# Psychophysical scaling of timbre-related audio descriptors

*Savvas Kazazis*

Music Technology Area, Department of Music Research
Schulich School of Music, McGill University
Montreal, Canada

August 2020

A thesis submitted to McGill University in partial fulfilment of the requirements for the degree of Doctor of Philosophy.



**2020/08/22**

to
Mom
Dad
&
Nona

# Abstract

Timbre has most often been studied through correlational analyses that associate perceptual dimensions (or in other words, the orthogonal axes of a timbre space), with a wide set of audio descriptors. These timbre-related descriptors have been widely used for tasks such as acoustical interpretation of perceptual dimensions, explaining listeners' dissimilarity ratings between pairs of sounds, content delivery in music information retrieval systems, investigating timbre semantics, quantifying affective responses to sound, predicting the amount of blend between instrumental sounds, investigating cognitive factors related to memory for timbre, and developing perceptually based audio processing strategies and sound effects. However, audio descriptors identified through correlational analyses do not necessarily causally relate to listeners' perceptions and when entered into statistical regression models may lead to false positive interpretations about their perceptual significance, especially when features also strongly covary.

The present thesis undertakes an investigation of timbre by establishing psychophysical correspondences between perception and several timbre-related audio descriptors. Three studies investigate whether listeners perceive spectral audio descriptors on perceptual ordinal, interval, and ratio scales, and temporal descriptors on an ordinal scale. The stimuli used in each of the presented experiments were constructed through specifically designed synthesis algorithms that enabled control of each audio descriptor independently of the rest and that therefore isolated as much as was feasible the effect of each descriptor on listeners' perceptions.

The first study perceptually validates spectral audio descriptors through an ordinal scaling experiment. The results indicate that listeners were overall able to rank order the stimuli of a particular feature set when presented with an appropriate spacing of descriptor values. The second study underpins the effect of amplitude envelope features and spectral envelope features on the perceptual relevance of temporal audio descriptors and inharmonicity, through another ordinal scaling experiment. In addition, the results of that experiment suggested a combination of acoustically independent descriptors that could potentially explain variation along a spectrotemporal perceptual dimension, which in previous timbre studies had been considered to be strictly temporal. The above hypothesis was confirmed by conducting a meta-analysis on the timbre spaces derived from previous studies. The third study comprises two experiments that provided interval scale and ratio scale measurements on spectral audio descriptors. The results of the first experiment indicated that listeners were overall able to estimate intervals of spectral descriptors. We therefore proceeded to a ratio scaling experiment the results of which indicated that listeners can also produce ratios of descriptor values and enabled the construction of psychophysical ratio scales of each descriptor tested.

The findings advance the current knowledge on timbre perception both by establishing cause-and-effect relations between audio descriptors and perceptual dimensions and by expanding previous research in which the acoustical interpretations of perceptual dimensions

were made solely under the prism of correlational analysis. The psychophysical correspondences between perception and audio descriptors reported in this thesis will hopefully serve as a basis for future research, which may attempt to study timbre as a phenomenon that emerges from a combination of audio features and explore its psychophysical attributes through perceptual dominance hierarchies of those features.

# Résumé

Le timbre a le plus souvent été étudié au moyen d'analyses corrélationnelles qui associent à des dimensions perceptives (ou en d'autres termes, les axes orthogonaux d'un espace de timbre) un large éventail de descripteurs audio. Ces descripteurs liés au timbre ont été largement utilisés pour des tâches telles que l'interprétation acoustique des dimensions perceptives, l'explication des dissemblances d'estimation d'indices entre paires de sons par des auditeurs, la distribution de contenu par des systèmes de recherche d'informations musicales, l'étude de la sémantique du timbre, la quantification des réponses affectives au son, la prédiction du niveau de fusion entre sons instrumentaux, l'étude des facteurs cognitifs liés à la mémoire du timbre, ainsi que le développement de stratégies de traitement audio et d'effets sonores fondés sur la perception. Cependant, les descripteurs audio identifiés par ces analyses corrélationnelles ne sont pas nécessairement liés de façon causale à la perception des auditeurs et, lorsque traités par des modèles de régression statistique, peuvent conduire à des interprétations faussement positives de leur signification perceptive, en particulier lorsque les caractéristiques sont également fortement covariants.

Cette thèse présente une étude sur le timbre qui établit des correspondances psychophysiques entre perception et descripteurs audio liés au timbre. Trois études examinent si les auditeurs perçoivent les descripteurs audio spectraux sur des échelles ordinales, d'intervalle et de rapport perceptifs, et les descripteurs temporels sur une échelle ordinale. Les stimuli utilisés dans chacune des expériences présentées ont été construits grâce à des algorithmes de synthèse spécialement conçus pour contrôler chaque descripteur audio indépendamment des autres, isolant autant que possible l'effet de chaque descripteur sur la perception par les auditeurs.

La première étude valide perceptivement les descripteurs audio spectraux par une expérience d'échelle ordinale. Les résultats indiquent que les auditeurs étaient globalement capables de classer de manière ordonnée les stimuli d'une même caractéristique spécifique lorsque présentés selon un espacement approprié des valeurs des descripteurs. La deuxième étude souligne l'effet des caractéristiques d'enveloppe d'amplitude et des caractéristiques d'enveloppe spectrale sur la pertinence perceptive des descripteurs audio temporels et sur l'inharmonicité, grâce à une autre expérience de mise à l'échelle ordinale. De plus, les résultats de cette expérience suggèrent une combinaison de descripteurs acoustiquement indépendants qui pourrait potentiellement expliquer une variation le long d'une dimension perceptive spectro-temporelle, qui a été considérée jusqu'à maintenant comme strictement temporelle selon de précédentes études sur le timbre. L'hypothèse ci-dessus a été confirmée par la réalisation d'une méta-analyse sur des espaces de timbre résultants d'études antérieures. La troisième étude consiste en deux expériences qui ont fourni des mesures d'échelle d'intervalle et d'échelle de rapport sur des descripteurs audio spectraux. Les résultats de la première expérience ont indiqué que les auditeurs étaient globalement capables d'estimer des intervalles de descripteurs spectraux. Nous avons donc procédé à une expérience de mise à l'échelle des rapports dont les résultats indiquent que les auditeurs

peuvent également produire des rapports de valeurs de descripteurs ce qui a permis la construction d'échelles de rapports psychophysiques pour chaque descripteur testé.

Les résultats de cette thèse ont fait progresser les connaissances actuelles sur la perception du timbre à la fois en établissant des relations de cause à effet entre descripteurs audio et dimensions perceptives, et en prolongeant les recherches antérieures dans lesquelles les interprétations acoustiques des dimensions perceptives étaient établies uniquement à travers le prisme de l'analyse corrélationnelle. Nous espérons que les correspondances psychophysiques entre perception et descripteurs audio rapportées dans cette thèse serviront de base à de futures recherches, qui viseraient à étudier le timbre en tant que phénomène émergeant d'une combinaison de caractéristiques audio et à explorer ses attributs psychophysiques selon des hiérarchies de dominance perceptives de ces caractéristiques.

## Acknowledgments

My first sincere gratitude is extended to Paul Berg, who was my supervisor during my studies at the Institute of Sonology. Paul soon realized that I was very interested in this "timbre business", as he might say, and encouraged me to approach Stephen McAdams, hoping that I could join his team at McGill University for a semester as a visiting student. If it wasn't for Paul, ...

Stephen McAdams fortunately accepted me in his team and gave me the opportunity to collaborate with many brilliant timbre researchers, who also warmly welcomed me to Music Perception and Cognition Lab (MPCL). My thanks go to Meghan Goodchild, Chelsea Douglas, Sven-Amin Lembke, Kai Siedenburg, Yinan Tsao, and David Sears. In particular, I would like to thank Cecilia Taher for her kindness, support, and friendship, and Charalampos Saitis who was the first person to introduce me to Montreal.

Special thanks go to Bennett Smith for his kindness, technical assistance, and sense of humor. Bennett soon became my favorite person among MPCLers. I recall us standing together on McAdams' rooftop, during MPCL's Christmas party being held downstairs and sharing some cigarettes along with McAdam's homemade glöggi drinks, while peacefully watching the stars over snowy Montreal and standing mostly quiet at –20 C degrees. Finally, we exchanged a warm hug, as I was about to soon leave Montreal. This memory will always hold a special place in my heart – besides, that was the moment where I made a wish to be accepted as a PhD student at McGill after realizing that my research interests could not be nourished elsewhere, not without McAdam's supervision.

My wish came true, and about nine months later I was back in MPCL, the lab of my beloved advisor Stephen McAdams. I have no words to express my deepest gratitude and love to Stephen for his unfailing support and generosity during the various stages of my student career. Without his constant encouragement I wouldn't have been able to deliver this thesis. His tremendous research output has always been a great source of inspiration for me. He taught me how to do research and helped me to find the right questions before seeking their answers: "Eddy, no... You are making a leap of faith here!" Without his methodological rigor, such "leaps of faith" would not have been avoided in my work. Stephen has been one of my greatest teachers, and it has been a great honor to work with and learn from him.

I would also like to extend my gratitude to Philippe Depalle, who soon became my thesis co-supervisor and indeed, this work wouldn't have been possible without his contribution. I am thankful to Philippe for the many discussions we had on digital signal processing and most importantly for his genuine interest, support, and advice on solving some practical issues that at some point occurred during my academic career as a lecturer. I am honored to have been given the chance to collaborate with him and to TA one of his courses has significantly helped me to improve as a lecturer.

During my "service" at MPCL, I had the chance to meet and collaborate with another group of timbre researchers, which I would like to thank for their support and friendship.

My thanks go to Jason Noble, Moe Touizrar, Etienne Thoret (for more gin-gin tonics to come), Kit Soden, Lena Heng, and Max Henry. In particular, special thanks go to Erica Huynh for her help in recruiting and running participants for my experiments, Iza Korsmit for introducing me to misophonia and putting together a CIRMMT research project on that topic, and Aurélien Antoine for translating the abstract of this thesis in French. I would like to again acknowledge Bennett for programming the user interfaces of my experiments and for his extreme patience and tolerance to my capricious behavior during the various stages of my research. I thank him with all my heart.

I would also like to thank Robert Hasegawa for his co-supervision on a CIRMMT research project on which I collaborated with Jason Noble. At this point, I have to acknowledge the Schulich School of Music and CIRMMT for student scholarships, student awards, and travel grants over the years. I would also like to thank the school's administrative unit, and particularly Hélène Drouin and Lena Weman, for addressing formal matters concerning my doctoral degree.

Much love and many thanks to my friends, and to my sister Terpsichori Kazazi, for their continual support and company, which helped me to endure during rough times, as well as to celebrate success: Katerina Momitsa, my "younger siblings" Theodora and Katerina Kokore, Svetlana Jovanovich, Theodore Papageorgiou, and Myrto Kalouptsidi.

This thesis is dedicated to my dear parents and to my beloved nona, to whom I'll be forever indebted. I have no words to express my love, gratitude and respect toward them.

## Contribution of authors

This is a *manuscript*-based thesis. Its core chapters are formatted for publication in scientific journals and have their own reference list.

- Chapter 2: Kazazis, S., Depalle, P. and McAdams, S. (in preparation). Ordinal scaling of timbre-related audio descriptors: Spectral envelope features. *Journal of the Acoustical Society of America*.
- Chapter 3: Kazazis, S., Depalle, P. and McAdams, S. (in preparation). Ordinal scaling of timbre-related audio descriptors: Amplitude-envelope features and inharmonicity. *Journal of the Acoustical Society of America*.
- Chapter 4: Kazazis, S., Depalle, P. and McAdams, S. (in preparation). Interval and ratio scaling of spectral audio descriptors. *Frontiers in Psychology*.

The Appendices have been published in conference proceedings and have their own reference list.

- Appendix A: Kazazis, S., Esterer, N., Depalle, P. and McAdams, S. (2017). A performance evaluation of the Timbre Toolbox and the MIRtoolbox on calibrated test sounds. In *Proceedings of the 2017 International Symposium on Musical Acoustics (ISMA2017)*, Montreal, Canada, June 18–22, 2017
- Appendix B: Kazazis, S., Depalle, P. and McAdams, S. (2016). Sound morphing by audio descriptors and parameter interpolation. In *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx16)*, Brno, Czech Republic, September 5–9, 2016

Stephen McAdams and Philippe Depalle were the thesis supervisors and their contributions concern all stages of research including the experimental design, discussion of analysis approaches, and interpretation of results. Stephen McAdams is also the director of the laboratory in which all of the research was conducted and financed the research with regard to technical facilities and the remuneration of experimental participants. Nicholas Esterer, a former Master's student in the Music Technology program, contributed in programming parts of the Timbre Toolbox related to Appendix A. My contribution as a principal author involves the conception and design of all experiments, conducting the statistical analyses, and authoring all parts of this thesis.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Most of the past research on timbre psychophysics has focused on determining acoustic correlates of perceptual dimensions derived from multidimensional scaling of dissimilarity ratings, in order to quantify the ways in which we perceive sounds to differ. However, there is only little empirical evidence to date demonstrating that acoustic features derived from correlational analysis causally correspond to psychological dimensions. Most importantly, even for cases in which the causality has been verified, there is almost no research on understanding how the sensation magnitudes of such acoustic features are apprehended, which is the core subject of this dissertation.

This chapter begins with an overview of timbre representation models along with a brief comparison between models that rely on (the parsimonious) audio descriptors against models that rely on modulation spectra. In order to situate the contributions of this thesis in a broader context, Section 1.2 presents some practical applications of timbre models focusing on timbral blend and sound morphing guided by audio descriptors. Section 1.3 gives an introduction to psychophysical scaling and presents some terms and methodologies that will be used in subsequent chapters. Finally, the research aims of this thesis conclude this chapter.

## 1.1 Timbre Representations

The majority of timbre studies postulate that timbre relies on a limited number of orthogonal perceptual dimensions. Listeners make paired comparisons between stimuli by judging their similarity on a numerical scale ranging from “very similar” to “very dissimilar”. The presented sounds are usually equalized in pitch, loudness and perceived duration, so that listeners’ judgments are based exclusively on a set of timbre attributes. The (hidden) perceptual structure of the dissimilarity data is most often revealed through multidimensional scaling techniques (MDS). MDS achieves that by finding a configuration of the data points in a low-dimensional space, such that the distances between points in the low-dimensional

space adequately represent the ratings between stimuli in the full-dimensional space. As a result, perceptually similar sounds are spaced close together and dissimilar sounds are spaced farther apart. The space defined from the orthogonal MDS axes is commonly referred to as a *timbre space*, the first visualization of which was given by Plomp (1970). The final step of the analysis is the psychophysical interpretation of the MDS dimensions, which in early studies used to rely heavily on the researcher's intuition. As the number of MDS dimensions increases, the model will represent listeners' ratings with higher fidelity, but a psychological interpretation of the axes becomes more difficult, and the resultant space is also harder to visualize. On the other hand, as the number of dimensions decreases, the possibility of having *metameric matches* between sounds increases, because there can be an infinite number of different sounds (at least in theory) that map to the same location in the timbre space.

### 1.1.1 Acoustic correlates of timbre spaces

In the early timbre studies, the MDS axes were *qualitatively* interpreted according to rather general acoustical descriptions of the stimuli such as, the number of harmonics present, the overall shape of the spectral and amplitude envelope, or the onset asynchrony of the partials (Grey, 1977; Miller & Carterette, 1975). Grey and Gordon (1978) were the first to *quantitatively* interpret perceptual dimensions, by inspecting the correlations of the points on the MDS axes with a set of algebraic models derived after performing an acoustic analysis on the stimuli. The model that correlated most strongly with the dimension that was (qualitatively) associated with the global energy distribution of the partials in Grey's (1977) study, was the spectral centroid (also known as the spectral center of gravity), which corresponds to timbral brightness and is a weighted average of the partials' energies. In addition, when they performed a new MDS analysis on the dissimilarity ratings of a spectrally modified stimulus set with respect to Grey's original set, they observed that pairs of synthesized sounds that had exchanged spectral envelopes also exchanged orders on the MDS axis that correlated most strongly with spectral centroid.

The approach of Grey and Gordon (1978) for quantitatively interpreting perceptual dimensions according to a set of acoustic parameters was extended and systematized by Krimphoff, McAdams, and Winsberg (1994) based on Krumhansl's (1989) three-dimensional timbre space. Krimphoff et al. examined the correlations of each axis of that timbre space with a set of unidimensional acoustic parameters extracted from the stimuli. The dimensions that were referred to as "Spectral Envelope" and "Temporal Envelope" in Krumhansl's study, correlated strongly ($r = 0.94$) with spectral centroid and log-attack time respectively, the latter parameter being useful for distinguishing between impulsive (e.g., struck, plucked) and continuant instruments (e.g., bowed, blown). Although subsequent studies employing different MDS models and stimulus sets confirmed the importance of these two acoustic parameters, there seems to be no consensus about the acoustic correlate of the third MDS dimension, which in most cases also exhibits weaker correlations than the other

two parameters. Throughout the various studies, the third dimension has been associated with: the harmonic odd-to-even ratio, and spectral deviation (also known as spectral irregularity), which is a measure of the spectral envelope's jaggedness Krimphoff et al. (1994); spectral flux (also known as spectral variation), which measures the spectral fluctuation over a tone's duration McAdams, Winsberg, Donnadieu, Soete, and Krimphoff (1995); harmonic odd-to-even ratio, and to a lesser extent, spectral flux Caclin, McAdams, Smith, and Winsberg (2005). In addition, probably due to no significant correlations with any of the above-mentioned parameters, some researchers would step back to qualitative interpretations such as in Lakatos's (2000) study, in which the generic term "timbral richness" was used to characterize the third dimension.

An exception related to the robustness of attack time in explaining dissimilarity ratings is the study of Kendall, Carterette, and Hajda (1999), who explored the timbral similarities between physical instrument tones and emulations of them generated by frequency modulation (FM) synthesis, sampling, and a mixture of the two. Similar to McAdams et al.'s (1995) study, two of the dimensions of the three-dimensional MDS space were correlated with spectral centroid and spectral variation. Interestingly, the third dimension did not correlate with log-attack time, probably due to the homogeneity of the synthetic stimuli with respect to that parameter. Instead, the authors observed that the emulated sounds did not match the time-varying characteristics of the authentic sounds, such as spectral flux or amplitude modulation during the steady state and therefore, this dimension was qualitatively associated with spectrotemporal parameters. Nonetheless, Kendall et al. (1999) suggested that in their study, attack time was a perceptual cue used by listeners for distinguishing the real instruments from their emulations. A similar observation was made by Iverson and Krumhansl (1993), who argued that attack time may not be as important for dissimilarity judgments as it is for instrument identification, or near categorical separation between impulsive and continuant instruments.

The incongruence of the reported results in relation to the acoustical interpretation of a third perceptual dimension, may indicate the possibility that psychological dimensions result from a combination of acoustic parameters rather than a single parameter. However, currently there is no research that has systematically investigated the possible ways and conditions under which such parameters collapse onto single perceptual dimensions. Another explanation could be that the MDS solution depends on stimulus context, which is different from study to study. However, there is some research that counteracts this hypothesis. For instance, McAdams and Giordano (2006), and McAdams (2015) observed that the distances between sounds in a particular timbre space remain invariant in the presence of new stimulus sets, which suggests that timbre relations do not depend on stimulus context. This invariance could be partly explained from the fact that timbre perception is strongly linked to the identification and categorization of sound sources, as pointed out from the previous discussion.

### 1.1.2 Audio Descriptors

Since Krimphoff et al.'s (1994) study, the set of acoustic parameters started expanding with the appearance of the MPEG-7 standard, according to which such parameters would be termed "audio descriptors" (ISO/IEC, 2002), and which in the field of music information retrieval (MIR) fall under the umbrella term "features". These two terms will be used interchangeably in the present manuscript.

Some descriptors are computed on the temporal energy envelope of the signal and are therefore called "global". This family of descriptors includes attack and decay time, temporal centroid, attack slope, and the frequency and amplitude of the modulation of the temporal envelope. Another set of audio descriptors is known as "time-varying" because they are computed on analysis frames derived from the short-term Fourier transform (STFT). This set of descriptors includes spectral centroid, spread, skewness and kurtosis, spectral slope, harmonic spectral deviation, harmonic odd-to-even ratio, and spectral flux, which is the only spectrotemporal descriptor. In general, global descriptors capture some aspects of the evolution of the time-domain waveform, whereas time-varying descriptors capture some aspects of the spectral envelope. The values of the time-varying descriptors are often further compressed to a single number by computing summary statistics over the analysis frames (e.g., mean, median, and interquartile range). In fact, it is the summary statistics that are used when computing the correlations of MDS axes with audio descriptors.

A large set of descriptors is offered in MATLAB toolboxes such as the Timbre Toolbox (Peeters, Giordano, Susini, Misdariis, & McAdams, 2011), the MIRtoolbox (Lartillot & Toiviainen, 2007) and more recently in MATLAB's native Audio Toolbox (although the set of descriptors is limited compared to the other two toolboxes). However, it is customary with most audio analysis tools, the user has to specify a set of analysis parameters, which may have an impact on the accuracy of computed descriptors with respect to the type of sound being analyzed. Such specifications may be related to the parameters used for computing the amplitude envelope before extracting the related descriptors, the shape and length of the analysis STFT window, whether spectral descriptors will be computed on the magnitude or on the power values of the FFT bins, or related to the parameters that will be used for extracting harmonics from the raw spectrogram values, in order to derive descriptors that relate only to the harmonic content of the signal. A study on how such analysis parameters affect the accuracy of descriptors is given in Appendix A, which evaluates the performance of the Timbre Toolbox and the MIRtoolbox when using their default settings for extracting descriptors. In addition, Nymoen, Danielsen, and London (2017) evaluated the calculation of some global descriptors in the Timbre Toolbox and the MIRtoolbox, and proposed parameter settings that lead to descriptor values that are in closer agreement with empirical results.

### 1.1.3 Audio features extracted from modulation power spectra (MPS)

In contrast to the approaches presented in section 1.1.1, a different class of timbre representations favors the notion that timbre emerges from an indivisible high-dimensional structure characterized by modulation power spectra (MPS). MPS is derived from a two-dimensional Fourier transform of the spectrogram, and unifies the spectral and temporal domains through a *scale-rate* representation from which spectrotemporal features may be extracted. The temporal modulations, also called *rates*, indicate the amount of amplitude periodicity within each channel, whereas the spectral modulations, also called *scales*, represent the amount of spectral periodicity, which conceptually can be thought of as a measure of spectral density.

Patil, Pressnitzer, Shamma, and Elhilali (2011) showed that the MPS can be efficiently used as a data source in machine learning tasks for musical instrument identification regardless of pitch and playing style. Most importantly, the machine-learning model was able to reproduce human dissimilarity judgments, which may suggest that listeners rely on similar features for instrument identification. The MPS representation was averaged over time and initially consisted of 128 frequencies, 22 rates, and 11 scales that resulted in a total of 30,976 features. The dimensionality of the features was greatly reduced through singular value decomposition (SVD) to 420 features with 21 eigen-frequencies, 4 eigen-rates, and 5 eigen-scales. The correlation between the human dissimilarity matrix and the one generated from the machine-learning model was $r = 0.94$, whereas the strongest correlations between audio descriptors and a two-dimensional MDS space were $r = 0.97$ for log-attack time and $r = 0.62$ for spectral centroid.

Hemery and Aucouturier (2015) examined in depth different possible ways to process MPS representations and their ability to compute perceptual distances between pairs of environmental sounds. The results showed that processing the data as time-series is no more effective than models that rely on summary statistics along time, but processing the data in series that are organized along frequency, scale, or rate, does give better results than processing their summary statistics. Contrary to the study of Patil et al. (2011), there were no systematic differences between the processing of the scale-rate representation, and the processing of just the frequency dimension. However, as also noted by the authors, most environmental sounds are stationary and therefore do not exhibit strong spectrotemporal modulations as acoustic instruments do, and for which a scale-rate representation could be more appropriate.

Elliott, Hamilton, and Theunissen (2013) suggested that combining the results from an acoustic analysis based on MPS with an analysis based on audio descriptors offers complementary insights for acoustically interpreting perceptual dimensions. In their study, a five-dimensional MDS space was derived from dissimilarity judgments between tones from physical instruments. Four out of five dimensions were significantly correlated with MPS features. The dimensionality of the MPS was first reduced through principal component analysis (PCA). Similar results were obtained from a second regression analysis that used as

predictors the attack time, spectral centroid, temporal centroid, spectral spread, skewness, kurtosis, and entropy. The complementary viewpoint resulting from these two different analyses is supported from the fact that the fifth dimension was only correlated with MPS features (albeit weakly) but not with the classical descriptors, whereas the third dimension was purely spectral and only correlated with spectral centroid and spectral spread. However, from a statistical point of view, both analyses led to very similar results: the MPS features accounted for 73%, 59%, 60%, and 10%, of the variance along the first, second, fourth, and fifth MDS dimensions, respectively, whereas audio descriptors accounted for 70%, 57%, 40%, and 22%, of the variance along the first, second, third, and fourth MDS dimensions, respectively.

From the above discussion, it can be concluded that features derived from MPS have similar explanatory power along perceptual dimensions to the classical descriptors. In addition, listeners' dissimilarity ratings are hard to interpret when modeled through MPS features that have undergone a data-reduction step, especially because the transformed features (e.g., the eigen-rates and eigen-scales used in Patil et al., 2011) no longer reflect the physical aspects of the initial feature space.

## 1.2 Timbral blend and feature-based sound synthesis

Timbre spaces along with audio descriptors have been widely used for explaining the amount of timbral blend between sound sources, which is of primary importance in several orchestration treatises (Lembke, 2014), as well as in computer aided orchestration environments in order to create (composite) instrumental sounds that exhibit a particular spectromorphology (Carpentier & Bresson, 2010). In addition, audio descriptors have been used to indirectly control sound synthesis parameters related to audio morphing, or more generally for synthesizing a (target) sound that exhibits spectrotemporal characteristics that match a set of descriptor values, as well as for synthesizing realistic sound textures.

### 1.2.1 Timbral blend

Kendall and Carterette (1993) found that sounds in close proximity within a two-dimensional timbre space blend better than sounds that are spaced farther apart, which indicates that the amount of perceived blend is proportional to the distances between the constituent sounds located within a timbre space. Although they did not offer a quantitative interpretation of the MDS dimensions, they made some general observations related to the spectrotemporal contrasts between the lower and the upper partials of each sound.

Sandell (1995) attempted to identify which acoustic parameters contribute to ratings of blend between dyads of instrument sounds when played in unison, and when separated by a minor third. The audio descriptor that correlated the most with the ratings was the spectral centroid: blend increased when the composite spectral centroid was low, or when the difference between the centroids (differential centroid) of the two sounds was small.

