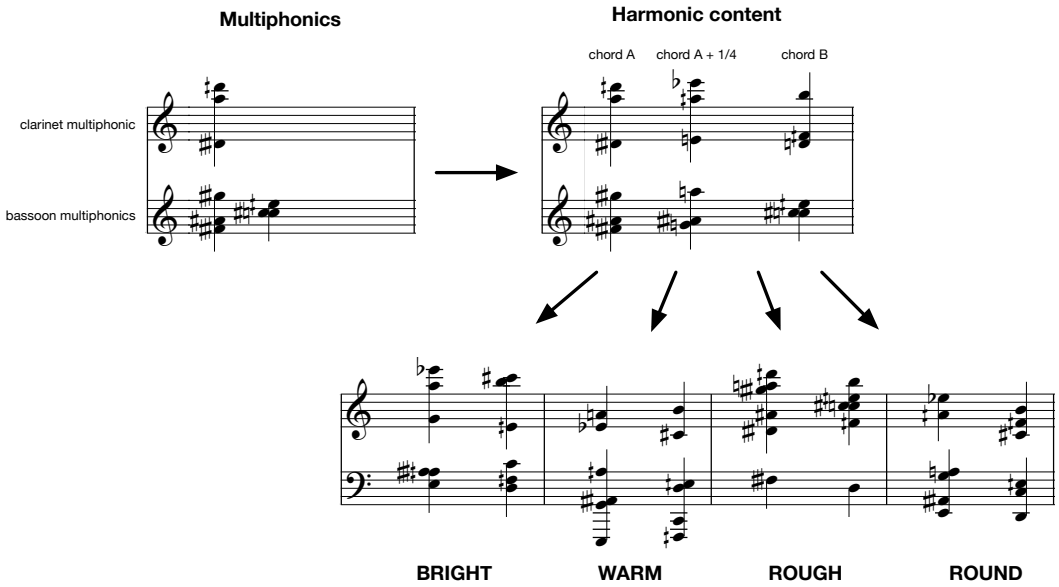


Figure 5.3: The harmonic content of *Quadrangulation* based on bassoon and clarinet multiphonics.

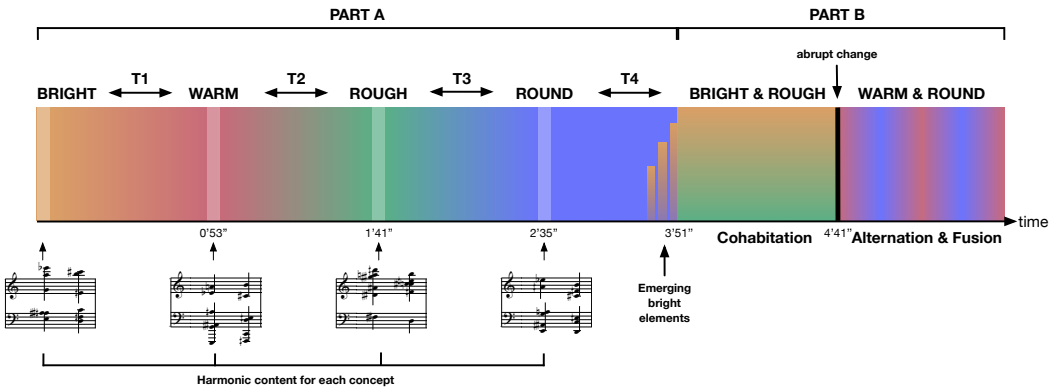


Figure 5.4: Schematic representation of the temporal structure of *Quadrangulation*. T1,T2,T3,T4 are transitions from one concept to another.

5.2.3 Temporal structure

Figure 5.4 is a schematic representation of the temporal structure of *Quadrangulation*. It is organised in two main parts. The first part (PART A in Figure 5.4) corresponds to the sequence of the four concepts in this order: $bright - warm - rough - round - bright$. The concepts first appear in prototypical forms (light vertical lines in Figure 5.4) that kept the same harmonic content. They alternate with interpolations from one concept to another. The second part (PART B in Figure 5.4) of the piece presents typical interactions between pairs of concepts, namely bright/rough that show no significant correlation and warm/round that appeared as strongly correlated (Figure 5.1-B). See appendix C.3 for the spectrogram of *Quadrangulation*.

Part A

Here is the detailed organization of the first part of *Quadrangulation*, including the transitions from one concept to another.

T1/ Bright-Warm (bar 1-14) : After the presentation of brightness, the transition to warmth is made in one movement. It starts with a "*hoquet*" (hiccup). This mode of composition is based on the superimposition of binary (4 sixteenth notes) and ternary (3 triplet eighth notes) metrics of each beat.

Each of the 6 time slots created is thus played by a single instrument at a time on each beat (see Figure 5.5).

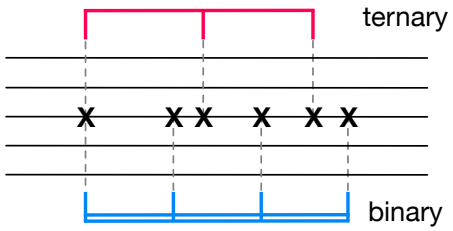


Figure 5.5: Example of the compositional element called "hoquet" (hiccup). Each corresponds to a note played by a single musician.

This process goes along with an interpolation on the harmonic and dynamic level. Thereafter, the duration of the notes gradually lengthens and the register keeps falling. As the winds stop playing (bar 11), the strings and the marimba keep descending gradually in range until the presentation of warmth (bar 14).

T2/ Warm-Rough (bar 14-28): Warmth persists with musicians playing in homorhythmic style until they reach the general silence at the second eighth note of bar 17. The transition consists of two stages. From bar 17 onwards, the sounds gradually shift out of phase while maintaining the same harmonic reference as before. Rough sounds (i.e., *flatterzunge*, *tremolo*, multiphonics) appear little by little to prepare the arrival of roughness.

T3/ Rough-Round (bar 28-44): After the presentation of roughness (four bars), the transition to roundness takes three phases. First of all, the sounds are shortened and played in fast rhythms to give a feeling of decelerating roughness. Second, playing techniques stay the same for two bars. Time seems to stop. This sensation serves as a pivot for the third stage. Third, the winds lengthen the notes they play while the strings play *pizzicato* and let the notes resonate. In addition, harmonics slide towards a low-medium register of roundness.

T4/ Round-Bright (bar 44-65): Prototypical roundness is presented for four bars. After this, strings keep playing *pizzicato* on the forefront, while the winds produce layers of sound in the background, with slow dynamic effects. The harmony is frozen. Then brightness suddenly appear three times with a fast trumpet crescendo with a harmonic mute and fast glockenspiel phrases. Roundness is hence contaminated by brightness. At the end (bar 65-66), the harmonic and rhythmic pattern of brightness we heard at the beginning is presented once more, but played much less loudly.

Part B

In this second part, the piece explores the specific relations between bright and rough, then between round and warm.

Cohabitation/ Bright-Rough (bar 66-80): This first phase corresponds to the cohabitation of brightness and warmth. It is a demonstration that the two concepts can coexist without impacting their respective key characteristics. Rough and bright playing modes alternate and even end up superimposed, yielding a feeling of saturation between bars 76 and 80. This phase ends in an abrupt stop of all instruments except the bass clarinet playing pianissimo and transitioning to the next phase.

Alternation and Fusion/ Warm-Round (bar 80-end): The bass clarinet enters alone in its lowest register with a concealed breathy sound (bar 80). This last section represents the similarities and subtle distinctions between warmth and roundness. They are primarily expressed in the timbres of prototypical instruments. We hear the warmest bass clarinet sound in the whole dataset, then the strings follow with the double bass rubbing its bow on the bridge and the cello in unison with the bass clarinet. Roundness is then expressed with a single trombone played with a bowl mute in its upper register, imitating a tuba. After a few chords that embody the characteristics of both concepts, the piece ends with a long piano chord and a final pizzicato from the double bass.

5.3 Recording session

The recording of *Quadrangulation* took place on April 10^{th} 2022 in IRCAM's studio 5 in the presence of the seven musicians, a sound engineer, and the composer Bertrand Plé. Because of the complexity of some parts of the piece, I took the responsability of conducting the musicians both for the rehearsals and the recording. Figure 5.6 shows the disposition of the instrumentalists for the recording session. The recording, mixing and mastering were done by Jérémie Bourgogne (IRCAM). The group of seven musicians was constituted of Hélène Richard (clarinet and bass clarinet), Diane Mugot (bassoon), Simon Philippeau (trombone), Luce Perret (trumpet), Iris Plaisance-Godey (double bass), Solène Chevalier (cello) and Quentin Broyart (marimba and glockenspiel). The recording session consisted of two rehearsal services, and we did three takes of the piece.

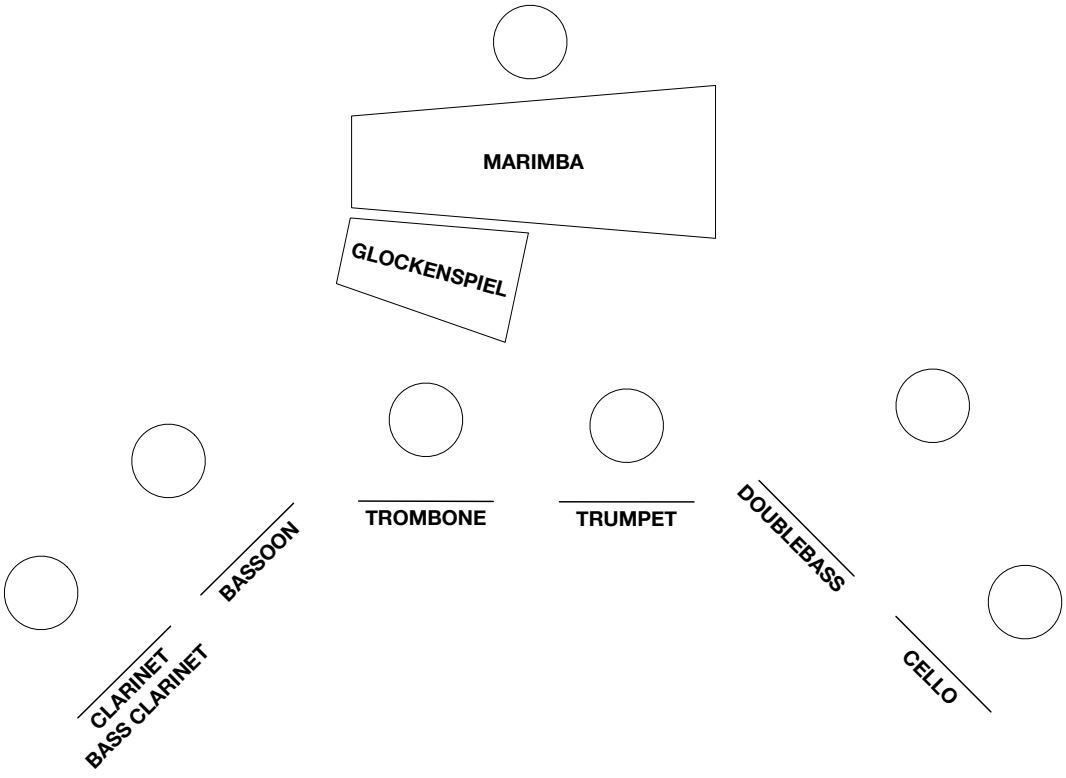


Figure 5.6: Disposition and of the musicians during the recording session at IRCAM's studio 5.

Figure 5.7: Pictures from the recording session – credits Léane Salais.

5.4 Experimental & pedagogical perspectives

Creating *Quadrangulation* was a challenge for the transmission of the knowledge gathered through our two studies (Chapter 2 and 4). We had to state the definition of the concepts and their interactions clearly, beyond the composer's personal beliefs.

This musical work could both be seen as an experimental material to assess the possibility of transmitting the identity of the four sound concepts – and a pedagogical material for semantically informed composition in orchestration or composition classes.

Validating the piece requires comparing it with the empirical results of this thesis. We could present the semantic definitions of each concept with or without a few sound examples to a group of naive participants, and then ask them to annotate excerpts from *Quadrangulation* in an experiment involving pairwise comparison or BWS. Alternatively, a continuous evaluation method could help monitor a specific perceptual/cognitive concept as the stimulus unfolds in time. Such continuous rating methods have already been used to measure emotional responses to music. The interested reader is referred to research by Schubert (2001), and by McAdams et al. (2004) for a description of the methodological issues associated with the continuous evaluation of complex musical materials.

Other options would include annotating the whole piece using automation curves. In the end, we would compare the results of the aforementioned group of participants who were not told any definition. On the one hand, if the definition-aware group performs better at identifying the concepts, it would mean that this collaboration was coherent enough. On the other hand, if both groups identify the concepts in the same way, it could mean that the piece is a sufficient learning material (if the results are good), or that the concepts were poorly rendered by the piece otherwise.

On another note, the second part of the piece should be examined with attention because it questions the nature of the interaction of certain concepts.

For example, we are curious to see whether listeners can identify the cohabitation of brightness with roughness and the alternations between roundness and warmth.

A less formal experimental option could be to make the participants listen to the piece with specific listening instructions. We could then conduct interviews to find out how the participants understood the concepts and their relations.

5.5 Chapter conclusion

Shared mental representations underlie metaphorical sound attributes

In this chapter, I presented the creation and structure of *Quadrangulation*, an explanatory musical work for four metaphorical concepts related to timbre. To produce it, we transmitted different types of knowledge related to brightness, warmth, roundness, and roughness to a composer who wielded them for his musical creation. Eventually, we recorded this piece describing each of these concepts and their interactions. An immediate scientific perspective for *Quadrangulation* is to design a validation experiment with the recording, to evaluate the piece's capacity to convey the meaning of the four concepts.

Quadrangulation is the quintessence of our empirical findings that is in line with the art/science vision at the origin of IRCAM, which Pierre Boulez poetically described as "the utopian marriage of fire and water" (Boulez, 1986). It invites listeners to discover sound concepts as story-characters who converse, attract, or repel each other. *Quadrangulation* can be simply listened to for the pleasure of the ears, or for educational purposes, when trying to learn the meaning of metaphorical sound concepts.

6. Discussion

6.1 Summary

This thesis explored the semantic and acoustic portraits associated with well-known metaphorical sound attributes wielded by sound professionals: brightness, warmth, roundness, and roughness. The results therefore provide access to the mental representations associated with these attributes, as well as a reflection on the cognitive pathways involved in their formulations:

1. Through interviews with sound professionals and a qualitative analysis of the verbal data, we detailed the meaning of the four metaphorical attributes of sounds. Specifically, we established a corpus characterizing the attributes with acoustic, metaphoric, and source-related descriptions and linked them to prototypical orchestral sound samples (section 2.1). We then refined the corpus of descriptions through an online survey that allowed us to build a definition for each attribute based on the most relevant description (section 2.2). Our results show that, despite rich and fine-grained verbal data, there is some circularity in the definitions and references to other metaphorical and ambiguous descriptors.

2. To determine the appropriate measurement method to obtain acoustic portraits of the metaphorical descriptors, we benchmarked methods for the subjective evaluation of sound qualities. Specifically, we compared the performances of the new Best-Worst Scaling (BWS) method with the classical rating scale method (section 3.2). For the first time, we validated the use of BWS for the evaluation of sound perceptual quali-

ties. Importantly, our results confirmed its accuracy and highlighted its efficiency.

3. Using the newly validated BWS method on a dataset of orchestral instruments, we obtained acoustic portraits for all concepts/attributes based on Machine Learning technologies. It allowed us to characterize the sound experts' shared mental representation of the four sound concepts (sections 4.1, and 5.1.2). Eventually, we reported on the relations existing between them in an understandable way.

4. By comparing the results of three populations that took part in the experiment (i.e., sound engineers, conductors, and non-experts), we revealed the specific and shared aspects associated with the understanding of concepts across populations (section 4.1). First, we observed a high consistency of results between expert populations, as well as better reliability and agreement among sound engineers. Second, unlike other attributes, the well-known brightness appeared to have the least shared mental representation across populations, highlighting a potential cultural evolution of brightness' meaning in expert populations.

5. Thanks to the artistic collaboration with the composer Bertrand Plé, our scientific results brought forth a musical work called *Quadrangulation* (chapter 5). The objective of this piece was to channel the expert knowledge gathered through our different approaches (chapters 2 and 4) to illustrate the four concepts and their atypical relations. This rich process fed an artistic strategy for communicating semantic and acoustic data of metaphorical sound concepts. In the end, we created a piece of music that is semantically informed by experimental results, enabling a new way of making the ambiguous essence of sound description more explicit.

Overall, these different studies highlighted the multiplicity of levels of understanding of the four concepts studied, underpinning their complexity. We have however shown that humans, and more particularly sound experts, rely on overall similar acoustic features when describing sounds. Strangely

enough, there is a clear antisymmetry in the mental representations between the less verbally defined term (roughness) and the best defined one (brightness). The musical work allowed us to materialize the results of our studies, emphasizing the shared mental representation that we generally have of brightness, warmth, roundness and roughness.

This general discussion is organized in two main parts. First, I will present Deep Learning-based perspectives for our Best-Worst Scaling results that would promise new horizons for the visualization and the analysis of sound attributes (section 6.2). Second, I will compare the information obtained for each concept throughout our study. I will question the importance of the acoustic and semantic characteristics as well as the concept denoted by each attribute. I will propose to identify the nature of the relations between each concept and propose a geometrical representation. Finally, I will discuss possible cognitive pathways bridging the gap between sounds, metaphorical descriptions, and sound mental representations.

6.2 Beyond BWS scores: A semantic timbre space

In this section, we wish to highlight the modularity of our experimental method and its potential application to other subjects. In chapters 3 and 4, we discussed the ergonomics and versatility of the experimental procedure. Here we propose a new representation of the labelled sounds, invoking a Deep Learning classification task.

6.2.1 A semantically informed timbre latent space

We saw in chapters 3 and 4 that Best-Worst Scaling offers an atypical response format that provides different ways to understand the judgments made by the participants (Hollis, 2018). Going further, we thought of developing a Deep Learning tool for the generation of a latent timbral space based on Best-Worst Scaling judgments. The usefulness of this latent space would then lie in the

visualization of the data it offers and in the perceptual labelling of new sounds – i.e., sounds that had not been used for the experiment.

To design such a tool, we took inspiration from state of the art recognition/classification systems based on Machine Learning. A proposal by Chen et al. (2018) dedicated to Speech Emotion Recognition raised our interest. Therefore, we chose this method as a starting point for this work. The inputs of our classification system were 3D log-mel spectrograms of the sounds of our dataset (see appendix D). It consists of an encoder which produces high level timbral salient features when fed with these representations. The information is then temporally summarized and condensed in fixed size vectors. The architecture of such system allows us to generate a latent space in which each sound lies. By constraining this space to meet certain properties, we expect to turn it into a perceptual timbral space.

Re-interpreting BWS judgments for a perceptual loss

We had to choose a training criterion reflecting perceptual judgments to train our model. To do this, we first considered using the BWS scores obtained for each sound sample from our dataset (see chapter 4) in each dimension of interest (i.e., roundness, warmness, roughness and brightness) as groundtruth in a multi-class classification scope. However, this approach would not make use of most of the richness of raw judgments – i.e., the paired comparisons between sounds (see section 3.1.2). Hence and for better classification performances and structuring of the latent timbral space, we focused on using the raw judgements as groundtruth. In practical terms, we considered the judgments made by participants on sounds in terms of distances between sound-related vectors in the latent space. For example, in trial X (A, B, C, D), if sound A is chosen as the best and sound D as the worst, we expect the distance between them to be greater than their respective distances to sounds B and C. Thus we propose a learning criterion inspired by the classic triplet-loss (Schroff et al., 2015). Figure 6.1 presents the interpretation of raw judgments in terms of distance relations between sound embeddings.

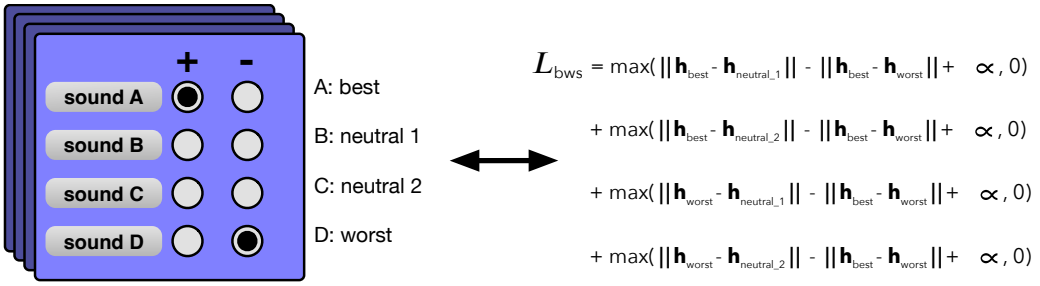


Figure 6.1: Formulation of the perceptual loss, based on Best-Worst Scaling judgments. **h**: sound embedding vector corresponding to a sound in the latent space. and α a non-zero constant

By training a model that satisfies these distance constraints, we wish to arrange the latent space so that it incorporates a perceptual structure. Figure 6.2 displays a schematic representation of the model (right) and depicts the configurations of the latent space at training (right-top) and inference (right-bottom) phases. At training time, the embedded sounds are moved through the space so as to comply with the constraints. See appendix D for the presentation of the model and a more detailed explanation on its implication with the perceptual data.

6.2.2 Uses of the timbral latent space

An analysis of the timbral latent space can bring information regarding the shape and boundaries of timbral clusters and promote the emergence of sound prototypes. On a semantic level, it approaches the prototypical understanding of categorizations seen in the introduction (see section 1.2.1). Consequently, concepts for which mental representations are very similar, such as roundness and warmth, would be intertwined in this latent space. Finally, one could imagine mapping acoustic features to the dimensions of the latent space, to understand its configuration.

We will also be able to plunge new sounds into this latent space to obtain their semantic portrait on the basis of their 3-D mel spectrograms. The position of a given sound in the latent space and its relative distance to timbral clusters of of the four concepts will inform the semantic nature of its timbre.

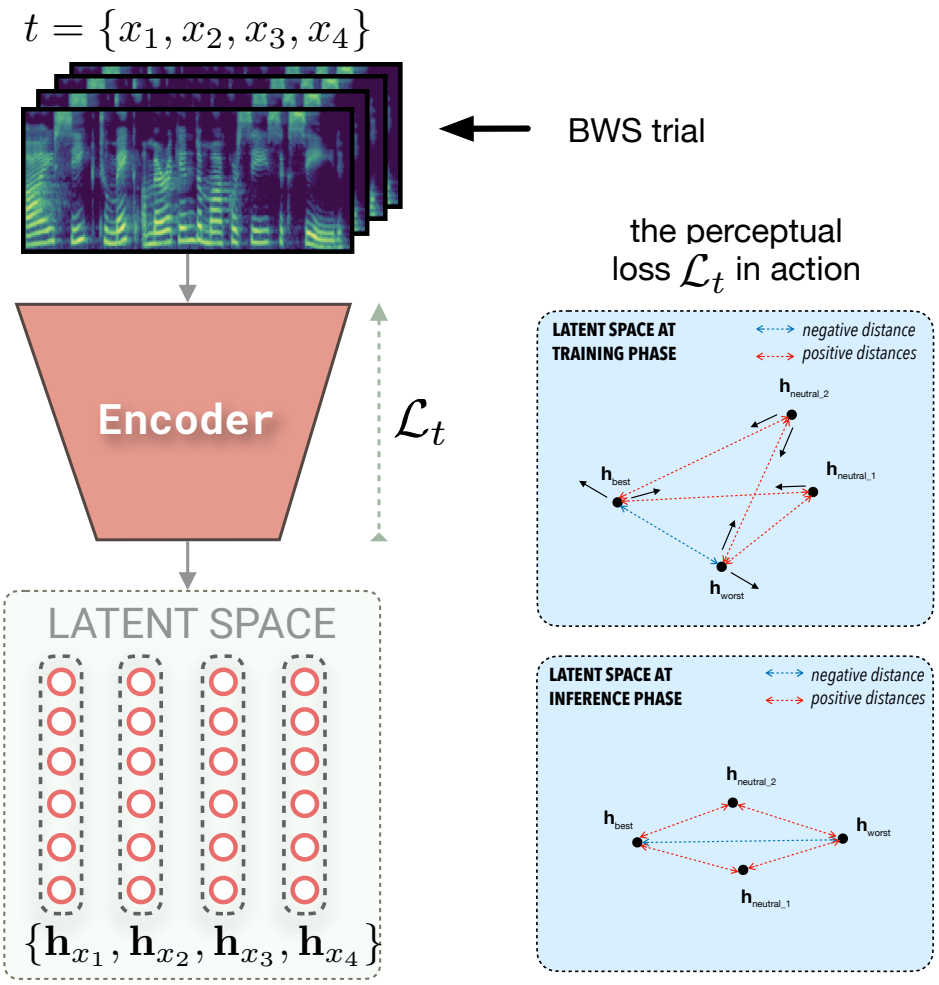


Figure 6.2: Simple schematic representation of the classification system, along with an expression of the action of the perceptual loss. **x**: a sound in a BWS trial. **h**: sound embedding vector in the latent space.

Overall, and just like our experimental design, the use of the technology we provided is not restricted to timbre. It can be applied to other issues related to sounds of different natures like the results of the BWS experiment on vocal attitudes introduced in section 3.3.1 of chapter 3.