More specifically, when sounds were played in unison, the composite spectral centroid, followed by attack time, and loudness envelope accounted for 51% of the variance. When the sounds were separated by a minor third, the composite centroid, differential centroid, attack time, and offset synchrony, accounted for 63% of the variance.

Tardieu and McAdams (2012) attempted to identify factors that influence the ratings of blend between dyads consisting of a continuant and an impulsive tone. The results were similar to those reported by Kendall and Carterette (1993): sounds with low spectral centroids, and thus darker timbres, along with long attack times, lead to higher degrees of blend. The contribution of the acoustic characteristics of the impulsive sound to the perceived blend was greater than the characteristics of the continuant sounds. However, in a second experiment where they examined factors that influence the dissimilarity ratings among dyads, they found that that the contribution of the continuant sound was greater than that of the impulsive sound. The first MDS dimension correlated with the attack time of the dyad. Interestingly, the second dimension was occupied by two different dyad-clusters, one of which correlated with spectral spread, and the other one with spectral flatness.

Lembke and McAdams (2015) using a step-wise regression model found that the formants of the spectral envelope are stronger predictors than the spectral descriptors for explaining the amount of timbral blend between wind instruments. The regression model explained about 87% of the total variance in listeners' blend ratings with the strongest predictor being the upper bound of a formant. The other two predictors were the differential spectral centroid of the dyad and a binary contrast factor related to the energy between the lower and upper formants. However, the contribution of the last two predictors in the regression model was about five times less than the contribution of the first predictor.

### 1.2.2 Feature-based sound synthesis

Hoffman and Cook (2006) proposed a general framework for feature-based synthesis according to an optimization scheme that maps synthesis parameters to target feature values. The results are very preliminary: the source sound consists of stationary sinusoids, and white noise that is spectrally shaped through mel-frequency cepstral coefficients (MFCC); the target features are limited to spectral centroid, spectral roll-off, and fundamental frequency histograms. Park, Biguenet, Li, Conner, and Travis (2007) treat single features as modulation signals that are applied to a source harmonic sound. The feature set includes the overall shape of the amplitude envelope, spectral centroid, spectral spread, spectral flux, and inharmonicity. According to their proposed synthesis scheme, the imposed constraints can only control one feature at a time and therefore, the combination of multiple target features leads to unpredictable results. Furthermore, treating the residual part of the signal (i.e., the noisy part) is left for future work.

Caetano and Rodet (2010a) investigate spectral envelope representations, which lead to linearly varying values of audio descriptors when linearly interpolated according to a

morphing factor. In a subsequent study (Caetano & Rodet, 2010b), the authors use optimization techniques based on genetic algorithms, in order to obtain morphed spectral envelopes that approximate target audio descriptor values. Olivero, Depalle, Torrésani, and Kronland-Martinet (2012) propose a sound morphing technique that relies on the interpolation of Gabor masks (i.e., time-frequency filters), and in which the imposed constraints force the morphing intermediates to exhibit a predesigned temporal sequence of centroids. In relation to the above-mentioned studies, Appendix B presents a morphing strategy that relies on the interpolation of synthesis parameters related to the signal model and the independent control of several audio descriptors.

Audio descriptors have also proven to be useful in producing compelling and realistic sound textures. McDermott, Schlemitsch, and Simoncelli (2013) were able to resynthesize realistic textures using the time averaged statistics of centroid, spread, skewness, kurtosis, as well as cross-band correlations computed from the envelopes of cochlear-like filterbank responses extracted from real textures. Schwarz and O'Leary (2015) used granular synthesis for extending the duration of environmental sound texture recordings, by controlling the perceptual similarity between successive grains according to metrics based either on differences between audio descriptor values or MFCCs. The set of descriptors included the loudness of the grain, fundamental frequency, spectral centroid, spectral spread, and spectral slope. Formal listening tests indicated that grain concatenation according to audio descriptor values rendered more natural sounding textures than the textures synthesized by concatenating grains according to MFCC distances.

## 1.3 Psychophysical scaling

The most common measurement scales are the ordinal scale, which indicates whether listeners are able to rank order the stimuli; the interval scale, which indicates whether they can judge the relative size of intervals between stimuli; and, the ratio scale, which indicates whether ratios between stimuli can be perceived, e.g., whether a given interval is perceived as being twice the size of another interval (Stevens, 1946). According to Luce and Krumhansl (1988), psychophysics can be classified into two broad categories. The first category is *local psychophysics*, in which the focus is on stimulus changes that are small enough to cause confusion among stimuli. For this reason, the methods used to construct perceptual scales from experiments that belong to this category are referred to as *confusion scaling* methods (Gescheider, 1997). The scales derived from such methods are indirectly constructed through the discrimination responses of the observer on stimuli that are close in magnitude and result in *interval measurements*, because the observer has to indicate the differences rather than the ratios of perceived magnitudes. Such scales are useful when it is more convenient to specify the differences of stimulus intensities in a number of discriminable steps through just noticeable differences (JND), rather than using the actual stimulus intensity units. On the other hand, the main drawback of scales constructed with confusion

scaling methods is that the constructed intervals capture only 'local' effects, because the data are derived from the discrimination of neighboring stimuli, whereas the stimuli in the real world are widely distributed. Nonetheless, the methods related to local psychophysics are particularly useful when one wishes to study a class of psychoacoustic phenomena such as those related to detection, discrimination, masking effects, and time-intensity tradeoffs.

The second category is *global psychophysics* in which the focus is toward understanding listeners' responses on stimulus changes over the full dynamic range of the signal, and not on stimulus changes that are small enough and which therefore cannot be easily detected (Luce & Krumhansl, 1988). The measurements from global psychophysics experiments may lead to ordinal, interval or ratio scales, depending on both the experimental method and the stimulus properties. Since audio descriptors capture global properties (rather than the fine structure) of the spectral and temporal envelopes of a sound event, in order to establish psychophysical correspondences between perception and several timbre-related audio descriptors, this thesis has only considered experimental procedures that are part of global psychophysics, the basic methods of which are presented in the next two subsections.

### 1.3.1 Partition scaling methods

In *partition scaling* methods the main objective is the partitioning of the psychological continuum into equal sensory intervals, which leads to either interval or ordinal scale measurements. There are two main approaches used for constructing such scales, one based on *equisection scaling* and the other on *category scaling* (Gescheider, 1997). In equisection scaling, the main task of the observer is to report whether the sensory distance between a pair of stimuli is equal to, greater than, or less than the distance between a different pair. The observer is presented with the upper and lower limits of a physical continuum and is instructed to choose a number of stimuli in order to create a prescribed number of equidistant sensory steps between those limits. This technique is referred to as the *simultaneous solution* because the observer has to estimate all the scale values at once. Another technique is the *progressive solution*, in which the observer is presented with the highest and lowest stimulus values and is instructed to bisect the given sensory distance by choosing a single stimulus. The process is progressively repeated using as the highest or lowest limits of the previously chosen stimulus and terminates when the desired number of successive equal sensory intervals is reached.

In *category scaling*, the observer's task is to distribute a set of stimuli in a number of specified categories. The final scale is derived by treating the assigned category values as interval values, under the basic assumption that the observer is able to keep the intervals between category boundaries equal during the assignment of stimuli to each category. The main issue with this approach is that observers tend to make equal use of all the categories, which in the best-case scenario could lead to an ordinal scale. Nonetheless, if the stimulus set is perceptually uniform, the chances of deriving interval measurements are increased (Gescheider, 1997).

### 1.3.2 Direct estimation methods

This set of methods is known as *direct*, because the observer makes direct estimations of sensation magnitudes on presented stimuli. This set includes the methods of *ratio estimation*, *ratio production* and generalizations of them known as *magnitude estimation* and *magnitude production* (Gescheider, 1997). In ratio production, the task of the observer is to adjust according to a prescribed ratio the intensity of a variable stimulus with respect to a reference stimulus (also known as the *standard stimulus*). The most apprehensible ratio, and the one most often used, is the 2-to-1 along with its complement, which serves as a validity check on the observer's judgments. Any biasing effects due to intensity's upward or downward adjustments (also known as *hysteresis* effects) can be lessened by averaging the two scales derived from each direction. In ratio estimation, the observers do not make any adjustments on the stimuli, but instead are asked to estimate their apparent ratios. One of the two methods can also be used to validate the other.

The methods of magnitude estimation and magnitude production are generalizations of the ratio estimation and ratio production methods, in which observers make direct numerical estimations of the sensory magnitudes according to a given numerical value (also known as modulus) of a reference stimulus, or by arbitrarily choosing their own reference value for the reference stimulus. In another variant, called *absolute magnitude estimation* (Hellman & Zwislocki, 1961), the observers match numbers to stimuli without the presentation of the standard, and for each new trial, they are instructed to do so independently of their previous matches. The final scale can be computed from the data of the subjects using either the geometric mean or the median since the arithmetic mean may distort the scale in the presence of a few unrepresentative and extremely high judgments. As in the previous case, one method can be used to test the validity of the other, but some times these methods lead to small but systematic differences in the psychometric functions (Gescheider, 1997). Observers are often reluctant to report extremely low or high judgments although their perceptions may be correct and thus, a psychometric function derived from magnitude production will have a steeper curve than the function derived from magnitude estimation. The results from both methods are usually combined under the assumption that the unbiased function lies somewhere in between.

### 1.3.3 Scale types and practical considerations

The type of scale is determined by the numerical rules that best model the empirical data, and which therefore provide the most information about the invariance of scale values across a variety of conditions (Baird & Noma, 1978). In other words, the scale invariance across different trials ensures that the *pattern* of response magnitudes remains the same.

In an *ordinal scale*, the order of scale values is constant over trials, but the ratios and interval sizes may change from trial to trial. This scale can be modeled according to: $x' = monf(x)$, where the values in trial $x'$ are a monotonic transformation of the values in

$x$. For instance, if $u > v$ in trial $x$, then $u' > v'$, in trial $x'$, because $f(x)$ is a transformation that maintains the order of $u$ and $v$.

In an *interval scale* the relative size of intervals is retained over trials, or in other words, the ratio of two intervals remains the same. It is modeled according to the affine transformation: $x' = ax + b$, where $a$ is a positive multiplicative constant, and $b$ is an additive constant. The interval scale also satisfies the conditions for the ordinal scale.

The *ratio scale* is the only type of scale in which the concept of "times as much" has a meaning and is modeled according to: $x' = ax$, where $a > 0$. One of its distinctive characteristics, and what differentiates it from the interval scale, is the existence of the so-called absolute zero, which means that a stimulus attribute will always remain zero throughout different trials regardless of the multiplicative constant. The ratio scale satisfies all the criteria for the interval, and ordinal scales, and is therefore the most informative type of scale. Furthermore, the operation of addition is invariant under all possible representations within the scale only for ratio scale measurements. For instance, the transformation $f$ from one scale value to another must be increasing and satisfy the constraint $f(u + v) = f(u) + f(v)$, for $u, v > 0$, which is true for the ratio scale but not for the interval scale (Luce & Krumhansl, 1988).

Another type of scale is the *log-interval scale*, which remains invariant up to a power transformation: $\psi = k\phi^{\beta}$, where $k$ is a constant which depends on the units of measurement, and the exponent $\beta$ is an index of perceptual sensitivity (Baird & Noma, 1978). This type of scale has been extensively studied by S. S. Stevens (Stevens, 1975) and is also referred to as *Stevens' power law*. By taking the logarithms, the above equation becomes: $log\psi = logk + \beta log\phi$, and when plotted in log-log coordinates it describes a straight line, where the exponent $\beta$ becomes the slope of the line. The practical utility of this function can be summarized into the following sentence (Stevens, 1975): *Equal stimulus ratios produce equal subjective ratios*. In practice, however, there are cases in which the simple form of the power law does not always accurately describe the data either because the zero of the subjective scale $\psi$ does not coincide with the absolute threshold on the stimulus scale $\phi$ (Stevens, 1975), or simply because the observers' responses depart from linearity (in log-terms) near the low end of the scale (Baird & Noma, 1978). In such cases, the fit of the power function on the data can be improved by subtracting (or adding) a constant ($\phi_0$) from the stimulus values, in which case the power function becomes $\psi = k(\phi – \phi_0)^{\beta}$.

In the previous subsections, it was mentioned that partition scaling methods lead to interval scale measurements whereas direct estimation methods lead to ratio scale measurements. The power law can be used to model both these two types of responses although the exponent derived from partition scaling will in most cases be lower than the exponent found from direct estimation methods (Baird & Noma, 1978; Stevens, 1975). These observations were made after comparing the results of the two methods on a class of continua that S. S. Stevens called prothetic, because he assumed that discrimination along such continua is mediated by additive processes at the physiological level (e.g., loudness). A different

class of continua is what S. S. Stevens called methathetic, because he assumed that for these continua, the discrimination is mediated by substitutive processes at the physiological level (e.g., pitch: the progression along the continuum is achieved by changing the locus of excitation along the basilar membrane). Experiments on this type of continua have indicated that the exponents from partition scaling and direct estimation methods almost coincide, meaning that the results of these two experimental methods lead to almost identical perceptual scales (Stevens, 1975).

## 1.4 Research Aims

This thesis investigates whether, how, and to what extent listeners perceive magnitude differences along acoustic features, some of which have been arbitrarily used in the field of music information retrieval (Siedenburg, Fujinaga, & McAdams, 2016) and have been appraised as physical correlates of perceptual dimensions in several timbre studies (Section 1.1.1). Besides the topics presented in this chapter, audio descriptors have also been widely used for investigating the role of timbre in areas such as timbre semantics (Zacharakis, Pastiadis, & Reiss, 2015), quantifying affective responses to sound (Farbood & Price, 2017; McAdams, Douglas, & Vempala, 2017), investigating cognitive factors related to memory for timbre (Siedenburg, 2016), and psychomechanics (McAdams, Roussarie, Chaigne, & Giordano, 2010). However, audio descriptors identified through correlational analyses do not necessarily causally relate to listeners' perceptions and when entered into statistical regression models may lead to false positive interpretations about their perceptual significance, especially when features also strongly covary.

The aim of this thesis is to test whether listeners perceive audio descriptors on perceptual ordinal, interval, and ratio scales. Chapter 2 presents an experiment that tested the ordinal scalability of the following spectral descriptors: spectral centroid, spectral spread, spectral skewness, harmonic odd-to-even ratio, spectral deviation, and spectral slope.

Chapter 3 presents a similar experiment that tested the ordinal scalability of temporal descriptors, and the extent to which their scalability depends on amplitude envelope features and spectral envelope features. The following descriptors were tested: attack time, decay time, temporal centroid with fixed attack or decay time, and inharmonicity. In addition, with respect to the results of the ordinal scaling experiment, a meta-analysis on previously reported timbre spaces was conducted in order to test the hypothesis that variation along a spectrotemporal perceptual dimension may be explained by a combination of acoustically independent descriptors.

Chapter 4 presents two experiments that provided interval scale, and ratio scale measurements, respectively, of all the spectral descriptors tested in the second chapter. In addition, it presents the psychophysical ratio scales of each descriptor that were constructed after taking into account the absolute zero of the stimulus scale, and by extrapolating the subjective scale of the ratio measurements outside the tested range.

Chapter 5 concludes the dissertation by summarizing the main results and the experimental procedures used in each study. It also provides suggestions for future research, and discusses how the findings of this thesis contribute to our understanding of timbre.

# Chapter 2

# Ordinal scaling of timbre-related audio descriptors: Spectral envelope features

This chapter is based on the following research article:

Kazazis, S., Depalle, P. and McAdams, S. (in preparation). Ordinal scaling of timbre-related audio descriptors: Spectral envelope features. Manuscript prepared for submission to *Journal of the Acoustical Society of America*.

**Abstract** A psychophysical experiment was conducted to perceptually validate several spectral audio features through ordinal scaling: spectral centroid, spectral spread, spectral skewness, odd-to-even harmonic ratio, spectral slope, and harmonic spectral deviation. Several sets of stimuli per audio feature were synthesized at different fundamental frequencies and spectral centroids by controlling (wherever possible) each spectral feature independently of the others, thus isolating the effect that each feature had on the stimulus rankings within each sound set. Listeners were overall able to order stimuli varying along all the spectral features tested when presented with an appropriate spacing of feature values. For specific cases of stimuli in which the ordering task partially failed, we provide psychophysical interpretations to explain listeners' confusions. The results of the ordinal scaling experiment outline trajectories of spectral features that correspond to listeners' perceptions and suggest a number of sound synthesis parameters that could carry contour information.

## 2.1 Introduction

Musical timbre has often been studied by evaluating listeners' dissimilarity ratings between pairs of instrumental sounds (Caclin, McAdams, Smith, & Winsberg, 2005; Grey, 1977; Grey & Gordon, 1978; Iverson & Krumhansl, 1993; Lakatos, 2000; McAdams, Winsberg, Donnadieu, Soete, & Krimphoff, 1995). Researchers then employ a nonlinear dimensionality reduction technique on the data, presuming that timbre relies on a limited bundle of perceptual dimensions, usually ranging from two to three. The techniques most often used are variants of Multidimensional Scaling (MDS), which project the dissimilarity data onto a lower-dimensional "timbre space" by maximizing the fit between average dissimilarities and mathematical distances in the model. A psychophysical meaning is then sought for the MDS dimensions by examining how strongly each axis correlates with a set of audio features, which are most often extracted within time frames from traditional spectrograms, and the time-varying values are further compressed to a scalar value with summary statistics representing central tendency and variability over time. The systematic development of audio features for quantitatively interpreting perceptual dimensions started with the work of Grey and Gordon (1978), and this approach was extended and systematized by Krimphoff, McAdams, and Winsberg (1994) who conducted acoustic analyses of Krumhansl's (1989) sound set and examined the correlations of various features with each axis of her three-dimensional timbre space. The feature set started expanding with the appearance of the MPEG-7 standard according to which audio features would be termed audio descriptors (ISO/IEC, 2002). These two terms will be used interchangeably in the present paper.

There is only limited empirical evidence to date demonstrating that features derived from correlational analysis, or arbitrarily used for example in music information retrieval, causally correspond to psychological dimensions. On the contrary, recent research indicates that even some well-established features, such as spectral flatness, do not correspond to listeners' perceptions (Agus, Anderson, Chen, Lui, & Herremans, 2018). One exception is the study of Grey and Gordon (1978) who observed that pairs of synthesized sounds that had exchanged spectral envelopes also exchanged orders on the MDS axis which correlated most strongly with the spectral centroid weighted by the loudness function of Zwicker and Scharf's (1965) model. Another exception is the confirmatory study of Caclin et al. (2005), who tested and confirmed with synthesized stimuli the saliency of attack time, spectral centroid, and the odd-to-even harmonic ratio, but not spectral flux (i.e., the change in shape of a spectral envelope over time) for explaining dissimilarity ratings. However, we find that for the experiments of Caclin et al. (2005), the interpretation of one dimension being associated only with spectral centroid could be problematic because the stimuli were actually varying in spectral slope, which in the special case where the spectral components monotonically increase (or decrease) in magnitude is linearly dependent on spectral centroid but also affects the spectral spread and skewness. As such, the authors were directly evaluating the perception of spectral slope, although this descriptor does covary strongly with centroid in these stimuli. In conclusion, the aforementioned timbre studies support

a unidimensional or two-dimensional spectral representation of timbre, along with a third temporal dimension associated with attack time.

However, when the experimental paradigm is switched from judging the dissimilarities between pairs of sounds to discrimination or identification tasks, features other than centroid and attack time become also prevalent. Although Caclin et al. (2005) did not confirm the saliency of spectral flux in dissimilarity judgments (albeit computed over the first 100 ms), McAdams, Beauchamp, and Meneguzzi (1999) had previously identified it as one of the most salient spectrotemporal features when listeners were asked to discriminate instrument sounds resynthesized with the full parameters of an additive synthesis model from sounds resynthesized with simplified synthesis parameters (in this particular case, spectral flux was eliminated in the simplified versions). In addition, the discrimination was also very good between sounds that exhibited a certain amount of harmonic spectral deviation and the simplified versions, in which the spectral deviation was minimized (for each time frame each harmonic amplitude was replaced by the average of itself and its two harmonic neighbors), which indicates that harmonic spectral deviation had been another salient spectral feature for this particular task. Horner, Beauchamp, and So (2011) used additive synthesis for synthesizing sounds with matched spectral centroids, and reported that the relative amplitudes of the first five harmonics could account for 85% of the variance in predicting spectral discrimination performance. Nonetheless, spectral centroid also plays a role in discrimination and identification tasks. Wun, Horner, and Wu (2014) used synthesized instrument sounds based again on additive synthesis, and showed that discrimination levels were above 75% when the spectral centroid was increased by 40% or decreased by 24%, and that instruments started to lose their identity when the spectral centroid was increased by 64% or decreased by 48%. McDermott, Schlemitsch, and Simoncelli (2013) measured the discrimination of various sound textures, which were synthesized according to the time-averaged summary statistics of the respective pre-analyzed textures. Their results indicate that the combination of summary statistics (including the mean, variance, and skewness) greatly accounts for the categorical discrimination among different textures, but limits the ability to discern temporal detail.

Based on the above discussion, it can be concluded that the relative perceptual salience of each audio feature depends on the experimental task (e.g., dissimilarity judgments, discrimination, identification), the range of feature values within a stimulus set (e.g., matched spectral centroid, minimum spectral deviation, etc.), and whether clear categorical boundaries exist or not (e.g., instrument and texture categories). The aim of this study is to perceptually validate a number of spectral descriptors by testing whether, and in some cases the extent to which, independently controlled features can each be perceived on an ordinal scale. The following descriptors related to spectral shape were tested: spectral centroid, spread, skewness, odd-to-even harmonic ratio, spectral slope, and harmonic spectral deviation, which measures the average amplitude deviation of harmonics from a global spectral envelope smoothed by the average amplitude of three consecutive harmonics. The

Tristimulus values (T1 being the normalized amplitude of the fundamental frequency, T2 the mean normalized amplitude of the second, third and fourth harmonics, and T3 the mean normalized amplitude of all the upper partials) were not tested, but the T2 value was used as a criterion for constructing stimuli for spectral deviation. A mathematical formulation of these descriptors can be found in Peeters, Giordano, Susini, Misdariis, and McAdams (2011). The above descriptors were tested in an experiment that consisted of two sessions, and for clarity of presentation the results of each session are presented in different sections. Discussion for both sessions is included in Section 2.4.

## 2.2 Experimental Session A

### 2.2.1 Method

#### Participants

Twenty-five participants, 10 female and 15 male, with a median age of 23 years (range: 18–40) were recruited from the Schulich School of Music, McGill University. All of them were self-reported amateur or professional musicians with formal training in various disciplines such as performance, composition, music theory, and sound engineering. Participants who were not affiliated with the authors' lab were compensated with 10 Canadian dollars.

#### Stimuli

Several sound sets consisting of synthetic sounds were created by independently controlling the values of spectral centroid, spectral spread, and spectral skewness in the synthesis process. All of the spectral manipulations described in this section were applied to a flat harmonic spectrum (harmonics set at equal amplitude) with a fundamental frequency ($f_0$) of 120 Hz and harmonics up to 21,960 Hz (99.6% of the Nyquist frequency). The choice of using an $f_0$ at 120 Hz was based on that fact that it was found to be high enough to construct complex periodic tones without audible roughness that can be induced by the relative phases of the partials, and low enough to provide an adequate resolution of the harmonics when shaping spectra according to probability distributions for achieving specific values of statistical moments (described below).

The stimuli were synthesized in MATLAB version R2015b (The MathWorks, Inc., Natick, MA) using additive synthesis at a sampling frequency of 44.1 kHz with 16-bit amplitude resolution. The peak amplitude of the waveforms was normalized to 0.5 and the duration was set to 600 ms, gated with 10-ms raised-cosine ramps. Due to substantial differences among the spectral envelopes the stimuli were loudness-normalized according to the algorithm of Moore, Glasberg, and Baer (1997) as implemented in the Loudness Toolbox v.1.2 (Genesis S. A., 2009), and further adjusted by the authors who observed that the algorithm overestimated the loudness of sounds that had most of their energy in the higher frequen-

cies, an observation in line with the results of Schlittenlacher, Ellermeier, and Hashimoto (2015).

**Spectral centroid** Two obvious strategies that can be used to modify the spectral centroid of harmonic sounds while preserving their overall harmonic structure and fundamental frequency are: i) preserving the spectral slope and increasing or decreasing the number of harmonics, and ii) preserving the number of harmonics and altering the spectral slope as in Caclin et al. (2005). In the former case, changing the number of harmonics will cause changes in spectral spread whereas in the latter case, changes in amplitude slope will cause changes in spectral spread and skewness. In order to counteract such side effects, which would not allow for the independent control of centroid from spread and skewness, the stimuli were constructed by shaping the flat harmonic spectrum described above to follow a normal probability mass function. The normal distribution is a two-parameter family of curves and as such enables the construction of spectra with different centroids (means) for a given spread (standard deviation, $\sigma$) and zero skewness.

To set a fixed spread throughout this sound set there had to be a compromise between: i) the level of difficulty of the ordering task (smaller spreads would make the task easier as in the extreme case the spectrum would consist of just one harmonic), and ii) the lowest centroid to be included in the sound set, which is constrained by the $f_0$ used in the initially flat spectrum. After these considerations, the spread was set to 480 Hz (four times the $f_0$), which allows for a minimum centroid of 1640 Hz and a minimum bandwidth (or full width at half maximum of the distribution) of nine harmonics for each stimulus spectrum. The number of harmonics $H$ that fall inside the bandwidth of the normal distribution is given by:

$$H = \left\lfloor 2\sigma\sqrt{2\ln 2}/f_0 \right\rfloor \tag{2.1}$$

In total, the stimulus set consisted of 15 centroids at {1640, 1800, 2000, 2280, 2560, 2880, 3240, 3680, 4160, 4760, 5400, 6200, 7120, 8200, 9560} Hz chosen to be centered approximately at one-$\mathrm{ERB_N}$ steps on the $\mathrm{ERB_N}$-number scale (Moore & Glasberg, 1983) {19.31, 20.08, 20.95, 22.05, 23.01, 23.98, 24.95 25.98, 26.96, 28.02, 28.99, 30.01, 31.01, 31.99, 33.00} respectively, as shown in Fig. 2.1. It should be noted that for all stimuli, the harmonic spacing of the components ensured a (virtual) pitch percept at the $f_0$.