6.3 From semantics to acoustics

We studied four different attributes across two differing approaches, leading to several types of characterization. Thus, we want to answer the following question: "Do we talk about these concepts in the same way that we perceive or represent them?". Like Faure (2000) who observed a gap between the use of a descriptor and its relevance, we are curious to gather and compare semantic and acoustic information.

To answer that question, I will focus here on results obtained from the population of sound professionals, just like for the conception of *Quadrangulation* (see section 5.1.2).

First, I will compare the verbal and acoustic descriptions of each concept. In the first study, descriptions were of three kinds (i.e., acoustic, metaphorical and source-related). With a different perspective for each concept, I will discuss the links existing between these descriptions and the acoustic portraits obtained by the expert population of the experiment in chapter 4. Second, I will question the importance and definition of the attack for timbre descriptions. Third, I will propose a semantic explanation of the relations between concepts along with a 3D representation of them. Finally, I will comment on possible cognitive mechanisms and pathways mediating the use of metaphorical descriptors for sounds.

6.3.1 Brightness: Why not so clear?

Across the thesis, brightness seems to have a stable acoustic denotation of its meaning while presenting differences at the level of its mental representation. Looking at the two most prominent descriptions of its definition, there are "high-frequency spectral content" and "high pitch". The former is validated by the relation of the brightness scores obtained in chapter 4 with the spectral centroid and the spectral bandwidth, and the latter by the importance of fundamental frequency (F0). All of this is coherent with the literature and show once more that brightness is mostly defined by spectral features.

Although our results promote a strong link between brightness and the spectral centroid, we were surprised to see that such an important timbre descriptor is not the most consensual. As argued in the discussion of chapter 4, we proposed that it may be the manifestation of a dead metaphor phenomenon, meaning that its technical use in an expert domain has specified its meaning (see section 4.1.4). But the fact is that even within populations, this term is not the most consensual. In other words, it suggests that everyone has a personal view of brightness compared to other concepts, or at least a different way of representing it in the sound domain.

In spite of these differences, we observed consistency in the choices of sounds made during the interviews with the experts and the top sounds of the BWS experiment (see Table 5.1). Specifically, the sound of glockenspiel that has the highest spectral centroid seems to stand out from the rest of the dataset. Therefore, we wondered whether the denotation of brightness is something prototypical, namely, that it is based on the reference to specific source types. In an experiment evaluating the impact of source information on the evaluation of the timbral brightness of musical instruments, Saitis and Siedenburg (2020) argued that having additional information on the type of source did not improve in any way the brightness model they proposed. Similarly and to assess the influence of the source type on brightness representation, we fed the model presented in chapter 4 with information about instruments and playing techniques. But this additional information led to a very poor improvement of the model's performance. Even if all glockenspiel sounds occupy the top of the brightness rankings, this may support the idea that brightness is an attribute that emerged from *reduced listening*, and is not attached to a specific source type.

6.3.2 Roughness: Poorly expressed but clearly represented

The psychoacoustic definition of roughness is rather straightforward (see section 1.5.2). However, the not-so-acoustic verbal description given by sound professionals suggested a fuzzy mental representation of roughness. Yet, it is pretty clear that the expert (and even non-expert) participants of the different

experiments were all having the same representation of auditory roughness (see Figures 4.2, 5.1). We are in the opposite case of brightness which had a very clear verbal definition and a more ambiguous mental representation that is specific to each individual. Conversely, roughness' definition was vague, but ended up being the most consensual and best modeled concept according to the experience of chapter 4.

As explained in its definition (see Figure 5.1), the roughness of a sound is estimated in reference to the sound of a friction, or more specifically, of a friction with a rough surface. It is thus possible that participants of the perceptual experiment have tried to imagine how to restore this friction sound regardless of the instrument. This could be the reason why no particular source was estimated as rough in our results (see Table 2.2). Participant focused on some playing techniques like *flatterzunge* and multiphonics that offer temporal irregularity at different speeds, mediating a sensation of roughness.

The acoustic features involved in the modeling of roughness are mainly related to noise. In a coherent way, a metric inspired by Arnal et al. (2015) – who studied voice roughness – was the most salient feature to describe roughness. However, according to the contributions of other features like Harmonic-to-Noise Ration (HNR) or spectral crest, the sound concept of roughness may not be reduced to a simple amplitude modulation created by the proximity of sounds in the critical bands.

With a stable mental representation but a definition that is not acoustically explicit, it is difficult to account for the cognitive process involved in defining sound roughness. And to this day, we wonder whether a rough sound is the reproduction of friction against a rough surface or an imitation of the sound created by any friction.

6.3.3 Warmth and roundness: same but different

Throughout this study, it always felt rather unnatural to define warmth without mentioning its relation to roundness.

However, and if we take a step back, the intuitive connection that people make between these two descriptions for sound is not obvious. For instance,

the two concepts are very different from a semantic point of view: warmth designates a tactile sensation while roundness comes from the observation of the shape of an object. Here, we compare one last time the results grouped in Figure 5.1 on warmth and roundness.

Let's start with some commonalities. First, the definitions obtained during the interviews seemed to put them on an equal footing in terms of level of understanding and the results of the perceptual experiment confirm this trend (see *compliance* Fig. 4.2).

Second, we notice that for both a round or a warm sound must have a low spectral centroid. Moreover, it is important to note that this information is largely expressed verbally as such by sound professionals.

Third, although descriptions of the attack were more salient for roundness than for warmth, its depiction no longer seems relevant in the results of the BWS experiment for either concepts. We will discuss the trajectory taken by the importance of attack in section 6.3.4.

Finally, the two concepts shared multiple emotional descriptions during the interviews. Indeed, warm and round sounds were considered "pleasant" and were always opposed to "aggressive" sounds. This suggests the intervention of an emotional processing of sound as presented by Schön et al. (2010) in his theory of conceptual processing of sound (see Figure 1.4). In other words, to access the concepts of warmth and roundness, we may go through a step of emotional judgment when listening to a sound.

We also brought out some subtle differences between warmth and roundness thanks to their acoustic portraits. First, a warm sound will necessarily be a rather low-pitched sound, which is not always true for a round sound. Second, a round sound seems to have hardly any noise (whatever the noise's nature), while a warm sound is less impacted by noisiness.

We can speculate on how this difference in the importance of noise is articulated in the verbal descriptions of each of the concepts (see Figure 2.2). First, the non-noisy aspect of round sounds can be embodied by descriptions such as "homogeneous", "smooth" and "full" (all three opposed to roughness). Second, descriptions such as "rich", and "non-pure" will be emphasized for warmth. Further experiments similar to the one described in Chapter 4 would

be necessary to evaluate the meaning of these descriptions, which are just as metaphorical as the ones treated in this thesis.

6.3.4 Where did the attack go?

Attack morphology is a complex parameter for sound characterization that covers a variety of forms. For instance, in his *Traité des objets musicaux*, Schaeffer (1966) proposed multiple attack shapes associated with different semantic labels (see Figure 6.3).

TIMBRE DYNAMIQUE		1	2	3	4	5	6	7
GENRES D'ATTAQUES	Tracé bathygraphique							
	Nature d'attaque	ABRUPTE ou explos.	RAIDE	MOLLE	PLATE	DOUCE	SFOR-ZANDO ou appui	NULLE ou très progress.
	Symbole conventionnel							
		(choc ou plectre) sans résonance appréciab.	(marteau feutré) avec forte résonance liée	(pizz ou mailloche douce) avec résonateur	(pseudo attaque) ou mordant	son posé sans attaq. apparente	ou crescendo rapide	perception du profil
PRÉDÉTER-MINATION DU PROFIL en fonction du genre d'attaque	Profil dynamique	pointe dynamique (choc)	régulier dégressif	renforcem. du résonateur	nul, sauf la pseudo-attaque	profil nul	Profil caractérist. sons en général courts	seul cas de seuil émergence du profil
	Profil harmonique	son double (2 timbres)	appau-vrissement	réponse du résonateur	nul dans les instruments tels que l'orgue varié en musique électr. ou cordes	Profils souvent progressifs	Timbrage caractérist.	Profils le plus souvent liés ou artificiellement indépendants.

Figure 6.3: Types of attacks described by Schaeffer (1966)

It has been suggested in the literature that the attack has a role of acoustic correlate of the secondary dimension of timbre (Lakatos, 2000; McAdams et al., 2017). However, and despite being an important factor for three of the attributes during the interview, it was barely present in the survey results and non-existent in the BWS experiment results.

The discrepancies regarding the role of the attack between studies 2.1 and 2.2 are likely due to a change in the presentation of the sounds. During

the interview, we saw very diverse descriptions of attacks (e.g., "soft attack", "round attack", "strong attack", "dry attack"). However, and considering the analysis of the BWS results, there is a good chance that our attack features (i.e., log-attack time, attack slope) did not properly account for the diversity of attacks in our database. For instance, as mentioned above, a soft or slow attack was very important for the definition of roundness during the interviews. However, none of the attack features significantly contributed to roundness scores in the analysis of the BWS scores. Even the double bass *pizzicato*'s samples – the roundest sounds in the dataset according to participants – did not reveal an especially slow attack. In consequence, it is likely that our model lacks information about the diversity of attack morphology, because of our descriptors or the sound dataset itself.

Alternatively, a concept may be connected to an attribute on a lexical level without being similarly associated to it on a perceptual level (Goldstone et al., 2013). For instance, a sound with a soft attack may not be perceived as being round even though a soft attack is considered as an important component for defining roundness, implying some kind of asymmetry in the concept's representation.

Further studies are needed to reveal the temporal or spectro-temporal properties of attack morphology, in order to understand the perceptual representation of semantic descriptions of the attack. A first lead would be to reveal the profile of a specific kind of attack using reverse correlation (Ahumada Jr and Lovell, 1971), a method that allowed to successfully reveal the spectro-temporal profiles of social traits in the voice (Ponsot et al., 2018), and crucially, loudness perception (Ponsot et al., 2013). Ultimately, such a study would allow to report on the specificities of each attack and the existing links between attack and timbre.

6.3.5 Semantic relations between metaphorical concepts

In our quest to shed light on the concepts denoted by metaphorical descriptions, we incidentally explored the nature of their semantic relations. Crucially, this information is intimately connected to one of the main goals of our study:

to unravel the meaning of warmth and roundness. Previous timbre semantics research had already pointed out the similarity between warmth and roundness and their rather opposition with brightness and roughness. Our study allowed us to deepen prior understandings of the relations between concepts in order to enrich their definitions, while accounting for atypical relations in all pairs of concepts. The relations expressed here are based on the chapter *Meaning and Logic* in Löbner (2013).

Orthogonality: brightness and roughness

As noticed in our different studies with expert participants, brightness and roughness are expressed on two different dimensions both on the semantic level and on the acoustic level. The two concepts do not seem to interact and can cohabit without impacting each other. This is something that we illustrated in the second part of *Quadrangulation* and that we hope to verify in the future (see section 5.2).

However, we noted that similarly to the definition of brightness, the relation between brightness and roughness is not necessarily consistent for all people. Thus the non-expert participants of the study of chapter 4 tended to oppose brightness and roughness in terms of tonal strength (i.e., HNR, spectral crest).

Synonyms: warmth and roundness

In section 6.3.3, we brought out some strong semantic and acoustic links between warmth and roundness. We cannot decently consider them as synonyms because of the fine difference in terms of acoustic portraits (i.e., importance of the F0 for warmth, importance of noisiness for roundness) and meaning. Moreover, we uncovered that this difference is also expressed in their relation to brightness and roughness. Therefore, we describe them as partial synonyms, i.e., as two words that may have one meaning component in common. Obviously, the similarity between warmth and roundness is such that they may have a very high degree of partial synonymy.

At this point, we can speculate about a possible relation of subordination (i.e., $A \Rightarrow B$) between the two. In other words: "Is a warm sound necessarily round, or is a round sound necessarily warm?". While we cannot answer that question, the importance of the F0 for warmth and of the quantity of noise for the roundness let us think that such a relation does not exist.

Incompatibility

Based on their descriptions and on the correlations of their BWS scores, four of the relations between concepts are oppositions, namely brightness-warmth, brightness-roughness, warmth-roughness and roundness-roughness. We generally count two types of logical opposition for words: complementarity and incompatibility (Löbner, 2013). On the one hand, the relation of complementarity imposes a perfect opposition between two concepts or expressions A and B:

> Two terms A and B are **logically complementary** if and only if their denotations (categorizations) have no elements in common and together exhaust the set of possible cases.

Semantically, the two words can either be directional opposition or antonyms.

On the other hand, the incompatibility relation is less restrictive and expresses the fact that only the combination of A and B is impossible, without restriction regarding the combination of non-A and non-B.

> Two terms A and B are called **logically incompatible** if and only if their denotations have no elements in common.

An incompatibility between descriptions is the manifestation of a heteronymia, (e.g., the seven principal notes of music). This is precisely what seems to be expressed by our different negative correlations. Thus, while we had sounds that could be both rough and bright or round and warm, they could not be bright and warm, or bright and round. Crucially, a sound that is not round is not necessarily bright or rough.

Although all are heteronyms to each other, their opposition can be quantified thanks to the results presented in Figure 5.1. Here are the relations of incompatibility, from the strongest to the weakest:

1. roundness-roughness
2. roundness-brightness
3. warmth-brightness
4. warmth-roughness

While these relations are verified at lexical and perceptual levels, it is interesting to speculate that it may be due to the physical nature of the instrumental sound sources which prevents any spectral or temporal incongruity. Thus, through the synthesis of acousmatic sounds, it is maybe possible to obtain sounds that will be judged both round and bright, or warm and bright.

6.3.6 Geometric representation of the four concepts

In order to finalize the reflection on the relations between concepts, we propose a spatial representation showcasing these relations, based on experts BWS judgments. In accordance with the previous section's discussion, we name P_{BR} the plane defined by the orthogonal dimensions of brightness and roughness. Thus, the 3D-representation is formed of P_{BR} and a third orthogonal axis. Thanks to that, we can draw the roundness and warmth vectors according to their projection on P_{BR}. To do this, we performed two multiple-linear regressions (i.e., ordinal least square regression) for warmth and roundness, using brightness and roughness values as inputs. We reported on the degree of prediction (R^2) and the coefficients associated with brightness for both task in Table 6.1.

Table 6.1: Multilinear regression results of warmth, and roughness. $c_{bright.}$ and $c_{round.}$ denote respectively the coefficients of brightness and roughness for the regression. R^2 is the coefficient of determination.

	$c_{bright.}$	$c_{rough.}$	R^2
Roundness	-0.57	-0.60	0.58
Warmth	-0.48	-0.31	0.30

The coefficients of brightness and roughness are consistent with the observed correlations of their BWS scores with those of warmth and roundness. We see that roundness is closer to P_{BR} than warmth. Therefore, it seems the remaining variance of warmth and roughness is explained by one or several additional dimensions. Figure 6.4 shows a possible schematic representation of each concept. It features the degree of explanation of warmth and roundness by brightness, roughness, and a potential third axis. Similarly to studies that evaluated the semantic axes of timbre, the continuity of this work would be to reveal the one or multiple latent concepts behind this axis. For instance, Zacharakis et al. (2014) evaluated three semantic concepts of sound: luminance, texture and mass. Therefore, if we were to relate luminance to our brightness, texture to our roughness, the last dimension could be related to something close to mass.

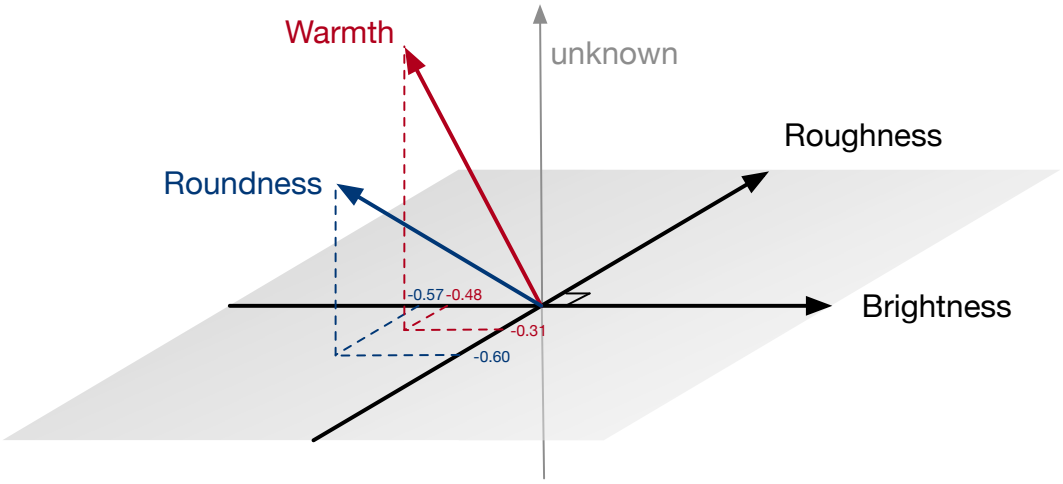


Figure 6.4: Schematic representation of warmth and roundness in the 3D space (*x: brightness, y: roughness, z: unknown dimension*).

6.3.7 The conceptual processing of sensory metaphors

At the beginning of the thesis, we considered different proposals regarding the cognitive processes attached to the origin and development of the metaphorical sound descriptions (see 1.3). Specifically, the discussions revolve around the idea that it would be based either on semantics and cultural aspects (e.g.,

sound expertise), or on a distinct cognitive pathway that does not involve a semantic representation of a concept (e.g., crossmodal correspondence). Although this is not the object of our research and we have not in any way questioned this idea in our experiments, it is interesting to note that the results suggest a specific cognitive processing of certain concepts based on the results of brightness and roughness.

Let's first consider the case of brightness. The behavioral and acoustical results for brightness from chapter 4 exposed a cultural difference in its representation based on sound expertise profiles. We have notably evoked the possibility of the evolution of its metaphorical use, towards a term whose meaning is endogenous to the technical environment within which it is used (see section 4.1.4). Thus, even if the meaning of brightness is coherent with various findings in the literature, this difference suggests the intervention of a cultural and thus semantic channel. In other words, while the brightness of a sound may originally be due to a crossmodal correspondence or an amodal representation, it may have been largely impacted by its verbal description in experts circles.

Roughness seems to have completely opposite characteristics. First, the mental representation of sound roughness is largely consensual, regardless of sound expertise. Second, the corresponding verbal descriptions are vague, metaphorical, and barely related to the acoustic description. Again, we are not sure what cognitive process is involved in defining a rough sound. However, it is rather likely that because of its vague definition and strong shared representation, roughness is a clear expression of a cross-modal correspondence independent of any semantic pathway.

As for roundness and warmth, the process involved in their use for sound description remains mysterious. We made a connection with an eventual emotional processing step (see section 6.3.3), but their semantic complexity and acoustic similarities makes it difficult to account for a unique and clear conceptual process. A potential conclusion would then be that these sound descriptions borrow from several different cognitive strategies to access a concept which would transcend sensory modalities.

7. Conclusion

Metaphors, Sound culture, and Music

In this inter-disciplinary study, we explored the sometimes incongruous pathways from rich semantic descriptions to the acoustic aspects of four metaphorical sound attributes : brightness, warmth, roundness, and roughness. Through two different approaches involving verbal definitions and perceptual experiments we formulated definitions and revealed acoustically-based mental representations for all attributes. We proposed a great overview of the attributes and their relations with each other. In this way, we reported on the partial synonymy of warmth and roundness, and the orthogonality of brightness and roughness. Taking a step further, we reflected on the cognitive processes that would explain the uses of such sensory metaphors. Eventually, our findings on brightness and roughness will be a great material for future work willing to uncover the unanswered question: "Is the sound projection of a metaphorical sound concept the result of a cultural reinforcement, or of a more direct connection with the source sensory modality?"

On a different note, we showcased our portrayals of the four attributes in a semantically informed musical creation named *Quadrangulation*. In the future, this piece will be an opportunity to transmit but also to question our conclusions with possible cross-cultural perspectives. *Quadrangulation* is a fair return to the musical world from where the metaphorical vocabulary of sound is usually heard. Despite its explanatory purpose of sound metaphors, *Quadrangulation* remains a piece of music that can be discovered without a word, only with the ears.

A. Supplementary materials of chapter 2

This document contains supplementary materials for chapter 2, and the article named *Investigating the shared meaning of metaphorical sound attributes: bright, warm, round and rough*.

A.1 Study 1: Interviews

A.1.1 Questionnaire of the interviews in French

ID : ATTRIBUT :

FORMULAIRE

1. Quelle est le contexte et la fréquence d'utilisation de l'attribut ?
2. Comment définissez-vous l'attribut étudié ?
3. Pouvez-vous trouver des exemples sonores associé à cette définition (au moins trois) ?
4. Pouvez-vous trouver des exemples sonores opposé à cette définition (au moins trois) ?
5. Comment définissez-vous l'opposé de l'attribut étudié ?
6. Y-a-t'il un affect rattaché à cet attribut et/ou son opposé ?

Figure A.1

A.1.2 Professional profiles of the participants

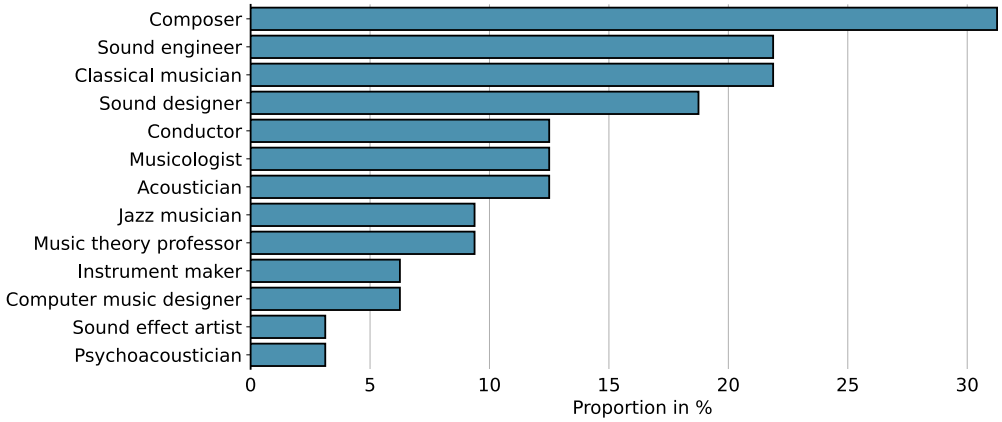


Figure A.2: Percentage of professional profiles in the population for Study 2.1 (1 participant = 2%)

A.1.3 Description of lemmas for the four attributes

Table A.1 reports the number of lemmas for the four studied attribute for the two questions.
Table 2 and Table 3 reports the top lemmas for the four studied attribute for the two questions.

- Question 1: How do you define the studied attribute?
- Question 5: How do you define the opposite of the studied attribute?

Table A.1: Number of lemmas evoked at least by three experts both for question 2 and question 5 of the interviews for each of the four attributes.

Attribut	Question 2	Question 5
Bright	41	50
Warm	68	63
Round	64	45
Rough	37	25

Table 2. Top 100 lemmas for question 2 of all attributes.

Lemma	Nb participants	Lemma	Nb participants
brillant	32	agréable	7
chaud	32	variation	7
rond	32	couleur	7
aigu	29	équilibré	7
attaque	27	parasite	7
harmonique	27	espace	7
rugueux	26	archet	7
grave	24	classique	7
spectre	21	large	6
doux	19	présence	6
médium	19	humain	6
fréquence	17	naturel	6
timbre	17	flatterzunge	6
fort	17	net	6
corde	17	trompette	6
musique	16	gratte	6
piano	14	mezzoforte	6
souffle	14	nature	6
plein	14	vibrato	6
riche	13	contrebasse	6
frotter	13	orchestre	6
bas	13	enveloppe	6
lisse	12	bois	6
bruit	12	aspérité	6
matière	12	froid	6
basse	12	impact	6
chaleureux	11	ensemble	6
clair	11	affect	5
haut	11	violoncelle	5
résonant	10	vie	5
agressif	10	intensité	5
pur	10	gros	5
homogène	9	dense	5
voix	9	irrégulier	5
dur	9	timbré	5
proche	9	organique	5
long	8	défini	5
enveloppant	8	métal	5
bruité	8	propre	5
hertz	8	tessiture	5
temporel	8	registre	5
dynamique	8	confortable	5
clarinette	8	franc	5
percussion	8	creux	5
cuivre	8	volume	5
entretenu	8	généreux	5
vibration	7	surface	5
stable	7	sonner	5
nuance	7	lumière	5
vibrer	7	court	5

Table 3. Top 100 lemmas for question 5 of all attributes.

Lemma	Nb participants	Lemma	Nb participants
rond	31	terne	7
brillant	30	baguette	7
chaud	30	spectre	7
rugueux	29	cuivre	7
aigu	27	bas	7
attaque	25	marimba	6
harmonique	24	vibration	6
lisse	21	musique	6
froid	21	intéressant	6
grave	20	agréable	6
résonant	20	vibraphone	6
fort	18	propre	6
agressif	15	matière	6
doux	15	tessiture	6
corde	15	xylophone	6
timbre	14	bois	6
piano	14	aspérité	6
pur	14	espace	6
fréquence	12	précis	5
contrebasse	12	long	5
cuivré	11	proximité	5
trompette	11	violoncelle	5
clarinette	11	gros	5
mat	11	flatterzunge	5
étouffé	10	voix	5
stable	10	bruité	5
riche	10	vibrer	5
bruit	10	sec	5
court	10	chaleureux	5
flûte	10	variation	5
pizzicato	9	mou	5
dur	9	saturer	5
vibrato	9	métal	5
métallique	9	sombre	5
souffle	9	sul	5
homogène	8	péjoratif	5
violon	8	tuba	5
sourd	8	accordéon	5
cor	8	bartok	5
plein	8	hautbois	5
glockenspiel	8	droit	5
clair	8	bonne	5
médium	8	basse	5
piccolo	7	haut	5
pauvre	7	contrebasson	5
basson	7	archet	5
sourdine	7	précision	4
mezzoforte	7	plat	4
ponticello	7	intensité	4
franc	7	percussif	4

A.2 Study 2: Online survey

A.2.1 Example of the interface of the survey in French

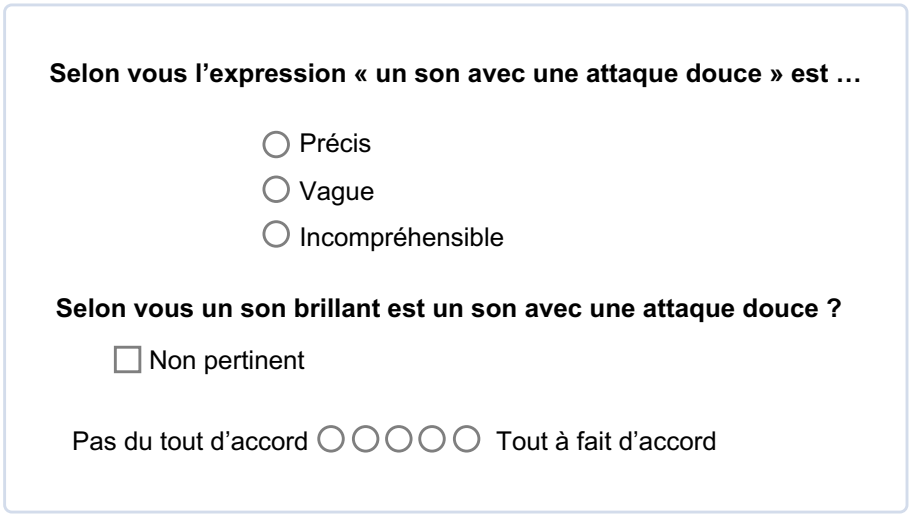


Figure A.3: Interface of the survey from study 2.2

A.2.2 Professional profiles of the participants

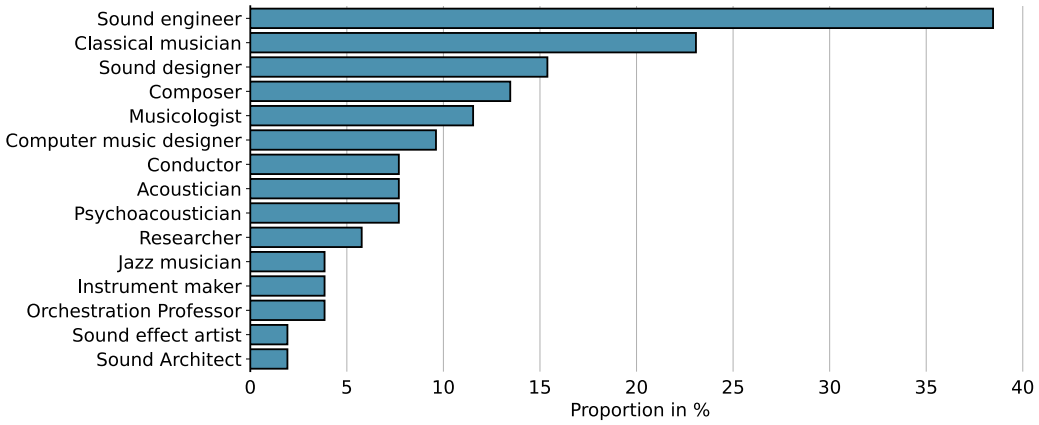


Figure A.4: Percentage of professional profiles in the population for Study 2.2 (1 participant = 2%)

A.2.3 Statistical analysis of the survey results

This section presents the statistical tests performed on the answers for the survey of study 2. All the translations of the descriptions are presented along with the tests.

Tables 1, 2, 3 and, 4 present the tests results for the descriptions of the four attributes. In addition, we give some details on the nature of the tests, the criterion for selecting the descriptions based on the tests results, and the choice made regarding the management of multiple comparisons. The English translations of the French descriptions were formulated by the authors.

Statistical tests

In the tables, the columns Familiarity and Relevance group the results of two Chi-squared tests (χ^2) evaluating the familiarity (question 1) and the relevance (question 2A) of a description with the test value and p-value. The third column presents the test for tendencies on Likert scales (question 2B). It is a Wilcoxon signed-rank test (T, p), with an assessment of the sign of the tendency on Likert scales (sign) and an estimate of the effect size by calculating a pseudo-median (Mdn).

Selection of descriptions

The three tests were performed for each sentence in the corpus with R[1]. Following the order of the questions, if a description was not significantly familiar, relevant, or with a clear tendency, then it was discarded. For example, the description *fermé* (closed) for bright was not considered significantly familiar and was therefore rejected. On the other hand, the description *résonant* (resonant) was found to be significantly familiar but also not significantly relevant and was therefore discarded.

Italicized p-values for familiarity and relevance tests mean that the description was rated as significantly unfamiliar or irrelevant. For example, the description *court* (short) for bright was evaluated as significantly irrelevant ($p = .001$).

[1]https://www.r-project.org/

Note on multiple comparison

For each description we performed three tests for the three questions that allow to filter the descriptions sequentially with regards to: 1. Familiarity, 2. Relevance, 3. Tendency.

Therefore, a description with a p-value below an alpha level of .05 for all three tests would be equivalent to an overall p-value being below a alpha level resulting of the product of the three test alpha levels ($.05^3 = .0001$). The probability of type 1 error is therefore very low.

If we applied a Bonferroni correction to those alpha levels, the chances of type 1 errors would remain very low. For instance, the most numerous descriptions concerned roundness. Here is an example for the 68 descriptions of bright:

$$\begin{aligned} \alpha_{corrected} &= 68 * .05^3 \\ &= 0.009 (< .05) \end{aligned} \tag{A.1}$$

The p-values of the selected descriptions are below the criterion of .05. Therefore, no correction was applied for the selection of the descriptions.

TABLE 1. BRIGHT *BRILLANT*

	χ^2 (2, N=51) *Familiarity*		χ^2 (2, N=51) *Relevance*		Wilcoxon signed-rank test *Tendency*			
Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
closed *fermé*	1.58	.208						
resonant *résonant*	47.08	**<.001**	0.49	.110				
with a soft attack *avec une attaque douce*	51.0	**<.001**	0.96	.327				
with little precision/definition *avec peu de précision/définition*	43.31	**<.001**	18.84	**<.001**	23.0	**<.001**	-	1.5
soft *doux*	39.71	**<.001**	21.35	**<.001**	47.0	**<.001**	-	2.0
with a lot of high harmonics *avec beaucoup d'harmoniques aigus*	43.31	**<.001**	32.96	**<.001**	899.0	**<.001**	+	5.0
low *grave*	51.0	**<.001**	4.41	**.036**	29.50	**<.001**	-	1.5
matte *mat*	51.0	**<.001**	21.35	**<.001**	0	**<.001**	-	1.0
round *rond*	51.0	**<.001**	24.02	**<.001**	116.0	**<.001**	-	1.5
short *court*	39.71	**<.001**	10.37	***.001***				
medium *medium*	1.40	**<.001**	7.80	**.008**	63.0	**.001**	-	2.0
warm *chaud*	47.08	**<.001**	16.50	**<.001**	68.0	**<.001**	-	1.5
high-medium *médium-aigu*	43.31	**<.001**	16.49	**<.001**	447.0	**<.001**	+	4.0
with a lot of harmonics *avec beaucoup d'harmoniques*	51.0	**<.001**	29.82	**<.001**	818.0	**<.001**	+	4.5
with attack *avec de l'attaque*	39.70	**<.001**	0.96	.327				
dull *sourd*	43.31	**<.001**	29.82	**<.001**	0	**<.001**	-	1.0
with a lot of low harmonics *avec beaucoup d'harmoniques graves*	36.25	**<.001**	16.49	**<.001**	70.0	**<.001**	-	1.5
pure *pur*	39.70	**<.001**	0.49	.484				
with little attack *avec peu d'attaque*	39.70	**<.001**	2.37	.123				
metallic *métallique*	51.0	**<.001**	26.84	**<.001**	560.0	**<.001**	+	4.0
with little harmonics *avec peu d'harmoniques*	51.0	**<.001**	21.35	**<.001**	40.0	**<.001**	-	1.5
low-medium *bas-médium*	43.31	**<.001**	4.41	**.036**	9.0	**<.001**	-	1.5
rough *rugueux*	36.25	**<.001**	0.18	.674				

BRIGHT

Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
wooden *boisé*	32.96	**<.001**	3.31	.069				
with a posed attack *avec une attaque posée*	0.96	0.327						
muffled *étouffé*	51.0	**<.001**	39.70	**<.001**	0	**<.001**	-	1.0
with a clear attack *avec une attaque franche*	51.0	**<.001**	0.49	.484				
with clarity in the attack *avec de la clarté dans l'attaque*	43.31	**<.001**	3.31	.069				
dark *sombre*	24.01	**<.001**	21.35	**<.001**	0	**<.001**	-	1.0
high *aigu*	51.0	**<.001**	12.25	**<.001**	459.50	**<.001**	+	4.0
with little high harmonics *avec peu d'harmoniques aigus*	43.31	**<.001**	29.82	**<.001**	58.50	**<.001**	-	1.5
that stands out *qui ressort*	12.25	**<.001**	0.49	.484				
with a precise attack *avec une attaque précise*	51.0	**<.001**	0.02	.889				
loud *fort*	39.70	**<.001**	2.37	.123				
piano/pianissimo	51.0	**<.001**	0.18	.674				
with shine *avec de l'éclat*	21.35	**<.001**	21.35	**<.001**	861.0	**<.001**	+	4.5
with a dull attack *avec une attaque molle*	43.31	**<.001**	0.18	.674				
with projection *avec de la projection*	18.84	**<.001**	0.02	.889				
with impact *avec de l'impact*	12.25	**<.001**	2.37	.123				
with clarity of articulation *avec une clarté d'articulation*	12.25	**<.001**	2.37	.123				
inharmonic *inharmonique*	14.29	**<.001**	4.41	***.036***				
clear *clair*	51.0	**<.001**	32.96	**<.001**	739.50	**<.001**	+	4.2
dull *terne*	26.84	**<.001**	18.84	**<.001**	0	**<.001**	-	1.0
with clarity *avec de la clarté*	47.08	**<.001**	29.82	**<.001**	848.50	**<.001**	+	4.5
velvety *feutré*	39.70	**<.001**	26.84	**<.001**	0	**<.001**	-	1.0

TABLE 2. WARM *CHAUD*

	χ^2 (2, N=51) *Familiarity*		χ^2 (2, N=51) *Relevance*		Wilcoxon signed-rank test *Tendency*			
Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
comfortable *confortable*	10.37	**.001**	4.41	.036	0485	**<.001**	+	4.5
with little high harmonics *avec peu d'harmoniques aigus*	43.31	**<.001**	24.01	**<.001**	349.0	.172		
unpleasant *désagréables*	29.82	**<.001**	1.59	.207				
generous *généreux*	18.84	**<.001**	14.29	**<.001**	703.0	**<.001**	+	4.0
with a dense timbre *avec un timbre dense*	29.82	**<.001**	4.41	**.036**	332.0	**<.001**	+	4.0
loud *fort*	32.96	**<.001**	14.29	***<.001***				
with a loud attack *avec une attaque forte*	39.70	**<.001**	3.31	.069				
piano	43.31	**<.001**	10.37	***.001***				
low *grave*	51.0	**<.001**	5.67	**<.001**	304.0	**<.001**	+	4.0
sharp / net *net*	21.35	**<.001**	1.59	.208				
with a lot of low harmonics *avec beaucoup d'harmoniques graves*	39.71	**<.001**	21.35	**<.001**	570.0	**<.001**	+	4.0
wide *large*	36.25	**<.001**	4.41	**<.001**	331.0	**<.001**	+	4.0
with clarity *avec de la clarté*	32.96	**<.001**	7.08	**.008**	82.0	**.003**	-	2.0
with a hard attack *avec une attaque dure*	39.71	**<.001**	0.49	.484				
expressive *expressif*	8.65	**<.001**	0.18	.674				
with little harmonics *avec peu d'harmoniques*	51.0	**<.001**	18.84	**<.001**	52.5	**<.001**	-	2.0
mezzoforte-forte	51.0	**<.001**	12.25	***<.001***				
that emit breadth *qui dégage de l'ampleur*	18.84	**<.001**	2.37	.123				
high *aigu*	51.0	**<.001**	10.37	**<.001**	0	**<.001**	-	1.5
with low frequencies in the attack *avec du grave dans l'attaque*	26.84	**<.001**	0.49	.484				
breathing *avec du souffle*	43.31	**<.001**	0.49	.484				
with a lot of high harmonics *avec beaucoup d'harmoniques*	47.08	**<.001**	26.84	**<.001**	49.50	**<.001**	-	2.0
that vibrates *qui vibre*	21.35	**<.001**	0.96	.327				

WARM Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
hard *dur*	29.82	**<.001**	10.37	**.001**	0	**<.001**	-	1.0
narrow in frequency *resserré en fréquence*	26.84	**<.001**	3.31	.069				
rich *riche*	39.71	**<.001**	12.26	**<.001**	465.0	**<.001**	+	4.5
full *plein*	39.71	**<.001**	24.02	**<.001**	665.5	**<.001**	+	4.0
with a lot of medium harmonics *avec bcp d'harmoniques médium*	39.71	**<.001**	21.35	**<.001**	360.0	**<.001**	+	4.0
with little low-medium harmonics *avec peu d'harmoniques bas-méd.*	32.96	**<.001**	12.26	**<.001**	60	**<.001**	-	2.0
with a lot of low-medium harmonics *avec bcp. d'harmoniques bas-méd.*	39.71	**<.001**	16.49	**<.001**	691.0	**<.001**	+	5.0
not vibrated *non vibré*	16.49	**<.001**	2.37	.123				
without variation *sans variation*	24.02	**<.001**	5.67	***.017***				
pleasant *agréable*	32.96	**<.001**	18.24	**<.001**	728.5	**<.001**	+	4.5
with little low harmonics *avec peu d'harmoniques aigus*	39.70	**<.001**	14.29	**<.001**	50.5	**<.001**	-	1.5
metallic *métallique*	51.0	**<.001**	12.25	**<.001**	0	**<.001**	-	1.0
cold *froid*	18.84	**<.001**	5.67	**.017**	0	**<.001**	-	1.0
poor *pauvre*	36.25	**<.001**	4.41	**.036**	0	**<.001**	-	1.0
organic *organqiue*	8.65	**.003**	0.49	.484				
rough *rugueux*	39.71	**<.001**	.020	.889				
balanced *équilibré*	26.84	**<.001**	5.67	**.035**	153.5	.318		
brassy *cuivré*	47.08	**<.001**	16.49	**<.001**	237.0	.299		
pure *pur*	43.31	**<.001**	5.67	**.017**	35.0	**<.001**	-	1.5
smooth *lisse*	32.96	**<.001**	0.49	.888				
soft *doux*	36.25	**<.001**	24.02	**<.001**	585.5	**<.001**	+	4.0
with a presence *avec une présence*	5.67	**.017**	0.02	.889				
resonant *résonant*	47.08	**<.001**	4.41	**.036**	219.50	**.005**	+	4.0
with life *avec de la vie*	4.41	**.036**	0.96	0.33				
straight *droit*	0.49	.484						

WARM Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
aggressive *agressif*	43.31	**<.001**	16.49	**<.001**	6.50	**<.001**	**-**	1.5
a little rough *un peu rugueux*	39.71	**<.001**	0.49	.484				
with a soft attack *avec une attaque douce*	51.0	**<.001**	3.31	.069				
neutral *neutre*	0.96	.327						
sustained *entretenu*	32.96	**<.001**	2.37	.123				
bright *brillant*	51.0	**<.001**	29.82	**<.001**	106.0	**<.001**	**-**	2.0
medium-high *médium-aigu*	43.31	**<.001**	4.41	**.036**	34.0	**<.001**	**-**	2.0
stable *stable*	29.82	**<.001**	5.67	***.017***				
enveloping *enveloppant*	36.26	**<.001**	21.35	**<.001**	741.0	**<.001**	+	4.5
with little medium harmonic *avec peu d'hamroniques médium*	36.26	**<.001**	16.49	**<.001**	22.50	**<.001**	**-**	1.5
which gives a proximity sensation *qui donne une sensation de proximité*	24.02	**<.001**	0.18	.674				
distant *distant*	32.96	**<.001**	0.02	.889				
short *court*	32.96	**<.001**	8.65	***.003***				
defined *défini*	36.25	**<.001**	0.18	.674				
with a human touch *avec une dimension humaine*	1.59	.208						
low-medium *bas-médium*	36.25	**<.001**	12.26	**<.001**	493.50	**<.001**	+	4.0
medium *médium*	39.71	**<.001**	10.37	**.001**	267.0	**.006**	+	4.0
with a lot of harmonics *avec beaucoup d'harmoniques*	51.0	**<.001**	12.26	**<.001**	327.0	**.006**	+	3.5
with a wide spectrum *avec un spectre large*	39.70	**<.001**	8.65	**.003**	270.5	**.005**	+	4.0

TABLE 3. ROUND *ROND*

	χ^2 (2, N=51) *Familiarity*		χ^2 (2, N=51) *Relevance*		Wilcoxon signed-rank test *Tendency*			
Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
rich *riche*	43.31	**<.001**	8.65	**.003**	278.5	**.038**	+	3.5
little rich *peu riche*	32.96	**<.001**	8.65	**.003**	78.0	**<.001**	-	2.0
comforting *réconfortant*	10.37	**.003**	1.59	.208				
with a dry attack *avec une attaque sèche*	26.84	**<.001**	0.96	.327				
with a bouncy attack *avec une attaque rebondie*	0.02	.889						
with little attack *avec peu d'attaque*	43.31	**<.001**	4.41	**.04**	346.0	**.002**	+	4.0
hard *dur*	29.82	**<.001**	21.35	**<.001**	0	**<.001**	**-**	1.5
pleasant *agréable*	29.82	**<.001**	10.37	**.001**	561.0	**<.001**	+	4.5
high *aigu*	51.0	**<.001**	51.0	**<.001**	8	**<.001**	**-**	1.5
sour *aigre*	4.41	**.036**	0.17	.674				
uncomfortable *inconfortable*	10.37	**.001**	0.96	.327				
with emergin harmonics *avec des harmoniques qui ressortent*	36.25	**<.001**	5.67	**.017**	68.0	**.013**	**-**	2.0
with a resonance close to the f0 *avec une resonance proche de la f0*	18.84	**<.001**	2.37	.123				
homogeneous *homogène*	39.71	**<.001**	7.08	**.008**	433.5	**<.001**	+	4.0
noisy *bruité*	29.82	**<.001**	1.59	.208				
with little high harmonics *avec peu d'harmoniques aigus*	39.71	**<.001**	16.49	**<.001**	242.0	.332		
dry *sec*	32.96	**<.001**	8.65	**.003**	0	**<.001**	**-**	1.50
with a clear attack *avec une attaque franche*	47.08	**<.001**	0.96	.327				
with a lot of attack *avec beaucoup d'attaque*	47.08	**<.001**	7.08	**.008**	44.0	**<.001**	**-**	1.50
with a slow decrease *avec une décroissance lente*	29.82	**<.001**	10.37	***.001***				
piano	39.71	**<.001**	10.37	***.001***				
sharp *pointu*	0.96	.327						
brassy *cuivré*	47.08	**<.001**	2.37	.123				

ROUND Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
laid *posé*	1.59	.208						
with a clean attack *avec une attaque nette*	51.0	**<.001**	0.02	.889				
short *court*	43.31	**<.001**	0.490	.484				
with a lot of low harmonics *avec bcp. d'harmoniques graves*	36.25	**<.001**	24.02	**<.001**	509.0	**<.001**	+	4.0
with a soft decrease *avec une décroissance douce*	26.84	**<.001**	3.31	.069				
precise *précis*	14.29	**<.001**	0.18	.674				
with a lot of harmonics *avec bcp. d'harmoniques*	51.0	**<.001**	12.25	**<.001**	196.0	.438		
forte	51.0	**<.001**	8.65	***.003***				
unstable *instable*	21.35	**<.001**	0.02	.889				
with a dull attack *avec une attaque molle*	39.71	**<.001**	0.96	.327				
aggressive *agressif*	39.71	**<.001**	26.84	**<.001**	0	**<.001**	-	1.0
sustained *entretenu*	32.96	**<.001**	0.17	.674				
breathing *avec du souffle*	47.08	**<.001**	1.59	.208				
vibrating *qui vibre*	26.84	**<.001**	0.02	.886				
with a short envelope *avec une enveloppe courte*	7.08	**.02**	3.31	.069				
bright *brillant*	51.0	**<.001**	16.49	**<.001**	25.5	**<.001**	-	1.5
soft *doux*	39.71	**<.001**	24.02	**<.001**	725.0	**<.001**	+	4.0
resonant *résonant*	47.08	**<.001**	8.64	**.003**	329.0	**.001**	+	4.0
expressive *expressif*	10.37	**.003**	3.31	.069				
with a soft attack *avec une attaque douce*	51.0	**<.001**	8.64	**.003**	370.0	**.001**	+	4.0
pure *pur*	39.71	**<.001**	0.96	.327				
with low-medium spectral balance *avec un équilibre spectral bas-méd.*	21.35	**<.001**	7.08	**.008**	457.5	**<.001**	+	4.0
stable *stable*	32.96	**<.001**	0.49	.484				
with little low harmonics *avec peu d'harmoniques graves*	36.25	**<.001**	18.84	**<.001**	97.0	**<.001**	-	2.0
with a lot of medium-high harmonics *avec bcp. d'harmoniques méd.-aigu*	39.71	**<.001**	16.49	**<.001**	184.0	.054		

ROUND Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
low *grave*	47.08	**<.001**	7.08	**.008**	316.0	**<.001**	+	4.0
with a varying dynamic *avec une dynamique qui varie*	39.71	**<.001**	4.41	**.036**	16.5	**.002**	-	1.5
with a lot of low-medium harmonics *avec bcp. d'harmoniques bas-méd.*	39.71	**<.001**	21.35	**<.001**	674.0	**<.001**	+	4.0
smooth *lisse*	29.82	**<.001**	7.08	**.008**	223.0	**.012**	+	4.0
with a lot of high harmonics *avec beaucoup d'harmoniques aigus*	39.71	**<.001**	18.84	**<.001**	41.0	**<.001**	-	1.5
soft *doux*	4.41	.035	0.490	0.484				
with a precise attack *avec une attaque précise*	47.08	**<.001**	0.02	.887				
without asperities *sans aspérités*	18.84	**<.001**	10.37	**.001**	457.5	**<.001**	+	4.0
with a hard attack *avec une attaque dure*	43.31	**<.001**	4.41	**.046**	6.5	**<.001**	-	1.5
rough *rugueux*	36.25	**<.001**	8.65	**.003**	4.5	**<.001**	-	1.0
without noise *sans bruit*	5.67	**.051**	4.41	***.035***				
with a long attack *avec une attaque longue*	39.71	**<.001**	0.02	.887				
timbred *timbré*	29.82	**<.001**	8.64	**.003**	287.0	**.023**	+	3.5
screaming *criard*	36.25	**<.001**	24.02	**<.001**	0	**<.001**	-	1
metallic *metallique*	51.0	**<.001**	12.25	**<.001**	0	**<.001**	-	1
full *plein*	36.25	**<.001**	21.35	**<.001**	498.0	**<.001**	+	4.5
with little harmonics *avec peu d'harmoniques*	51.0	**<.001**	14.29	**<.001**	111.0	**.003**	-	2.0
low-medium *bas-médium*	32.96	**<.001**	14.29	**<.001**	502.0	**<.001**	+	4.0
wide *large*	29.82	**<.001**	5.67	**.017**	245.0	**.002**	+	4.0

TABLE 4. ROUGH *RUGUEUX*

	χ^2 (2, N=51) *Familiarity*		χ^2 (2, N=51) *Relevance*		Wilcoxon signed-rank test *Tendency*			
Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
that rasps *qui râpe*	16.49	**<.001**	16.49	**<.001**	771.0	**<.001**	+	4.5
friction *de frottement*	39.71	**<.001**	18.84	**<.001**	615.0	**<.001**	+	4.5
bright *brillant*	51.0	**<.001**	1.59	.208				
not spectrally pure *pas pur spectralement*	12.25	**.001**	0.96	.327				
organic *organique*	10.37	**.004**	5.67	***.017***				
with little attack *avec peu d'attaque*	47.08	**<.001**	5.67	***.017***				
with a lot of attack *avec beaucoup d'attaque*	47.08	**<.001**	1.59	.208				
with little harmonic *avec peu d'harmoniques*	47.08	**<.001**	0.02	.887				
resonant *résonant*	43.31	**<.001**	0.96	.327				
raw *brut*	8.64	**.007**	0.02	.887				
with parasites *avec des parasites*	43.31	**<.001**	14.29	**<.001**	564.5	**<.001**	+	4.0
rugged *accidenté*	2.37	.123						
round *rond*	47.08	**<.001**	18.84	**<.001**	7.0	**<.001**	-	1.5
sustained *entretenu*	32.96	**<.001**	2.37	.123				
harsh / bitter *âpre*	10.37	**.001**	8.65	**.003**	386.0	**<.001**	+	4.0
with frequency variations *avec des variations fréquentielles*	21.35	**<.001**	0.49	.483				
noisy *bruité*	32.96	**<.001**	5.67	**.017**	320.0	**<.001**	+	4.0
soft *doux*	39.71	**<.001**	10.37	**.001**	5.0	**<.001**	-	1.0
with little vibrations *avec peu de variations*	18.84	**<.001**	0.02	.887				
full *plein*	39.71	**<.001**	0.02	.887				
unstable *instable*	21.35	**<.001**	0.96	.327				
with pulses *avec des battements*	36.25	**<.001**	5.67	**.017**	177.5	.484		
stable *stable*	32.96	**<.001**	0.02	.887				
itchy *qui gratte*	5.67	**.035**	5.67	**.017**	457.0	**<.001**	+	4.5

ROUGH Phrase	χ^2	p	χ^2	p	T	p	sign	Mdn
with fast variations *avec des variations rapides*	24.02	**<.001**	0.18	.674				
smooth *lisse*	39.71	**<.001**	32.96	**<.001**	0	**<.001**	-	1.0
with grain *avec du grain*	43.31	**<.001**	36.25	**<.001**	848.0	**<.001**	+	4.5
with asperities *avec des aspérités*	24.02	**<.001**	21.35	**<.001**	848.5	**<.001**	+	4.5
that rubs *qui frotte*	29.82	**<.001**	21.35	**<.001**	693.5	**<.001**	+	4.5
aggressive *agressif*	43.31	**<.001**	2.37	.123				
pure *pur*	39.71	**<.001**	8.65	**.003**	42.0	**<.001**	-	1.5
with a harsh attack *avec une attaque âpre*	2.37	.123						
grainy *granuleux*	39.71	**<.001**	29.82	**<.001**	847.5	**<.001**	+	4.5
which gives an impression of non-homogeneous material *qui donne une impression de matière non-homogène*	7.08	**.008**	2.37	.123				

bcp. : beaucoup
méd. : médium

A.2.4 Translations for the results of study 2

Four figures presenting the relevant results of the online survey (Fig. 2.2), with the associated French translation, as they were presented to the participants during the survey.