**Spectral spread** The normal distribution was again used to construct stimuli with fixed centroids, zero skewness, and variable spreads. The stimuli were constructed based on the rationale that when the spectrum is normalized at unit amplitude, it can be considered as a probability distribution of the harmonic amplitudes, therefore reducing the spread will cause an increase in the relative amplitude of the centroid. More formally, we consider that a particular normalized spectrum $i$ defines the probability distribution of the harmonic

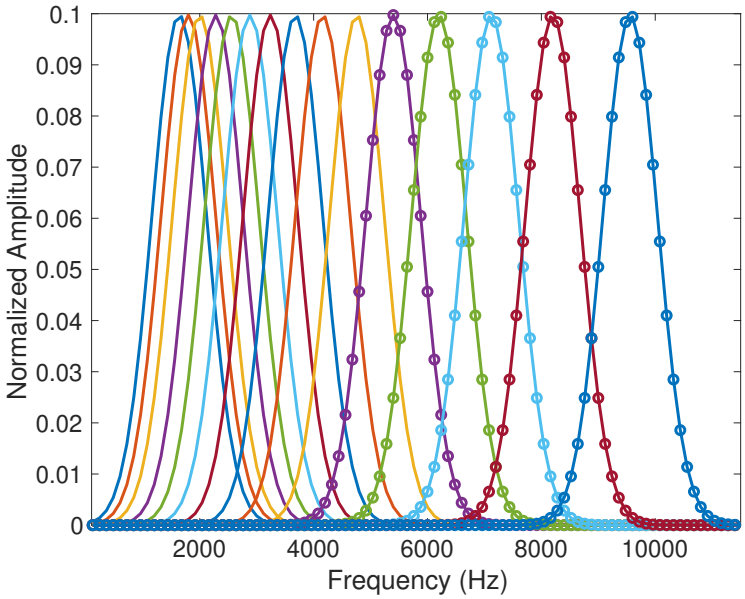


**Fig. 2.1** Spectral envelopes of stimuli in the spectral centroid set. The dots on the last five Gaussian curves indicate the positions of harmonics. The harmonics within the rest of the Gaussian curves are omitted for display purposes.

amplitudes:

$$A_i(f_h) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(f_h - f_c)^2}{2\sigma_i^2}\right) \tag{2.2}$$

where $f_h$ and $f_c$ denote the harmonics and spectral centroid, respectively. We can then compute the desired spread $\sigma_{i+1}$ according to the ratios of (desired) amplitude values and in relation to a given spread $\sigma_i$:

$$\frac{A_{i+1}(f_c)}{A_i(f_c)} = \frac{\sigma_i}{\sigma_{i+1}} \tag{2.3}$$

This approach was constrained by: i) the number of maximum audible successive differences in decibels of the centroid that could be achieved between successive spreads ($\sigma_{i+1}$), ii) the choice of the initial spread ($\sigma_1$), which was further constrained by the centroid and $f_0$, and iii) the spacing resolution of the harmonics.

After taking into account these restrictions, three sound sets were constructed with centroids of 1640, 5600, and 7800 Hz, and initial spreads of 480, 1440, and 1800 Hz, respectively. The choice of the initial spreads allowed for ten stimuli per sound set for which the centroids' relative amplitudes differed by approximately 2 dB in succession. An example of this process is shown in Fig. 2.2 for the 7800 Hz centroid sound set.

**Spectral skewness** The Skew-normal distribution (Azzalini, 2005) is a three-parameter family of curves and was employed for constructing stimuli with different skewness while

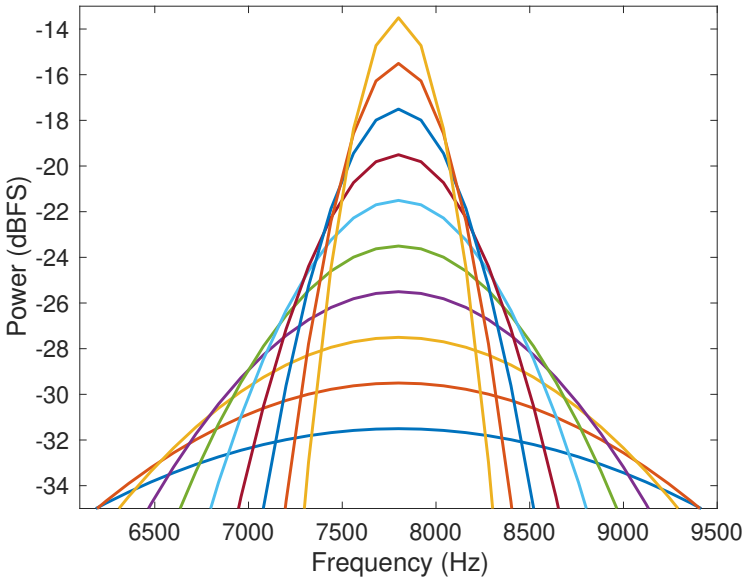


**Fig. 2.2** Spectral envelopes of the stimuli in the spectral spread set at a spectral centroid of 7800 Hz. Line segments connect harmonics to keep the figure readable.

the centroid and spread were being kept constant. The probability density function of the Skew-normal distribution with shape parameter $\alpha \in \mathbb{R}$, scale $\theta \in \mathbb{R}^+$, and location $\xi \in \mathbb{R}$, is given by:

$$f(f_h; \xi, \theta, \alpha) = \frac{2}{\theta}\phi\left(\frac{f_h - \xi}{\theta}\right)\Phi\left(\alpha\left(\frac{f_h - \xi}{\theta}\right)\right), f_h \in \mathbb{R} \tag{2.4}$$

where $\phi$ is the normal probability density function and $\Phi$ its cumulative distribution function. The restrictions that were taken into account with respect to centroids and spreads were similar to the ones mentioned above, with the additional constraint that skewness in the Skew-normal distribution only vary within a range of (–0.9953, 0.9953). For testing both positive and negative skewness separately, three sets of nine stimuli were constructed for each condition with centroids spaced at 1640, 5600 and 7800 Hz and spreads at 360, 1080, and 1440 Hz, respectively. After informal listening tests, the authors concluded that it was easier to distinguish between successive positive skewness value when increased logarithmically, whereas for negative skewness a linear spacing seemed to work better. Given the spacing of harmonics and spreads, it was also noticed that skewness values close to the extremes of |0.995| caused the first or last harmonic (depending on whether the distribution was positively or negatively skewed) to clearly stand out of the harmonic complex, something that might confuse listeners in the ordering task and was therefore avoided. Based on these observations, the following sets of values were used for positive and negative skewness, respectively: {0, 0.2496, 0.4428, 0.5924, 0.7082, 0.7979, 0.8674, 0.9211, 0.9628}, and {0, –0.1106, –0.2211, –0.3317, –0.4422, –0.5528, –0.6633, –0.7739, –0.8844}. Example spec-

tral envelopes with negative skewness and a centroid of 7800 Hz are shown in Fig. 2.3. Note that the values reported above were computed in continuous frequency and were only used as references to construct the stimuli. The measured values, which were computed in discrete frequency and after synthesizing the stimuli, differed slightly due to the spacing resolution of the harmonics and the parameter estimation of the Skew-normal distribution (see Appendix).

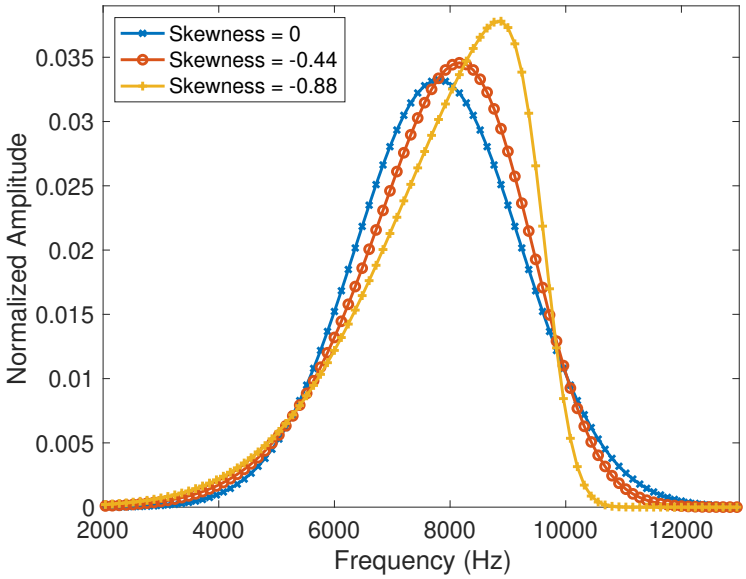


**Fig. 2.3** (Color online) Spectral envelopes with three values of negative skewness (low and high anchors and mid-point value) and a 7800 - Hz spectral centroid. Markers represent harmonics.

## Procedure

Before the experiment, participants signed an informed-consent form. Afterwards, they passed a pure-tone audiometric test at octave-spaced frequencies from 125 Hz to 8 kHz (ISO 389-8, 2004; Martin & Champlin, 2000) and were required to have thresholds at or below 20 dB HL to proceed to the experiment. The instructions that described the task and user interface were presented on paper and were further explained by the experimenter. Any questions had to be asked during the practice trials for which the experimenter was also present in the booth.

In each trial, participants were presented two sounds, which served as anchors and which had the minimum and maximum values of a given audio feature. The task was to order the rest of the sounds in between those anchors according to "any criteria that differentiate them the most." Any verbal labeling of the anchors was intentionally avoided. The stimuli were presented in the form of sound boxes on which participants could click to

hear each stimulus and then drag them to the desired position for the ordering task. The user interface consisted of two main panels. In each trial the top panel contained only two stimuli (i.e., the anchors), which had the minimum and maximum values of a particular audio feature and between which the ordering of the rest of the stimuli would take place. The rest of the stimuli were presented in randomized order in the lower panel. The task was completed when all of the stimuli in the lower panel were dragged and re-arranged according to the desired order in the top panel.

In each trial and for each participant, the stimuli were presented in random order, but the sound sets were presented in the following (fixed) order: 1) practice trial, 2) centroid, 3) spread with centroid at 5600 Hz, 4) positive skewness with centroid at 5600 Hz, 5) negative skewness with centroid at 5600 Hz, 6) negative skewness with centroid at 1640 Hz, 7) positive skewness with centroid at 1640 Hz, 8) negative skewness with centroid at 7800 Hz, 9) positive skewness with centroid at 7800 Hz, 10) spread with centroid at 7800 Hz, 11) spread with centroid at 1640 Hz. The order of presentation was intentionally kept fixed, following an increasing and then decreasing level of difficulty (empirically estimated) so that trials including stimuli that were harder to order would be presented around the middle of the experiment. The practice trial consisted of ordering stimuli that had the same spectral centroids with the ones used in the main experiment, but which had half the amount of spectral spread, and thus exemplified the experimental task by making the ordering easier. This session took approximately 40 minutes to complete.

### Apparatus

The experimental session was run with the PsiExp computer environment (Smith, 1995). Sounds were amplified through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA) and presented diotically over Sennheiser HD600 headphones (Sennheiser Electronic GmbH, Wedemark, Germany). The sound pressure levels had a range of 54.4–74 dB SPL (A-weighted) as measured with a Brüel & Kjær Type 2205 sound-level meter with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark). Listeners were seated individually in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY).

### Data analysis

Because of the ordering task, nonparametric tests were used on participants' stimulus rankings. For each stimulus set, separate Friedman tests were used to evaluate the main effect of each audio descriptor. To account for the nonsphericity present in the data, which can transmit to Friedman ranks (Beasley & Zumbo, 2009), the main effects of each audio feature were also tested with a proportional-odds mixed model (McCullagh, 1980), which had a full random effects structure with random intercepts for each participant and random by-participant slopes for the fixed factor of sound set (Barr, Levy, Scheepers, & Tilly, 2013).

The main effect was evaluated by a likelihood ratio test (in which the maximum likelihood was estimated by the Laplace approximation) between the full model and a reduced model, which had the same random effects structure but excluded the effect of interest from the fixed factors. The main trends of the data were identified through forward stepwise ordinal regression with orthogonal polynomials constructed with the modified Gram-Schmidt algorithm (Hoffman, 1989) on ranked stimulus values. In cases where the dependent variable could be perfectly determined by the predictors (i.e., complete separation: Albert and Anderson (1984)), a linear regression model was used instead. Two-tailed *post hoc* Wilcoxon signed-rank tests were used to examine whether the rank of each stimulus was significantly different from the rest and thus, to identify stimulus combinations that were confused by the listeners. Due to the large number of multiple comparisons within each stimulus set, the *post hoc* tests were corrected with the Holm-Bonferroni method (critical $\alpha = 0.05$)), which controls the family-wise error rate (Holm, 1979). Although the aligned-rank transform (Higgins & Tashtoush, 1994) and its variants allow for nonparametric analyses of variance to test for interaction effects on ranked data, the transform was originally developed for continuous dependent variables and recent studies suggest that it is not appropriate for ordinal responses (Luepsen, 2017). As such, a proportional odds mixed model, which had a nested random effects structure with random intercepts for each participant and the subsets of stimuli nested within each participant, was used to examine the interaction effects between the ranking of the stimuli along a given descriptor and the subsets with different values of a parameter, such as spectral centroid or fundamental frequency, used in each subset of the same descriptor (e.g., the ranking of spectral spread between the sound sets centered at three different centroids). The interactions were examined after fitting the model using sum coding for the predictor variables and performing an ANOVA on the fixed effects (Barr et al., 2013). All the statistical analyses were done in MATLAB. Although analyses are conducted on ranks, in all data graphs, the actual stimulus values are plotted on the $x$ axis, at times appearing concave or convex even though a linear relation may exist between physical ranks and mean response ranks.

### 2.2.2 Results

Both the Friedman and likelihood ratio tests shown in Table 2.1 confirmed the main effect of each audio descriptor. Fig. 2.4 and Fig. 2.5 show the mean rankings for spectral centroid and the three sound sets used for spectral spread for which only a linear trend between the ranked stimulus values and the mean rankings was found to be significant (Table 2.2). Fig. 2.6 shows the ratings for the sound sets of negative and positive skewness. For negative skewness, the sets with centroids at 1640 and 7800 Hz showed a linear trend, whereas for the set at 5600 Hz both linear and quadratic terms significantly described the trend of the data (Table 2.2). For positive skewness with a centroid at 1640 Hz, linear and quadratic terms both significantly described the patterns of the mean rankings of the stimuli, whereas for the other two sets, only a linear term was found to be significant (Table 2.2).

**Table 2.1** Friedman ($\chi^2_F$) and likelihood ratio tests ($\chi^2_{LRT}$) for the stimulus sets of Session A. $sc$ = spectral centroid, $df$ = degrees of freedom; $^*p < 0.001$

| Stimulus sets | $sc$ (in Hz) | $df$ | $\chi^2_F$ | $\chi^2_{LRT}$ |
|---|---|---|---|---|
| centroid | - | 12 | $296.30^*$ | $113.12^*$ |
| spread | 1640 | 7 | $171.13^*$ | $86.28^*$ |
| spread | 5600 | 7 | $172.71^*$ | $86.13^*$ |
| spread | 7800 | 7 | $173.77^*$ | $87.05^*$ |
| neg. skewness | 1640 | 6 | $70.11^*$ | $41.51^*$ |
| neg. skewness | 5600 | 6 | $105.69^*$ | $67.00^*$ |
| neg. skewness | 7800 | 6 | $109.80^*$ | $61.77^*$ |
| pos. skewness | 1640 | 6 | $114.31^*$ | $54.30^*$ |
| pos. skewness | 5600 | 6 | $129.19^*$ | $75.92^*$ |
| pos. skewness | 7800 | 6 | $141.98^*$ | $76.37^*$ |

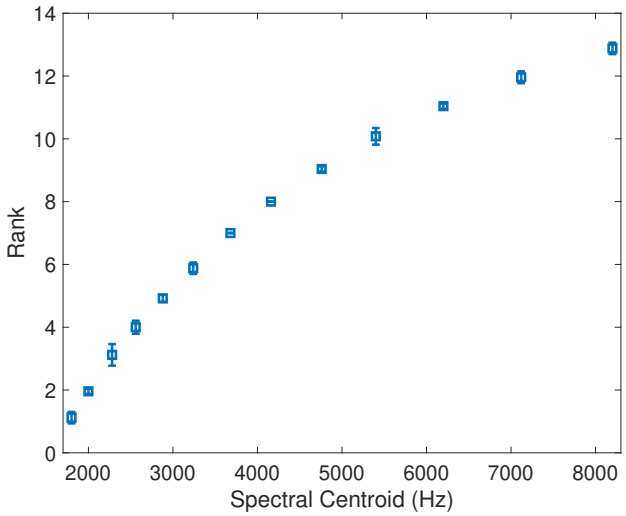


**Fig. 2.4** Spectral centroid: mean rankings of spectral centroid stimuli between the anchors. Error bars represent 95% confidence interval (CI).

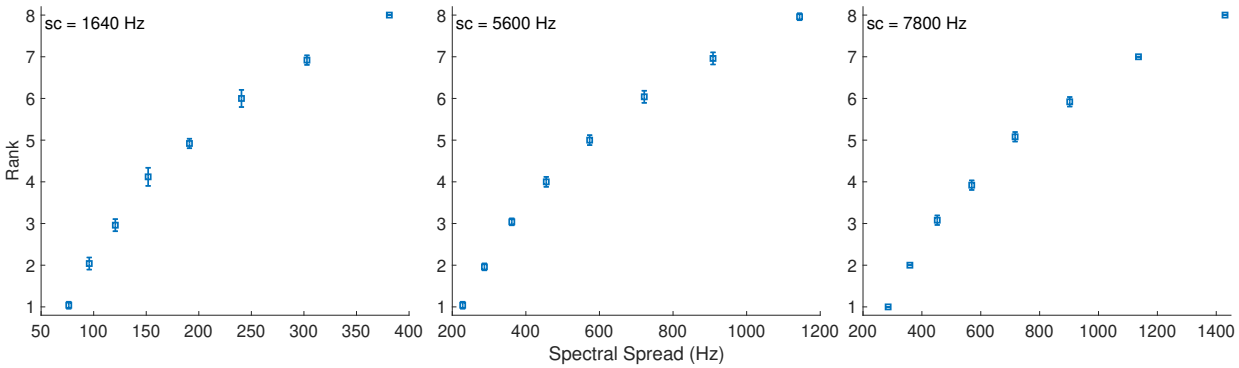


**Fig. 2.5** Spectral spread: mean rankings of spectral spread stimuli between the anchors. sc = spectral centroid. Error bars represent 95% confidence interval (CI).

**Table 2.2** Ordinal regression coefficients for the stimulus sets of Session A. *sc* = spectral centroid; L, Q: linear and quadratic terms respectively.

| Stimulus sets | *sc* (in Hz) | Term | *b* | *t* | *p* < |
|---|---|---|---|---|---|
| centroid | - | L | –535.51 | –12.89 | .001 |
| spread | 1640 | L | –176.55 | –12.64 | .001 |
| spread | 5600 | L | –206.16 | –11.90 | .001 |
| spread | 7800 | L | –243.21 | –10.62 | .001 |
| neg. skewness | 1640 | L | –23.55 | –9.70 | .001 |
| neg. skewness | 5600 | L | –37.93 | –11.27 | .001 |
| | | Q | –11.45 | –3.27 | .010 |
| neg. skewness | 7800 | L | –45.48 | –11.94 | .001 |
| pos. skewness | 1640 | L | –58.53 | –11.97 | .001 |
| | | Q | –13.40 | 3.04 | .010 |
| pos. skewness | 5600 | L | –64.93 | –12.29 | .001 |
| pos. skewness | 7800 | L | –110.54 | –12.30 | .001 |

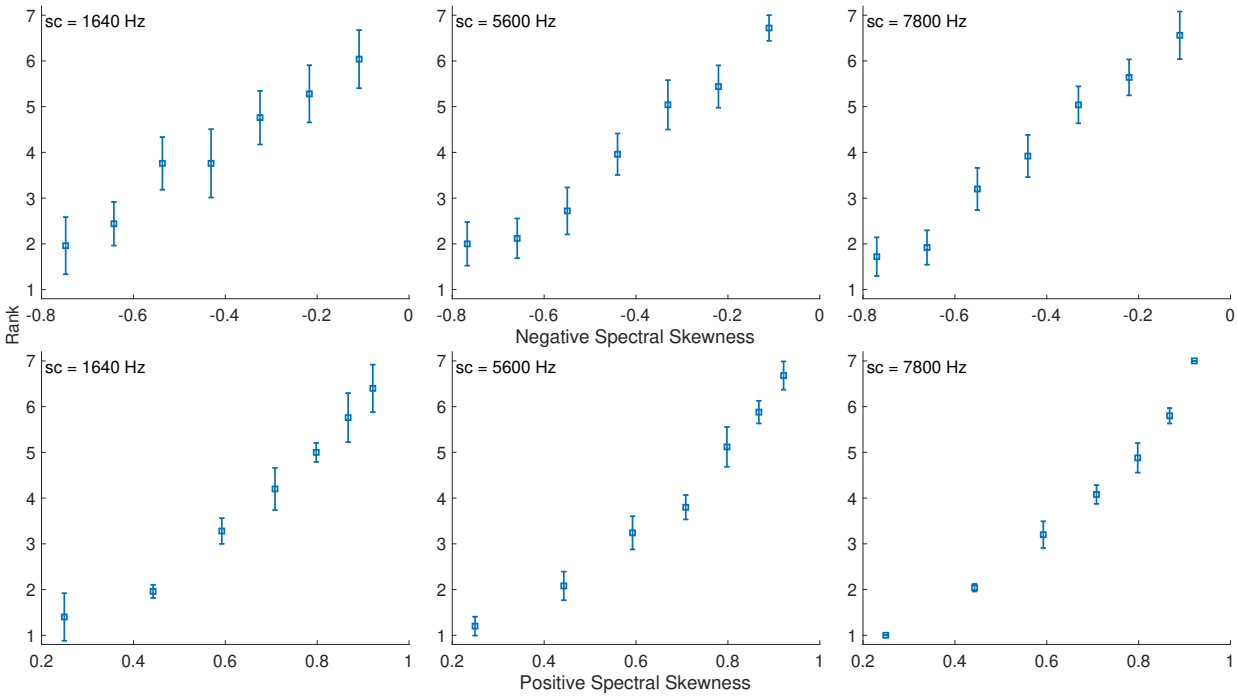


**Fig. 2.6** Spectral skewness: mean rankings of negative and positive spectral skewness stimuli between the anchors. sc = spectral centroid. Error bars represent 95% CI.

Interaction effects between the stimulus ranks and the fixed parameters of the stimuli used in the different subsets of each descriptor were only found to be significant for negative skewness, $F(12, 504) = 2.05, p = 0.02$. Table 2.3 lists pairs of stimuli from the negative skewness sound sets, the rankings of which were found to be nonsignificant in in the *post hoc* analysis and were thus confused by most listeners. The results indicate that confusion between stimuli within different sound sets (indicating potential difficulties in discrimination) decreased with increasing values of spectral centroid and thus with higher values of spectral spread, because in each stimulus set the amount of maximum allowable spread increased with increasing centroid (Section 2.2.1). For the rest of the descriptors, the *post hoc* tests were all significant ($|z| \geq 3.39$, $p_{adj} \leq 0.001$, $|z| \geq 4.12$, $p_{adj} \leq 0.001$, and $|z| \geq 2.03$, $p_{adj} \leq 0.042$ for the sound sets of spectral centroid, spread, and positive skewness, respectively), indicating that all stimuli within each sound set could be ordered correctly and were not confused with each other.

**Table 2.3** Stimulus pairs of nonsignificant rankings (critical $\alpha = 0.05$) for negative skewness stimuli. The values of stimulus pairs correspond to spectral skewness. *sc* = spectral centroid.

| Stimulus sets (*sc*) | Stimulus pairs |
|---|---|
| neg. skewness (1640 Hz) | −0.75, −0.64 |
| | −0.54, −0.43 |
| | −0.54, −0.32 |
| | −0.43, −0.32 |
| | −0.43, −0.22 |
| | −0.32, −0.22 |
| | −0.22, −0.11 |
| neg. skewness (5600 Hz) | −0.77, −0.66 |
| | −0.77, −0.55 |
| | −0.66, −0.55 |
| | −0.33, −0.22 |
| neg. skewness (7800 Hz) | −0.77, −0.66 |
| | −0.55, −0.44 |
| | −0.33, −0.22 |

## 2.3 Experimental Session B

The experimental procedure was the roughly same for both sessions (notably Participants and Apparatus), and therefore only differences with Session A are reported below.

### 2.3.1 Method

#### Stimuli

Several sound sets consisting of synthetic sounds were created by controlling the values of spectral slope, odd-to-even ratio, and spectral deviation in the synthesis process. For each audio feature, three $f_0$ at 120, 300, and 720 Hz were used to test whether different frequency regions would have an effect on the ratings. Loudness normalization and the general synthesis procedure following the spectral manipulations (described further below) were the same as described in Session A.

**Spectral slope** The stimuli of these sound sets were constructed by varying their spectral slopes. The spectral slope was controlled by reducing the differences between the amplitude levels of successive harmonics. The reduction was performed in nine logarithmic steps between the extremes of a flat spectrum and a $1/h^4$ or $1/((H+1)–h)^4$ harmonic (amplitude) spectrum for negative and positive slopes, respectively, where $h$ is the harmonic number and $H$ is the total number of harmonics. An example of this process is shown in Fig. 2.7. Note that the slopes are linear in log-frequency.

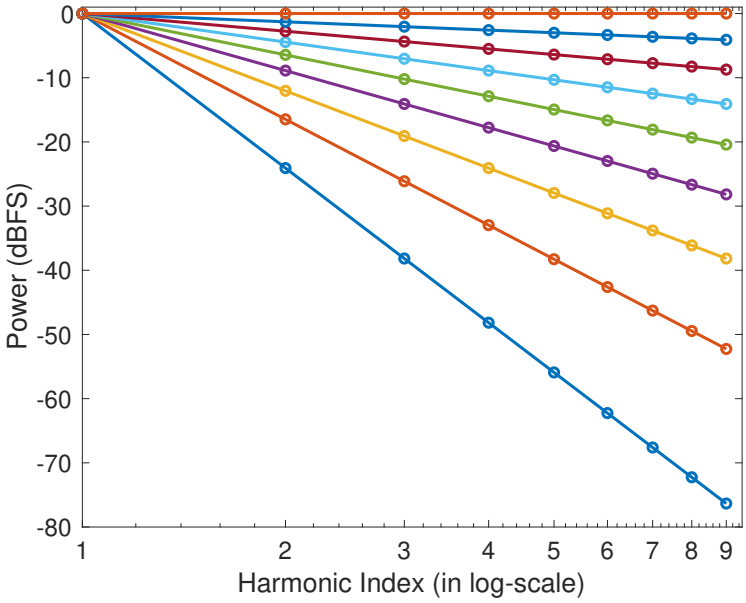


**Fig. 2.7** The negative spectral slopes used for synthesizing the stimuli of the spectral slope sound set.

In total, three sound sets with $f_0$ at 120, 300, and 720 Hz were constructed for both positive and negative slopes. Each set contained nine different slopes. In all conditions, nine harmonics were used for ensuring that roughness would not be a major factor in listeners' ratings, as the differences in $\mathrm{ERB_N}$ between the $8^{\mathrm{th}}$ and $9^{\mathrm{th}}$ harmonics were approximately

0.92, 0.97, and 0.87 for $f_0$ at 120, 300, and 720 Hz, respectively. At this point, it should also be noted that the computation of spectral slopes on the synthesized stimuli reported in Section 2.3.2 was performed using linear regression over the power in dB of log-spaced harmonics (Fig. 2.7), and not over linear amplitudes as in Peeters et al. (2011).