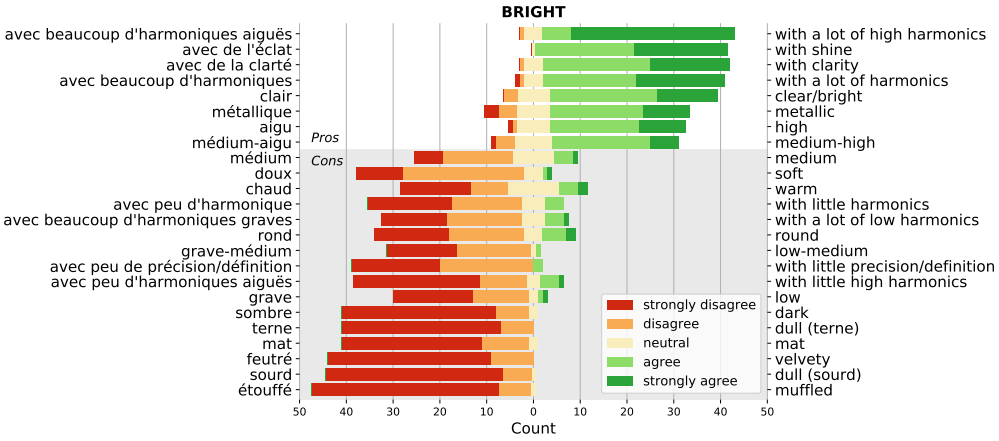


Figure A.5: Relevant descriptions and distribution of answers on the Likert scales obtained through the online survey for Bright with translations. The grey area gathers the descriptions in mismatch with the attribute.

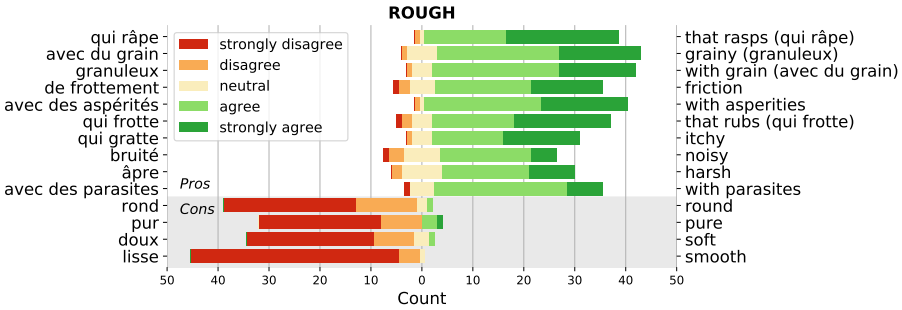


Figure A.6: Relevant descriptions and distribution of answers on the Likert scales obtained through the online survey for Rough with translations. The grey area gathers the descriptions in mismatch with the attribute.

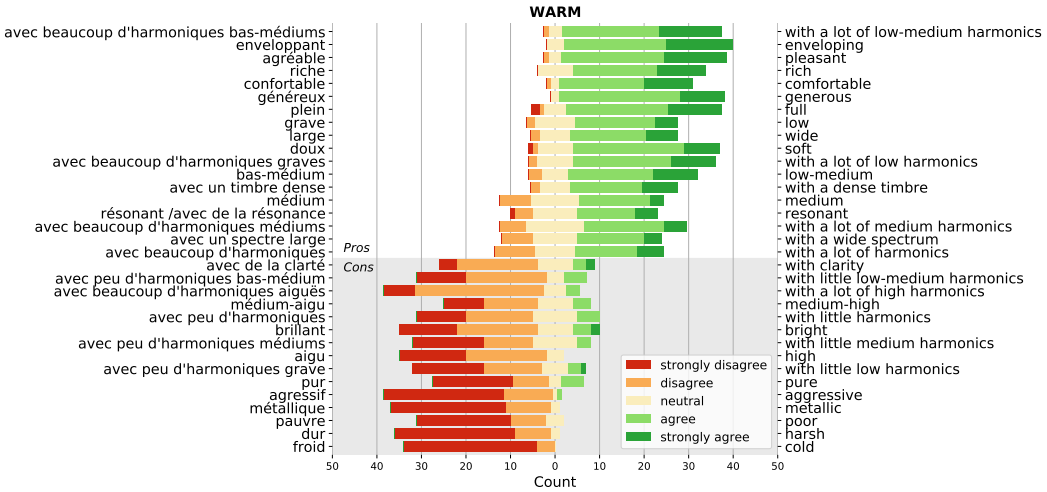


Figure A.7: Relevant descriptions and distribution of answers on the Likert scales obtained through the online survey for Warm with translations. The grey area gathers the descriptions in mismatch with the attribute.

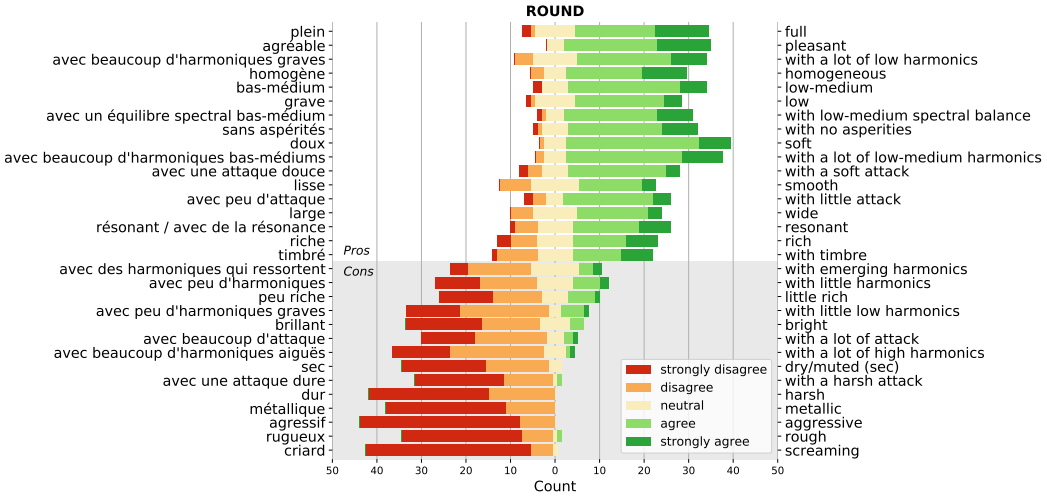


Figure A.8: Relevant descriptions and distribution of answers on the Likert scales obtained through the online survey for Round with translations. The grey area gathers the descriptions in mismatch with the attribute.

B. Supplementary materials of chapter 4

B.1 Acoustic and meta features

B.1.1 Acoustic features

Here are the acoustic features computed and used for the analysis of the experiment of chapter 4. We declined several features into median (med) values and interquartile ratio (iqr) values.

- Spectral centroid (med, iqr) - *Librosa*
- Spectral bandwidth (med, iqr) - *Librosa*
- Spectral contrast (med, iqr) - *Librosa*
- Spectral crest (med, iqr) - *Librosa*
- Spectral flatness (med, iqr) - *Librosa*
- Spectral rolloff (med, iqr) - *Librosa*
- Spectral flux (med, iqr) - *Librosa*
- Zero-crossing rate (med, iqr) - *Librosa*
- Log-attack time - *Timbre toolbox*
- Attack slope - *Timbre toolbox*
- Decrease slope - *Timbre toolbox*
- Release - *Timbre toolbox*
- Harmonic-noise ratio (med, iqr) - *Praat-Parselmouth*
- F0 - *in dataset*

The libraries used to extract the static features were *Librosa* (v0.8.1) (McFee et al., 2015), *Parselmouth* (v0.4.1) (Jadoul et al., 2018), and a python version of the *Timbre toolbox*[1] (Peeters et al., 2011).

We pruned down the features and selected only 16 features for the regression task by checking the collinearity between all features. This involved testing the performance of the model with different feature configurations for which we interchanged the most redundant. We thus chose the configurations allowing the best performances of the model for the prediction of the scores. Figures B.1 shows the selected features after the collinearity check. The "MPS roughness" metric is presented below in section B.1.3.

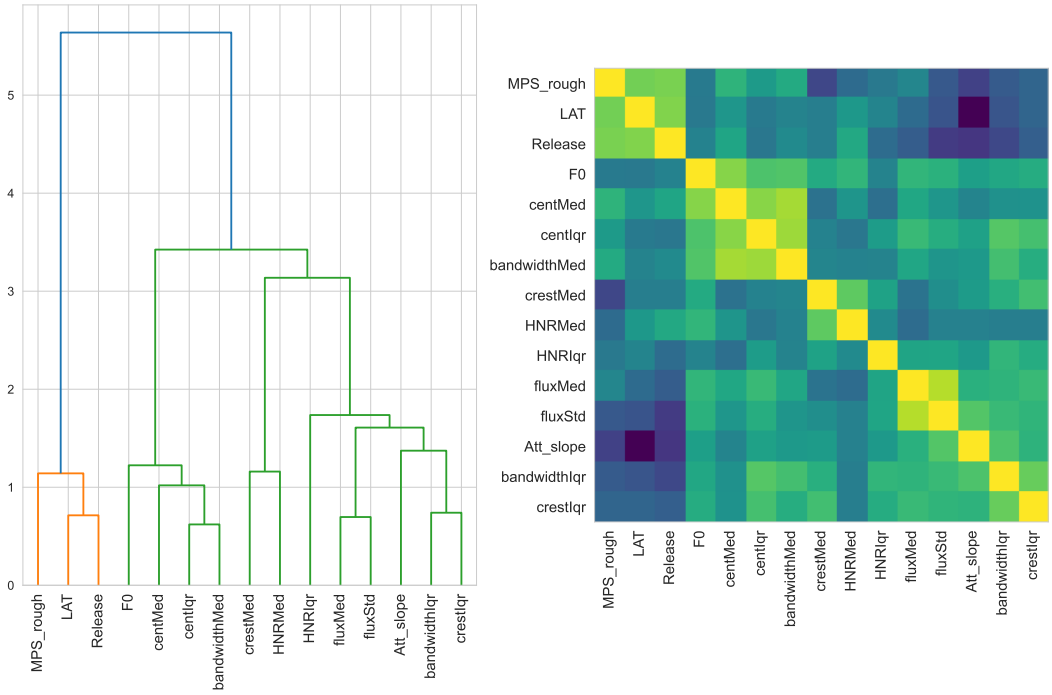


Figure B.1: Dendogram and collinearity matrix of the acoustic features.

[1] https://github.com/geoffroypeeters/imdABCDJhardfeatures

B.1.2 Meta features

Here we present the features corresponding to the playing techniques (Table B.1), and instruments (Table B.2) we gathered from the IRCAM's studio-online library (Ballet et al., 1999), and the Vienna Symphonic Library[2].

Instrument family	**Playing techniques**
All	Ordinario[3]
Strings	Non vibrato Pizzicato Pizzicato Bartók Sul ponticello Artifical harmonics
Winds	Flatterzunge - *woodwinds/brass* Staccato - *woodwinds/brass* Multiphonics - *woodwinds* Semi-aeolian - *flute/clarinet* Brassy - *brass* Straight/cup/harmon mute - *trumpet/trombone* Play & Sing - *tuba* Pedal tone - *trombone*
Mallets	Soft sticks Hard sticks arco - *vibraphone*
Plucked strings	harmonics

Table B.1: List of playing techniques in the dataset that were implemented as meta features in the experiment of chapter 4. Instruments with specific playing techniques are present in italic.

[2]http://www.vsl.co.at

Instrument family	**Instruments**
Strings	Violin
	Alto
	Cello
	Doublebass
Woodwinds	Flute
	Alto flute
	Piccolo flute
	Oboe
	English horn
	Clarinet
	Bass clarinet
	Bassoon
	Contrabassoon
	Alto saxophone
Brass	Trumpet
	Trombone
	French horn
	Tuba
Mallets	Glockenspiel
	Xylophone
	Vibraphone
	Marimba
Others	Harpa
	Guitar
	Accordion

Table B.2: List of instruments in the dataset that were implemented as meta features in the experiment of chapter 4

B.1.3 Modulation Power Spectrum Roughness

The **Modulation Power Spectrum** (MPS) is the 2D-Fourier transform of the spectrogram of a sound. It is often used to characterize and represent the spectro-temporal specificities of sounds (Arnal et al., 2015; Thoret et al., 2021). It depicts the power of modulation on both temporal and spectral dimensions.

On the MPS, the region between 30 and 150 Hz on the temporal modulation axis has been proven to be a good predictor of psychoacoustical roughness (Arnal et al., 2015).

Figure B.2 reports on two MPSs of saxophone sounds. The one on the left is an *ordinario* sound played pianissimo ($BWS_{rough.}rank : 13^th$), the one on the right is a multiphonic ($BWS_{rough.}rank : 496^th$). The rectangles indicate the region of interest regarding the rough content of sounds. On the multiphonic MPS, we clearly see modulation components (i.e., vertical lines) that contributes to the roughness of the sound. The MPS roughness metric introduced in this paper is the mean energy within the frequency range 30-150Hz. The MPS roughness values in the dataset range from 31.0 to 70.9 (arb. unit). Here, it equals 49.6 for the *ordinario* sound, and 60.9 for the multiphonic sound.

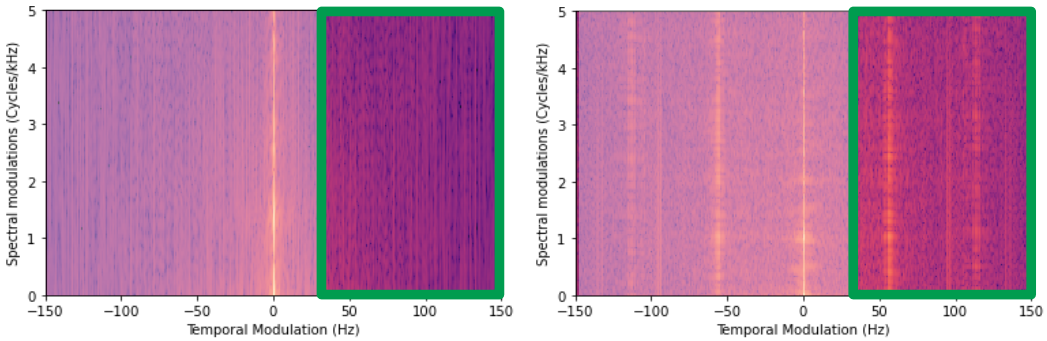


Figure B.2: Modulation power spectrums (MPS) of a *ordinario* saxophone sound (left) and a saxophone's multiphonic (right). The green frame indicates the region capturing the roughness of a sound.

Figure B.3 displays the correlation between BWS scores obtained for the expert population (i.e., sound engineers and conductors) for roughness and brightness. We see that while it highly correlates with roughness scores, it is not correlated at all with brightness scores.

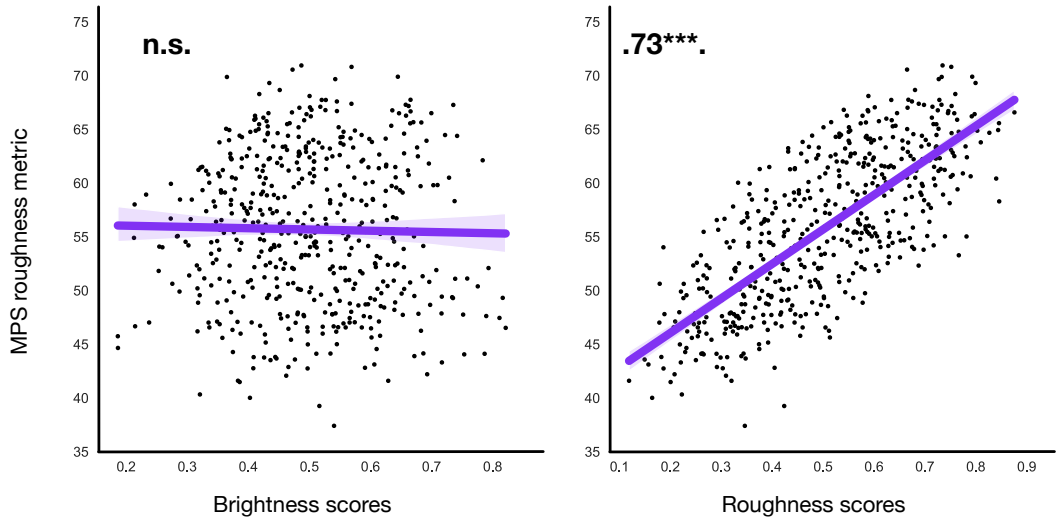


Figure B.3: Correlations of MPS with the brightness and roughness scores provided bu the expert population (sound engineers and conductors).

B.2 Correlations between compliance and the model's accuracy

When predicting BWS scores obtained in the experiment of chapter 4, we found that mean compliance scores (see Figure 4.2-A) were highly correlated with the R^2 values for the prediction of each concept ($r = 0.89, p < .001$). Figure B.4 report on this correlation.

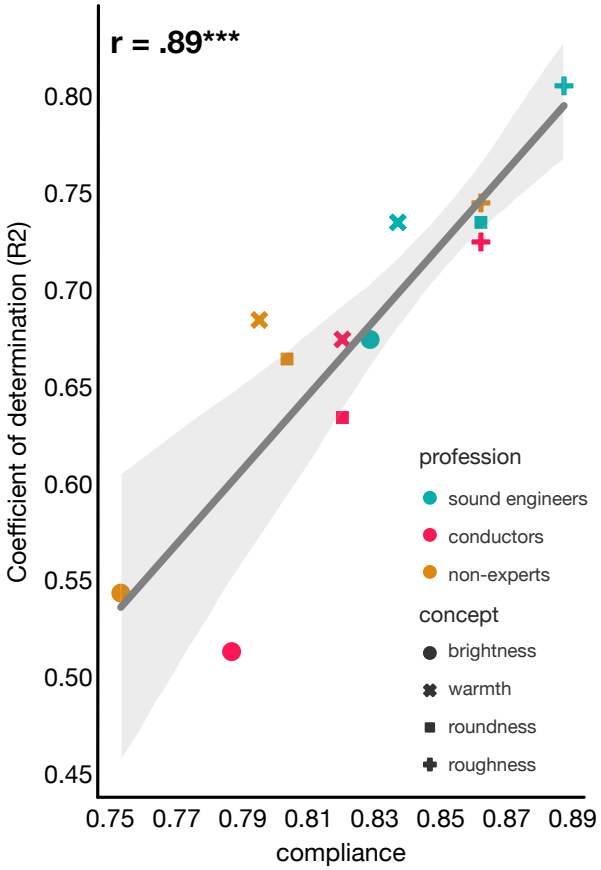


Figure B.4: Correlations between the mean compliance scores for each population (i.e., sound engineers, conductors, non-experts) and each concept, and the coefficients of determination (R^2) of the model predicting BWS scores.

C. Supplementary materials of chapter 5

C.1 Nature of the contributions of the main acoustic features

Here we present the contributions of the most important features of each attribute's acoustic portraits. Figures C.1 and C.2 depict the SHAP values (Lundberg and Lee, 2017) of these features that were used to predict the four attributes (see section 4.1.2 for the presentation of the SHAP values) based on judgments of the expert population, i.e., sound engineers and conductors. The shape of the contribution of each of these features makes it easier to visualize and understand the oppositions and similarities between the four attributes as they are discussed in the chapters 5 and 6. We therefore focused here on the contribution of four acoustic features: the spectral centroid, the F0 (Figure C.1), the HNR[1], and the spectral cres (Peeters et al., 2011) (Figure C.2).

[1]https://parselmouth.readthedocs.io/en/stable/

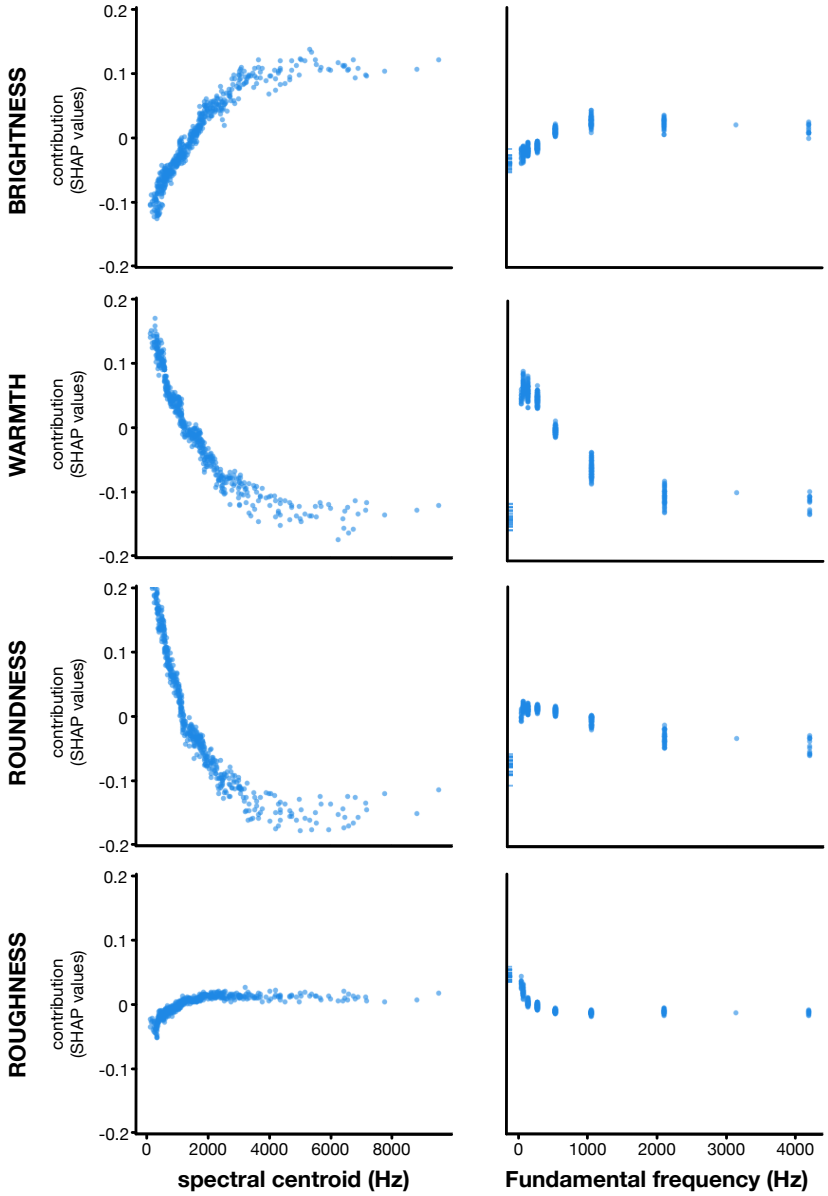


Figure C.1: Contribution (SHAP values) of spectral centroid and F0 for the prediction of each attribute and all sounds, based on expert participants judgments (i.e., sound engineers and conductors. Spectral centroid (left), Fundamental frequency (right)

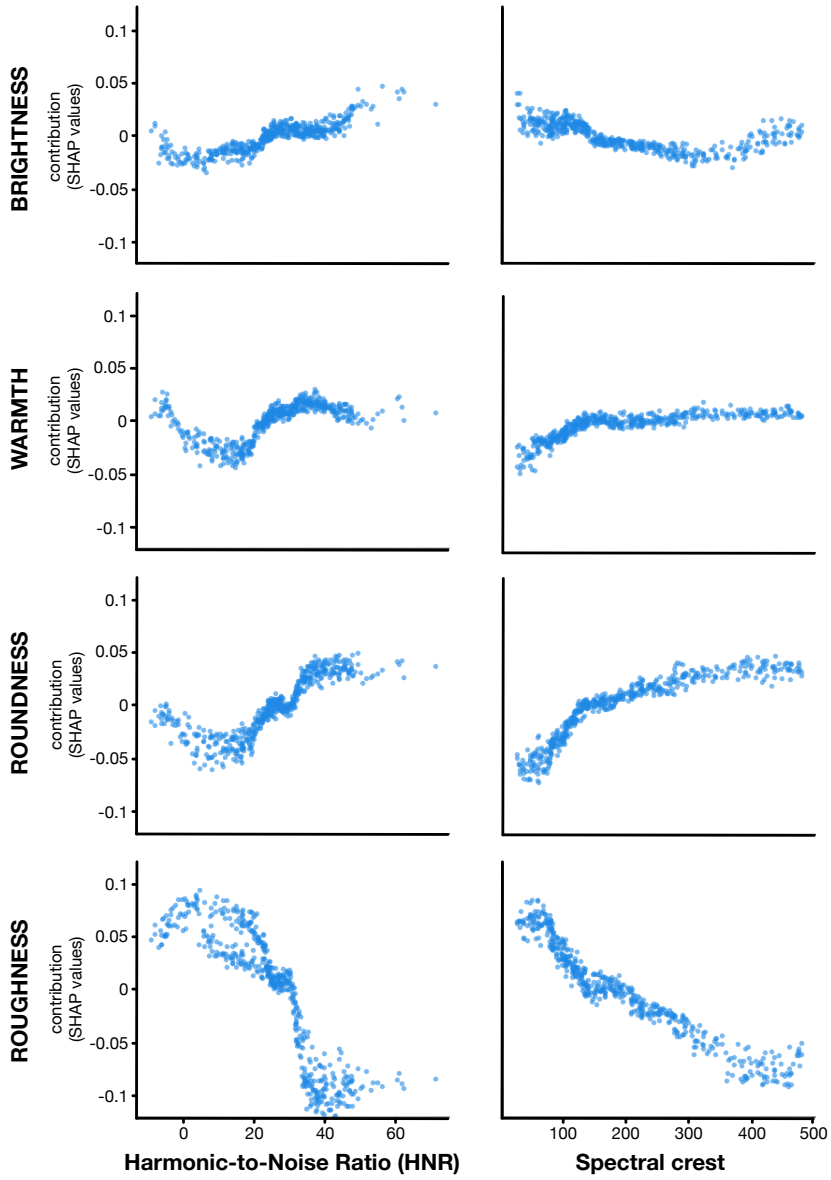


Figure C.2: Contribution (SHAP values) of Harmonic-to-Noise ratio (HNR) and spectral crest for the prediction of each attribute and all sounds, based on expert participants judgments (i.e., sound engineers and conductors. HNR (left), spectral crest (right)

C.2 Scores of meta features

Figure C.3 presents the BWS scores from the expert population (conductors and sound engineers) per instruments for each concept. Figure C.4 presents the BWS scores from the expert population (conductors and sound engineers), per technique, for each concept. See section B.1.2 for the collection of instruments and playing techniques.

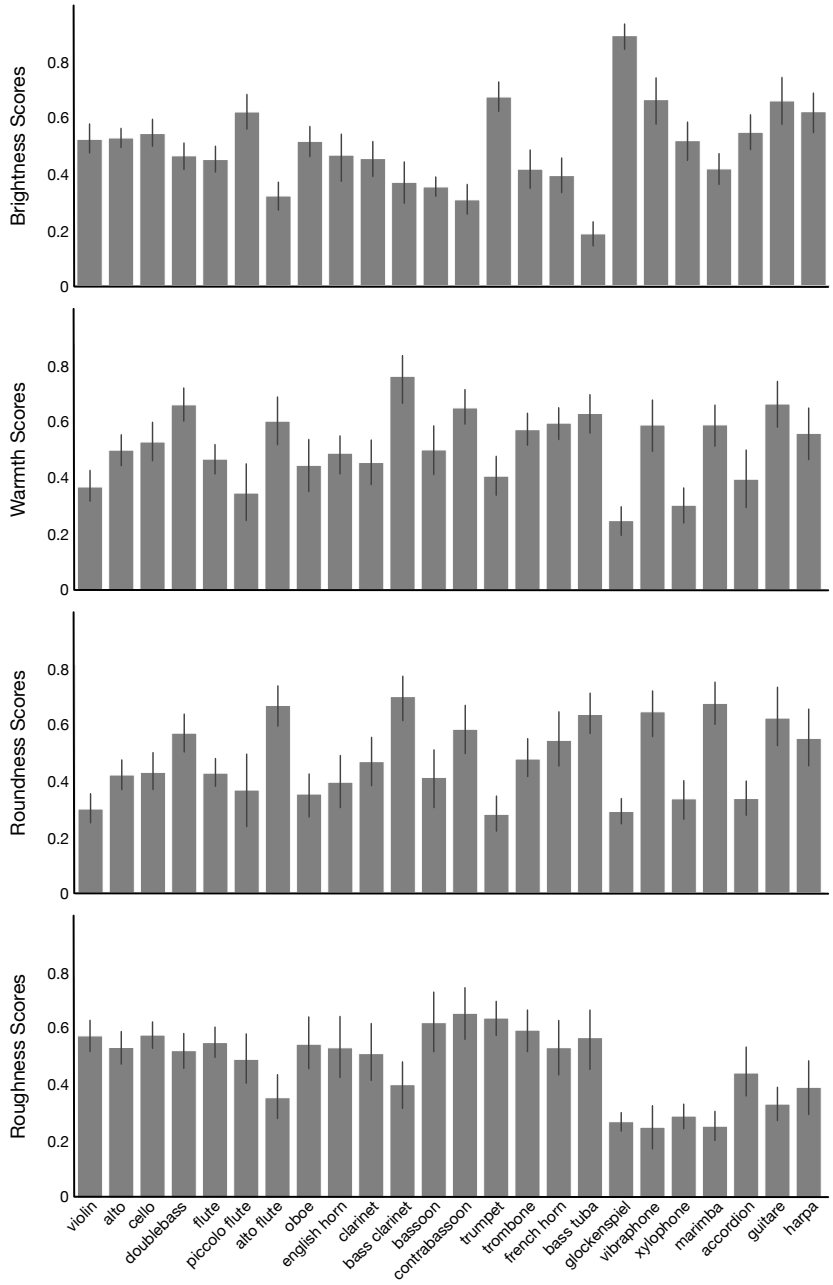


Figure C.3: Presentation of mean BWS scores from the expert population according to each instruments, for each concept: brightness (to), warmth (mid-to), roundness (mid-bottom), roughness (bottom)

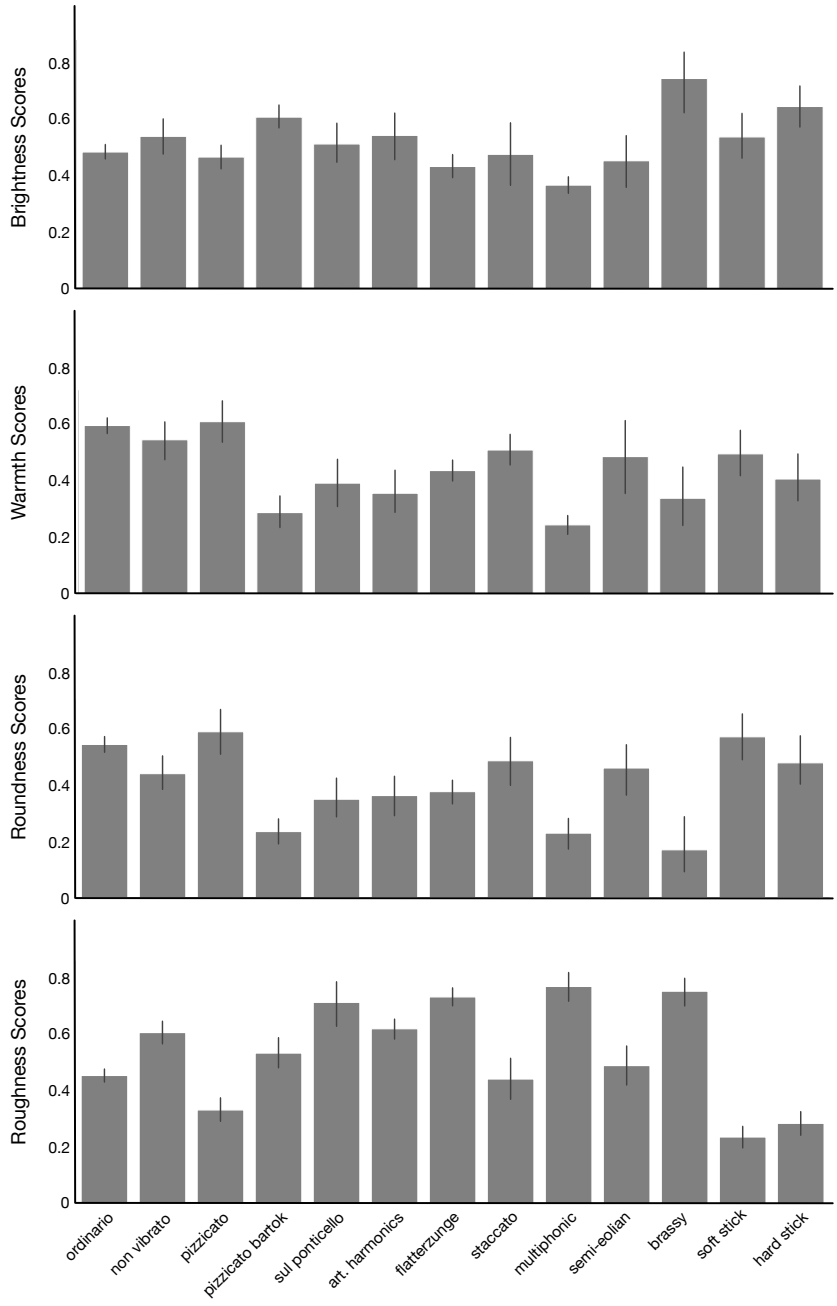
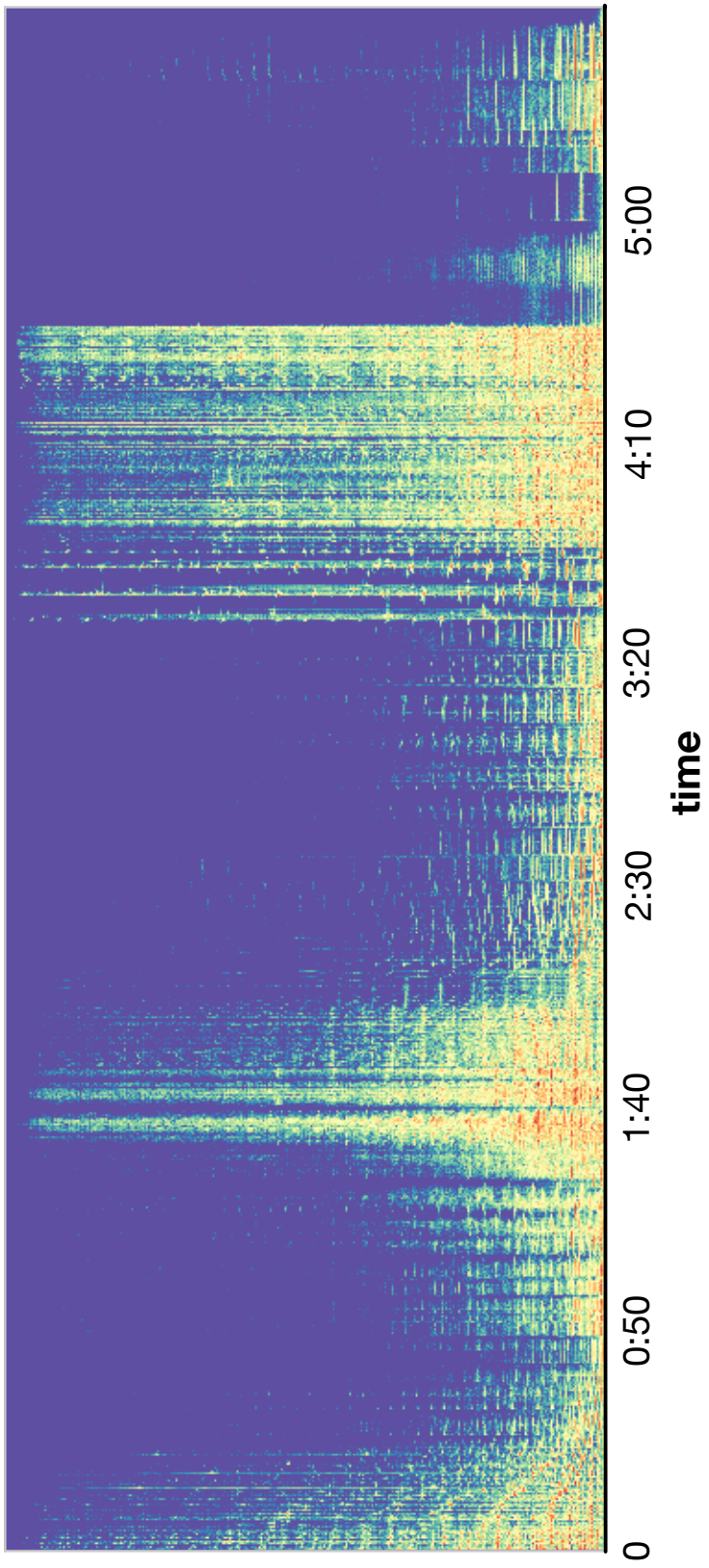
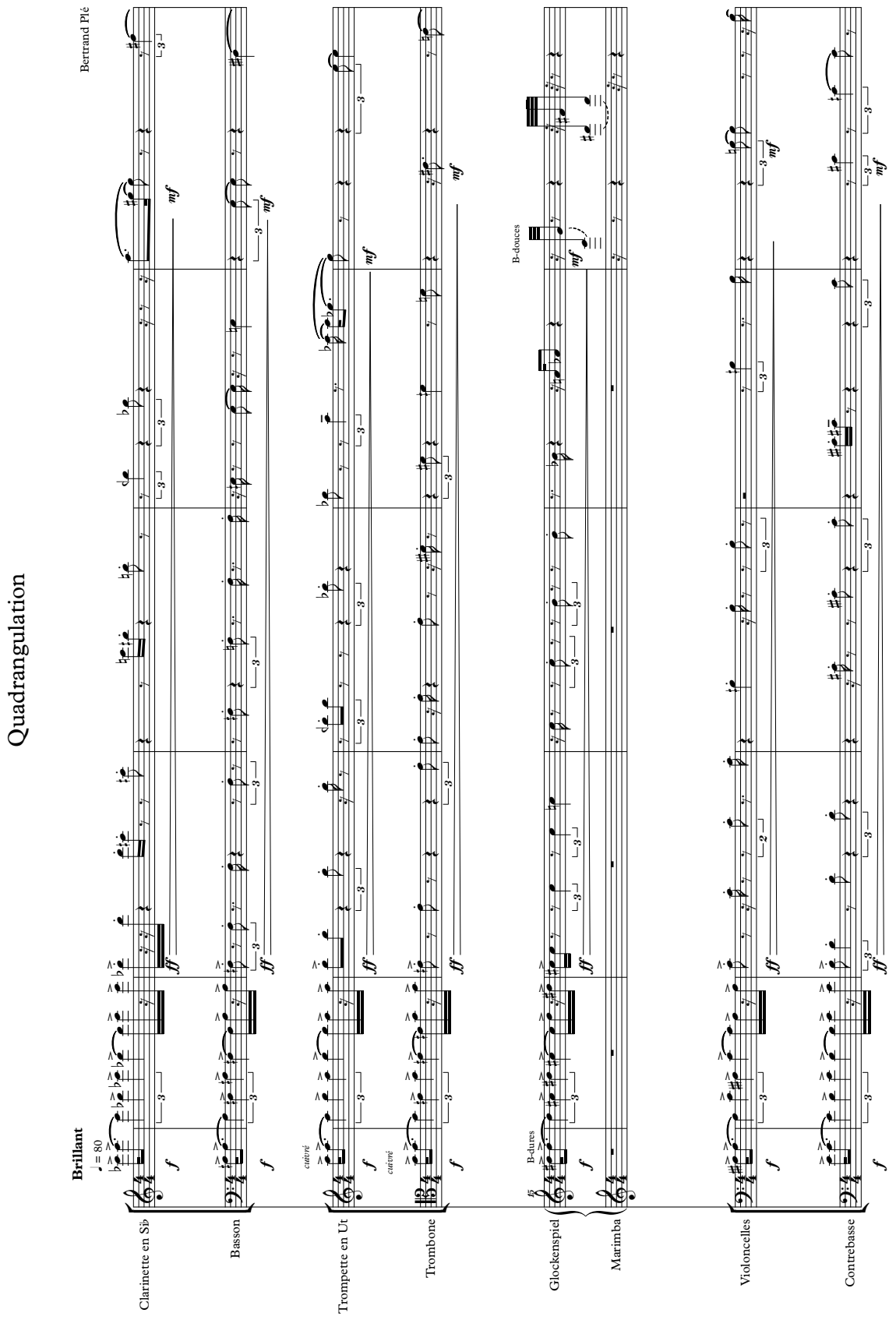
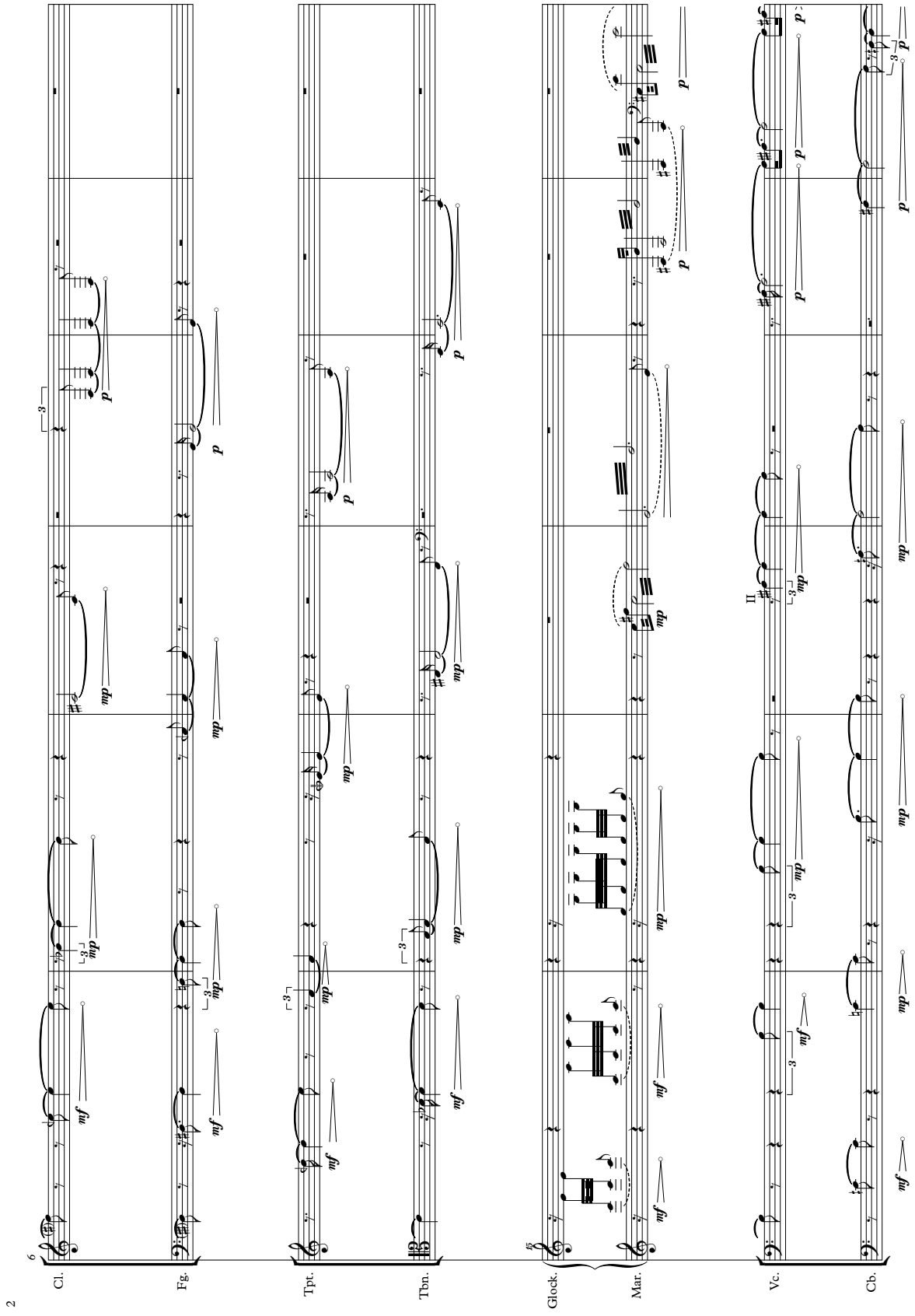


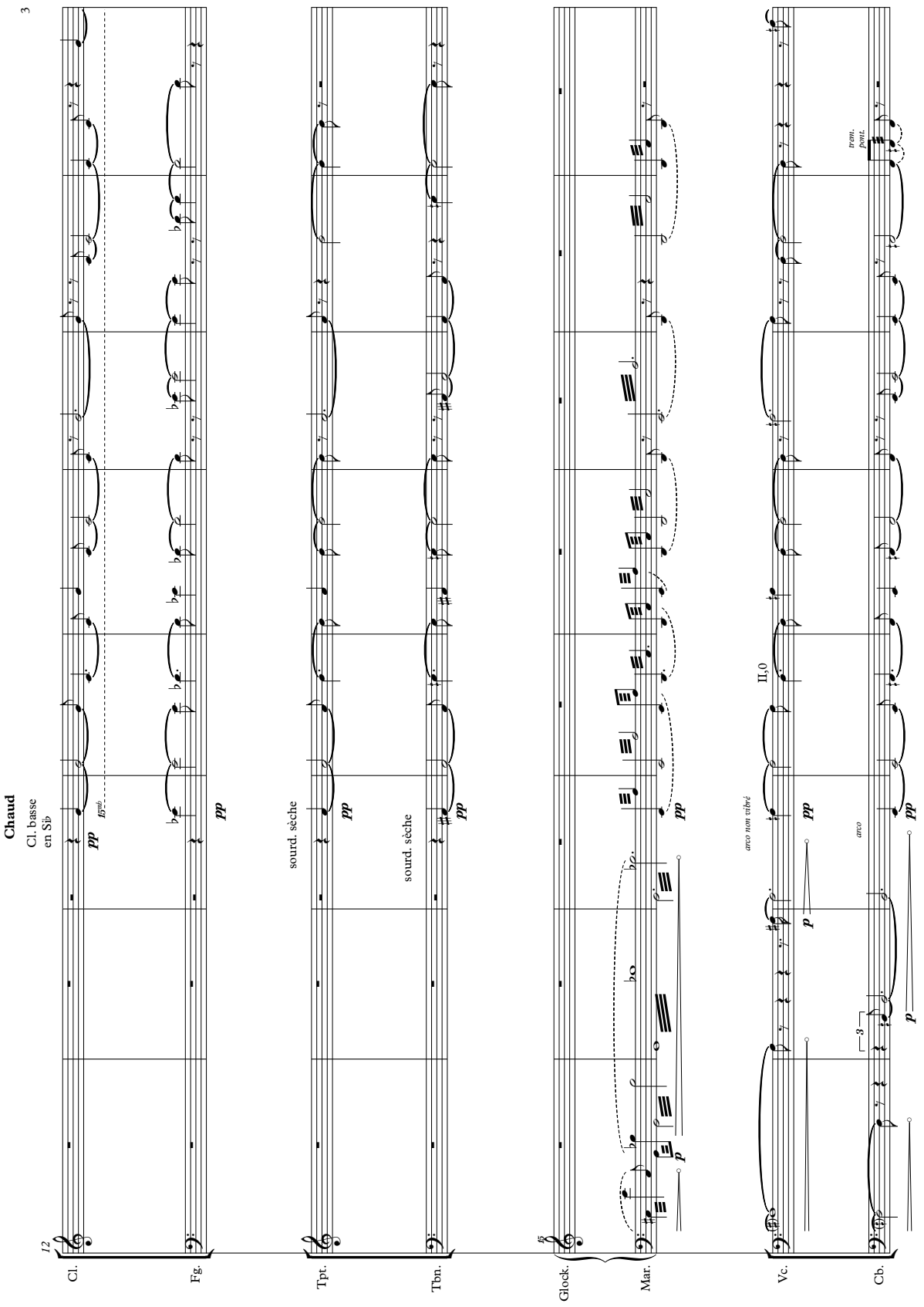
Figure C.4: Presentation of mean BWS scores from the expert population according to each playing technique, for each concept: brightness (to), warmth (mid-to), roundness (mid-bottom), roughness (bottom)