**Odd-to-even harmonic ratio** The odd-to-even ratio was controlled by equally attenuating the level in dB of the even harmonics while keeping the odd ones fixed at 0 dBFS (dB relative to digital full scale). In total, three sound sets with $f_0$ at 120, 300, and 720 Hz were constructed with the following attenuation levels for the even harmonics: {–60, –30, –20, –15, –10, –5 –1, 0}. The values of –30 and –1 dBFS were used to test whether listeners actually perceived differences between –60 and –20 dBFS, and between –5 and 0 dBFS respectively (in an otherwise limited set of only four stimuli to be ordered), by assuming that the differences between –1 and 0 dBFS, and between –60 to –30 dBFS would be imperceptible with respect to the presentation level.

Although ideally an equal number of odd and even harmonics should have been used for achieving an odd-to-even ratio of 1 when all harmonics have a value of 0 dBFS, only nine successive harmonics were used instead. This decision was based on the authors' informal listening tests during which it was noticed that the successive reductions in level of the last (even) harmonic were clearly audible, and might have been used as the major cue in listeners' ratings, preventing them from focusing on the overall reduction level of the even harmonics. This issue was mitigated by having the last harmonic to be odd, which offered the additional advantage of having a constant spectral centroid and skewness throughout the sound set and minimal successive differences between the spectral spreads.

**Spectral deviation** Spectral deviation was calculated as the average deviation of each harmonic amplitude from the average of itself and its two harmonic neighbors (Peeters et al., 2011). Although it is possible to achieve spectral deviations that vary monotonically across the stimuli by sampling spectra constructed by randomizing the amplitudes of the harmonics according to some probability distribution, this process does not guarantee that other perceptually important parameters will not also vary monotonically, such as the level of the $f_0$, the spectral centroid or the rest of spectral moments, to name a few. A nonmonotonic variation of these parameters would be confusing for listeners with respect to which parameter they should be focusing on when completing an ordering task on the stimuli.

To circumvent the nonmonotonic variations of parameters, a sample of one thousand amplitude distributions was generated by uniform randomizations of the levels of the harmonics in the range of [–25, 0] dBFS. The amplitude distribution, which had the greatest spectral deviation along with an odd-to-even ratio of approximately 1, and the greatest T2 tristimulus value below the level of the $f_0$, was then chosen as the reference for constructing stimuli with controlled deviations. The decision to choose an odd-to-even ratio

of approximately 1 ensured that this sound set did not vary predominantly according to that parameter (which was tested separately), whereas the choice of having the greatest possible T2 ensured that most of the deviation resulted from differences in levels among the upper harmonics.

In total, three sound sets were constructed consisting of nine stimuli each with $f_0$ at 120, 300, and 720 Hz. The reference distribution of amplitudes was rescaled to the range of [–60, 0] dBFS, and the deviation was controlled by reducing the differences in level between successive harmonics in nine logarithmic steps until all harmonics had reached a level of 0 dBFS. For these sound sets, the number of harmonics was increased to 16 (as opposed to 9), which facilitated the generation of a more uniform sample of amplitude distributions and the evaluation of a wider range of deviations among the higher harmonics. An example of this process is shown in Fig. 2.8.

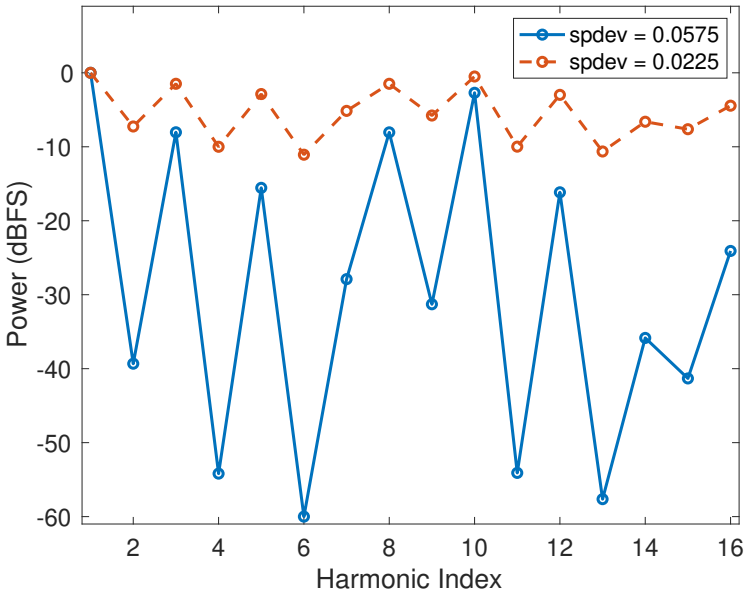


**Fig. 2.8** An example of high and low spectral deviation across harmonic amplitudes. spdev = spectral deviation.

### Procedure

In each trial and for each participant, the sounds in a given set were presented in random order but the sound sets were presented in the following (fixed) order for the same reason as described in Section 4.2.1: 1) spectral deviation, 2) odd-to-even harmonic ratio, 3) negative slope, and 4) positive slope. Each sound set was presented with three different $f_0$. The order of the three subsets was randomized within the sound set. This session took approximately 20 minutes to complete.

### 2.3.2 Results

As with the previous sets of descriptors, both the Friedman and likelihood ratio tests (Table 2.4) confirmed the main effect of the descriptors tested in this session. Fig. 2.9 shows the mean rankings for spectral deviation[1]. For the spectral deviation sets with fundamental frequencies at 120 and 300 Hz the trend analysis indicated that both linear and quadratic terms significantly described the pattern of the data on the ranked stimulus values (Table 2.5). For the set at 720 Hz, a linear term estimated by linear regression perfectly predicted the data.

**Table 2.4** Friedman ($\chi^2_F$) and likelihood ratio tests ($\chi^2_{LRT}$) for the stimulus sets of Session B. *df* = degrees of freedom; OER: odd-to-even ratio; $^*p < 0.001$

| Stimulus sets | $f_0$ (in Hz) | *df* | $\chi^2_F$ | $\chi^2_{LRT}$ |
|---|---|---|---|---|
| deviation | 120 | 6 | 143.73 | 82.71 |
| deviation | 300 | 6 | 146.19 | 86.43 |
| deviation | 720 | 6 | 150.00 | 79.95 |
| neg. slope | 120 | 6 | 150.00 | 79.95 |
| neg. slope | 300 | 6 | 148.87 | 83.09 |
| neg. slope | 720 | 6 | 147.67 | 82.40 |
| pos. slope | 120 | 6 | 148.77 | 82.55 |
| pos. slope | 300 | 6 | 149.18 | 79.12 |
| pos. slope | 720 | 6 | 149.59 | 79.85 |
| OER | 120 | 5 | 115.51 | 70.89 |
| OER | 300 | 5 | 119.56 | 72.29 |
| OER | 720 | 5 | 122.87 | 69.78 |

The same perfect linear relationship was found for the negative spectral slopes with $f_0$ at 120 Hz (Table 2.5), the mean rankings of which[1] are shown in Fig. 2.10. A linear trend was also observed for the set at 300 Hz, whereas for the set at 720 Hz both linear and quadratic terms were found to be significant. For positive slopes at 120 Hz, the linear and quadratic terms were also significant, whereas for the two other sets the analyses indicated a strong linear trend.

Fig. 2.11 shows the mean rankings for the sound sets of odd-to-even ratios. For the set with $f_0$ at 300 Hz the trend was linear, whereas for the other two sets both linear and quadratic terms were significant (Table 2.5). All the interaction effects between the stimulus

[1]Although anchoring the highest value of a particular feature at rank 1, and the lowest value at the highest rank might seem counterintuitive, we plot the results in the order that the anchor stimuli were presented to the listeners (i.e., the lowest rank corresponds to the left anchor, and the highest rank to the right anchor).

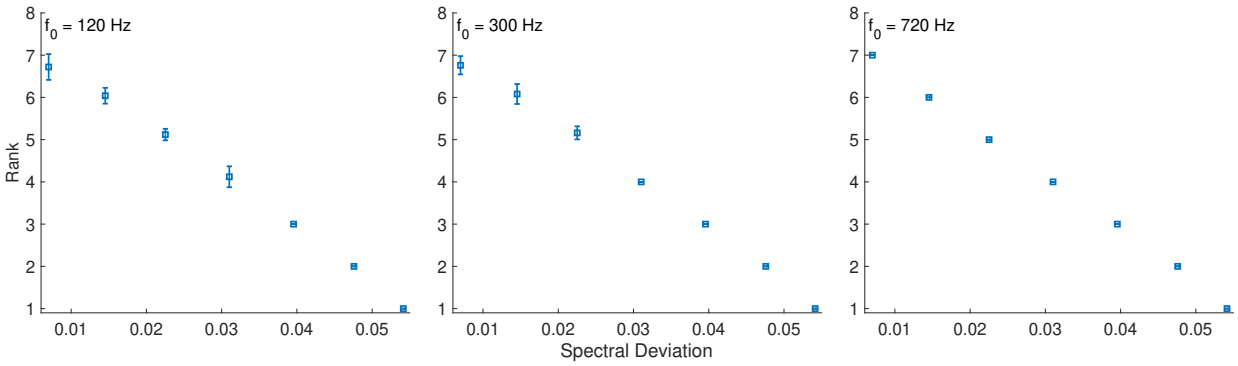


**Fig. 2.9** Spectral deviation: mean rankings of spectral deviation. Error bars represent 95% CI.

**Table 2.5** Ordinal regression coefficients for the stimulus sets of Session B. OER: odd-to-even ratio; L, Q: linear and quadratic terms respectively; -: perfect correlation.

| Stimulus sets | $f_0$ (in Hz) | Term | $b$ | $t$ | $p <$ |
|---|---|---|---|---|---|
| deviation | 120 | L | –188.32 | –7.68 | .001 |
| | | Q | 68.70 | 4.23 | .001 |
| deviation | 300 | L | –237.71 | –6.17 | .001 |
| | | Q | 96.71 | 4.04 | .001 |
| deviation | 720 | L | 26.46 | - | .001 |
| neg. slope | 120 | L | 26.46 | - | .001 |
| neg. slope | 300 | L | –206.31 | –9.51 | .001 |
| neg. slope | 720 | L | –219.82 | –6.94 | .001 |
| | | Q | 60.70 | 2.92 | .010 |
| pos. slope | 120 | L | –312.44 | –4.14 | .001 |
| | | Q | 100.56 | 2.26 | .050 |
| pos. slope | 300 | L | –227.82 | –8.60 | .001 |
| pos. slope | 720 | L | –264.73 | –7.08 | .001 |
| OER | 120 | L | –90.03 | –10.69 | .001 |
| | | Q | –15.58 | –2.28 | .050 |
| OER | 300 | L | –116.57 | –10.60 | .001 |
| OER | 720 | L | –175.30 | –6.48 | .001 |
| | | Q | –39.91 | –2.13 | .050 |

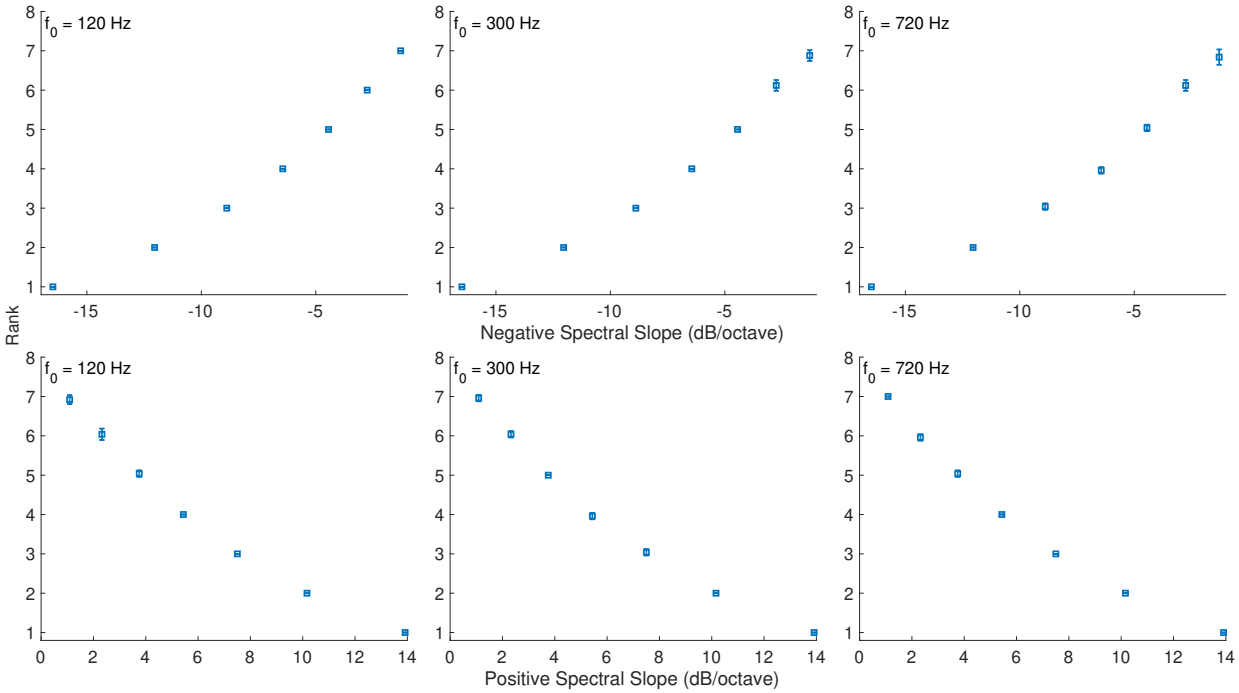


**Fig. 2.10** Spectral slope: mean rankings of negative and positive spectral slopes. Error bars represent 95% CI.

rank and $f_0$ were found to be nonsignificant ($p > 0.05$). Although the *post hoc* tests were all significant ($|z| \geq 2.76$, $p_{adj} \leq 0.006$ ), it can be clearly seen from the corresponding figures, that in the sets of spectral deviation and odd-to-even ratio, the variability of the rankings tended to decrease with increasing fundamental frequency.

## 2.4 Discussion and Conclusions

We synthesized stimuli that varied predominantly according to a given audio feature for testing whether listeners are able to perceive differences and order sounds varying according to one descriptor between the extreme values of each sound set. The synthesis strategies developed for constructing the stimuli were successful, because in each trial listeners were able to identify the attribute under study by exploring the range of feature values and without receiving any verbal explanation from the experimenters about the features on which the ordering task was to be based. The analyses indicated significant main effects of all descriptors tested, and in most cases, listeners could accurately order the stimuli over a wide range of descriptor values.

Skewness had a very narrow range restricted by the allowable values of the Skew-normal distribution, which allowed control of this parameter independently of spectral centroid and spread. This constraint was partly responsible for making this the hardest descriptor to or-

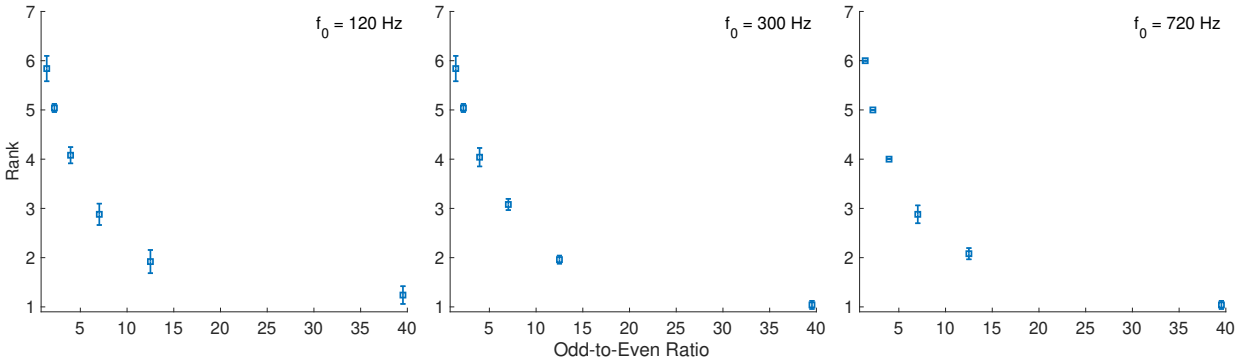


**Fig. 2.11** Odd-to-Even ratio: mean rankings of odd-to-even ratio. Error bars represent 95% CI.

der. The ordering was performed more accurately for higher centroids, but we assume that this is not due to the centroid values per se, because higher centroid values also allowed for greater spectral spreads in each sound set and thus greater discriminability among stimuli. We hypothesize that the increase in the number of spectral components led to more across-channel comparisons performed by listeners and made the task easier. The great amount of confusion found for negative skewness, when compared to the judgments for positive skewness, can be attributed to the asymmetry of auditory filters, which are negatively skewed with steep slopes on the higher-frequency sides (Glasberg & Moore, 1990). Fig. 2.12 shows the excitation patterns of stimuli exhibiting negative, zero, and positive skewness. From these patterns, it can be inferred that when the spectrum is progressively negatively skewed by a small amount (as in the present case), the resulting excitation patterns are very similar, leading to identical percepts. On the contrary, spectra that are progressively positively skewed lead to more drastic changes in the excitation patterns with steeper slopes on the low-frequency side of the envelope and make the differentiation between successive stimuli easier. For example, the slopes of the excitation patterns shown in Fig. 2.12 computed through a linear regression on the low-frequency sides of the envelopes, and below 6 dB of each envelope's maximum SPL down to 0 dB, have values at {0.07, 0.09, 0.14, 0.21, 0.43} per $ERB_N$, for stimuli with skewness of {–0.9, –0.4, 0, +0,4 +0.9}, respectively. At this point, it should be mentioned that before calculating the excitation patterns, the root mean squared amplitude values of the analyzed waveforms were calibrated to match the SPL (A-weighted) presentation levels of the respective stimuli.

For spectral deviation and odd-to-even ratio, the variability of the rankings decreased with increasing fundamental frequency, indicating that listeners were more sensitive in detecting spectral bumps at higher fundamental frequencies. This result is in general qualitative agreement with the results of Yost and Hill (1978) who found in their experiments using sinusoidal rippled noise spectra that sensitivity was a U-shaped function of the spacing between spectral peaks, and that best sensitivity occurred when the spacing was from 200 to 500 Hz, deteriorating severely below 200 and above 1000 Hz. The fact that in our

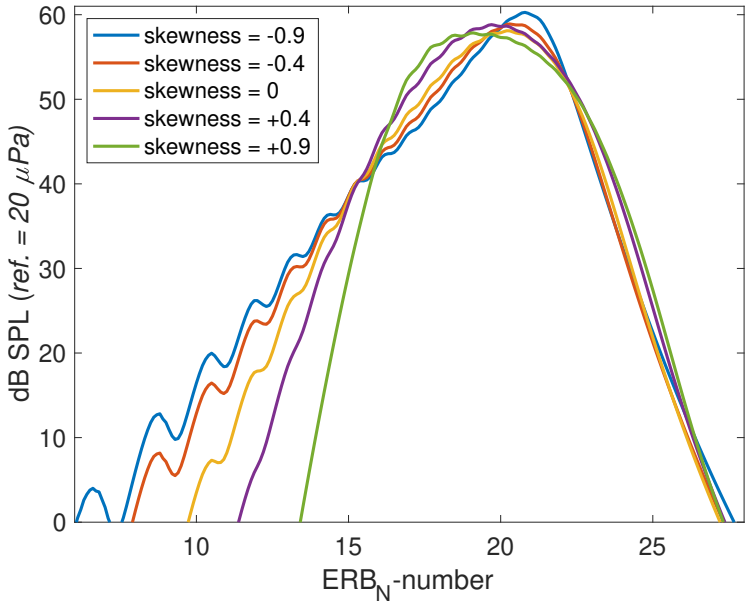


**Fig. 2.12** Auditory excitation patterns of stimuli with negative, zero, and positive skewness.

experiment sensitivity was best at a higher $f_0$ of 720 Hz, compared to their 500 Hz ripple, might be due to the differences in stimuli between the two studies (purely harmonic versus rippled noise). In order to also observe a U-shaped function, we would most likely need to test higher $f_0$.

The results of this experiment show that it is possible to proceed to subsequent psychophysical scaling experiments based on interval and ratio estimations, as there was no prior evidence that stimuli varying along the features tested here could be perceived on an ordinal scale, the existence of which is a prerequisite for constructing interval and ratio scales. Given the lack of prior knowledge on just noticeable differences of these parameters, the identification of values of a given feature that are not clearly distinguishable, as well as the trend analyses, also enable the construction of supraliminal stimuli which are required for deriving interval and ratio scale measurements.

The ordinal measurements derived from the present experiment indicate that spectral audio features carry perceptual contours, which can be used to compare, and group sounds according to their spectral shape. The contour is different from an interval because it contains only the signs of magnitude-changes and not the actual differences between magnitudes. McDermott, Lehr, and Oxenham (2008) provided evidence that listeners can recognize transpositions of contours in loudness and spectral centroid, and that the contours of these features are also useful for recognizing familiar melodies that are normally conveyed via pitch. In this study we have demonstrated the ordinal scalability of several spectral audio features, which suggests that listeners are also able to perceive contours

other than those in spectral centroid and loudness. In conclusion, the results of the ordinal scaling experiment provide evidence that all of the spectral features tested here are perceptually valid. In addition to the majority of previous timbre studies that have relied on correlational analysis, the present study has outlined trajectories of spectral features that causally correspond to listeners' perceptions.

## Appendix

If a random variable $Z$ follows the Skew-normal distribution with shape parameter $\alpha$ ($Z \sim SN(\alpha)$), and $Y = \xi + \theta Z$, then $Y \sim SN(\xi, \theta^2, \alpha)$ with scale parameter $\theta$, and location $\xi$. The mean, variance, and skewness ($\gamma_1$) of $Y$ are given by the following equations (Azzalini, 2005):

$$\mathbb{E}\{Y\} = \xi + \theta\mu_z \tag{2.5}$$

$$\text{var}\{Y\} = \theta^2(1 - \mu_z^2) \tag{2.6}$$

$$\gamma_1 = \frac{4-\pi}{2}\frac{\mu_z^3}{(1-\mu_z^2)^{3/2}} \tag{2.7}$$

where $\mu_z = \delta\sqrt{2/\pi}$, $\delta = \alpha/\sqrt{1+\alpha^2}, \in (-1, 1)$, and therefore $\alpha = \delta/\sqrt{1-\delta^2}$. In order to control skewness independently from a given mean and variance we estimate the parameters of the distribution sequentially by inverting the above equations. From the skewness equation we get the value of $\mu_z$:

$$\mu_z = \pm\frac{1}{\sqrt{1 + \left(\frac{(4-\pi)/2}{\gamma_1}\right)^{2/3}}} \tag{2.8}$$

The scale and location parameters for a given variance and mean can then be derived from Eqs. (2.5) and (2.6).

# References


Agus, N., Anderson, H., Chen, J., Lui, S., & Herremans, D. (2018). Perceptual evaluation of measures of spectral variance. *J. Acoust. Soc. Am.*, *143*, 3300–3311.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*, 1–10.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scand. J. Statist.*, *32*, 159–188.

Barr, D. J., Levy, R., Scheepers, C., & Tilly, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang*, *68*, 255–278.

Beasley, M. T., & Zumbo, B. D. (2009). Aligned rank tests for interactions in split-plot designs: Distributional assumptions and stochastic heterogeneity. *J. Mod. Appl. Stat. Methods*, *8*, 16–50.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *J. Acoust. Soc. Am.*, *118*, 471–482.

Genesis S. A. (2009). History and description of loudness models.
(s.l.: Loudness Toolbox for Matlab)

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, *47*, 103–138.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, *61*, 1270–1277.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am.*, *63*, 11493–1500.

Higgins, J. J., & Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World*, *1*, 201–211.

Hoffman, W. (1989). Iterative algorithms for Gram-Schmidt orthogonalization. *Computing*, *41*, 335–348.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, *6*, 65–70.

Horner, A. B., Beauchamp, J. W., & So, R. H. (2011). Evaluation of mel-band and MFCC-based error metrics for correspondence to discrimination of spectrally altered musical instrument sounds. *J. Audio Eng. Soc.*, *59*, 290–303.

ISO 389-8. (2004). *Acoustics – Reference Zero for the Calibration of Audiometric Equipment – Part 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones.* (Technical report (International Organization for Standardization, Geneva, Switzerland))

ISO/IEC. (2002). *MPEG-7: Information Technology – Multimedia Content Description Interface - Part 4: Audio.* ((ISO/IEC FDIS 15938–4:2002))

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *J. Acoust. Soc. Am.*, *94*, 2595–2603.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caracterisation du timbre des sons complexes. 2: Analyses acoustiques et quantification psychophysique. [Characterization of the timbre of complex sounds. 2: Acoustic analysis and psychophysical quantification]. *Journal de Physique*, *4*, 625–628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzén & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (Vol. 1989, pp. 43–53). Amsterdam: Excerpta Medica.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Percept. Psychophys.*, *62*, 426–1439.

Luepsen, H. (2017). The aligned rank transform and discrete variables: A warning. *Commun. Stat. – Simul. Comput.*, *46*, 6923–6936.

Martin, F. N., & Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *J. Am. Acad. Audiol*, *11*, 64–66.

McAdams, S., Beauchamp, J. W., & Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *J. Acoust. Soc. Am.*, *105*, 882–897.

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

McCullagh, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc. Series B*, *42*, 109–142.

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychol. Sci.*, *19*, 1263–1271.

McDermott, J. H., Schlemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nat. Neurosci.*, *16*, 493–498.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.*, *74*, 750–753.