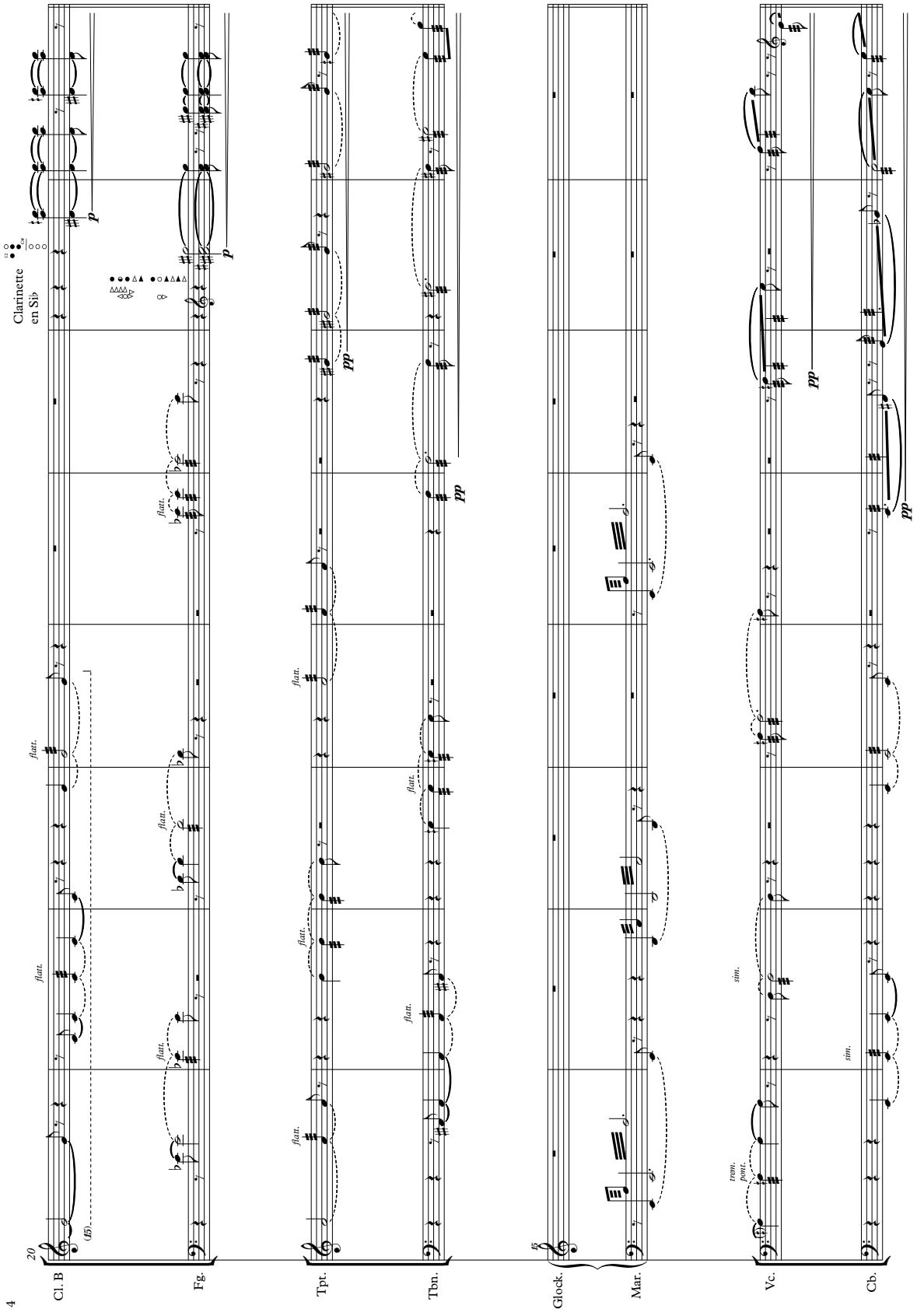
C.3 Spectrogram of *Quadrangulation*

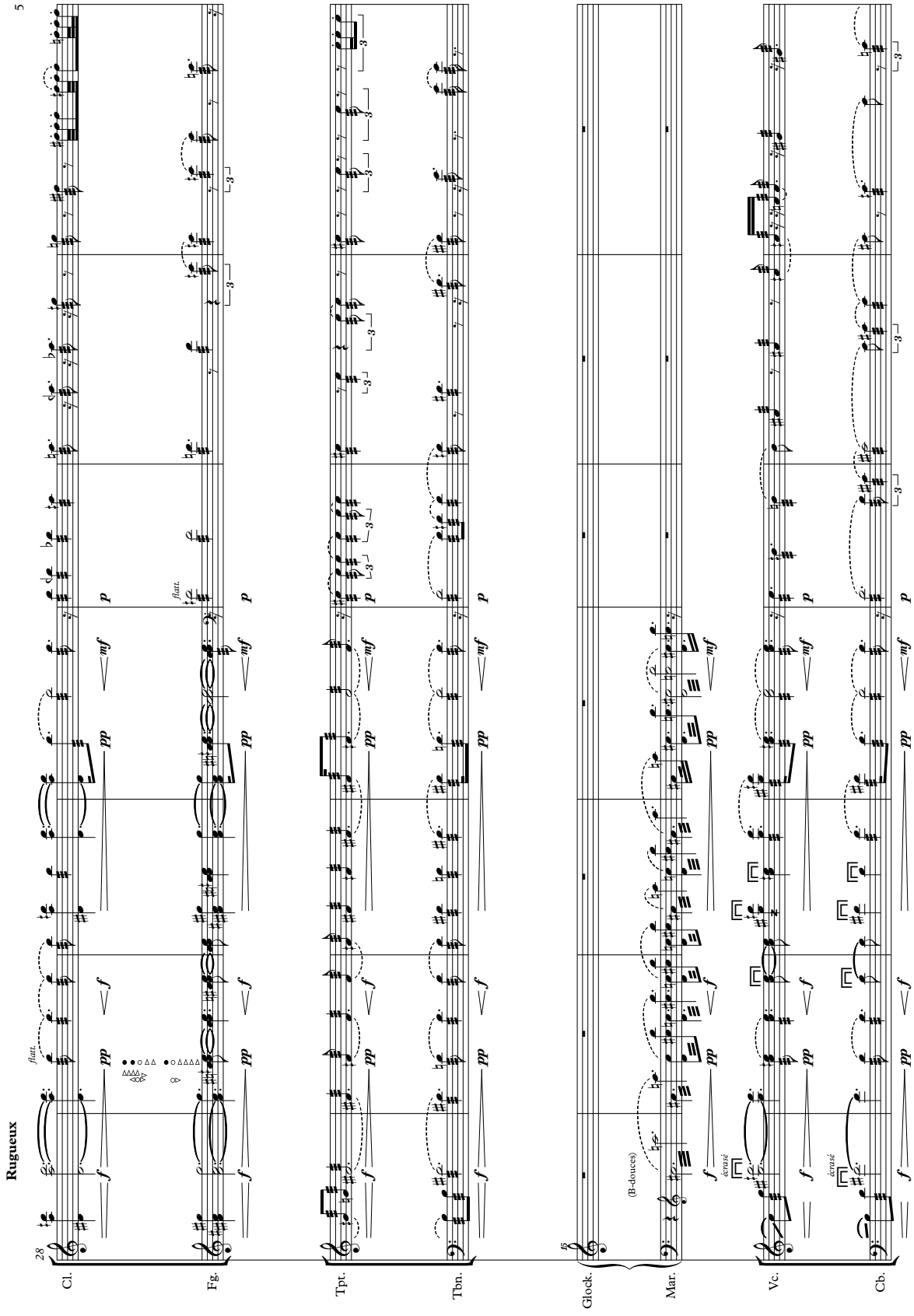


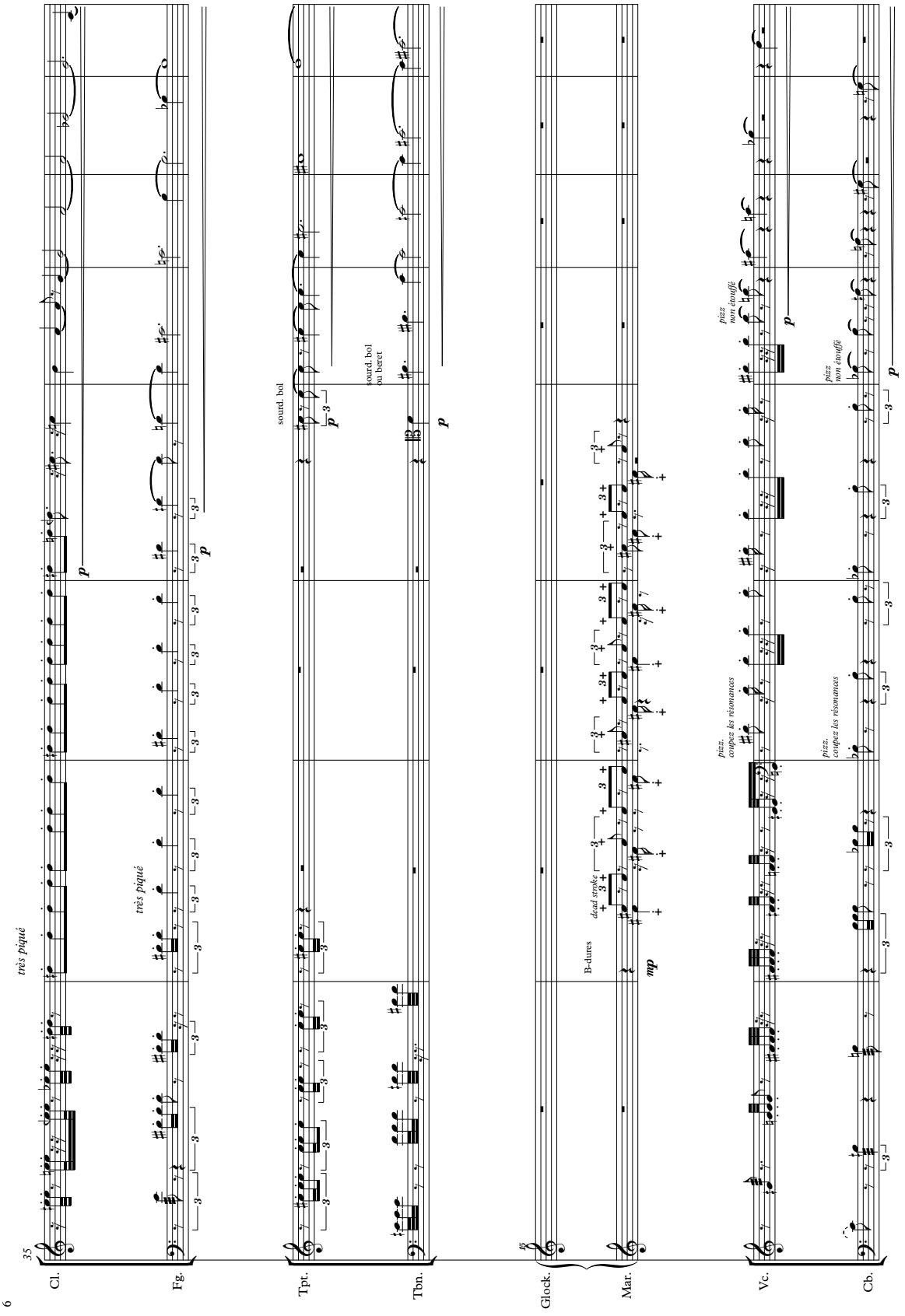

Quadrangulation
Bertrand Plé
Brillant
♩ = 80
Clarinette en Sb
Basson
Trompette en Ut
Trombone
Glockenspiel
Marimba
Violoncelles
Contrebasse
B-dures
B-douces
cuivré

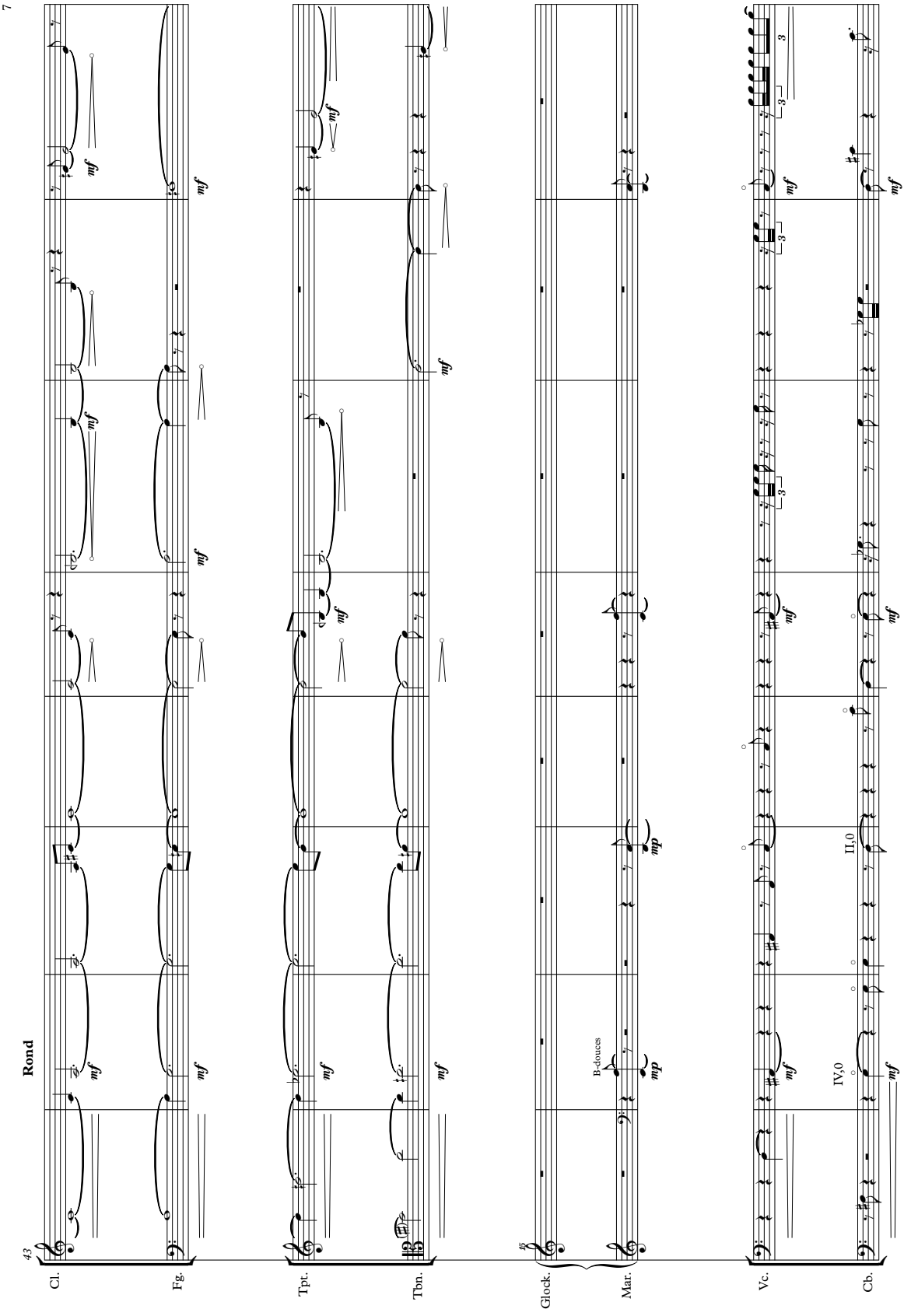

Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.

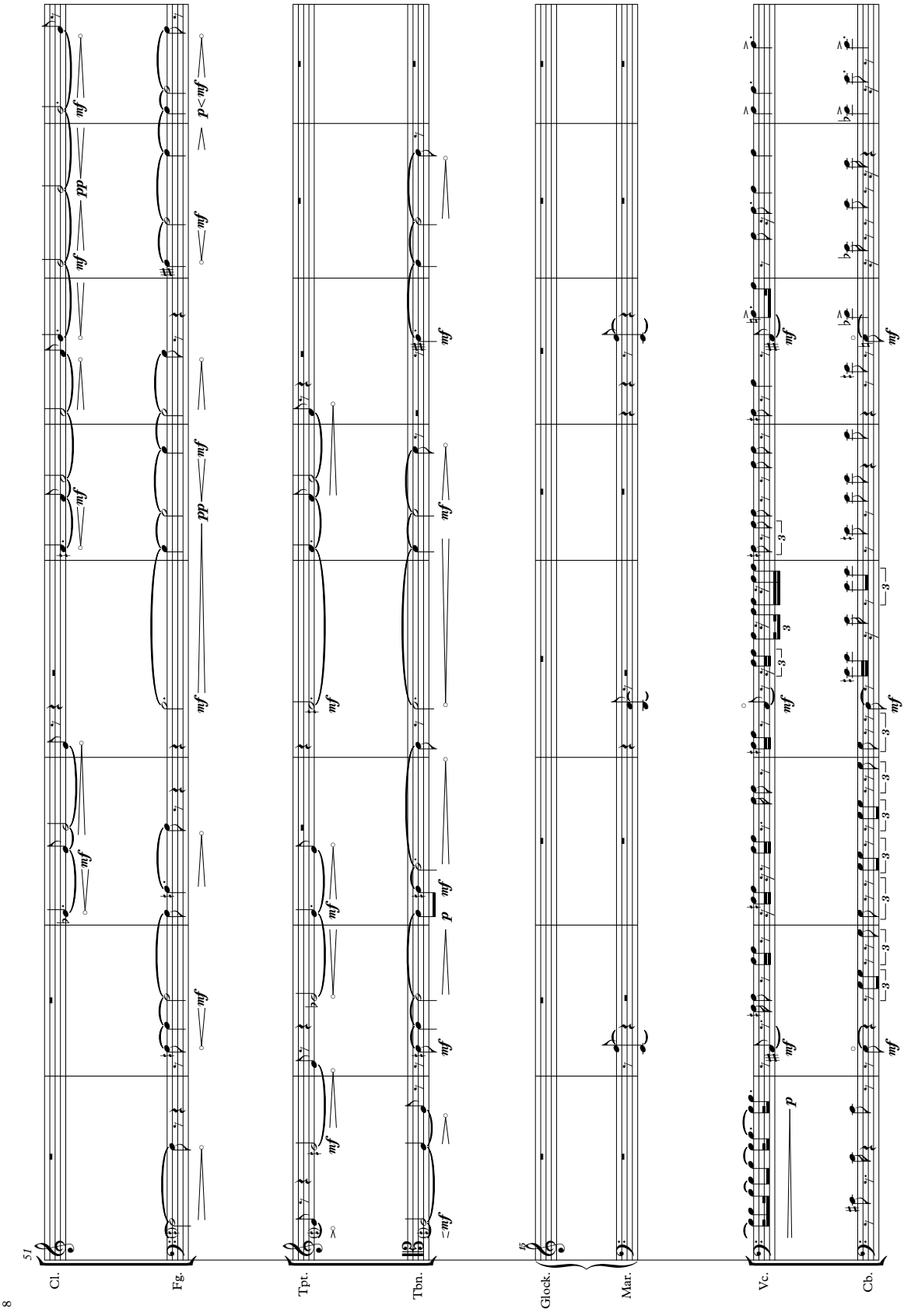

Chaud
Cl. basse en Si♭
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
sourd. sèche
sourd. sèche
arco non vibré
arco
tremolo
pont.
II.0

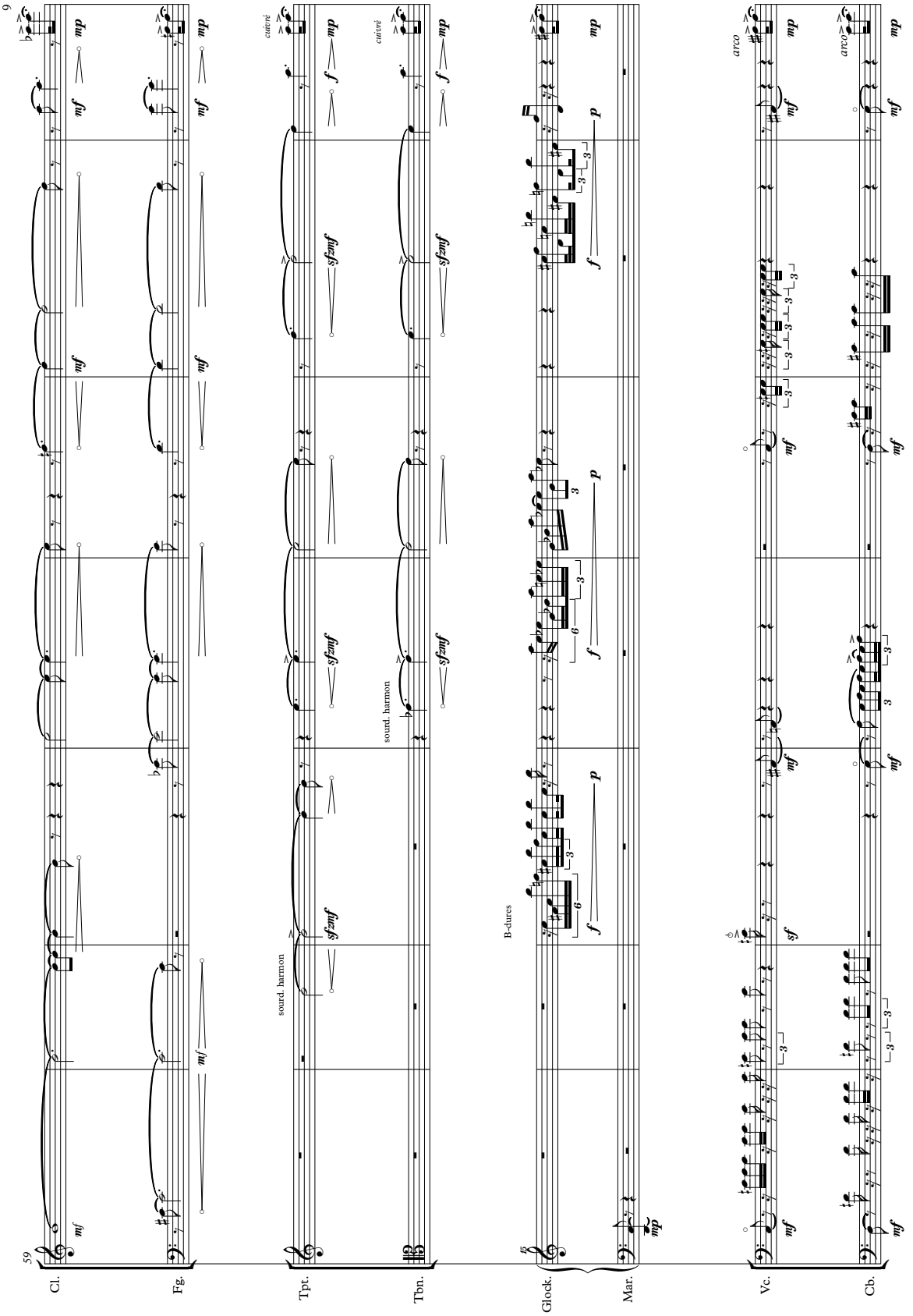

Clarinette en Sib
Cl. B
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
flatt.
sim.
trem.
pont.

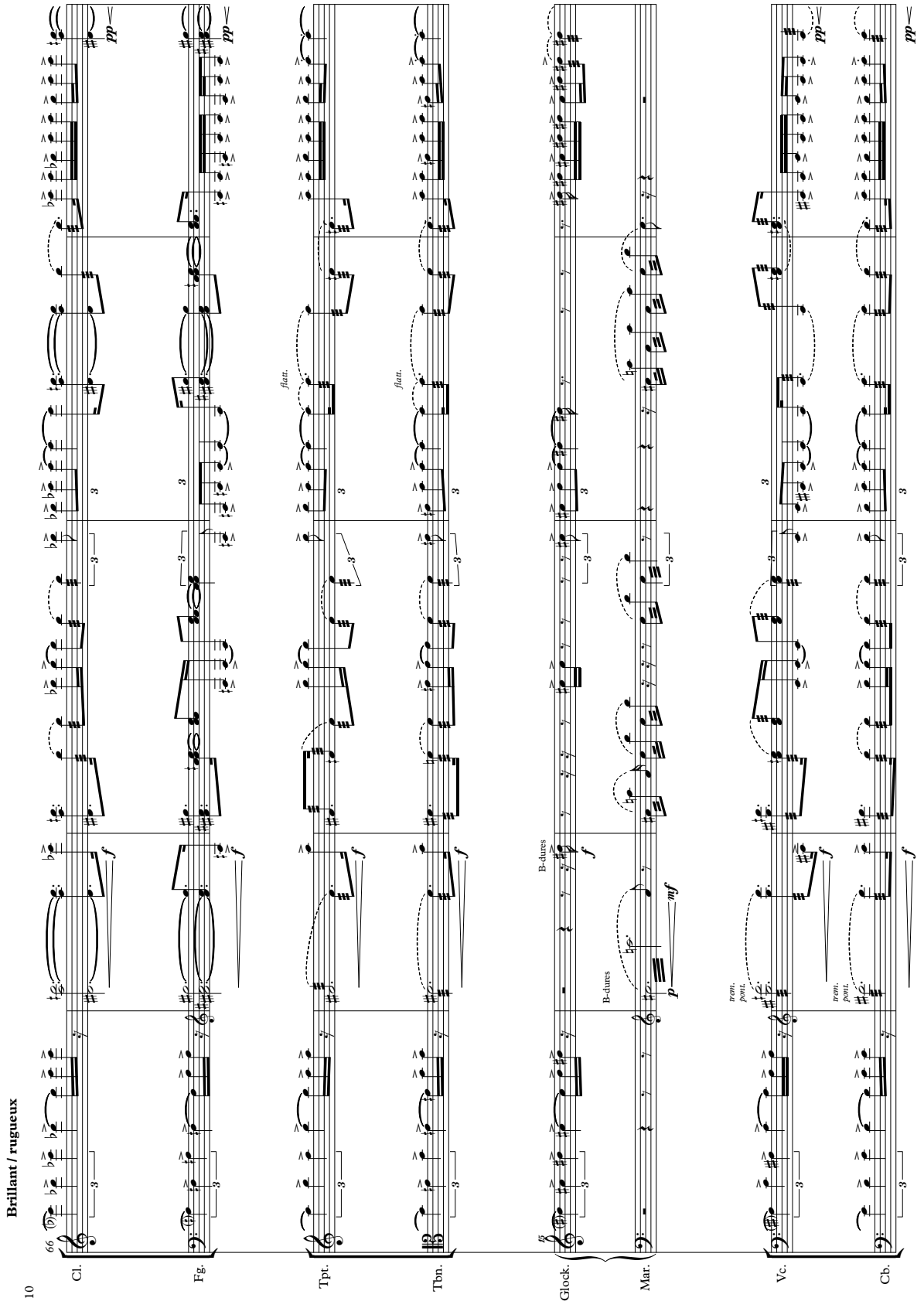

Rugueux
28
5
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
flatt.
écrasé
(B-douces)

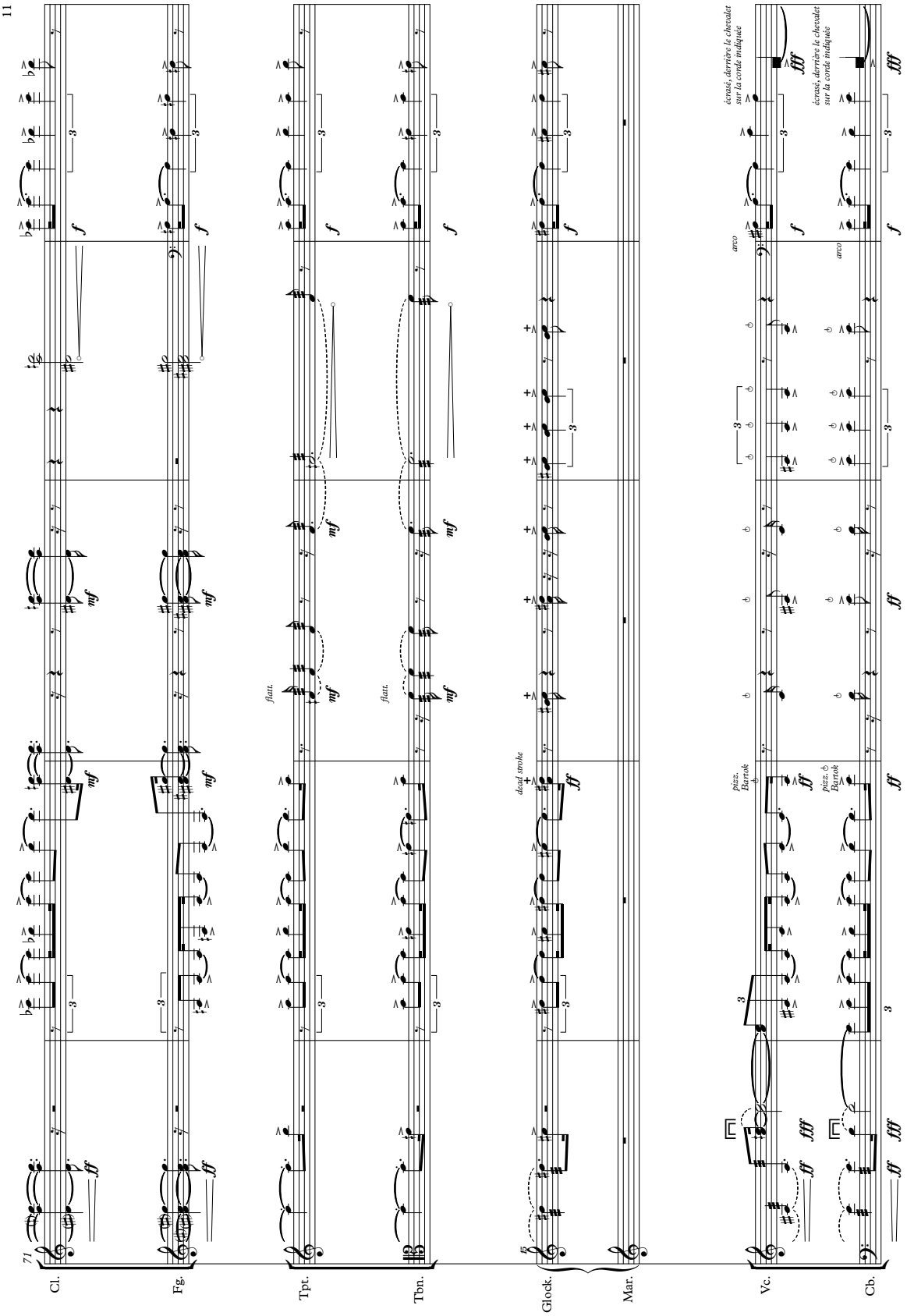

35
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
très piqué
très piqué
sourd. bol
sourd. bol
ou beret
B-dures
dead stroke
pizz.
coupez les résonances
pizz.
coupez les résonances
pizz
non étouffé
pizz
non étouffé
p
p
p
p
mp
p
p
6

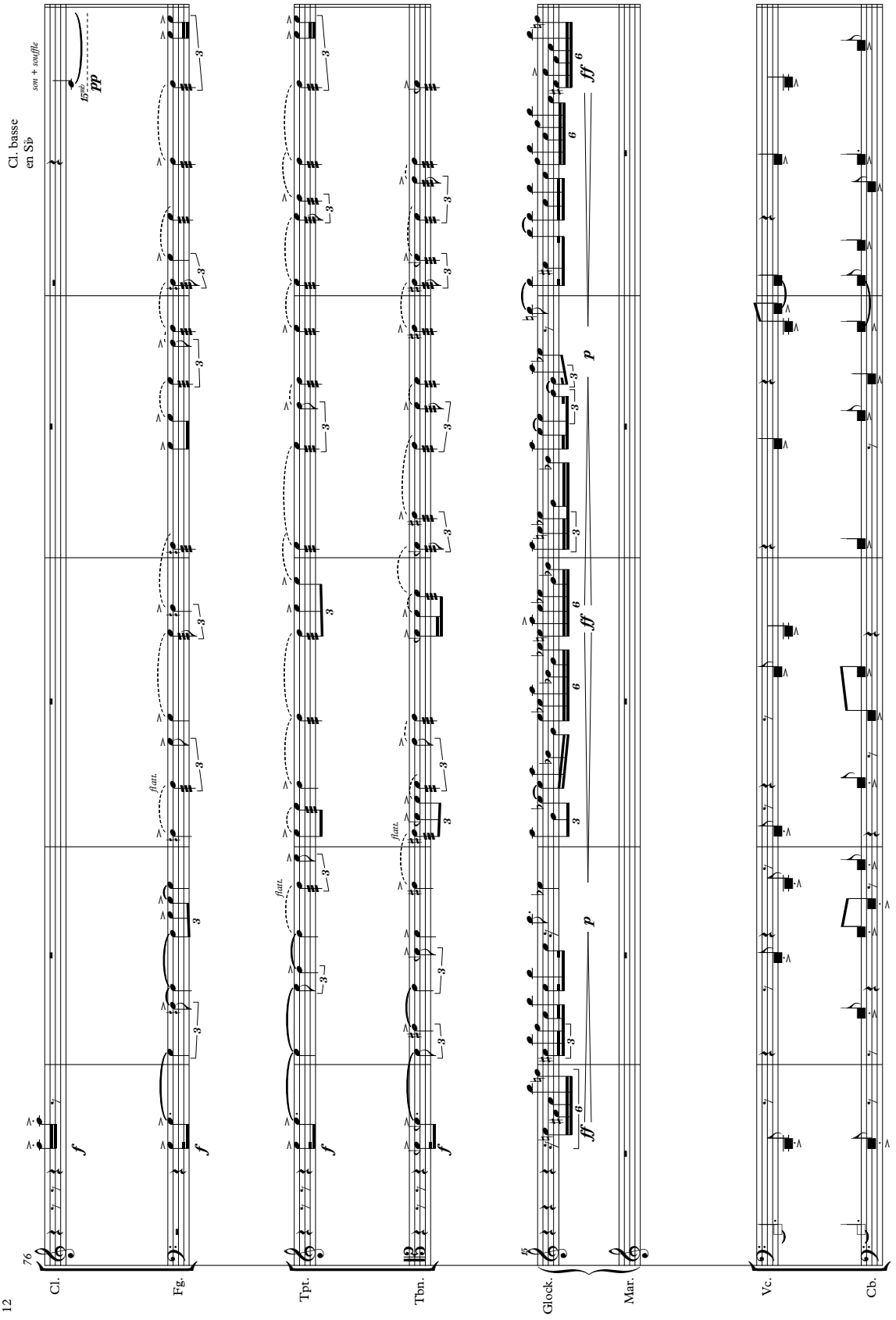

Rond
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
B-douces
IV,0
II,0
mf
mp

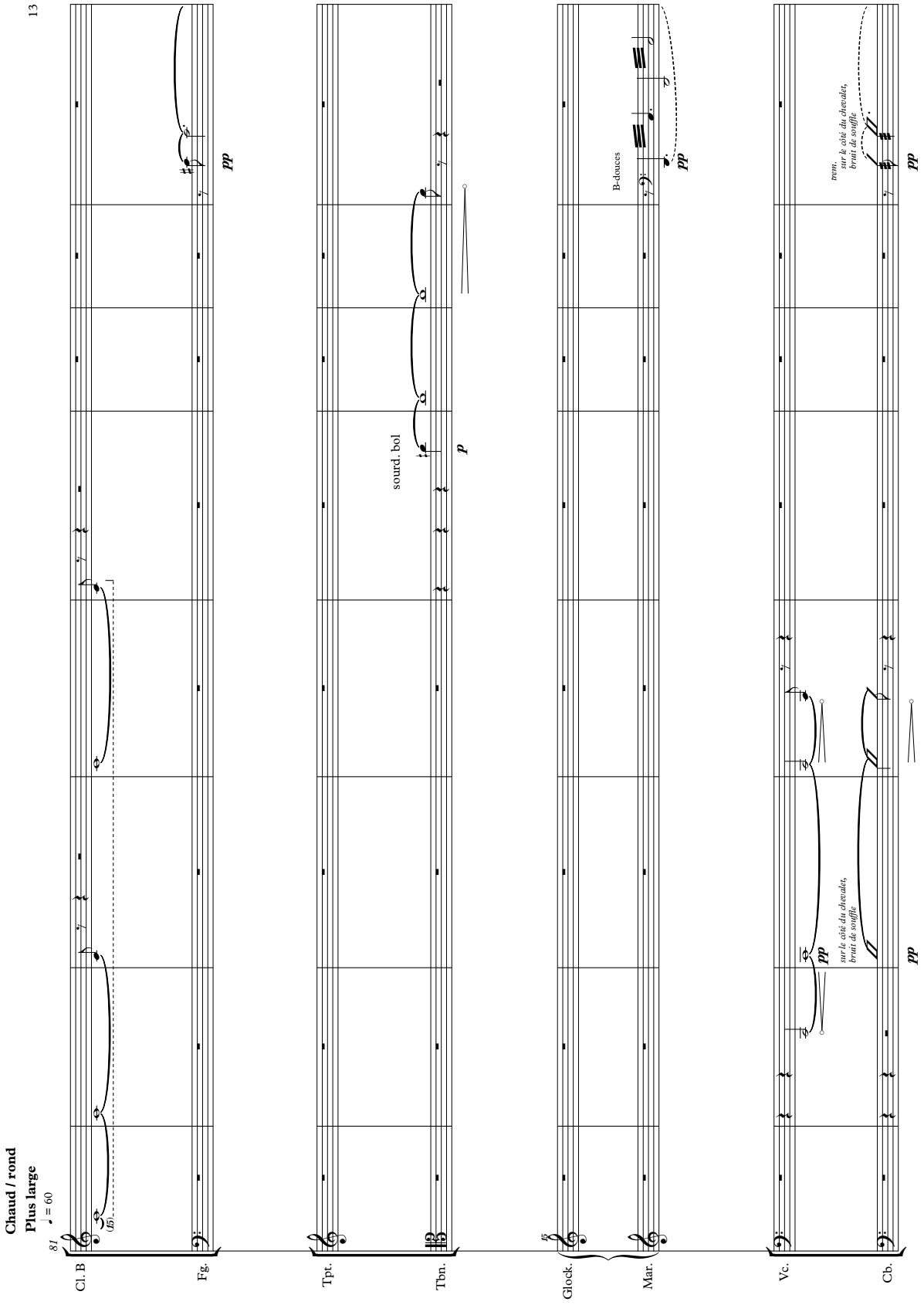

51
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
8

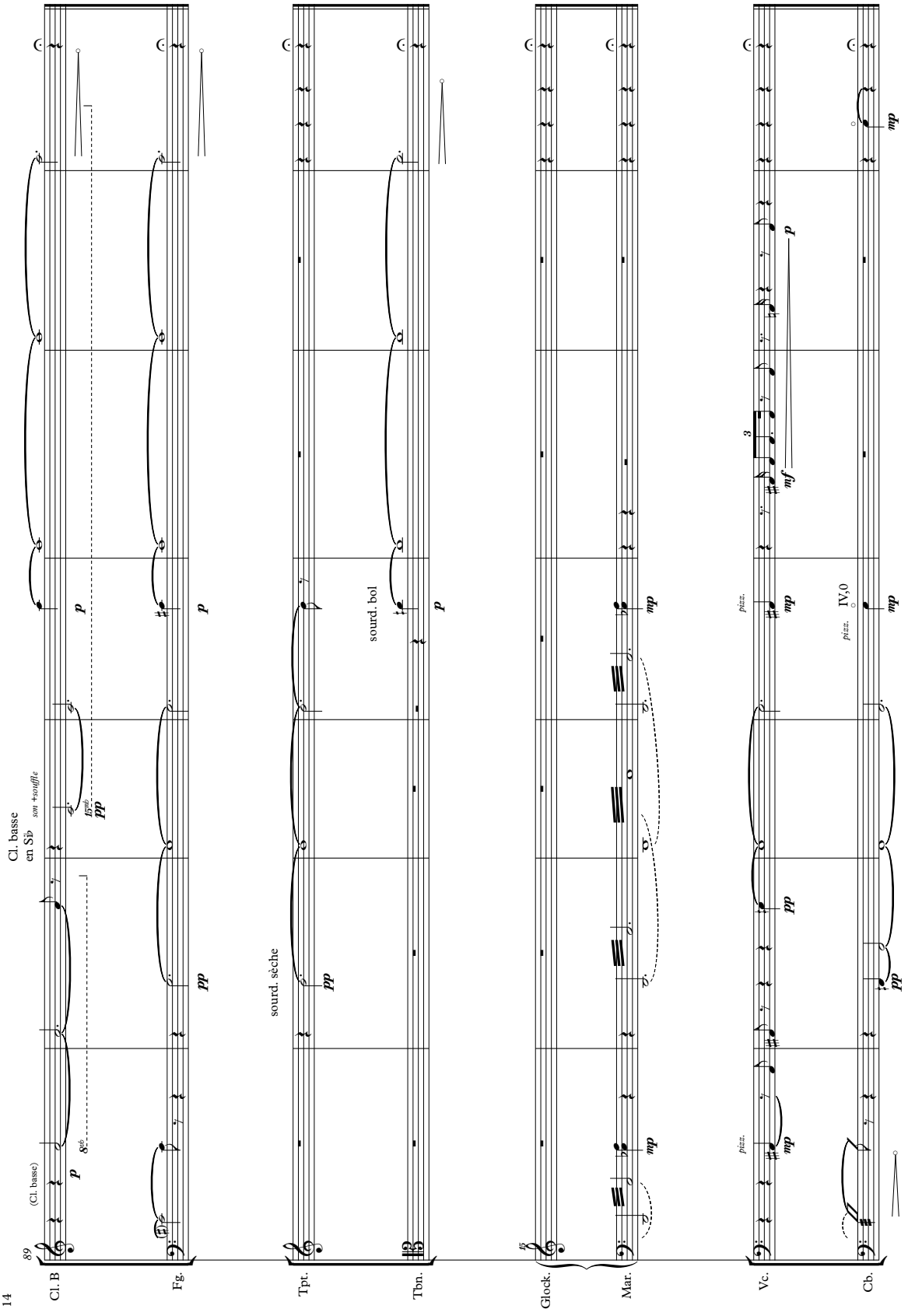

9
59
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
cuivré
sourd. harmon
B-dures
arco
mf
f
p
sf
sfzmf


Brillant / rugueux
10
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
flatt.
B-dures
trem.
pont.


11
71
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
flatt.
dead stroke
écrasé, derrière le chevalet
sur la corde indiquée
arco
pizz.
Bartok


Cl. basse
en Si♭
son + souffle
pp
f
flatt.
ff
p
Cl.
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
76
12


Chaud / rond
Plus large
♩ = 60
81
13
Cl. B
Fg.
Tpt.
Tbn.
Glock.
Mar.
Vc.
Cb.
sourd. bol
B-douces
sur le côté du chevalet, bruit de souffle
trem. sur le côté du chevalet, bruit de souffle
pp
p
ff


14
89
Cl. B
Cl. basse
en Si♭
son +souffle
(Cl. basse)
Fg.
Tpt.
sourd. sèche
Tbn.
sourd. bol
Glock.
Mar.
Vc.
pizz.
IV 0
Cb.

D. Timbre latent space based on BWS judgments

This appendix present the technical details of the deep learning experiment that aims to create a latent space based on the Best-Worst Scaling (BWS) judgment mentioned in the general discussion 6.2.1.

3-D Log-Mel spectrograms

3-D Log-Mel spectrograms (with delta and delta-deltas), already used as input features for various tasks, were introduced for Speech Emotion Recognition (SER) by Chen et al. (2018) as input of a ACRNN model and later used in Meng et al. (2019). Here, the Log-Mel spectrograms are computed as presented in Meng et al. (2019). The 3-D Log-Mel spectrogram consists of a three channel input. The first channel is the static of the Log-Mel spectrogram from 40 filterbanks, the second and third channels are respectively deltas and delta-deltas which can be considered as approximations of the first and second derivatives of the first channel. Once obtained, each 3-D input sample is normalized to have zero mean and unit variance across the entire dataset.

Architecture of (self) Attentive Convolutional Recurrent Neural Net (ACRNN)

Given 3-D log-Mel spectrograms, we used CRNN to extract timbre related high-level features for our specific task.

Architecture of CRNN

The CRNN architecture consists of several 3-D convolution layers, one 3-D maxpooling layers, one linear layer and one Long Short-Term Memory (LSTM) layer. Specifically, the first convolutional layer has 128 feature maps, while the remaining convolutional layers have 256 feature maps, and the filter size of each convolutional layer is 5×3 , where 5 corresponds to the time axis, and 3 corresponds to the frequency axis. A max-pooling is performed after the first convolutional layer with pooling size is 2×2. The 3D features are reshaped to 2D, keeping time dimension unchanged and passed to a linear layer for dimension reduction before reaching the recurrent layer. As precised in Chen et al. (2018), a linear layer of 768 output units is shown to be appropriate. These features are then processed through a bi-directional recurrent neural network with long short term memory cells (BLSTM) (Pan et al., 2020), with 128 cells in each direction, for temporal summarizing, which allows to get d-dimensional high-level feature representations (d = 256).

Self Attention (SA) mechansim

With a sequence of high-level representations, an attention layer is employed to focus on relevant features and produce discriminative utterance-level representations for classification, since not all frame-level CRNN features contribute equally to the representation of timbre.

Specifically, with the classifier's BLSTM output $\mathbf{H} = [\mathbf{h}^1, ..., \mathbf{h}^T] \in \mathbb{R}^{T \times d}$, a temporal vector $\alpha \in \mathbb{R}^T$, representing the contribution per frame to perceived timbre, is computed depending on learnt weights vector $W \in \mathbb{R}^d$. Then $\mathbf{alpha}$ is used to obtain an utterance-level representation by computing the weighted sum of temporal BLSTM internal states $\mathbf{c}$ often called context vector. The attention layer is followed by a fully connected layer that determines the embedding size.

$$\alpha = \frac{\exp(\mathbf{H}W)}{\sum_{t=1}^{T} \exp(W\mathbf{h}^t)} \in \mathbb{R}^{\mathbb{T}} \tag{D.1}$$

$$\mathbf{c} = \sum_{t=1}^{T} \alpha^t \mathbf{h}^t \tag{D.2}$$

The attention layer is followed by a fully connected layer that determines the embedding size.

Cost Function : Relative Contrastive Loss

Definition of the problem

A trial is a group of N sounds on which judgements are made with respect to an attribute $a \in \mathcal{A} = \{round, warm, bright, rough\}$, among those sounds, one has been judged best and one has been judged worst, others can be considered neutral. We denote $\mathcal{T}$ the set containing all the trials considered for the experiment and $\mathcal{J} = \{b, w, n\}$ the set of possible judgements made by participants, each sound has possibly been judged best, worst or neutral, respectively denoted b, w and n. We hence denote $\mathcal{T}^a$ the set containing all the trials related to attribute a considered for the experiment. Hence, for $t \in \mathcal{T}, t = \{x_1, ..., x_N\}$ and $x \in t, j_x^t \in \mathcal{J}$, where j_x^t corresponds to the judgement that has been made on x in the trial t.

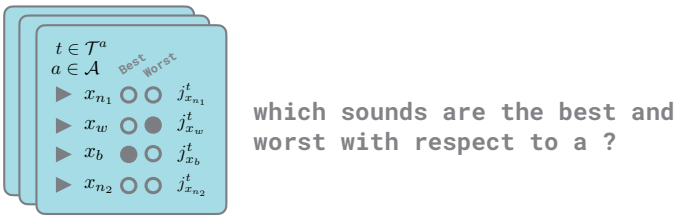


Figure D.1: A trial $t \in \mathcal{T}^a$ of $N = 4$ sounds judged with respect to the BWS paradigm

We denote $\mathbf{h}_x$ the embedding vector corresponding to the sound x, then for a given trial $t \in \mathcal{T}$ depicted in Figure D.1 we assess :

$$\exists x_b \in t | j^t_{x_b} = b \tag{D.3}$$

$$\exists x_w \in t | j^t_{x_w} = w \tag{D.4}$$

$$\exists \{x_{n_1}, ..., x_{n_{N-2}}\} \in t | \forall i \in \{1, ..., N-2\}, j^t_{x_{n_i}} = w \tag{D.5}$$

Therefore, the embedding vectors related to trial t can be denoted as follows

$$\mathbf{h}^t_b = \mathbf{h}_{x_b} \tag{D.6}$$

$$\mathbf{h}^t_w = \mathbf{h}_{x_w} \tag{D.7}$$

$$\forall i \in \{1, ..., N-2\}, \mathbf{h}^t_{n_i} = \mathbf{h}_{x_{n_i}} \tag{D.8}$$

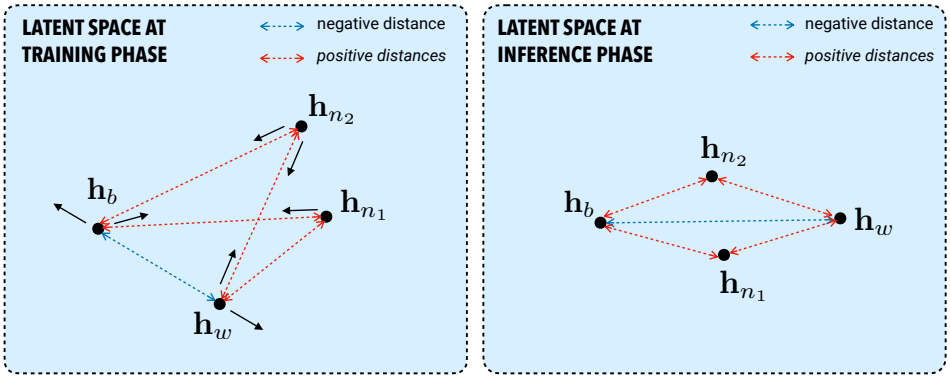


Figure D.2: Configurations of the latent space at training phase (left) and inference phase (right)

Loss design

The RC loss $\mathcal{L}_t$ can be defined for all trial $t \in \mathcal{T}$ as

$$\mathcal{L}_t = \sum_{n \in \{1,\dots,N-2\}} \max\left(\|\mathbf{h}_b^t - \mathbf{h}_{n_i}^t\| - \|\mathbf{h}_b^t - \mathbf{h}_w^t\| + \alpha, 0\right) + \max\left(\|\mathbf{h}_w^t - \mathbf{h}_{n_i}^t\| - \|\mathbf{h}_b^t - \mathbf{h}_w^t\| + \alpha, 0\right) \tag{D.9}$$

For all $x \in \mathcal{S}$ and $a \in \mathcal{A}$, we denote $\mathcal{T}_x^a = \{t^a \in \mathcal{T} | x \in t^a\}$ the subset of $\mathcal{T}^a$ in which each trial contains x and $\mathcal{T'}_x^a$ a subset containing N_t randomly picked elements of $\mathcal{T}_x^a$. From each sound x in the dataset $\mathcal{S}$ we can generate a batch $\mathcal{B}_x$ of size N_b built as follows

$$\mathcal{B}_x = \bigcup_{a \in \mathcal{A}} \{y \in \mathcal{S} | y \in \mathcal{T'}_x^a\} \tag{D.10}$$

where $N_t = \frac{N_b}{|\mathcal{A}| * N}$, and $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$.
Hence, for a given batch $\mathcal{B}_x$ the RC loss can be expressed as follows :

$$\mathcal{L}_{\mathcal{B}_x} = \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}_a} \mathcal{L}_t \tag{D.11}$$

References

Ahumada Jr, A. and Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, 49(6B):1751–1756. Publisher: Acoustical Society of America.