Moore, B. C. J., Glasberg, B. R., & Baer, T. (1997). A model for the prediction of thresholds, loudness and partial loudness. *J. Audio Eng. Soc.*, *45*, 224–240.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.*, *130*, 2902–2916.

Schlittenlacher, J., Ellermeier, W., & Hashimoto, T. (2015). Spectral loudness summation: Shortcomings of current standards. *J. Acoust. Soc. Am.*, *137*, EL26–EL31.

Smith, B. K. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. In *Proceedings of the meeting of the society for music perception and cognition.* Berkeley, CA: University of California, Berkeley.

Wun, S., Horner, A., & Wu, B. (2014). Effect of spectral centroid manipulation on discrimination and identification of instrument timbres. *J. Audio Eng. Soc.*, *62*, 575–583.

Yost, W. A., & Hill, R. (1978). Strength of the pitches associated with ripple noise. *J. Acoust. Soc. Am.*, *64*, 485–492.

Zwicker, E., & Scharf, B. (1965). A model of loudness summation. *Psych. Rev.*, *72*, 3–26.

# Chapter 3

# Ordinal scaling of timbre-related audio descriptors: Amplitude-envelope features and inharmonicity

This chapter is based on the following research article:

Kazazis, S., Depalle, P. and McAdams, S. (in preparation). Ordinal scaling of timbre-related audio descriptors: Amplitude-envelope features and inharmonicity. Manuscript prepared for submission to *Journal of the Acoustical Society of America*.

**Abstract** Temporal audio features play an important role in timbre perception and sound identification. An experiment was conducted to test whether listeners are able to rank order synthesized stimuli over a wide range of feature values restricted within the range of instrument sounds. The following features were tested: attack and decay time, temporal centroid with fixed attack and decay time, and inharmonicity. The spectral envelope played an important role when ordering stimuli with various inharmonicity levels, whereas the shape of the amplitude envelope was an important parameter when ordering stimuli with various attack and decay times. Linear amplitude envelopes made the ordering of various attack times easier and caused the least amount of confusion among listeners, whereas exponential envelopes were more effective when ordering various decay times. Although there were many confusions in ordering short attack and decay times, listeners performed well in ordering temporal centroids even at very short attack and decay times. A meta-analysis of six timbre spaces was therefore conducted to test the explanatory power of attack time versus the attack temporal centroid (ATC) along a perceptual dimension. The results indicate that ATC has greater overall explanatory power than attack time itself.

## 3.1 Introduction

The attack (or rise) time and decay time, temporal centroid, and inharmonicity, are all considered here as temporal features that play an important role in timbre perception. Vos and Rasch (1981) defined the *perceptual onset* (POT) of a musical tone as the moment in time at which the stimulus is first perceived. Based on experiments of isochronous adjustments between tones of different attack times and a prediction model, the authors concluded that the perceptual onset is related to the time at which the amplitude envelope crosses a relative threshold below the maximum level of the tone, and that this threshold depends on the presentation level of the stimulus. Gordon (1987) defined *perceptual attack time* (PAT) as the time a tone's moment of attack *or* rhythmic emphasis is perceived, and claimed that Vos and Rasch's definition of POT coincides with his own definition of PAT. Nevertheless, he made a distinction between these two terms and argued that although for some instruments (e.g., percussive) the perceptual attack and onset may coincide, for some other instruments (e.g., reeds) it is possible for listeners to hear both an onset and a later attack. Based on experiments of isochronous and synchronous judgments between synthetic tones, he proposed a model for predicting PAT that takes into account not only a relative threshold to the maximum of the amplitude envelope, but also a threshold based on the slope of the envelope through the attack portion of the sound. However, POT was not quantified, and its influence on predicting PAT was only indirectly evaluated comparing models that were solely based on attack time (i.e., the time the amplitude envelope reaches its maximum value) or on fixed amplitude thresholds (i.e., the time the amplitude envelope crosses an absolute amplitude threshold) against models that incorporated relative thresholds (Fig. 3.1).

The attack time is important in many discrimination tasks and has been shown to play an important role in dissimilarity ratings between pairs of sounds, which is a basic experimental method on which many timbre studies have relied (Grey, 1977; Grey & Gordon, 1978; Iverson & Krumhansl, 1993; Lakatos, 2000; McAdams et al., 1995). The dissimilarity ratings are often analyzed with multidimensional scaling (MDS) techniques, which aim to project the data onto a low-dimensional space such that the proximities between all pairs of sounds in the full-dimensional space are preserved as much as possible in the lower-dimensional space (usually two to three dimensions). The analysis result is commonly referred to as a timbre space. Except for Grey (1977), and Grey and Gordon (1978), the rest of the aforementioned timbre studies found that attack time correlated significantly with one of the MDS axes and concluded that attack time is a perceptually salient feature for explaining dissimilarity ratings between instrumental sounds.

The decay time, temporal centroid, and inharmonicity have received less attention but nevertheless have been shown to be important for discrimination tasks, especially when the experimental paradigm is switched from judging the dissimilarities between pairs of sounds to identification and tone-quality judgments between sounds of similar timbres. McAdams, Chaigne, and Roussarie (2004) found a two-dimensional MDS solution for dissimilarity rat-

ings of impacted bar sounds that were generated by physical modeling synthesis (Chaigne & Doutaut, 1997). The first dimension was related to mass density and bar length and correlated with fundamental frequency. The second dimension was related to damping parameters and correlated with a linear combination of decay time constant and spectral centroid computed on the ERB-rate scale (Moore & Glasberg, 1983). In a similar study, participants also had to identify the material of the struck object as being made of glass or aluminum, and the damping properties proved to be more reliable for material categorization (McAdams, Roussarie, Chaigne, & Giordano, 2010).

Temporal centroid (or the center of gravity of the energy envelope) has been proposed as a feature for distinguishing between percussive and sustained sounds in music information retrieval tasks (Peeters, Giordano, Susini, Misdariis, & McAdams, 2011), and more recently has been shown to be an important feature for quantifying the similarities between action categories (i.e., sound-producing events such as strike, drop, rattle) on different material categories. Hjortkjær and McAdams (2016) found a two-dimensional MDS solution on dissimilarity ratings between sounds produced with different actions on different material categories in which one of the dimensions was related to the identification of the action category and was correlated strongly with temporal centroid.

Inharmonicity is not only an essential feature directly related to the timbre of some instruments such as the piano (Galembo, Askenfelt, Cuddy, & Russo, 2004) and acoustic guitar (Järveläinen & Karjalainen, 2006), for example, but also plays an important role in emotional responses to sound. Farbood and Price (2017) found that sounds with increasing degree of inharmonicity elicited higher emotional tension even in the case where no other spectral feature was positively correlated with inharmonicity. In the context of the present study, inharmonicity is considered to be a temporal feature because slight mistuning of the harmonic components within a complex induces modulations in the temporal domain, which provide cues for detecting inharmonicity.

In the following, an experiment is presented that investigated whether listeners can rank order stimuli with varying attack and decay times, with temporal centroids having fixed attack and decay times, and with different levels of inharmonicity (Section 3.2). The potential effect of different amplitude and spectral envelopes, and fundamental frequency (for inharmonicity) on the ratings was also tested. The statistical analysis reveals listeners' confusions between stimuli within a sound set constructed for testing a particular audio feature. It also compares qualitatively and quantitatively the results between different sound sets of the same feature (e.g., the rankings of attack time between sound sets constructed with different amplitude envelopes) and across different features (e.g., the rankings of attack versus decay times). The results of that experiment suggested conducting a meta-analysis on the dissimilarity ratings from previous timbre studies, in which the degree to which attack temporal centroid (ATC) can explain variation along a given MDS dimension (Section 3.3) is estimated. For the sake of brevity, and to avoid confusion between the various terms used throughout this paper, the following definitions are provided: *attack time* refers

to the time the amplitude envelope of a waveform reaches its maximum value; *perceptual attack time* (PAT) refers to the time at which the amplitude envelope crosses a (relative) threshold level below its maximum; *temporal centroid* refers to the center of gravity of the amplitude envelope (the exact formulation is given further below); *attack temporal centroid* (ATC) refers to the temporal centroid computed only up to attack time or PAT, ignoring the rest of the amplitude envelope. Fig. 3.1 exemplifies the above definitions by displaying their respective metrics along the time-axis of an exponential amplitude envelope. Note that the PAT and its corresponding ATC are labelled with the threshold level given in dB.

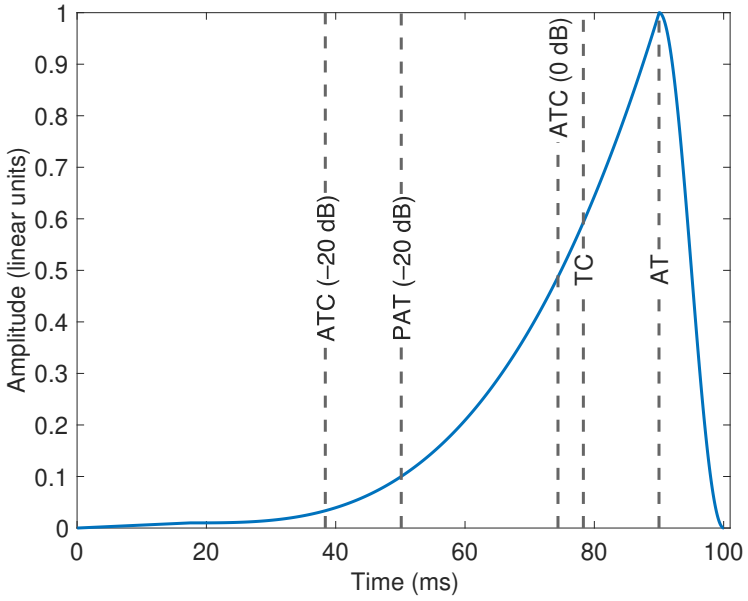


**Fig. 3.1** Computations of attack time (AT), temporal centroid (TC), perceptual attack time (PAT) and attack temporal centroid (ATC) on an exponentially increasing amplitude envelope. See the main text for definitions. Dashed lines indicate the respective metrics on the time-axis. ATC (0 dB): ATC computed up to attack time; PAT (–20 dB): PAT computed at 20 dB below the maximum level of the envelope; ATC (–20 dB): the corresponding ATC of PAT (–20 dB).

## 3.2 Experiment: Ordinal scaling

### 3.2.1 Method

#### Participants

Twenty-eight participants, 9 female, 18 male, and 1 "prefer not to answer", with a median age of 26.5 years (range: 18–34) were recruited from the Schulich School of Music, McGill

University. All of them were self-reported amateur or professional musicians with formal training in various disciplines such as performance, composition, music theory, and sound engineering. Only 25 out of the 28 participants had thresholds at or below 20 dB HL (Section 3.2.1) and were allowed to proceed to the experiment. Furthermore, the results of one participant were not included in the statistical analysis because the experimenters noticed that the participant was occupied with their mobile phone during testing. Participants who were not affiliated with the authors' lab and were allowed to proceed to the experiment and were compensated for their time.

### Stimuli

Several sound sets were constructed for testing the discernibility between successive attack and decay times, temporal centroids, and amounts of inharmonicity. The stimuli were synthesized using additive synthesis in MATLAB version R2015b (The MathWorks, Inc., Natick, MA) at a sampling frequency of 44.1 kHz with 16-bit amplitude resolution. The peak amplitude of the waveforms was normalized to 0.5.

**Inharmonicity** Inharmonicity was tested by constructing sound sets of ten inharmonic stimuli with nine components at equal or $1/n^2$ amplitude levels (for inharmonic components, $n$ is not integer), and at three $f_0$s of 120, 300, and 720 Hz. $1/n^2$ amplitude levels were preferred over $1/n$ because it was assumed that this spectral envelope would have a bigger influence on perceived inharmonicity when compared to the inharmonicity produced by a flat spectrum. All stimuli had a duration of 600 ms, gated with 10-ms raised-cosine ramps and were loudness-normalized according to the algorithm of Moore, Glasberg, and Baer (1997).

In a first attempt at constructing the stimuli, different inharmonicity levels were created by progressively increasing the amount of random, simultaneous mistuning of the harmonics. After several listening tests, the authors concluded that even slight random simultaneous mistuning of the eight components promoted multiple simultaneous entities (Hartmann, McAdams, & Smith, 1990). As such, within each sound set, there was no clear monotonic increase of perceived inharmonicity, but rather a percept of indeterminable "density" due to segregation effects.

A parametric model which had the following form and which is related to the inharmonicity of piano strings was found to be more appropriate (Fletcher, 1964):
$f_n = nf_0\sqrt{1+Bn^2}$, where $n$ is the component's rank and $B$ is the inharmonicity coefficient. For each sound set, the following ten inharmonicity coefficients were used ($\times 10^{-4}$): {0.10, 1.00, 2.16, 3.66, 5.60, 8.10, 11.34, 15.51, 20.91, 27.88}. The coefficients were spaced logarithmically (base $e$), and the maximum value was chosen so that the frequency of the last mistuned component would not exceed the frequency value of its next harmonic. This deterministic model led to no segregation effects and to a systematic monotonic increase in perceived inharmonicity.

**Attack and Decay times** All stimuli had a $f_0$ of 300 Hz with nine harmonics at equal amplitude. The attack and decay times were spaced approximately logarithmically (base 10) and ranged between 40 to 500 ms. The following 20 attack and decay times (in ms) were tested: {40, 47, 53, 60, 67, 77, 90, 100, 117, 133, 150, 173, 197, 227, 257, 293, 337, 383, 437, 500}. For these attack times, three sound sets were constructed based on three different types of amplitude envelopes for determining whether the shape of the envelope would have an effect on the ordering. The envelopes were constructed in a breakpoint fashion connecting three different segments (Fig. 3.2). The first segment was a linear ramp reaching –40 dBFS (dB relative to full digital scale) in the first 18 ms and was common to all envelopes. The second segment, which actually differentiates the three envelopes, had a duration of the attack time minus the duration of the first segment, and was constructed according to $y(t) = mt^c$, where $m$ is a constant that controls the slope of the attack time, and $c$ is the curvature constant, which controls the shape of the envelope. For each of the three sound sets, the curvature constant was given the values of $c = 1$ (linear), $c = 3$ (exponential) and $c = 10$ ("heavy" exponential). The last segment was a 10-ms raised-cosine decay gate and was also common to all envelopes.

The use of the first and last (fixed) segments ensured that all amplitude envelope manipulations were above –40 dBFS thereby minimizing the possibility that listeners' judgments would be based on the effective (or perceived) duration of the stimuli (Peeters et al., 2011). In all cases, the envelopes were applied to the waveforms so that the peak of the amplitude envelope matched the absolute peak of the waveform. Furthermore, the concatenation of the first linear ramp with the middle segment was made at a zero-crossing point of the waveform to minimize any amplitude discontinuities. The same amplitude envelopes in reverse direction were used for the sound sets of decay times. All stimuli had a total duration of attack time plus decay time.

**Temporal Centroids** The discriminability between temporal centroids was tested by creating sound sets in which the temporal centroids varied by way of changes in the amplitude envelopes while the attack and decay times were kept constant (Fig. 3.3). All stimuli had a fundamental frequency of 300 Hz with nine harmonics at equal amplitude. The temporal centroid ($tc$) was computed according to:

$$tc = \frac{\sum_{i=1}^{N} i \cdot e(i)}{f_s \sum_{i=1}^{N} e(i)} \tag{3.1}$$

where $i$ is the sample index, $e(i)$ the corresponding value of the amplitude envelope, $N$ is the total duration of the envelope in samples, and $f_s$ the sampling frequency of 44.1 kHz. In total, five sound sets, each with ten linearly spaced temporal centroids, were created for

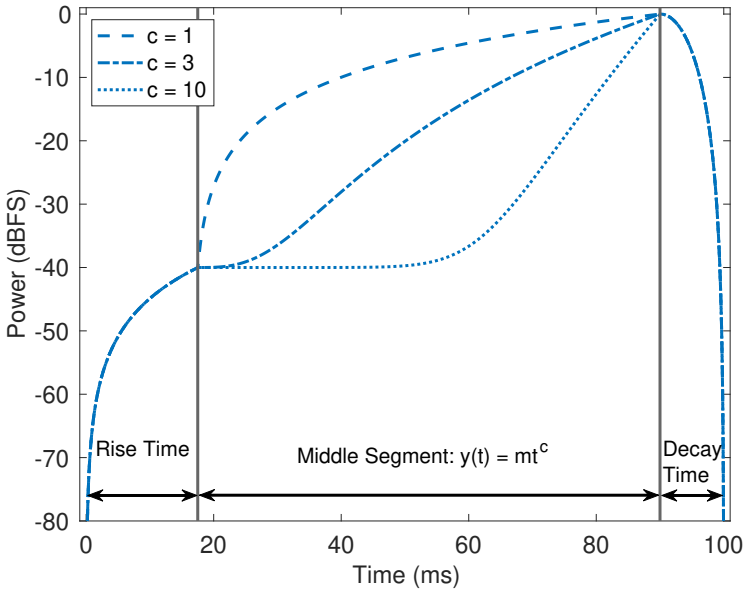


**Fig. 3.2** Amplitude envelopes (displayed in dB) of stimuli with 90-ms attack time. The initial rise time reaching –40 dBFS (1st segment) and the decay time (3rd segment) is common to all envelopes. $m$ = slope of attack time; $c$ = curvature value.

five attack times, the values of which are shown in Table 3.1. The maximum and minimum temporal centroids of each sound set were calculated first by applying amplitude envelopes constructed in the same way as described in the previous section, with curvature values of $c = 10$ and $c = 0.33$. The intermediate linearly spaced centroid values were achieved by calculating their corresponding curvature values, which define the shape of the amplitude envelope during the attack time. In all cases, the initial rise time over the first 18 ms and the decay time, as described in the previous subsection, were kept constant in duration and shape for all sounds. A similar procedure was used for generating five sound sets of temporal centroids based on decay times, which had the same value with the attack times shown in Table 3.1. All stimuli had a total duration of attack time plus decay time.

### Procedure

Before the experiment, participants signed an informed-consent form. Afterwards, they passed a pure-tone audiometric test at octave-spaced frequencies from 125 Hz to 8 kHz (ISO 389-8, 2004; Martin & Champlin, 2000) and were required to have thresholds at or below 20 dB HL to proceed to the experiment. The instructions that described the task and user interface were presented on paper and were further explained by the experimenter. Any questions had to be asked during the practice block for which the experimenter was also present in the booth.

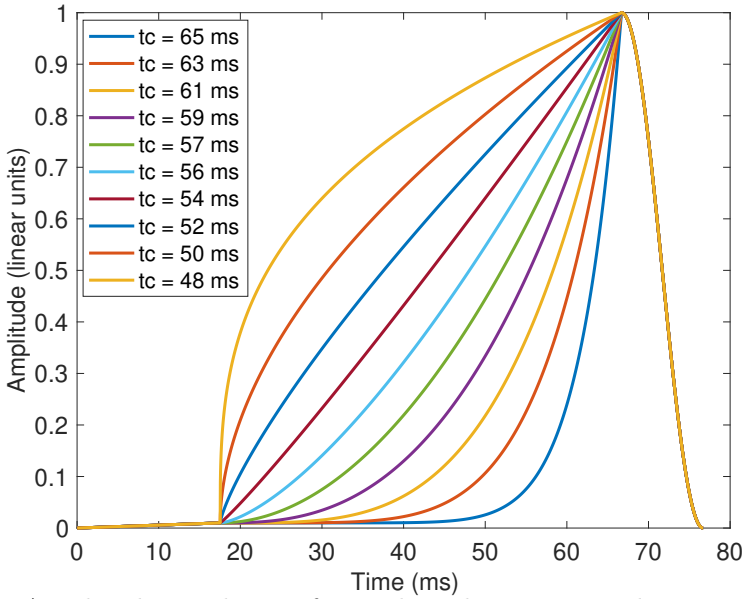


**Fig. 3.3** Amplitude envelopes of stimuli with 67-ms attack time in the temporal centroid set, and their corresponding temporal centroids. *tc* = temporal centroid.

**Table 3.1** Attack times, temporal centroids, and the curvature values that were used to generate the amplitude envelopes of stimuli in the temporal centroid set.

| Attack Time | Temporal Centroids (ms)<br>*Curvature Values* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 40 ms | 40.85 | 39.99 | 39.14 | 38.28 | 37.42 | 36.57 | 35.71 | 34.85 | 34.00 | 33.14 |
| | *10.00* | *5.55* | *3.69* | *2.63* | *1.93* | *1.44* | *1.06* | *0.77* | *0.53* | *0.33* |
| 67 ms | 64.67 | 62.86 | 61.04 | 59.23 | 57.41 | 55.60 | 53.78 | 51.96 | 50.15 | 48.33 |
| | *10.00* | *5.61* | *3.76* | *2.70* | *1.99* | *1.48* | *1.09* | *0.78* | *0.54* | *0.33* |
| 132 ms | 123.55 | 119.41 | 115.28 | 111.14 | 107.00 | 102.87 | 98.73 | 94.60 | 90.46 | 86.32 |
| | *10.00* | *5.56* | *3.75* | *2.70* | *1.99* | *1.48* | *1.09* | *0.79* | *0.54* | *0.33* |
| 257 ms | 231.81 | 223.46 | 215.10 | 206.74 | 198.39 | 190.03 | 181.67 | 173.32 | 164.96 | 156.61 |
| | *10.00* | *5.47* | *3.70* | *2.67* | *1.98* | *1.48* | *1.09* | *0.79* | *0.54* | *0.33* |
| 500 ms | 444.94 | 428.31 | 411.68 | 395.05 | 378.42 | 361.79 | 345.16 | 328.53 | 311.91 | 295.28 |
| | *10.00* | *5.39* | *3.66* | *2.65* | *1.97* | *1.47* | *1.09* | *0.78* | *0.54* | *0.33* |

The task of the participants was to order a set of stimuli according to "any criteria that differentiate them the most." Any verbal labeling of possible criteria was intentionally avoided. The stimuli were presented in the form of sound boxes on which participants could click to hear each stimulus and then drag them to the desired position for the ordering task. The user interface consisted of two main panels. In each trial the top panel contained only two stimuli (i.e., the anchors), which had the minimum and maximum values of a particular audio feature and between which the ordering of the rest of the stimuli would take place. The rest of the stimuli were presented in randomized order in the lower panel. The task was completed when all of the stimuli in the lower panel were dragged and re-arranged according to the desired order in the top panel. In each trial and for each participant, the stimuli were presented in random order, but the sound sets were presented in the following (fixed) order: 1) practice block, 2) attack time, 3) temporal centroid for a particular attack time, 4) inharmonicity, 5) decay time, 6) temporal centroid for a particular decay time. The practice block consisted of five trials, one trial for each of the afore-mentioned features with a maximum of five stimuli per feature set. The experiment took approximately 90 minutes to complete.

### Apparatus

The experimental session was run with the PsiExp computer environment (Smith, 1995). Sounds were amplified through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA) and presented diotically over Sennheiser HD600 headphones (Sennheiser Electronic GmbH, Wedemark, Germany). The sound pressure levels had a range of 55.6–66.9 dB SPL (A-weighted) as measured with a Brüel & Kjær Type 2205 sound-level meter with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark). Listeners were seated individually in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY).

### Data analysis

Because of the ordering task, nonparametric tests were used on participants' stimulus rankings. For each stimulus set, separate nonparametric Friedman tests were used to evaluate the main effect of each audio descriptor. To account for the nonsphericity present in the data, which can transmit to Friedman ranks (Beasley & Zumbo, 2009), the main effects of each audio feature were also tested with a proportional-odds mixed model (McCullagh, 1980), which had a full random effects structure with random intercepts for each participant and random by-participant slopes for the fixed factor of sound set (Barr, Levy, Scheepers, & Tilly, 2013). The main effect was evaluated by a likelihood ratio test (in which the maximum likelihood was estimated by the Laplace approximation) between the full model and a reduced model, which had the same random effects structure but excluded the effect of interest from the fixed factors. The main trends of the data were identified

through forward stepwise ordinal regression with orthogonal polynomials constructed with the modified Gram-Schmidt algorithm (Hoffman, 1989) on ranked stimulus values.

Two-tailed *post hoc* Wilcoxon signed-rank tests were used to examine whether the rank of each stimulus was significantly different from the rest and thus to identify stimulus combinations that were confused by the listeners. Due to the large number of multiple comparisons within each stimulus set, the *post hoc* tests were corrected with the Holm-Bonferroni method (critical $\alpha = 0.05$), which controls the family-wise error rate (Holm, 1979). A proportional-odds mixed model was used with a nested random effects structure having random intercepts for each participant and the subsets of stimuli nested within each participant. The aim of this analysis was to examine the interaction effects between the ranking of the stimuli along a given descriptor and the subsets with different values of a parameter, such as attack time or fundamental frequency, used in each subset of the same descriptor (e.g., the ranking of temporal centroid between the sound sets constructed at different attack times). The interactions were examined after fitting the model using sum coding for the predictor variables and performing an ANOVA on the fixed effects (Barr et al., 2013). Although analyses are conducted on ranks, in all data graphs, the actual stimulus values are plotted on the x-axis. As such, the graphs at times appear concave or convex even though a linear relation may exist between physical ranks and mean response ranks. All the statistical analyses were done in MATLAB (The MathWorks, Inc., Natick, MA).

### 3.2.2 Results

Both the Friedman and likelihood ratio tests shown in Table 3.2 confirmed the main effect of each temporal feature. All the interaction effects between the stimulus rank and amplitude envelopes, or $f_0$ and spectral envelopes (for the inharmonicity sets) were found to be nonsignificant ($p > 0.05$).

Fig. 3.4 shows the mean rankings for the inharmonicity sets at different $f_0$ and spectral envelopes, and Table 3.3 displays the results of ordinal regression. For the sound set at 120 Hz and with the $1/n^2$ spectral envelope, all terms up to cubic were significant, whereas the rest of the sound sets exhibited only a linear trend. Table 3.4 lists the pairs of stimuli that were confused by most listeners. The confusions generally increased with increasing $f_0$, and at 720 Hz the sound set with equal amplitudes had double the number of confusions compared to the $1/n^2$ spectral envelope. The differences between the mean rankings of the rest of stimuli were all significant ($|z| \geq 2.62$, $p_{adj} \leq 0.045$ and $|z| \geq 2.4$, $p_{adj} \leq 0.049$, for the sound sets with flat and $1/n^2$ spectral envelopes, respectively).

The mean rankings for the sound sets of attack and decay times are shown in Fig. 3.5. The results of ordinal regression (Table 3.3) confirmed a linear trend of these rankings with increasing value of the rank-ordered parameters. Table 3.5 lists the pairs of stimuli without significant differences in rankings as determined by the *post hoc* analysis and which were thus confused by most listeners. For the attack times, the confusion increased with

**Table 3.2** Friedman ($\chi^2_F$) and likelihood ratio tests ($\chi^2_{LRT}$). att = attack time (in ms); dec = decay time (in ms); c = curvature constant; tc: temporal centroid set; $\text{inh}^{\text{A}}$: inharmonicity set with spectral components at equal amplitude; $\text{inh}^{\text{B}}$: inharmonicity set with spectral components at $1/n^2$ amplitude; $f_0$ in Hz; $df$ = degrees of freedom; $^*p < 0.001$

| Stimulus sets | $df$ | $\chi^2_F$ | $\chi^2_{LRT}$ |
|---|---|---|---|
| att (c = 1) | 17 | $401.40^*$ | $125.02^*$ |
| att (c = 3) | 17 | $398.95^*$ | $124.40^*$ |
| att (c = 10) | 17 | $391.77^*$ | $124.67^*$ |
| tc (att = 40) | 7 | $138.10^*$ | $74.87^*$ |
| tc (att = 67) | 7 | $147.46^*$ | $83.26^*$ |
| tc (att = 132) | 7 | $150.46^*$ | $79.73^*$ |
| tc (att = 257) | 7 | $146.46^*$ | $72.78^*$ |
| tc (att = 500) | 7 | $161.51^*$ | $82.07^*$ |
| $\text{inh}^{\text{A}}$ ($f_0 = 120$) | 7 | $154.00^*$ | $82.59^*$ |
| $\text{inh}^{\text{A}}$ ($f_0 = 300$) | 7 | $148.74^*$ | $75.79^*$ |
| $\text{inh}^{\text{A}}$ ($f_0 = 720$) | 7 | $138.92^*$ | $73.95^*$ |
| $\text{inh}^{\text{B}}$ ($f_0 = 120$) | 7 | $150.93^*$ | $73.49^*$ |
| $\text{inh}^{\text{B}}$ ($f_0 = 300$) | 7 | $152.38^*$ | $75.58^*$ |
| $\text{inh}^{\text{B}}$ ($f_0 = 720$) | 7 | $144.24^*$ | $70.87^*$ |
| dec (c = 1) | 17 | $393.33^*$ | $124.13^*$ |
| dec (c = 3) | 17 | $395.68^*$ | $123.50^*$ |
| dec (c = 10) | 17 | $372.89^*$ | $120.04^*$ |
| tc (dec = 40) | 7 | $131.68^*$ | $77.70^*$ |
| tc (dec = 67) | 7 | $137.38^*$ | $76.36^*$ |
| tc (dec = 132) | 7 | $151.11^*$ | $81.01^*$ |
| tc (dec = 257) | 7 | $160.89^*$ | $82.68^*$ |
| tc (dec = 500) | 7 | $165.46^*$ | $86.30^*$ |

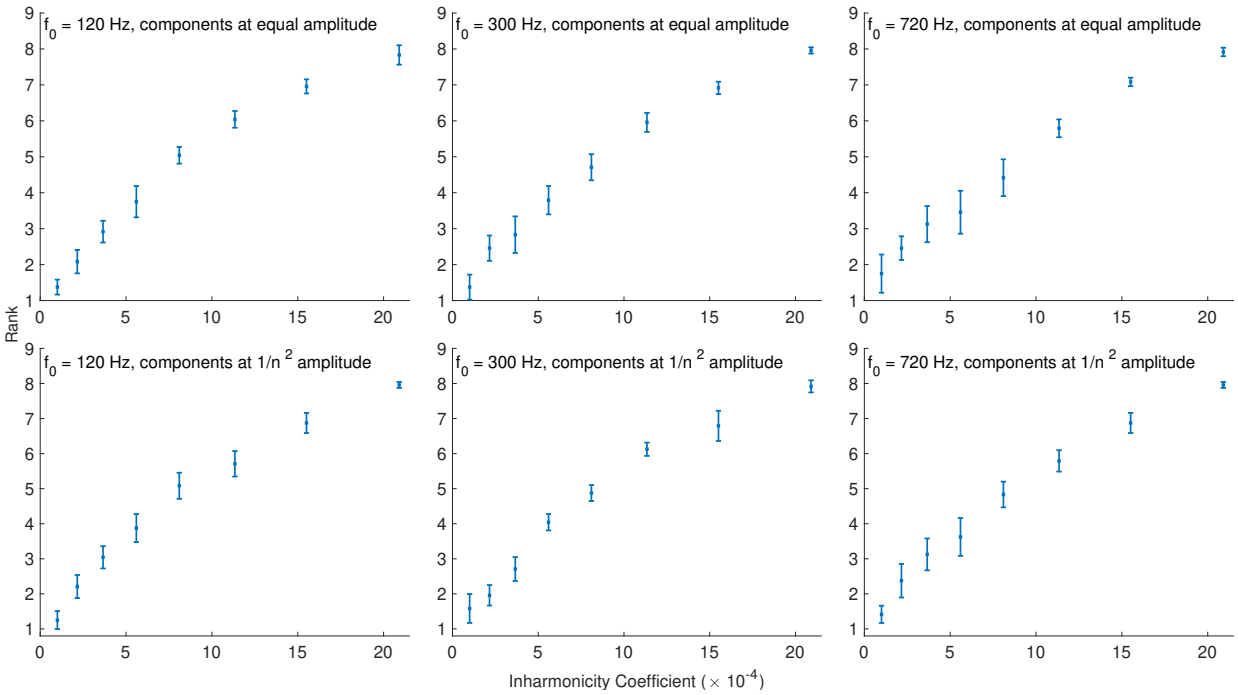


**Fig. 3.4** Mean rankings (sounds 2-9) of inharmonicity stimuli between the anchors. The top and bottom panels display the sound sets with flat and $1/n^2$ spectral envelopes, respectively. Error bars represent the 95% CI.

**Table 3.3** Ordinal regression coefficients. L, Q, C: linear, quadratic, and cubic terms, respectively; att = attack time (in ms); dec = decay time (in ms); c = curvature value; tc: temporal centroid set; $\text{inh}^{\text{A}}$: inharmonicity set with flat spectral envelope; $\text{inh}^{\text{B}}$: inharmonicity set with $1/n^2$ spectral envelope; $f_0$ in Hz.

| Stimulus sets | Term | $b$ | $SE$ | $t$ | $p <$ |
|---|---|---|---|---|---|
| att (c = 1) | L | –343.38 | 16.47 | –20.85 | 0.001 |
| att (c = 3) | L | –281.41 | 13.30 | –21.16 | 0.001 |
| att (c = 10) | L | –205.37 | 9.51 | –21.59 | 0.001 |
| dec (c = 1) | L | –218.37 | 10.15 | –21.51 | 0.001 |
| dec (c = 3) | L | –248.74 | 11.68 | –21.30 | 0.001 |
| dec (c = 10) | L | –158.95 | 7.31 | –21.74 | 0.001 |
| tc (att = 40) | L | 58.12 | 4.47 | 13.00 | 0.001 |
| tc (att = 67) | L | 72.47 | 5.57 | 13.02 | 0.001 |
| | Q | –9.61 | 4.79 | –2.01 | 0.045 |
| tc (att = 132) | L | 83.47 | 6.53 | 12.79 | 0.001 |
| | Q | –21.12 | 6.45 | –3.28 | 0.001 |
| | C | 6.86 | 2.98 | 2.30 | 0.021 |
| tc (att = 257) | L | 74.92 | 5.71 | 13.11 | 0.001 |
| tc (att = 500) | L | 137.62 | 10.60 | 12.99 | 0.001 |
| tc (dec = 40) | L | 50.64 | 3.95 | 12.82 | 0.001 |
| tc (dec = 67) | L | 54.81 | 4.24 | 12.93 | 0.001 |
| tc (dec = 132) | L | 82.68 | 6.36 | 13.00 | 0.001 |
| | Q | 135.93 | 10.89 | 12.48 | 0.001 |
| | Q | –23.72 | 8.28 | –2.86 | 0.004 |
| tc (dec = 500) | L | 190.61 | 15.89 | 12.00 | 0.001 |
| $\text{inh}^{\text{A}}$ ($f_0 = 120$) | L | –93.87 | 7.16 | –13.11 | 0.001 |
| $\text{inh}^{\text{A}}$ ($f_0 = 300$) | L | –73.19 | 5.59 | –13.10 | 0.001 |
| $\text{inh}^{\text{A}}$ ($f_0 = 720$) | L | –58.10 | 4.48 | –12.97 | 0.001 |
| $\text{inh}^{\text{B}}$ ($f_0 = 120$) | L | –92.51 | 7.34 | –12.60 | 0.001 |
| | Q | –24.56 | 7.51 | –3.27 | 0.001 |
| | C | –10.15 | 3.37 | –3.01 | 0.003 |
| $\text{inh}^{\text{B}}$ ($f_0 = 300$) | L | –93.01 | 7.11 | –13.09 | 0.001 |
| $\text{inh}^{\text{B}}$ ($f_0 = 720$) | L | –69.90 | 5.33 | –13.11 | 0.001 |

**Table 3.4** Stimulus pairs with nonsignificant differences in rankings for inharmonicity stimuli. $f_0$: the fundamental frequency (in Hz) that was used in each of the inharmonicity sound sets; $\text{inh}^{\text{A}}$: inharmonicity of stimulus pairs with flat spectral envelope; $\text{inh}^{\text{B}}$: inharmonicity of stimuli pairs with $1/n^2$ spectral envelope.

| $f_0$ | $\text{inh}^{\text{A}}$ | $f_0$ | $\text{inh}^{\text{B}}$ |
|---|---|---|---|
| 120 | – | 120 | – |
| 300 | 2.16, 3.66 | 300 | 1.00, 2.16 |
| 720 | 2.16, 3.66 | 720 | 2.16, 3.66 |
| | 2.16, 5.60 | | 3.66, 5.60 |
| | 3.66, 5.60 | | – |
| | 5.60, 8.10 | | – |

increasing curvature value. For the decay times, the overall amount of confusion was greater as compared to the attack times (mainly due to the sound set with linear envelopes), but it did not increase with increasing curvature value because the sound set with curvature value of $c = 3$ had about half the number of confusions compared to the sound sets with linear and $c = 10$ decay-envelopes. The differences between the mean rankings of the rest of the stimuli were all significant ($|z| \geq 2.76$, $p_{adj} \leq 0.041$ and $|z| \geq 2.87$, $p_{adj} \leq 0.049$, for the sound sets of attack and decay time, respectively).

Fig. 3.6 and Fig. 3.7 show the mean rankings for the sound sets of temporal centroids at different attack and decay times. For the sound sets with attack times at 40, 257, and 500 ms the trend was linear, for 67 ms both linear and quadratic terms were significant, and for 133 ms terms up to cubic were all significant (Table 3.3). For the sound sets with decay times at 40, 67, and 500 ms the trend was linear, whereas for the other two sets both linear and quadratic terms significantly described the pattern of the data. As can be seen from Table 3.6, the number of confusions between the stimuli of these sound sets did not decrease monotonically with increasing attack or decay times. For the sound sets at different attack times, most confusions mainly occurred between the middle curvature values. Confusions between the temporal centroids of decay times only occurred for the earliest decay times at 40 and 67 ms, and for the first few highest curvature values. The differences between the mean rankings of the rest of the stimuli were all significant ($|z| \geq 2.49$, $p_{adj} \leq 0.038$ and $|z| \geq 2.65$, $p_{adj} \leq 0.04$, for the sound sets of attack and decay temporal centroids, respectively).

### 3.2.3 Discussion

Overall, the experiment showed that listeners could rank order stimuli with various attack and decay times, temporal centroids with fixed attack and decay times, and inharmonicity levels, over a wide range of feature values. However, there were confusions in the stimulus

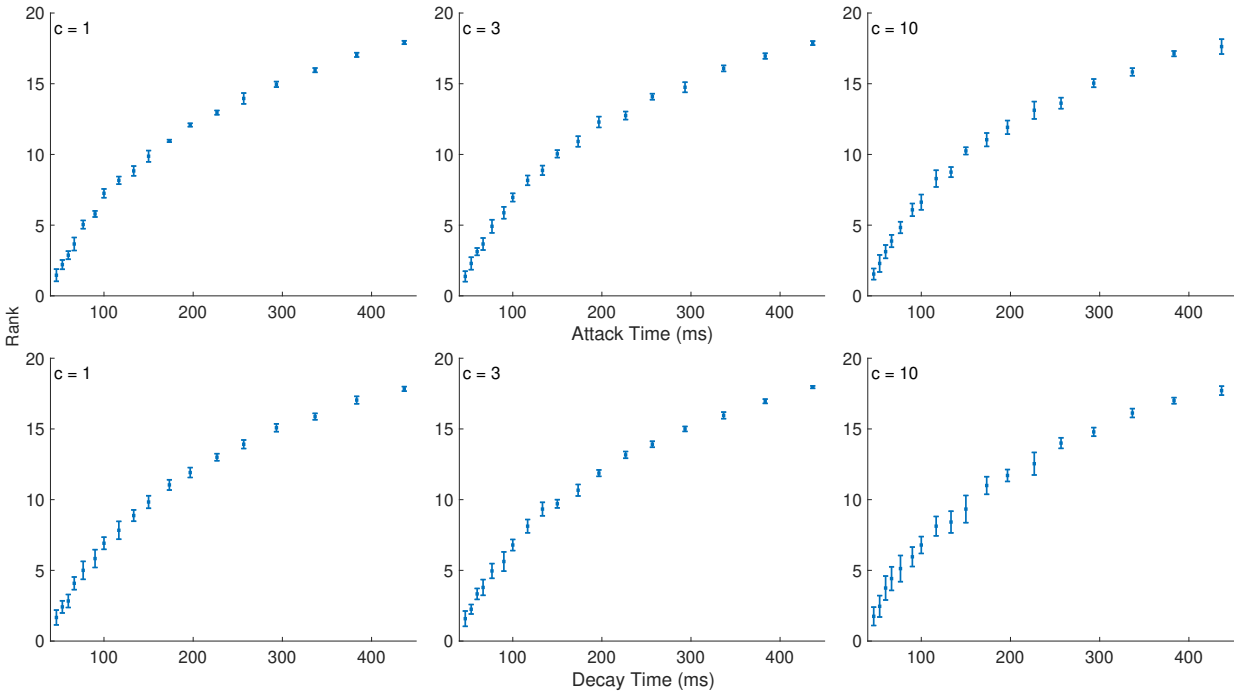


**Fig. 3.5** Mean rankings (sounds 2-19) of attack (top panel) and decay time (bottom panel) stimuli between the anchors. c = curvature value. Error bars represent the 95% CI.

**Table 3.5** Stimulus pairs with nonsignificant differences in rankings for attack and decay time stimuli. c: curvature values that were used to generate the amplitude envelopes of stimuli in the attack- and decay-time sound sets.

| c | Attack Times (ms) | c | Decay Times (ms) |
|---|---|---|---|
| 1 | 47, 53 | 1 | 47, 53 |
| | 53, 60 | | 47, 60 |
| | 60, 67 | | 53, 60 |
| | 117, 133 | | 67, 77 |
| | – | | 77, 90 |
| | – | | 90, 100 |
| | – | | 100, 117 |
| | – | | 117, 133 |
| | – | | 133, 150 |
| | – | | 173, 197 |
| 3 | 47, 53 | 3 | 47, 53 |
| | 53, 60 | | 60, 67 |
| | 60, 67 | | 77, 90 |
| | 77, 90 | | 90, 100 |
| | 117, 133 | | 133, 150 |
| | 197, 227 | | – |
| 10 | 47, 53 | 10 | 47, 53 |
| | 53, 60 | | 53, 60 |
| | 60, 67 | | 60, 67 |
| | 90, 100 | | 60, 77 |
| | 117, 133 | | 60, 90 |
| | 150, 173 | | 67, 77 |
| | 173, 197 | | 90, 100 |
| | 197, 227 | | 117, 133 |
| | 227, 257 | | 133, 150 |
| | – | | 173, 197 |
| | – | | 257, 293 |

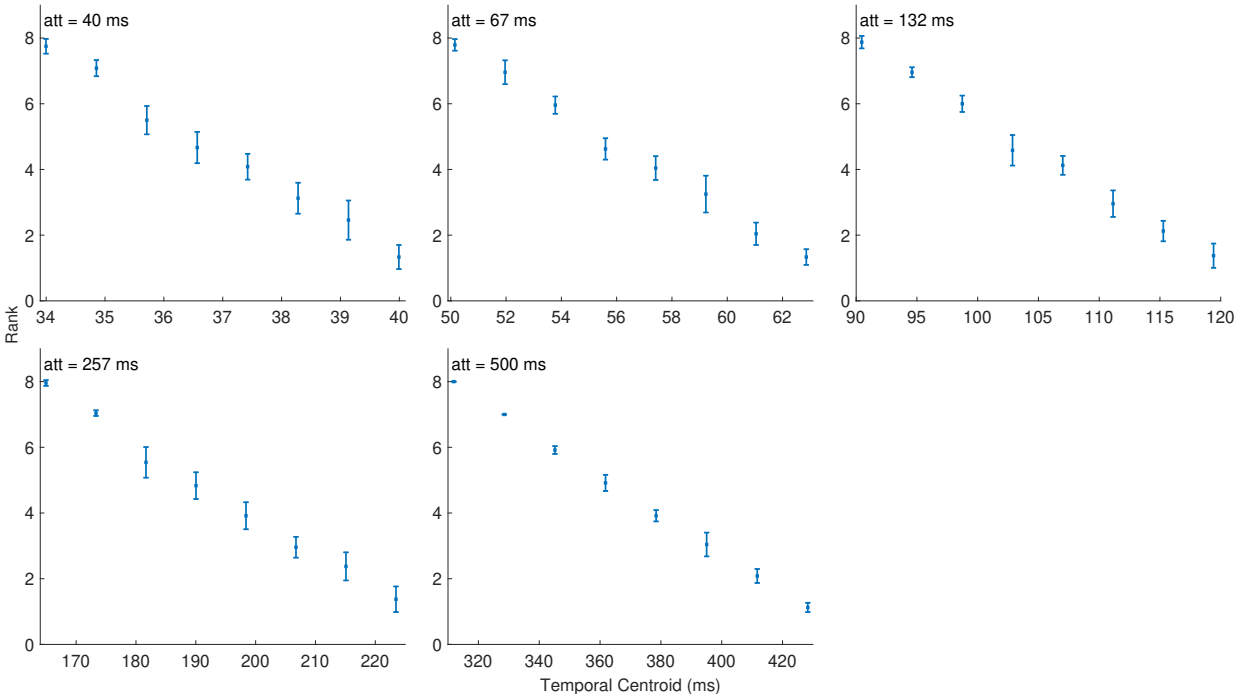


**Fig. 3.6** Mean rankings (sounds 2-9) of temporal centroid stimuli between the anchors. att = attack time. Error bars represent the 95% CI.

**Table 3.6** Stimulus pairs with nonsignificant differences in rankings for temporal centroid stimuli. att, dec: attack and decay times (in ms), respectively, that were used in each of the temporal centroid sound sets; tc: temporal centroids of stimulus pairs (in ms).

| att | tc | dec | tc |
|---|---|---|---|
| 40 | 36.57, 37.42 | 40 | 16.05, 15.19 |
| | 38.28, 39.14 | | 13.48, 12.62 |
| | – | | 13.48, 11.77 |
| | – | | 12.62, 11.77 |
| 67 | 57.41, 59.23 | 67 | 24.75, 22.93 |
| | – | | 22.93, 21.12 |
| | – | | 21.12, 19.30 |
| | – | | 19.30, 17.48 |
| | – | | 17.48, 15.67 |
| 132 | 102.87, 107.00 | 132 | – |
| 257 | 181.67, 190.03 | 257 | – |
| | 206.74, 215.10 | | – |
| 500 | – | 500 | – |

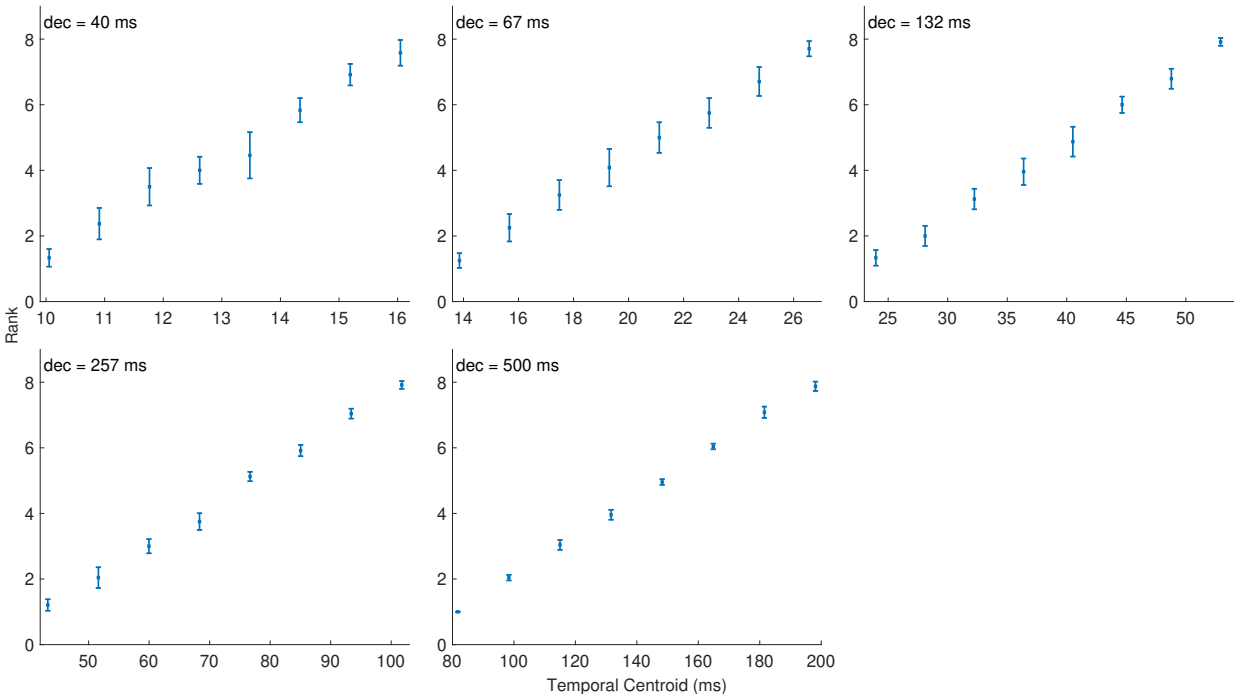


**Fig. 3.7** Mean rankings (sounds 2-9) of temporal centroid stimuli between the anchors. dec = decay time. Error bars represent the 95% CI.

rankings depending on the feature (attack versus decay), the type of the amplitude envelope used for each stimulus set, and the spectral envelope in the case of inharmonicity.

For inharmonicity, the sound sets with $1/n^2$ spectral envelopes caused less confusion than the sound sets with flat envelopes. With respect to the presentation levels of this stimulus set (in average: 63.3 dB SPL), this may be attributed to the detection of beats that are caused by the interaction of the mistuned components with combination tones produced by the ear. As advanced by Moore, Peters, and Glasberg (1985), because the combination tones occur at a level at least 15 dB below the level of the primary tones (Goldstein, 1967), the beats can be heard more distinctly if the levels of the mistuned components are reduced. In comparison to the flat spectral envelope, the $1/n^2$ envelope may have resulted in an increase in the modulation depth due to the interaction of the primary tones with the combination tones, thus giving rise to more distinct beats. This beating would then be used as a detection cue for the perceived inharmonicity.

For the attack times, the results showed that confusions increased with increasing curvature value. If listeners were detecting the perceptual attack time (PAT) according to a single threshold criterion such as the one proposed by Vos and Rasch (1981) and thus ignoring the shape of the amplitude envelope during the attack, then the results would have shown less confusions with increasing curvature value because the PATs of exponential envelopes occur later and are spaced farther apart than the PATs of the linear amplitude envelopes.

These results indicate that listeners were not using just a single amplitude threshold for ordering the attack time. They were also using a cue of how fast this threshold is reached with respect to the onset time, which is related to the shape of the amplitude envelope during the attack time and the importance of which is also underlined in Gordon's (1987) model.

There was a striking difference between the number of confusions that occurred in ordering the attack times with ramped linear amplitude envelopes versus the decay times with damped linear envelopes (4 versus 10 confusions, Table 3.5). This result provides further evidence for a temporal asymmetry in the auditory system(Moore, Glasberg, Plack, & Biswas, 1988; Patterson & Irino, 1998). One explanation for these results could be that damped sounds have been shown to be perceived as being shorter when compared to ramped sounds of equal duration (Schlauch, Ries, & DiGiovanni, 2001). Therefore, if sounds with short attack times (and a fixed decay) are poorly discriminated, then it can be expected to observe more confusions between sounds having the same short decay times while keeping the attack time fixed. However, this explanation does not seem to fully apply in these data mainly because, the number of confusions between damped and ramped sounds was about the same for the curvature value of $c = 3$ (in fact, there was less confusion for the damped sounds: 5 versus 6, Table 3.5). Most importantly, the overall amount of confusion between attack and decay times cannot be attributed to perceived duration because Schlauch et al. (2001) also found a weak coupling between subjective duration and duration discrimination between ramped and damped sounds. Furthermore, the subjective duration effects were controlled for by restricting the amplitude envelope manipulations above –40 dBFS.

Stecker and Hafter (2000) proposed a cognitive explanation for context effects of the temporal asymmetry on loudness judgments between ramped and damped sounds, which was also supported by Moore (2013). According to their viewpoint, sounds with damped envelopes are perceived to have two different segments, one being associated with the *source* sound and the other one with its *reverberant tail*. In the present study, the decay times with curvature values of $c = 3$ caused half the number of confusions when compared to the decay times with linear envelopes, which is opposite to the effect observed for attack times, for which the confusions increased with increasing curvature value (Table 3.5). If a damped sound is perceived to be a unitary and unmodulated source (Patterson, 1994), at least for short decay times, then according to Stecker and Hafter's (2000) explanation, the listeners in the present experiment must have been making judgments on the durations of the reverberant tails and the attributed reverberant quality due to context effects when ordering stimuli with a short and fixed attack time but of different decay times. In terms of attributing a reverberant quality to stimuli constructed with linear decay envelopes ($c = 1$) and with extreme exponential envelopes ($c = 10$), when the envelopes are viewed on a dB scale rather than in linear amplitude units, it becomes apparent that both envelops inhibit the reverberant tail of the stimulus: in the former case ($c = 1$), the source part is extended in time causing the stimulus to sound unreverberated, whereas in the latter case ($c = 10$),

the reverberant tail has faded out too quickly to be perceived and therefore, the stimulus sounds unreverberated as well. To the contrary, the exponential decay envelope with $c = 3$ leads to an almost linear decrease in dB per unit time, which enhances the difference in duration between the source sound and its reverberant tail and makes the discrimination between the tails of the stimuli easier to perceive. This may explain why decaying sounds with a curvature value of $c = 3$ were ordered more accurately when compared to linearly decaying envelopes that caused confusions comparable to those of the sounds with $c = 10$, in which the reverberant tail faded out too quickly to be perceived.