Allen, E. J. and Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *The Journal of the Acoustical Society of America*, 135(3):1371–1379.

Alluri, V. and Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, 27(3):223–242.

Amodio, D. M. (2019). Social Cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences*, 23(1):21–33. Publisher: Elsevier.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., and Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, 25(15):2051–2056. Publisher: Elsevier.

Auger, P., Devinney, T. M., and Louviere, J. J. (2007). Using best–worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of business ethics*, 70(3):299–326.

Austin, J. L. and Warnock, G. J. (1962). *Sense and sensibilia*, volume 83. Clarendon Press Oxford.

Ballet, G., Borghesi, R., Hoffmann, P., and Lévy, F. (1999). Studio online 3.0: An internet" killer application" for remote access to ircam sounds and processing tools. In *Journées d'Informatique Musicale*.

Baumgartner, H. and Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of marketing research*, 38(2):143–156. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Bellemare, M. and Traube, C. (2005). Verbal description of piano timbre according to advanced performers. *Proceedings of the European Society for the Cognitive Sciences of Music (ESCOM2005)*, page 40.

Bernays, M. and Traube, C. (2014). Investigating pianists' individuality in the performance of five timbral nuances through patterns of articulation, touch, dynamics, and pedaling. *Frontiers in Psychology*, 5:157. Publisher: Frontiers.

Bjork, E. (1985). The perceived quality of natural sounds. *Acustica*, 58:185–188.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR).

Bogner, A., Littig, B., and Menz, W. (2009). *Interviewing experts*. Springer.

Boulez, P. (1986). Technology and the Composer. In *The Language of Electroacoustic Music*, pages 5–14. Springer.

Burton, N., Burton, M., Rigby, D., Sutherland, C. A., and Rhodes, G. (2019). Best-worst scaling improves measurement of first impressions. *Cognitive research: principles and implications*, 4(1):1–10. Publisher: SpringerOpen.

Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America*, 118(1):471–482. Publisher: Acoustical Society of America.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319. Publisher: Springer.

Carron, M. (2016). *Méthodes et Outils pour Définir et Véhiculer une Identité Sonore*. PhD Thesis, Université Paris 6 (UPMC).

Carron, M., Dubois, F., Misdariis, N., and Susini, P. (2015). Définir une identité sonore de marque: méthodologie et outils. *Acoustique et Techniques*, 81:pp–20.

Carron, M., Rotureau, T., Dubois, F., Misdariis, N., and Susini, P. (2017). Speaking about sounds: a tool for communication on sound features. *Journal of Design Research*, 15(2):85–109. Publisher: Inderscience Publishers (IEL).

Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M. (2016). Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 619–623. IEEE.

Cheminée, P. (2009). Est-ce bien clair? Stabilité, instabilité et polysémie d'une forme lexicale en contexte. *Le Sentir et le Dire, Concepts et méthodes en psychologie et linguistique cognitives, Daniele Dubois (editor), L'Harmattan Ed*, pages 311–340.

Chen, M., He, X., Yang, J., and Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444. Publisher: IEEE.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chion, M. (2019). Audio-vision: sound on screen. In *Audio-Vision: Sound on Screen*. Columbia University Press.

Croijmans, I., Hendrickx, I., Lefever, E., Majid, A., and Van Den Bosch, A. (2020). Uncovering the language of wine experts. *Natural Language Engineering*, 26(5):511–530. Publisher: Cambridge University Press.

Dal Palù, D., Buiatti, E., Puglisi, G. E., Houix, O., Susini, P., De Giorgi, C., and Astolfi, A. (2017). The use of semantic differential scales in listening tests: A comparison between context and laboratory test conditions for the rolling sounds of office chairs. *Applied Acoustics*, 127:270–283. Publisher: Elsevier.

Darke, G. (2005). Assessment of timbre using verbal attributes. In *Conference on Interdisciplinary Musicology. Montreal, Quebec*. sn.

Deroy, O., Crisinel, A.-S., and Spence, C. (2013). Crossmodal correspondences between odors and contingent features: odors, musical notes, and geometrical shapes. *Psychonomic bulletin & review*, 20(5):878–896. Publisher: Springer.

Deroy, O. and Spence, C. (2013). Why we are not all synesthetes (not even weakly so). *Psychonomic bulletin & review*, 20(4):643–664. Publisher: Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Disley, A. C. and Howard, D. M. (2004). Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing*, 46:25–39. Publisher: Citeseer.

Disley, A. C., Howard, D. M., and Hunt, A. D. (2006). Timbral description of musical instruments. In *International Conference on Music Perception and Cognition*, pages 61–68. Citeseer.

Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive science quarterly*, 1(1):35–68.

Dubois, D., Guastavino, C., and Raimbault, M. (2006). A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. *Acta acustica united with acustica*, 92(6):865–874. Publisher: S. Hirzel Verlag.

Eerola, T., Ferrer, R., and Alluri, V. (2012). Timbre and affect dimensions: Evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. *Music Perception: An Interdisciplinary Journal*, 30(1):49–70. Publisher: University of California Press USA.

Eitan, Z. and Rothschild, I. (2011). How music touches: Musical parameters and listeners' audio-tactile metaphorical mappings. *Psychology of Music*, 39(4):449–467. Publisher: Sage Publications Sage UK: London, England.

Faure, A. (2000). *Des sons aux mots, comment parle-t-on du timbre musical?* PhD Thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS).

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29. Publisher: Taylor & Francis.

Giboreau, A., Dacremont, C., Egoroff, C., Guerrand, S., Urdapilleta, I., Candel, D., and Dubois, D. (2007). Defining sensory descriptors: Towards writing guidelines based on terminology. *Food quality and preference*, 18(2):265–274. Publisher: Elsevier.

Goldstone, R. L., Kersten, A., and Carvalho, P. F. (2013). Concepts and categorization. In *Handbook of psychology: Experimental psychology, Vol. 4, 2nd ed*, pages 607–630. John Wiley & Sons, Inc., Hoboken, NJ, US.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *the Journal of the Acoustical Society of America*, 61(5):1270–1277. Publisher: Acoustical Society of America.

Guarino, N. (1992). Concepts, attributes and arbitrary relations: Some linguistic and ontological criteria for structuring knowledge bases. *Data & Knowledge Engineering*, 8(3):249–261.

Guastavino, C. (2003). *Etude sémantique et acoustique de la perception des basses fréquences dans l'environnement sonore urbain*. PhD Thesis, Université Paris 6 (UPMC).

Guastavino, C. (2007). Categorization of environmental sounds. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 61(1):54–63. Place: Canada Publisher: Canadian Psychological Association.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120. Publisher: American Association for the Advancement of Science.

Guyot, F., Castellengo, M., and Fabre, B. (1997). *Chapitre 2. Étude de la catégorisation d'un corpus de bruits domestiques*. Éditions Kimé. Bibliographie_available: 0 Cairndomain: www.cairn.info Cite Par_available:

0 Pages: 41-58 Publication Title: Catégorisation et cognition : de la perception au discours.

Hajda, J. M., Kendall, R. A., Carterette, E. C., and Harshberger, M. L. (1997). Methodological issues in timbre research. In *Perception and cognition of music*, pages 253–306. Psychology Press/Erlbaum (UK) Taylor & Francis, Hove, England.

Helmholtz, H. L. (1954). On the Sensations of Tone, translated by AJ Ellis.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., and Hilbig, B. E. (2021). lab. js: A free, open, online study builder. *Behavior Research Methods*, pages 1–18. Publisher: Springer.

Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior research methods*, 50(2):711–729. Publisher: Springer.

Hollis, G. and Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior research methods*, 50(1):115–133. Publisher: Springer.

Houix, O. (2003). *Catégorisation auditive des sources sonores*. PhD thesis, Université du Maine.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., and Urdapilleta, I. (2012). A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52. Publisher: American Psychological Association.

Ilkowska, M. and Miśkiewicz, A. (2006). Sharpness versus brightness: A comparison of magnitude estimates. *Acta Acustica united with Acustica*, 92(5):812–819. Publisher: S. Hirzel Verlag.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., and Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244. Publisher: National Acad Sciences.

Jadoul, Y., Thompson, B., and De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15. Publisher: Elsevier.

Jamal, Y., Lacey, S., Nygaard, L., and Sathian, K. (2017). Interactions between auditory elevation, auditory pitch and visual elevation during multisensory perception. *Multisensory research*, 30(3-5):287–306. Publisher: Brill.

Jeon, J. Y., You, J., and Chang, H. Y. (2007). Sound radiation and sound quality characteristics of refrigerator noise in real living environments. *Applied acoustics*, 68(10):1118–1134. Publisher: Elsevier.

Jiang, W., Liu, J., Zhang, X., Wang, S., and Jiang, Y. (2020). Analysis and modeling of timbre perception features in musical sounds. *Applied Sciences*, 10(3):789. Publisher: Multidisciplinary Digital Publishing Institute.

Kendall, R. A. and Carterette, E. C. (1993). Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. *Music Perception*, 10(4):445–467. Publisher: University of California Press.

Kerbrat-Orecchioni, C. (2009). *L'énonciation: de la subjectivité dans le langage*. Armand Colin.

Kiritchenko, S. and Mohammad, S. M. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv preprint arXiv:1712.01765*.

Klapetek, A., Ngo, M. K., and Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics*, 74(6):1154–1167.

Klapuri, A. and Davy, M. (2007). *Signal processing methods for music transcription*. Springer New York, NY. Publisher: Springer New York, NY.

Klein, F. (1884). *Vorlesungen über das Ikosaeder und die Auflösung der Gleichungen vom fünften Grade*. BG Teubner.

Krimphoff, J., McAdams, S., and Winsberg, S. (1994). Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique. *Le Journal de Physique IV*, 4(C5):C5–625.

Krumhansl, Carol, L. (1989). Why is Musical Timbre so hard to understand? In Nielsen, S. and Olsson, O., editors, *Structure and perception of electroacoustic sound and music*, pages 43–53. Elsevier.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & psychophysics*, 62(7):1426–1439. Publisher: Springer.

Lakoff, G. (2007). Cognitive models and prototype theory. *The cognitive linguistics reader*, pages 130–167. Publisher: Equinox London.

Lakoff, G. and Turner, M. (2009). *More than cool reason: A field guide to poetic metaphor*. University of Chicago press.

Landau, S. I. (1984). *Dictionaries: The art and craft of lexicography*. Macmillan Reference USA.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284. Publisher: Taylor & Francis.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374. Publisher: JSTOR.

Lange, K., Kühn, S., and Filevich, E. (2015). " Just Another Tool for Online Studies"(JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS one*, 10(6):e0130834. Publisher: Public Library of Science San Francisco, CA USA.

Lavoie, M. (2014). *Conceptualisation et communication des nuances de timbre à la guitare classique*. PhD thesis, Université de Montréal. Accepted: 2014-03-24T16:06:42Z.

Le Moine, C. and Obin, N. (2020). Att-HACK: An expressive speech database with social attitudes. *arXiv preprint arXiv:2004.04410*.

Le Moine, C., Obin, N., and Roebel, A. (2021). Speaker attentive speech emotion recognition. *arXiv preprint arXiv:2104.07288*.

Lemaitre, G., Houix, O., Misdariis, N., and Susini, P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, 16(1):16. Publisher: American Psychological Association.

Lemaitre, G., Vartanian, C., Lambourg, C., and Boussard, P. (2015). A psychoacoustical study of wind buffeting noise. *Applied acoustics*, 95:1–12. Publisher: Elsevier.

Liu, B. and Udell, M. (2020). Impact of accuracy on model interpretations. *arXiv preprint arXiv:2011.09903*.

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Löbner, S. (2013). *Understanding semantics*. Routledge.

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):384. Publisher: American Psychological Association.

Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *The Journal of the Acoustical Society of America*, 114(5):2946–2957. Publisher: Acoustical Society of America.

McAdams, S., Douglas, C., and Vempala, N. N. (2017). Perception and modeling of affective qualities of musical instrument sounds across pitch registers. *Frontiers in Psychology*, 8:153. Publisher: Frontiers.

McAdams, S., Vines, B. W., Vieillard, S., Smith, B. K., and Reynolds, R. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception*, 22(2):297–350. Publisher: University of California Press USA.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3):177–192.

McCormick, K., Kim, J., List, S. M., and Nygaard, L. C. (2015). Sound to Meaning Mappings in the Bouba-Kiki Effect. In *CogSci*, volume 2015, pages 1565–1570.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python.

In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer.

Meng, H., Yan, T., Yuan, F., and Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE access*, 7:125868–125881. Publisher: IEEE.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Misdariis, N., Minard, A., Susini, P., Lemaitre, G., McAdams, S., and Parizet, E. (2010). Environmental sound perception: Metadescription and modeling based on independent primary studies. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:1–26. Publisher: Springer.

Misdariis, N., Susini, P., Houix, O., Rivas, R., Cerles, C., Lebel, E., Tetienne, A., and Duquesne, A. (2021). Mapping Sound Properties and Oenological Characters by a Collaborative Sound Design Approach – Towards an Augmented Experience. In Kronland-Martinet, R., Ystad, S., and Aramaki, M., editors, *Perception, Representations, Image, Sound, Music*, Lecture Notes in Computer Science, pages 504–516, Cham. Springer International Publishing.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Moore, B. C. and Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345. Publisher: S. Hirzel Verlag.

Moore, B. C., Glasberg, B. R., and Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240. Publisher: Audio Engineering Society.

Namba, S., Kuwano, S., Hashimoto, T., Berglund, B., Da Rui, Z., Schick, A., Hoege, H., and Florentine, M. (1991). Verbal expression of emotional impression of sound A cross-cultural study. *Journal of the Acoustical Society of Japan (E)*, 12(1):19–29. Publisher: Acoustical Society of Japan.

Nykänen, A., Johansson, , Lundberg, J., and Berg, J. (2009). Modelling perceptual dimensions of saxophone sounds. *Acta Acustica United with Acustica*, 95(3):539–549. Publisher: S. Hirzel Verlag.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning*. The measurement of meaning. Univer. Illinois Press, Oxford, England. Pages: 342.

Pan, Z., Luo, Z., Yang, J., and Li, H. (2020). Multi-modal attention for speech emotion recognition. *arXiv preprint arXiv:2009.04107*.

Parise, C. V., Knorre, K., and Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, 111(16):6104–6108. Publisher: National Acad Sciences.

Parizet, E., Hamzaoui, N., and Sabatie, G. (2005). Comparison of some listening test methods: a case study. *Acta Acustica united with Acustica*, 91(2):356–364. Publisher: S. Hirzel Verlag.

Parr, W. V., Heatherbell, D., and White, K. G. (2002). Demystifying wine expertise: Olfactory threshold, perceptual skill and semantic memory in expert and novice wine judges. *Chemical senses*, 27(8):747–755. Publisher: Oxford University Press.

Paté, A., Carrou, J.-L. L., Navarret, B., Dubois, D., and Fabre, B. (2015). Influence of the electric guitar's fingerboard wood on guitarists' perception. *Acta Acustica united with Acustica*, 101(2):347–359. Publisher: S. Hirzel Verlag.

Pawelec, A. (2006). The death of metaphor. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, (123).

Pearce, A., Brookes, T., and Mason, R. (2017). Timbral attributes for sound effect library searching. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society.

Pearce, A., Brookes, T., and Mason, R. (2019). Modelling timbral hardness. *Applied Sciences*, 9(3):466. Publisher: Multidisciplinary Digital Publishing Institute.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916. Publisher: Acoustical Society of America.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Plomp, R. (1970). Timbre as a multi-dimensional attribute of complex tones. *Frequency analysis and periodicity detection in hearing*, pages 397–414.

Ponsot, E., Burred, J. J., Belin, P., and Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences*, 115(15):3972–3977. Publisher: National Acad Sciences.

Ponsot, E., Susini, P., Saint Pierre, G., and Meunier, S. (2013). Temporal loudness weights for sounds with increasing and decreasing intensity profiles. *The Journal of the Acoustical Society of America*, 134(4):EL321–EL326. Publisher: Acoustical Society of America.

Porcello, T. (2004). Speaking of sound: language and the professionalization of sound-recording engineers. *Social Studies of Science*, 34(5):733–758. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological bulletin*, 86(4):777.

Pratt, R. and Doak, P. E. (1976). A subjective rating scale for timbre. *Journal of Sound and Vibration*, 45(3):317–328. Publisher: Elsevier.

Pressnitzer, D. and McAdams, S. (1999). Two phase effects in roughness perception. *The Journal of the Acoustical Society of America*, 105(5):2773–2782. Publisher: Acoustical Society of America.

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, pages 64–99. Publisher: Appleton-Century-Crofts.

Reymore, L. and Huron, D. (2020). Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia.

Psychomusicology: Music, Mind, and Brain, 30(3):124. Publisher: Educational Publishing Foundation.

Risset, J.-C. and Wessel, D. L. (1999). Exploration of timbre by analysis and synthesis. In *The psychology of music*, pages 113–169. Elsevier.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192. Publisher: American Psychological Association.

Rosi, V., Houix, O., Misdariis, N., and Susini, P. (2022a). Investigating the Shared Meaning of Metaphorical Sound Attributes: Bright, Warm, Round, and Rough. *Music Perception*, 39(5):468–483.

Rosi, V., Ravillion, A., Houix, O., and Susini, P. (2022b). Best-worst scaling, an alternative method to assess perceptual sound qualities. *JASA Express Letters*, 2(6):064404. Publisher: Acoustical Society of America.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.

Saitis, C., Fritz, C., and Scavone, G. (2019). Sounds like melted chocolate: How musicians conceptualize violin sound richness. In *International Symposium on Musical Acoustics*.

Saitis, C., Fritz, C., Scavone, G. P., Guastavino, C., and Dubois, D. (2017). Perceptual evaluation of violins: A psycholinguistic analysis of preference verbal descriptions by experienced musicians. *The Journal of the Acoustical Society of America*, 141(4):2746–2757. Publisher: Acoustical Society of America.

Saitis, C. and Siedenburg, K. (2020). Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories. *The Journal of the Acoustical Society of America*, 148(4):2256–2266. Publisher: Acoustical Society of America.

Saitis, C., Weinzierl, S., von Kriegstein, K., Ystad, S., and Cuskley, C. (2020). Timbre semantics through the lens of crossmodal correspondences: A new way of asking old questions. *Acoustical Science and Technology*, 41(1):365–368. Publisher: Acoustical Society of Japan.

Salais, L., Arias, P., Le Moine, C., Rosi, V., Teytaut, Y., Obin, N., and Roebel, A. (2022). Production Strategies of Vocal Attitudes.

Schaeffer, P. (1966). *Traité des objets musicaux*. Le Seuil.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Schubert, E. (2001). Continuous measurement of self-report emotional response to music. In *Music and emotion: Theory and research*, Series in affective science, pages 393–414. Oxford University Press, New York, NY, US.

Schubert, E. and Wolfe, J. (2006). Does timbral brightness scale with frequency and spectral centroid? *Acta acustica united with acustica*, 92(5):820–825. Publisher: S. Hirzel Verlag.

Schuman, H. and Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Schön, D., Ystad, S., Kronland-Martinet, R., and Besson, M. (2010). The evocative power of sounds: Conceptual priming between words and nonverbal sounds. *Journal of cognitive neuroscience*, 22(5):1026–1035. Publisher: MIT Press.

Shapley, L. (1953). QUOTA SOLUTIONS OP n-PERSON GAMES1. *Edited by Emil Artin and Marston Morse*, page 343.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140. Publisher: Springer.

Siedenburg, K., Jacobsen, S., and Reuter, C. (2021). Spectral envelope position and shape in sustained musical instrument sounds. *The Journal of the Acoustical Society of America*, 149(6):3715–3726. Publisher: Acoustical Society of America.

Smith, E. E. (1989). Concepts and induction. In *Foundations of cognitive science*, pages 501–526. The MIT Press, Cambridge, MA, US.

Solomon, L. N. (1958). Semantic approach to the perception of complex sounds. *The journal of the Acoustical Society of America*, 30(5):421–425. Publisher: Acoustical Society of America.

Spence, C. and Deroy, O. (2012). Crossmodal correspondences: Innate or learned? *i-Perception*, 3(5):316–318. Publisher: SAGE Publications Sage UK: London, England.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245. Publisher: American Psychological Association.

Stepánek, J. (2002). Evaluation of timbre of violin tones according to selected verbal attributes. In *32nd Int. Acoustical Conf., European Acoustics Association (EAA) Symp." Acoustics Banská Stiavnica*, pages 129–132.

Stepánek, J. (2006). Musical sound timbre: Verbal description and dimensions. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pages 121–126. Citeseer.

Strauss, A. and Corbin, J. (1994). Grounded theory methodology: An overview. In *Handbook of qualitative research*, pages 273–285. Sage Publications, Inc, Thousand Oaks, CA, US.

Susini, P., Lemaitre, G., and McAdams, S. (2011). Psychological measurement for sound description and evaluation. In *Measurement With Persons*. Psychology Press. Num Pages: 28.

Susini, P., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., and Rodet, X. (2004). Characterizing the sound quality of air-conditioning noise. *Applied acoustics*, 65(8):763–790. Publisher: Elsevier.

Suárez Toste, E. (2007). Metaphor inside the wine cellar: On the ubiquity of personification schemas in winespeak. *Metaphorik. de*, 12(1):53–64.

Terhardt, E. (1974). On the perception of periodic sound fluctuations (roughness). *Acta Acustica united with Acustica*, 30(4):201–213. Publisher: S. Hirzel Verlag.

Thoret, E., Caramiaux, B., Depalle, P., and Mcadams, S. (2021). Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre. *Nature Human Behaviour*, 5(3):369–377. Publisher: Nature Publishing Group.

Traube, C. (2004). *An interdisciplinary study of the timbre of the classical guitar*. PhD thesis, McGill University.

Van Audenhove, L. (2007). Expert interviews and interview techniques for policy analysis. *Vrije University, Brussel Retrieved May*, 5:2009.

Von Bismarck, G. (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica*, 30(3):146–159. Publisher: S. Hirzel Verlag.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., and Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, 21(1):21–25. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Wallmark, Z. (2019a). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, 47(4):585–605. Publisher: SAGE Publications Sage UK: London, England.

Wallmark, Z. (2019b). Semantic crosstalk in timbre perception. *Music & Science*, 2:2059204319846617. Publisher: SAGE Publications Sage UK: London, England.

Wallmark, Z., Frank, R. J., and Nghiem, L. (2019). Creating novel tones from adjectives: An exploratory study using FM synthesis. *Psychomusicology: Music, Mind, and Brain*, 29(4):188. Publisher: Educational Publishing Foundation.

Wallmark, Z., Nghiem, L., and Marks, L. E. (2021). Does Timbre Modulate Visual Perception? Exploring Crossmodal Interactions. *Music Perception: An Interdisciplinary Journal*, 39(1):1–20. Publisher: University of California Press.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer music journal*, pages 45–52. Publisher: JSTOR.

Winsberg, S. and De Soete, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika*, 58(2):315–330. Publisher: Springer.

Zacharakis, A., Pastiadis, K., and Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception: An Interdisciplinary Journal*, 31(4):339–358.

Zeitler, A. and Hellbrück, J. (2001). Semantic attributes of environmental sounds and their correlations with psychoacoustic magnitude. In *Proc. of the 17th International Congress on Acoustics [CDROM], Rome, Italy*, volume 28.

Zielinski, S., Hardisty, P., Hummersone, C., and Rumsey, F. (2007). Potential biases in MUSHRA listening tests. In *Audio Engineering Society Convention 123*. Audio Engineering Society.

Zwicker, E. and Fastl, H. (2013). *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media.

Zwicker, E. and Scharf, B. (1965). A model of loudness summation. *Psychological review*, 72(1):3. Publisher: American Psychological Association.

Özcan, E. and Egmond van, R. (2012). Basic Semantics of Product Sounds. *International Journal of Dsign*, 6(2).

Özcan, E. and van Egmond, R. (2005). Characterizing descriptions of product sounds. In *Proceedings of the 11th International Conference on Auditory Display*, pages 55–60.