Surprisingly, listeners performed well in ordering the differences between temporal centroids even at very short attack times with one to two confusions happening mainly between the middle curvature values of each sound set; no confusions were found for the longest attack time of 500 ms (Table 3.6). The fact that the number of confusions did not monotonically decrease with increasing attack time indicates that the judgments of the sound sets with short attack times were unrelated to the judgments of the sound sets with longer attack times. It is hypothesized that in the former case listeners were making judgments based on the spectral differences between stimuli, which were caused by the shapes of the short amplitude envelopes, whereas in the latter case they performed the ordering task by judging the temporal evolution of the stimuli during the attack time.

For the temporal centroids of decay times, the amount of confusion for the shortest decay times was much higher than that observed for corresponding attack times and occurred between the first few highest curvature values, which caused the reverberant tail to be heavily suppressed. According to the previous discussion, this indicates that the highest temporal centroids of short decay times had little perceptual effect on the source part of the stimulus when compared to the temporal centroids of short attack times, and that listeners were again using the temporal evolution of the reverberated part as a cue. Also, the fact that there were no confusions between the temporal centroids of longer decay times ($> 68$ ms) indicates that listeners could track the temporal evolution of the decay more accurately than that of the attack.

In conclusion, listeners were able overall to order the stimuli given the presented spacing of feature values. The interpretation of the reported results underpinned the importance of the spectral envelope shape in judgments of inharmonicity, and the shape of the amplitude envelope when ordering attack and decay times. Whereas there were many confusions in ordering short attack and decay times, listeners performed very well in ordering temporal centroids constructed with different amplitude envelopes of fixed duration. Based on these results, in the following section a meta-analysis is presented on the dissimilarity ratings of previous timbre studies, in which the explanatory power of attack time, which has been identified from past research as a primary temporal perceptual dimension, is compared to that of the attack temporal centroid (ATC).

## 3.3 The explanatory power of Attack Temporal Centroid in a Meta-Analysis of Timbre Spaces

The early studies of Grey (1977), and Grey and Gordon (1978) interpreted qualitatively the resultant dimensions of the MDS analysis. Grey (1977) associated the first dimension with the overall "energy distribution" and features similar to spectral centroid and spectral spread. The second dimension was related with spectrotemporal features such as "spectral fluctuation" and the synchronicity in the attack and decay times of individual harmonics. The third dimension was also related to the temporal patterns of the spectral distribution, and the presence of inharmonicity occurring during the attack. Grey and Gordon (1978) gave similar qualitative interpretations of their three-dimensional MDS solution to Grey's (1977), but also offered a quantitative interpretation for the purely "spectral" axis which correlated most strongly with the spectral centroid weighted by the loudness function of Zwicker and Scharf's (1965) model.

Later studies attempted to interpret the MDS dimensions with quantitative measures. Iverson and Krumhansl's (1993) MDS analysis on truncated instrumental tones with durations of 80 ms resulted in two dimensions, one of which was correlated with attack time. McAdams et al. (1995) found a three-dimensional timbre space in which one of the axes correlated strongly with log-attack time. Similarly, Lakatos (2000) found a two-dimensional timbre space of harmonic and non-percussive sounds in which one of the dimensions correlated strongly with log-attack time. Although Grey (1977), and Grey and Gordon (1978) did not directly associate any of their MDS dimensions with attack time, they interpreted two of the axes according to spectrotemporal features that occur *during* the attack. The rest of the abovementioned studies directly associated one dimension of the resultant timbre space with attack time. Based on the results reported in the previous section, according to which short attack times were poorly discriminated, one might argue that the observed correlations between attack time and a given MDS dimension might have been coincidental: if one of the MDS dimensions relates to a temporal feature but listeners cannot reliably judge differences between similar attack times, then there must have been another feature that listeners used in their dissimilarity ratings and which causally relates to the temporal dimension of a given timbre space. To test this hypothesis, a meta-analysis on the aforementioned timbre spaces was conducted in which the explanatory power of attack time was compared to that of the ATC.

### 3.3.1 Methods

The stimuli and dissimilarity ratings from the previous studies were available in the lab and had been previously analyzed with the same MDS algorithm for a different research project (McAdams & Giordano, 2006). More precisely, the dissimilarity ratings were analyzed with the extended CLASCAL algorithm following the procedures described in McAdams et al. (1995). Iverson and Krumhansl (1993) and Lakatos (2000) used different sets of stimuli

throughout their experiments. For the purposes of the current study, only the stimulus set of "complete tones" and "onsets" used in Iverson and Krumhansl's (1993) study, and the "harmonic set" of Lakatos's (2000) study were analyzed. The rest of the aforementioned studies used only a single stimulus set for deriving the timbre spaces, which were reanalyzed here. The results of the present analysis are expected to be different than the results reported in the previous studies not only because of the different MDS algorithms used but also due to differences in the computation of attack time[1].

For some stimuli, especially for non-synthetic and recorded sounds, computing the attack time based on the absolute maximum value of the waveform resulted in very long attack times on the order of hundreds of milliseconds, which were way longer than the previously reported values. Although these overestimations could be attributed to playing techniques (e.g., long vibratos) or miking artifacts, it remains unclear how the previously reported values were derived when these artifacts had been taken into account. A more robust alternative was to compute the attack times and their respective ATC through the squared amplitude envelope (i.e., the power envelope) of the waveform. The amplitude envelopes of the stimuli were extracted by computing the root-mean-squared (RMS) values of the positive and negative parts of the waveform, over a sliding window approximately equal to the period of the $f_0$, and then averaging the two parts together. Power envelopes were preferred over amplitude envelopes because the sum of squared amplitude values used in the calculation of temporal centroid relates in physical terms to the energy of the signal, and also because power envelopes were shown to correlate more strongly than the amplitude envelopes with PAT in Gordon's (1987) study. The envelopes were then normalized relative to their maximum values (0 dBFS) and were truncated below a threshold of –60 dBFS to avoid including background noise and the silent segments of some sound files occurring before the beginning of each tone. The ATC was computed by Eq. (3.1) where $N$ was confined to the attack time (in samples).

Vos and Rasch (1981) proposed that the PAT depends on presentation level and can be estimated using a threshold of about 6 to 15 dB below the maximum level of the tone, whereas Gordon's (1987) model used a much lower threshold of 22 dB. In order to relate the results of this analysis not only to attack time but also to PAT and its respective ATC, the following thresholds for computing PATs were also used (in dB): {6, 10.5, 15, 22}. In this case, the corresponding ATC was computed according to Eq. (3.1), but $N$ was confined to the duration of each PAT. The attack times, PATs, and their respective ATCs of each stimulus set were then correlated with each axis of the MDS analysis. A *percentile bootstrap* (Efron & Tibshirani, 1993) was used to test the statistical significance between the differences of the correlation coefficients of attack time or PATs, and the respective ATCs. The .95 confidence intervals were adjusted using Wilcox and Muska's (2001) correction,

[1]For McAdams et al.'s (1995) study, only the ratings of 24 out of 98 participants, who were assigned to the group of "professional musicians", were used in the meta-analysis in order to compensate for the fact that in the rest of the studies all participants were musically sophisticated.

which is suitable for small sample sizes.

### 3.3.2 Results and Discussion

The number of MDS dimensions using the extended CLASCAL algorithm in the meta-analysis was the same with the original studies. Table 3.7 summarizes the results for the MDS dimension of each study that correlated significantly ($p < 0.05$) with attack time and its temporal centroid. With the exception of the dimension for Iverson and Krumhansl's (1993) "onsets" sound set, the rest of the MDS dimensions of other studies correlated more strongly with the logarithm of the attack time and its temporal centroid. Although in general, PATs correlated more strongly than attack time [except for the Lakatos's (2000) study], the thresholds at which PAT exhibited the strongest correlations were different across studies.

In almost all cases, the ATCs (computed over the attack time and PATs) correlated with the MDS dimensions equally or stronger than the attack times and PATs *per se*. Although the sample sizes were relatively small, ranging from 16 to 18 sounds across studies, the differences between the ATCs' and attack times' or PATs' bootstrapped correlation coefficients were significant ($p < 0.05$) in six out of 30 cases (Table 3.7). In cases of sounds with short attack times, the corresponding PATs are of even shorter duration (Fig. 3.1). According to the discussion of the previous section, it is unlikely that listeners would have been able to discriminate between those sounds if their dissimilarity ratings were only based on that feature. The observed correlations indicate that when they are significantly different, the ATC is a more robust feature than attack time for explaining dissimilarity ratings between instrumental sounds. That is of course only true if the shape of the amplitude envelopes between stimuli is sufficiently different, otherwise the ATC has the same explanatory power as the attack time. In cases of sounds with similar but longer attack times, in which their amplitude envelopes usually differ (at least for instrumental sounds), and assuming that attack time is perceived on a logarithmic scale, it is expected that the attack temporal centroid would have a greater explanatory power than attack time per se, because it is related to the overall shape of the amplitude envelope during the attack time.

## 3.4 Conclusions

The ordinal scaling experiment showed that in general listeners could rank order stimuli with varying attack and decay times, temporal centroids of differently shaped amplitude envelopes with fixed attack and decay times, and inharmonicity levels. Furthermore, the trend analysis confirmed a linear trend of the rankings with increasing parameter values. The shape of the spectral envelope was an important parameter when ordering inharmonicity levels, because there were less confusions in the stimulus set constructed with $1/n^2$ spectral envelope compared to the set constructed with flat spectral envelope at a $f_0$ of 720 Hz. The

**Table 3.7** Significant correlations ($p < 0.05$) and *percentile bootstrap* analysis of attack time (AT), perceptual attack time (PAT) with the threshold used for each computation given inside parenthesis, and attack temporal centroid (ATC), with MDS dimensions. Grey: Grey's (1977) sound set; GreyGor: Grey and Gordon's (1978) sound set; IvKrOn, IvKrWh: Iverson and Krumhansl's (1993) "onset" and "whole" sound sets, respectively; Lakatos: Lakatos's (2000) "harmonic set"; McAdams: McAdams et al.'s (1995) sound set; $r$: correlation coefficient between a particular MDS dimension with AT or PAT; $r_{ATC}$: correlation coefficient between a particular MDS dimension with ATC computed up to AT or the corresponding PAT; $r^*$ and $r^*_{ATC}$: correlation coefficients computed on log-transformed values of AT, PAT and ATC; diffCI: 95% CI of the bootstrapped differences between $r$ and $r_{ATC}$; $p$: probability of the absolute difference between $r$ and $r_{ATC}$ being zero. The $p$-value of significant correlations ($p < 0.05$) are indicated in boldface; –: nonsignificant correlation ($p > 0.05$).

| Sound sets | Statistics | AT | PAT(–6 dB) | PAT(–10.5 dB) | PAT(–15 dB) | PAT(–25 dB) |
|---|---|---|---|---|---|---|
| Grey | $r^*$ | 0.70 | 0.73 | 0.73 | 0.76 | 0.74 |
| | $r^*_{ATC}$ | 0.72 | 0.72 | 0.74 | 0.76 | 0.74 |
| | diffCI | (–0.07, 0.02) | (–0.05, 0.03) | (–0.05, 0.03) | (–0.05, 0.04) | (–0.05, 0.05) |
| | $p$ | 0.23 | 0.47 | 0.63 | 0.83 | 0.92 |
| GreyGor | $r^*$ | – | 0.58 | 0.60 | 0.62 | 0.65 |
| | $r^*_{ATC}$ | – | 0.60 | 0.61 | 0.63 | 0.66 |
| | diffCI | – | (–0.04, 0.01) | (–0.04, 0.02) | (–0.04, 0.03) | (–0.05, 0.04) |
| | $p$ | – | 0.06 | 0.39 | 0.69 | 0.72 |
| IvKrOn | $r$ | – | 0.54 | 0.56 | 0.55 | 0.59 |
| | $r_{ATC}$ | – | 0.58 | 0.59 | 0.57 | 0.60 |
| | diffCI | – | (–0.13, 0.01) | (–0.10, 0.02) | (–0.10, 0.03) | (–0.06, 0.03) |
| | $p$ | – | 0.07 | 0.16 | 0.25 | 0.57 |
| IvKrWh | $r^*$ | 0.59 | 0.68 | 0.65 | 0.60 | 0.61 |
| | $r^*_{ATC}$ | 0.61 | 0.68 | 0.65 | 0.61 | 0.61 |
| | diffCI | (–0.07, 0.00) | (–0.06, 0.03) | (–0.06, 0.05) | (–0.07, 0.06) | (–0.05, 0.05) |
| | $p$ | **0.01** | 0.39 | 0.69 | 0.99 | 0.97 |
| Lakatos | $r^*$ | 0.61 | 0.51 | 0.48 | 0.52 | 0.49 |
| | $r^*_{ATC}$ | 0.62 | 0.54 | 0.51 | 0.54 | 0.52 |
| | diffCI | (–0.02, 0.05) | (–0.01, 0.05) | (0.00, 0.07) | (–0.01, 0.09) | (–0.01, 0.10) |
| | $p$ | 0.48 | 0.10 | **0.03** | 0.12 | 0.08 |
| McAdams | $r^*$ | 0.68 | 0.75 | 0.75 | 0.79 | 0.83 |
| | $r^*_{ATC}$ | 0.73 | 0.80 | 0.79 | 0.82 | 0.84 |
| | diffCI | (0.01, 0.10) | (0.01, 0.10) | (0.01, 0.10) | (0.00, 0.07) | (–0.01, 0.05) |
| | $p$ | **0.00** | **0.00** | **0.00** | **0.04** | 0.24 |

shape of the amplitude envelope was found to be an important parameter when ordering the stimulus sets of attack and decay times. For attack time, listeners' confusions increased with increasing curvature value, which produced linear, exponential and "extreme" exponential amplitude envelopes. For decay time, the sound set constructed with an exponential envelope caused the lowest confusion in the ordering task because according to Stecker and Hafter's (2000) cognitive explanation, both the linear and extreme exponential decaying envelopes used in this study suppress the perception of the reverberant tail of the stimulus (when viewed on a dB scale), and lead to similar percepts.

Although there were many confusions in the ordering of short attack and decay times, listeners performed very well in ordering the temporal centroids with fixed attack and decay times, which indicates that they were sensitive to the differences between the shapes of the amplitude envelopes even at very short durations. However, for short durations (e.g., for attack times of 40 or 67 ms), we hypothesize that these differences perceptually manifest as spectral, because as advanced by Hartmann and Wolf (2009), there are cases in which the amplitude envelope contributes importantly to the power spectrum of the signal presented to a listener, especially for stimuli of short duration. This hypothesis is also supported by the fact that the confusions between temporal centroids did not decrease with increasing attack time, which led to the conclusion that at short attack times listeners' judgments were based on the spectral differences between stimuli, whereas for longer attack times their judgments were based on tracking the temporal evolution of the amplitude envelopes. The results also indicate that listeners could track the temporal evolution of the envelopes during the decay more accurately than during the attack time because there were no confusions between temporal centroids at decay times above 68 ms.

In the ordinal scaling experiment, many confusions were observed between attack times occurring at least below 133 ms, a value which encompasses the attack time of most instrument sounds, but good performance in ordering temporal centroids even at very short attack times was found. A meta-analysis of six timbre spaces was therefore conducted in which the explanatory power of attack time was compared to that of the temporal centroid of the attack. The analysis showed that the ATC correlated more strongly with a given MDS dimension than did attack time or PATs per se, and that in a few cases the differences between the two correlation coefficients were significant. The meta-analysis of timbre spaces indicates that ATC is a robust feature for explaining dissimilarity ratings along a perceptual dimension. However, based on the discussion of the previous paragraph, this dimension might be referred to as *spectrotemporal* rather than *temporal* because differences between short ATCs most likely relate to perceived spectral differences, whereas longer ATCs relate to perceived differences between the temporal evolutions of the respective amplitude envelopes *during* the attack time.

# References


Barr, D. J., Levy, R., Scheepers, C., & Tilly, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang*, *68*, 255–278.

Beasley, M. T., & Zumbo, B. D. (2009). Aligned rank tests for interactions in split-plot designs: Distributional assumptions and stochastic heterogeneity. *J. Mod. Appl. Stat. Methods*, *8*, 16–50.

Chaigne, A., & Doutaut, V. (1997). Numerical simulations of xylophones. I. Time-domain modeling of the vibrating bars. *J. Acoust. Soc. Am.*, *101*, 539–557.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. New York: Chapman and Hall.

Farbood, M. M., & Price, K. C. (2017). The contribution of timbre attributes to musical tension. *J. Acoust. Soc. Am.*, *141*, 419–427.

Fletcher, N. H. (1964). Normal vibration frequencies of a stiff piano string. *J. Acoust. Soc. Am.*, *36*, 203–209.

Galembo, A., Askenfelt, A., Cuddy, L. L., & Russo, F. A. (2004). Perceptual relevance of inharmonicity and spectral envelope in the piano bass range. *Acta Acust.*, *90*, 528–536.

Goldstein, J. L. (1967). Thresholds for the detection of inharmonicity in complex tones. *J. Acoust. Soc. Am.*, *41*, 676–689.

Gordon, J. W. (1987). The perceptual attack time of musical tones. *J. Acoust. Soc. Am.*, *82*, 88–105.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, *61*, 1270–1277.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am.*, *63*, 11493–1500.

Hartmann, W. M., McAdams, S., & Smith, B. K. (1990). Hearing a mistuned harmonic in an otherwise periodic complex tone. *J. Acoust. Soc. Am.*, *88*, 1712–1724.

Hartmann, W. M., & Wolf, E. M. (2009). Matching the waveform and the temporal window in the creation of experimental signals. *J. Acoust. Soc. Am.*, *126*, 2580–2588.

Hjortkjær, J., & McAdams, S. (2016). Spectral and temporal cues for perception of material and action categories in impacted sound sources. *J. Acoust. Soc. Am.*, *140*, 409–420.

Hoffman, W. (1989). Iterative algorithms for Gram-Schmidt orthogonalization. *Computing*,

*41*, 335–348.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, *6*, 65–70.

ISO 389-8. (2004). *Acoustics – Reference Zero for the Calibration of Audiometric Equipment – Part 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones.* (Technical report (International Organization for Standardization, Geneva, Switzerland))

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *J. Acoust. Soc. Am.*, *94*, 2595–2603.

Järveläinen, H., & Karjalainen, M. (2006). Perceptibility of inharmonicity in the acoustic guitar. *Acta Acust. United Ac.*, *92*, 842–847.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Percept. Psychophys.*, *62*, 426–1439.

Martin, F. N., & Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *J. Am. Acad. Audiol.*, *11*, 64–66.

McAdams, S., Chaigne, A., & Roussarie, V. (2004). The psychomechanics of simulated sound sources: Material properties of impacted bars. *J. Acoust. Soc. Am.*, *115*, 1306–1320.

McAdams, S., & Giordano, B. L. (2006). Generalizing timbre space data across stimulus contexts: The meta-analytic approach. *J. Acoust. Soc. Am.*, *119*, 3395.

McAdams, S., Roussarie, V., Chaigne, A., & Giordano, B. L. (2010). The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *J. Acoust. Soc. Am.*, *128*, 1401–1413.

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

McCullagh, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc. Series B*, *42*, 109–142.

Moore, B. C. J. (2013). Revisiting the loudness of sounds with asymmetric attack and decay. *J. Acoust. Soc. Am.*, *134*, 4195.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.*, *74*, 750–753.

Moore, B. C. J., Glasberg, B. R., & Baer, T. (1997). A model for the prediction of thresholds, loudness and partial loudness. *J. Audio Eng. Soc.*, *45*, 224–240.

Moore, B. C. J., Glasberg, B. R., Plack, C. J., & Biswas, A. K. (1988). The shape of the ear's temporal window. *J. Acoust. Soc. Am.*, *83*, 1102–1116.

Moore, B. C. J., Peters, R. W., & Glasberg, B. R. (1985). Thresholds for the detection of inharmonicity in complex tones. *J. Acoust. Soc. Am.*, *77*, 1861–1867.

Patterson, R. D. (1994). The sound of a sinusoid: Time-interval models. *J. Acoust. Soc. Am.*, *96*, 1419–1428.

Patterson, R. D., & Irino, T. (1998). Modeling temporal asymmetry in the auditory system. *J. Acoust. Soc. Am.*, *104*, 2967–2979.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.*, *130*, 2902–2916.

Schlauch, R. S., Ries, D. T., & DiGiovanni, J. J. (2001). Duration discrimination and subjective duration for ramped and damped sounds. *J. Acoust. Soc. Am.*, *109*, 2880–2887.

Smith, B. K. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. In *Proceedings of the meeting of the society for music perception and cognition.* Berkeley, CA: University of California, Berkeley.

Stecker, G. C., & Hafter, E. R. (2000). An effect of temporal asymmetry on loudness. *J. Acoust. Soc. Am.*, *107*, 3358–3368.

Vos, J., & Rasch, R. (1981). The perceptual onset of musical tones. *Percept. Psychophys.*, *29*, 323–335.

Wilcox, R. R., & Muska, J. (2001). Inferences about correlations when there is heteroscedasticity. *Brit. J. Math. Stat. Psy.*, *54*, 39–47.

Zwicker, E., & Scharf, B. (1965). A model of loudness summation. *Psych. Rev.*, *72*, 3–26.

# Chapter 4

# Interval and ratio scaling of spectral audio descriptors

This chapter is based on the following research article:

Kazazis, S., Depalle, P. and McAdams, S. (in preparation). Interval and ratio scaling of spectral audio descriptors. Manuscript prepared for submission to *Frontiers in Psychology*.

**Abstract** Two experiments were conducted for the derivation of psychophysical scales of the following audio features: spectral centroid, spectral spread, spectral skewness, odd-to-even harmonic ratio, spectral deviation, and spectral slope. The stimulus sets of each audio feature were synthesized and (wherever possible) independently controlled through appropriate synthesis techniques. Partition scaling methods were used in both experiments, and the scales were constructed by fitting well-behaving functions to the listeners' ratings. In the first experiment, the listeners' task was the estimation of the relative differences between successive levels of a particular audio feature. The median values of listeners' ratings increased with increasing feature values, which confirmed listeners' abilities to estimate intervals. However, there was a large variability on the reliability of the derived interval scales depending on the stimulus spacing in each trial. In the second experiment, listeners had control over the stimulus values and were asked to divide the presented range of values into perceptually equal intervals, which provides a ratio scale. For every feature, the reliability of the derived ratio scales was excellent. With the exception of spectral centroid, for which the zero point of the scale was assigned empirically, the rest of the scales were assigned a zero point that also has a physical meaning. The unit of a particular scale was assigned empirically as well, so as to facilitate comparisons between the derived perceptual ratio scales of all audio features. The construction of psychophysical scales based on univariate stimuli, allowed for the establishment of cause and effect relations between audio features and perceptual dimensions, contrary to past research that has relied on multivariate stimuli and has only examined the correlations between the two.

## 4.1 Introduction

Audio features have been widely used in timbre research for explaining quantitatively the dimensions of timbre spaces (Grey & Gordon, 1978; Iverson & Krumhansl, 1993; Lakatos, 2000; McAdams, Winsberg, Donnadieu, Soete, & Krimphoff, 1995), affective ratings (Farbood & Price, 2017; Laurier, Lartillot, Eerola, & Toivianen, 2009; McAdams, Douglas, & Vempala, 2017), and the perceptual similarity of short music clips (Siedenburg & Müllensiefen, 2017). Most often, the spectral features are derived from statistical computations on a spectrogram, whereas the temporal features are usually extracted from the raw waveform. The time-series of these features, derived from a frame-by-frame analysis on the spectrogram, are then compressed through summary statistics into single numbers, which are presumed to serve as the spectrotemporal imprint of a stimulus and thereby designate its timbre. The systematic development of such features started with the work of Krimphoff, McAdams, and Winsberg (1994) for explaining quantitatively the perceptual dimensions of Krumhansl's (1989) timbre space. The set of features started expanding with the appearance of the MPEG-7 standard (ISO/IEC, 2002)according to which the audio features would be termed as *audio descriptors*.

Timbre research has mainly relied on correlational analysis between audio features and listeners' perception, and there have only been a few attempts for establishing causal relations between psychological and acoustic dimensions. One of these attempts is the confirmatory study of Caclin, McAdams, Smith, and Winsberg (2005), who validated with synthesized stimuli the saliency of attack time, spectral centroid, and the odd-to-even harmonic ratio, but not spectral "flux" for explaining dissimilarity ratings. Another study is from Almeida, Schubert, Smith, and Wolfe (2017) who attempted with synthesized stimuli to derive a ratio scale of brightness as a function of spectral centroid, albeit within a limited range of 1.46 octaves, and at a single fundamental frequency of 500 Hz. In fact, neither of the aforementioned studies evaluated directly the spectral centroid, but rather the spectral slope, which co-varies with spectral spread, skewness, and under certain circumstances is linearly dependent on spectral centroid. More recently, Kazazis et al. (in preparation) validated through ordinal scaling with synthesized stimuli several audio features by controlling each spectral feature independently of the rest, thus isolating the effect that each feature had on the stimulus rankings. The results of those experiments have served as the basis for the present study, because there was no prior evidence that stimuli varying along a particular audio feature could be perceived on an ordinal scale, the existence of which is a prerequisite for constructing perceptual interval and ratio scales.

Different experimental procedures are needed for testing different scales: an ordinal scale indicates whether listeners are able to rank order the stimuli; an interval scale, whether they can judge the relative size of intervals between stimuli; and, a ratio scale, whether ratios between stimuli can be perceived. However, the most informative scale is the ratio scale, which satisfies all the criteria of an interval scale, but also enables the derivation of ratios between stimuli. In other words, the ratio scale subsumes the interval scale and the

experimental procedure should devise operations for determining the following relations among stimuli (Stevens, 1946): equality; rank order; equality of intervals; and, equality of ratios.

There are some important methodological considerations that need to be taken into account before designing experiments for deriving either interval or ratio scales of audio features. Interval and ratio scaling methods are part of "global psychophysics" rather than "local psychophysics", where the aim is usually to derive just noticeable differences (JNDs) among stimuli, which do not predict the results of global psychophysical experiments. A psychophysical experiment is said to be global if the extreme stimuli of a stimulus set are almost perfectly identified in a 2-stimulus absolute identification design (Luce & Krumhansl, 1988). This has certain implications in the construction and selection of appropriate stimuli, which will be discussed further below. The methods used for constructing psychophysical interval and ratio scales, based on direct estimations of subjective magnitudes, can be classified in two general categories: *magnitude estimation* (or *production*), where listeners are instructed to assign numbers of their choice to stimuli so as to reflect subjective ratios in relation to a reference stimulus (or *standard*), which is usually located in the middle of the presented range of values; and, *category Scaling* (or difference estimation) where the lower and upper limits of the response scales are defined, and listeners are instructed to assign scale values along the continuum between the extremes so as to preserve subjective differences (or psychological distances) between stimuli. Irrespective of the method used, in most scaling experiments, the physical attributes of study can be easily explained and identified by the listeners and are often associated with a perceptual correlate such as loudness or pitch. One of the issues and challenges that arise in psychophysical scaling of audio features is that the experimenter cannot explain with clarity and in a simple manner the attribute of study to the listeners, without resorting to a purely technical formulation of a particular audio feature, and which in most cases will not be understandable by "naïve" participants (e.g., musicians without a physics background). In Stevens's (1946) terms, this difficulty often arises because audio features are measured on *derived* physical scales constructed by mathematical functions of certain magnitudes derived from *fundamental scales* for which a perceptual correlate can be more easily found (e.g., loudness for intensity). In the present case which deals with audio features, the fundamental scales are represented by fundamental frequency, and the amplitude-frequency pairs of spectral components.

### 4.1.1 The Present Study

In a first attempt to derive ratio scales, we designed a pilot experiment based on magnitude estimation, in which the largest effects are produced by the range of stimuli, the distance from threshold (if a threshold exists for a particular feature), and the degree of freedom given to listeners for choosing the lowest and the highest number for their responses (Poulton, 1968). The standard was positioned in the middle of a particular stimulus set and listeners were limited to one judgement per stimulus, which reduces the biases due to range

and spacing of stimuli (Stevens, 1971). In each trial, listeners were presented with a single stimulus along with the standard (reference stimulus) and were instructed to assign a number of their choice that reflected the subjective magnitude ratio between the two. Throughout the trials, the standard was positioned always in the middle of a particular stimulus set, and the presentation of just a single stimulus in each trial (instead of asking listeners to make simultaneous judgments on two or more stimuli with respect to the standard) was done in order to reduce the biases due to range and spacing of stimuli (Stevens, 1971). In addition, there were no limitations on the available set of numbers used as responses for both the stimulus and the standard other than the requirement of being positive, because any such limitations have shown to increase the bias in magnitude estimation experiments (Hellman & Zwislocki, 1961). We consider this experiment to be unsuccessful, because the sensation magnitudes were poorly apprehended which can be mainly attributed to the experimental design. First of all, for some audio features, the stimuli were hardly discriminable leading to poor independent judgments when each stimulus was presented in isolation from the rest. In some other cases, stimuli were very distant from the standard, which is considered to be a source of bias in magnitude estimation experiments, because although stimuli near to the standard are judged relative to the standard, stimuli far from the standard are not (Gescheider & Hughson, 1991). Most importantly, listeners were not given any indication of which attribute they were judging between presented stimuli other than the written instruction "... according to any criteria that differentiate them the most." Due to the nature of the stimuli and because of the reasons related to the psychophysical scaling of audio features described above, the attribute of study could only have been identified by the listeners if the experimental design had allowed for the discovery of invariances among stimuli through the exploration of a particular stimulus set, instead of a presentation of stimuli in isolation. Finally, experiments based on magnitude estimation place a heavy load on listener's memory and given the unfamiliarity of the listeners with the presented stimuli, this was considered to be an additional reason for which the magnitudes in this experiment were poorly apprehended. This last implication could have been avoided if the method of *absolute magnitude estimation* (Hellman & Zwislocki, 1961) had been used instead, in which listeners match numbers to stimuli without the presentation of a standard, and independently of the previous matches. However, an experimental design based on this method would not have been able to overcome the above-mentioned hurdles and provide positive results, mainly because it is difficult to make absolute judgments on timbral attributes.

In a second pilot experiment, the listeners' task was the same with the experiment described in the beginning of the previous paragraph except that in this one, in each trial listeners were making simultaneous judgments with respect to a standard on the whole stimulus set of a particular audio feature. As in the previous experiment, they could assign numbers of their choice to both the standard and the stimuli. From a methodological point of view, this design is a compromise between category scaling (or difference estimation) and

magnitude estimation, but it has also been shown to lead to a compromise between the derived scales of the two (Montgomery, 1975). Although this design allowed listeners to explore a range of stimuli, and thus identify the attribute of study more effectively than the previous experiment, this design had its own pitfalls. The most important bias resulted from the spacing of stimuli, which magnitude estimation methods aim to control for by restricting listeners to perform one judgment per stimulus. In other words, the derived scales may not be generalizable, in a sense that a different spacing of stimulus values might have led to different scales. However, there were also some practical issues, which limited the credibility of the results. Given that the order of presentation was random (i.e., stimuli were not presented in ascending or descending series of physical magnitude), and the large number of stimuli presented in a single trial, some listeners might only have focused on rank ordering, when verifying their judgments by listening to the stimuli sequentially and indexed according to their magnitude estimations, instead of making more accurate judgments between the stimuli and the standard. In addition, some listeners complained about the difficulty of the task, and reported that they were performing comparisons across all the presented stimuli, although it was clearly stated in the instructions that the comparisons should only be performed between a single stimulus and the standard. Such operations considerably increased the cognitive load of the listeners, which might have had a strong impact on the accuracy of their judgments.

Because of all the above-mentioned issues of each experimental design, we resorted to partition scaling methods (Stevens, 1975) for constructing both interval and ratio scales of audio features. These methods have been successfully used in the past, such as for the derivation of the Mel scale for pitch (Stevens & Volkmann, 1940). In the following sections, we present the experiments used to derive psychophysical scales of the following audio features: spectral centroid, spectral spread, spectral skewness, odd-to-even harmonic ratio, spectral deviation, and spectral slope. A mathematical formulation along with a detailed description of each feature can be found in (Peeters, Giordano, Susini, Misdariis, & McAdams, 2011). This manuscript is organized as follows. In Section 4.2 we present the synthesis processes used to construct each audio feature's stimulus set and the experiment for the derivation of interval scale measurements. Given the lack of previous knowledge of JNDs on audio features, this experiment could be considered as a confirmatory experiment on whether listeners are actually able to perceive intervals before proceeding to the next experiment with its additional operations needed for deriving ratio scales (described in Section 4.3 ). Finally, in Section 4.4 we present concluding remarks on the validity of the obtained results and implications for timbre perception.

## 4.2 Experiment 1: Interval Estimation

The aim of this experiment was to investigate whether listeners perceive intervals of audio features and the construction of interval scales. The listeners' task was the estimation of

the relative differences between successive levels of a particular audio feature, and thus this experiment provided interval scale measurements.

### 4.2.1 Method

#### Participants

Twenty-five participants, 11 female, 13 male, and 1 "prefer not to answer", with a median age of 23 years (range: 18–40) were recruited from the Schulich School of Music, McGill University. All of them were self-reported amateur or professional musicians with formal training in various disciplines such as performance, composition, music theory, and sound engineering. Participants were compensated for their time. One participant fallen asleep during the trials but after taking a short break he was able to complete the experiment. The study was certified for ethical compliance by the McGill University Research Ethics Board II. Before the experiment, participants had to sign an informed consent form. Afterwards, they passed a pure-tone audiometric test at octave-spaced frequencies from 125 Hz to 8 kHz (ISO 389-8, 2004; Martin & Champlin, 2000) and were required to have thresholds at or below 20 dB HL to proceed to the experiment.

#### Stimuli and Presentation

Several sets consisting of synthetic sounds were created by independently controlling the values of the above-mentioned features in the synthesis process. For spectral centroid, spread, and skewness, all the spectral manipulations were applied to a flat harmonic spectrum (harmonics set at equal amplitude) with a fundamental frequency ($f_0$) of 120 Hz and harmonics up to Nyquist limit. For the rest of the features, separate sound sets were synthesized at $f_0$s of 120, 300, and 720 Hz with the number of harmonics ranging from 9 to 16, depending on the feature. Following the spectral manipulations, the stimuli were synthesized in MATLAB version R2015b (The MathWorks, Inc., Natick, MA) using additive synthesis at a sampling frequency of 44.1 kHz with 16-bit amplitude resolution. The peak amplitude of the waveforms was normalized to 0.5 and the duration was set to 600 ms, gated with 10-ms raised-cosine ramps. All stimuli were loudness normalized according to the algorithm of Moore, Glasberg, and Baer (1997) as implemented in the Loudness Toolbox v.1.2 (Genesis S. A., 2009) and further adjusted by the authors because it was observed that the algorithm overestimated the loudness of sounds that had most of their energy centered at high frequencies.

For each feature, and for a particular spectral centroid or $f_0$, the stimuli were presented in three different sequences of feature values, under the constraint that the values of two successive stimuli should be different for each sequence. The spacing of stimulus values presented in each sequence was based on the results of Kazazis' et al. (in preparation) ordinal scaling experiment, in which listeners' confusions between successive stimuli were identified. This allowed for a supraliminal stimulus set within each sequence. In the

following subsections we present the synthesis methods used for the construction of stimuli that led to independent control of the values of each feature. Table 4.1 lists the ranges of all feature values computed on a number of sounds generated for a testing a particular feature. From this table can be inferred that within a particular stimulus set, most feature values remain relatively constant or vary within a very narrow range compared to the ranges of feature values according to which the stimulus set was generated. The most notable exceptions are the stimulus sets of spectral slope and odd-to-even ratio. This is because spectral slope covaries with spectral centroid, spread, and skewness and resulted in a greater skewness range than the stimulus set of spectral skewness due to the constraints imposed in the sound synthesis process that are outlined in the following subsections. In a similar way, the odd-to-even ratio is directly related to the computation of spectral deviation, and its respective stimulus set had a greater range of spectral deviation than the dedicated stimulus set of spectral deviation.

**Table 4.1** Ranges of feature values within designated stimulus sets. The ranges of feature values according to which each stimulus set was generated are shown in bold. The number of sounds on which the feature values were computed are shown inside parenthesis. The reported ranges for the spectral spread and skewness stimulus sets were computed on stimuli with 5600-Hz spectral centroid. -: linear regression over normally distributed spectral amplitudes is futile.

| Stimulus Sets (# sounds) | Feature Ranges | | | | | |
|---|---|---|---|---|---|---|
| | Centroid (Hz) | Spread (Hz) | Skewness | Odd-to-Even Ratio | Deviation | Slope (dB/octave) |
| Centroid (505) | **[1642, 9560]** | [479, 480] | [0.00, 0.02] | [1.00, 1.00] | [0.00, 0.00] | - |
| Spread (100) | [5600, 5600] | **[181, 1439]** | [0.00, 0.00] | [1.00, 1.00] | [0.00, 0.00] | - |
| Skewness (97) | [5600, 5600] | [1079, 1080] | **[–0.88, 0.96]** | [1.00, 1.00] | [0.00, 0.00] | - |
| Odd-to-Even Ratio (349) | [1260, 1500] | [768, 848] | [0.00, 0.21] | **[0.25, 1250.00]** | [0.00, 0.11] | [–11.67, –2.62] |
| Deviation (265) | [1723, 2550] | [1292, 1396] | [0.00, 0.28] | [1.00, 1.19] | **[0.00, 0.06]** | [0.00, –5.04] |
| Slope (349) | [332, 2082] | [134, 785] | [–1.04, 6.68] | [1.25, 15.05] | [0.00, 0.03] | **[–24.00, 5.44]** |

**Spectral Centroid** The stimuli of these sound sets were constructed by shaping the flat harmonic spectrum described above to follow a normal probability mass function that enabled the construction of spectra with different centroids (means), for a given spread (standard deviation) and zero skewness[1](Figure 4.1). The amount of spectral spread was set to 479 Hz (four times the $f_0$), which for the $f_0$ at 120 Hz allowed for a minimum centroid of 1640 Hz and a fixed bandwidth of 9 harmonics for each stimulus's spectrum. It should also be noted that the harmonic spacing of the components ensured a (virtual) pitch percept at the $f_0$. The spectral centroid values (in kHz) presented in each sequence (*Seq.*) were the following:

[1]The actual computed values (in discrete frequency) differ slightly from the theoretical values (calculated on continuous frequency) used in the synthesis processes due to round-off errors.

*Seq.1*: {1.64, 2.28, 3.68, 6.20, 9.56}
*Seq 2*: {1.64, 2.88, 3.68, 4.76, 9.56}
*Seq.3*: {1.64, 1.80, 3.68, 8.20, 9.56}

**Spectral Spread** The normal distribution was again used for constructing stimuli with fixed centroids, zero skewness and variable spectral spreads, by precisely controlling its bandwidth (Figure 4.1). The range of the allowable spreads in the synthesis process was constrained by the centroid and $f_0$ used in each stimulus set, as well as the spacing resolution of the harmonics. Three sound sets were constructed with centroids centered at 1640, 5600, and 7800 Hz that allowed for maximum spreads (with respect to the $f_0$) of 479, 1439, and 1800 Hz, respectively. For each of these sound sets (i.e., for each of the three spectral centroids), the spectral spread values (in Hz) where presented in the following three sequences:
*Seq.1*: {62, 96, 152, 241, 479}, {181, 287, 455, 722, 1439}, {227, 359, 569, 902, 1800},
for spectral centroids at 1640, 5600, and 7800 Hz respectively;
*Seq.2*: {62, 121, 191, 303, 479}, {181, 362, 573, 909, 1439}, {227, 452, 717, 1136, 1800}
*Seq.3*: {62, 76, 191, 381, 479}, {181, 228, 573, 1144, 1439}, {227, 285, 717, 1430, 1800}

**Spectral Skewness** The Skew-normal distribution (Azzalini, 2005) is a three-parameter family of curves and was employed for constructing stimuli with different skewness while the centroid and spread were being kept constant (Figure 4.1). The restrictions in the synthesis process that were taken into account with respect to the selection of centroids and spreads were similar to the ones mentioned above, with the additional constraint that skewness in the Skew-normal distribution can only vary within a range of $(-0.9953, 0.9953)$. Three sound sets were constructed with centroids spaced at 1640, 5600, and 7800 Hz, and spreads at 360, 1080, and 1439 Hz, respectively. The spectral skewness values presented in each sequence for a particular centroid were the following:
*Seq.1*: {–0.88, –0.33, 0.00, 0.71, 0.96}
*Seq.2*: {–0.88, –0.11, 0.25, 0.60, 0.96}
*Seq.3*: {–0.88, –0.55, 0.00, 0.87, 0.96}

**Odd-to-Even Ratio** The stimuli of these sound sets were constructed with 9 harmonics for ensuring that roughness would not be a major factor in listeners' ratings. The odd-to-even ratio was controlled by equally attenuating the level in dB of the even harmonics while keeping the odd harmonics fixed at 0 dBFS (dB relative to full scale), and by attenuating the level of the odd harmonics while keeping the even harmonics fixed. In each of those cases, the $f_0$ level was kept fixed at 0 dBFS (Figure 4.1). Three sound sets with the same

attenuation levels were constructed for each of the three $f_0$s at 120, 300, and 720 Hz. The odd-to-even ratio values presented in each sequence for a particular $f_0$ were the following:
*Seq.1*: {0.251, 0.648, 1.25, 3.14, 1250}
*Seq.2*: {0.251, 0.501, 1.25, 4.98, 1250}
*Seq.3*: {0.251, 0.881, 1.25, 1.98, 1250}

**Spectral Deviation** The reference spectrum was selected from a sample of one thousand amplitude distributions that were generated by randomly choosing the level of each harmonic from a uniform distribution covering the range of $[-60, 0]$ dBFS. The amplitude distribution that had the greatest spectral deviation along with an odd-to-even ratio of approximately 1 and the greatest T2 tristimulus value (Peeters et al., 2011) below the level of the $f_0$ was chosen as the reference spectrum for constructing stimuli with controlled deviations. The decision to choose an odd-to even ratio of approximately 1 (Table 4.1) ensured that this sound set did not vary predominantly according to that parameter (which was tested separately), whereas the choice of having the greatest possible T2 ensured that most of the deviation resulted from the differences in level among the upper harmonics. The spectral deviation was then controlled by reducing the differences in level between the successive harmonics of a reference spectrum until all harmonics had reached a level of 0 dBFS (Figure 4.1). In total, three sound sets with $f_0$'s at 120, 300, and 720 Hz were constructed. For these sound sets, the number of harmonics was increased to 16, which enabled the generation of a more uniform sample of amplitude distributions and the evaluation of a wider range of deviations that occur between the higher harmonics. The spectral deviation values presented in each sequence for a particular $f_0$ were the following ($\times 10^2$):
*Seq.1*: {0, 2.42, 4.76, 5.45, 5.75}
*Seq.2*: {0, 3.40, 5.17, 5.62, 5.75}
*Seq.3*: {0, 1.68, 4.58, 5.52, 5.75}
At this point it should be mentioned that the odd-to-even ratio stimulus set had a greater range of spectral deviation (Table 4.1) because all the odd or even harmonic components had a minimum level of –60 dBFS, whereas in the spectral deviation sets only one out of the sixteen components had a level at –60 dBFS (Figure 4.1).

**Spectral Slope** The spectral slope of each stimulus was controlled by reducing, or increasing, the levels in dB of 9 successive harmonics between the extremes of a flat and $1/n^4$ (i.e., 24 dB/octave), or $1/((N+1)-n)^4$, harmonic amplitude spectra for negative and positive slopes respectively, where $n$ is the harmonic number and $N$ is the total number of harmonics (Figure 4.1). In total, three sound sets with $f_0$'s at 120, 300, and 720 Hz were constructed for both positive and negative slopes the values of which were computed using linear regression over the power in dB of log-spaced harmonics. The spectral slope values (in dB/octave) presented in each sequence for a particular $f_0$ were the following:

*Seq.1*: {–24, –12, –1, 2, 5}
*Seq.2*: {–24, –16, –4, 0, 5}
*Seq.3*: {–24, –6, 0, 4, 5}

### Procedure

In every trial, listeners were first presented with a sequence of five stimuli that varied (monotonically) along an audio feature. Then, they had to adjust the spacing of five markers presented on screen (the first and last markers corresponding to the first and last stimulus were kept fixed) according to their perception of the relative spacing of the five stimuli in terms of differences between their successive audio feature levels. In other words, the separation between the markers reflected how far apart from each other the stimuli were perceived to be. The order presentation of features and the corresponding $f_0$'s or centroids of a given feature were randomized. For each feature, the three different sequences of stimulus values were presented randomly and in both ascending and descending orders. The experiment took approximately 60 minutes to complete.

The user interface was programmed in PsiExp (Smith, 1995). Sounds were amplified through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA) and presented diotically over Sennheiser HD600 headphones (Sennheiser Electronic GmbH, Wedemark, Germany). The range of the presentation levels over all stimuli was 53.4 – 74 dBA as measured with a Brüel & Kjær Type 2205 sound-level meter with a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled (Brüel & Kjær, Nærum, Denmark). Listeners were seated individually in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY).

### 4.2.2 Results

The interval scales were constructed by fitting a proper function on the median values of the ratings computed by first averaging each participant's ratings on the ascending and descending sequences of stimulus values in order to control for any hysteresis effects, which occur when the order of presentation affects the judgment of successive intervals between stimuli (Stevens, 1975). The criteria used for choosing the form of the function were that of monotonicity and maximum explained variance ($R^2$). The best fitting function in relation to the above criteria was determined after evaluating the performance of exponential, power, and polynomial functions which ranged from linear to the maximum allowable degree. In cases where the best fitting function was a power function, it was necessary to first transpose the feature values of the stimuli to strictly positive by adding an (offset) constant before applying the fitting algorithm. The reliability of listeners' ratings on a particular sequence of feature values was estimated according to *Cronbach's alpha* ($\alpha$), which was computed on the averaged ratings of each participant between the ascending and descending conditions of

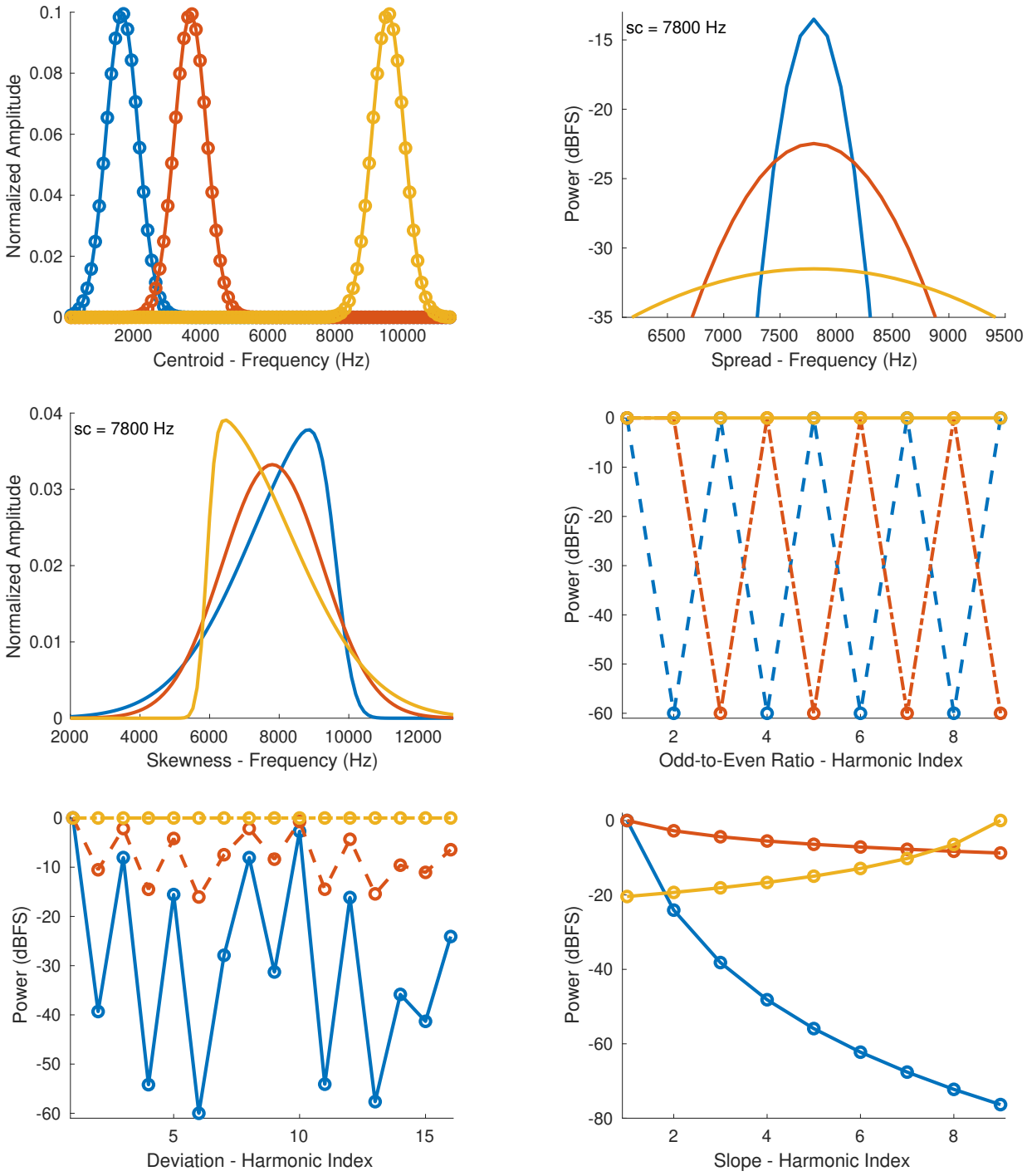


**Fig. 4.1** Spectral envelopes of anchor stimuli used in Experiments 1 and 2, and of mid-point stimuli used in the rating scale of Experiment 2. The spectral envelopes of the mid-point stimuli correspond to the middle sound of each stimulus set reported in Table 1. Dots indicate harmonics (for the spread and skewness plots the dots are omitted for display purposes). sc = spectral centroid.

each sequence (Stevens, 1975). *Cronbach's alpha* was estimated separately for each sequence rather on the combined ratings of all sequences, because some stimuli were presented in more than one sequence whereas some other where presented in just a single one. For qualitatively interpreting *Cronbach's alpha* we used the rule of thumb proposed by George and Mallery (2003): less than 0.5 – Unacceptable; between 0.5 and 0.59 – Poor; between 0.6 and 0.69 – Questionable; between 0.7 and 0.79 – Acceptable; between 0.8 and 0.89 good; above 0.9 – Excellent. The final median and interquartile ranges of listeners' ratings, and the fitting functions that are shown in the respective plots, were computed on each stimulus value after combining its ratings from all the sequences in which was presented and after averaging each participant's ratings for the ascending and descending conditions. For cases in which the fitting function had to be applied on transposed stimulus values, the respective plots display the actual stimulus values on the x-axis. All the analyses were done in MATLAB and R (R Core Team, 2013).

### Spectral Centroid

Figure 4.2 shows the combined median ratings and interquartile ranges calculated over all stimulus sequences, and the shape of the fitting function, from which it can be seen that the ratings increase monotonically for increasing spectral centroid. The second sequence in which the intermediate stimulus values were clustered around the middle one, had a good reliability score ($\alpha = 0.82$). The first sequence in which the intermediate values were more equidistant had a considerably lower but still acceptable reliability score ($\alpha = 0.72$). The third sequence in which the intermediate values were clustered around the edges near the first and last stimuli had the lowest score which indicated a lower than poor reliability ($\alpha = 0.41$). The fitting function was a power function of the form:

$$f(x) = -1284 \cdot x^{-0.1824} + 337.2, \qquad R^2 = 0.99 \tag{4.1}$$

### Spectral Spread

Figure 4.3 shows the combined median ratings and interquartile ranges calculated over all stimulus sequences (top panel), and the shape of the fitting function (bottom panel). The median ratings increased monotonically for increasing spectral spread with the exception of the sound set at a centroid of 7800 Hz, in which the second stimulus in the combined spacing was overestimated by most listeners. Overall, the reliability was good to marginally acceptable for the interval estimations of spectral spread, with $\alpha$ ranging from 0.86 to 0.69. The only exceptions were the third sequences which had their intermediate values clustered around the edges of the first and last stimuli with centroids at 5600 and 7800 Hz, and which exhibited questionable to less than poor reliability with $\alpha$'s at 0.6 and 0.34 respectively.

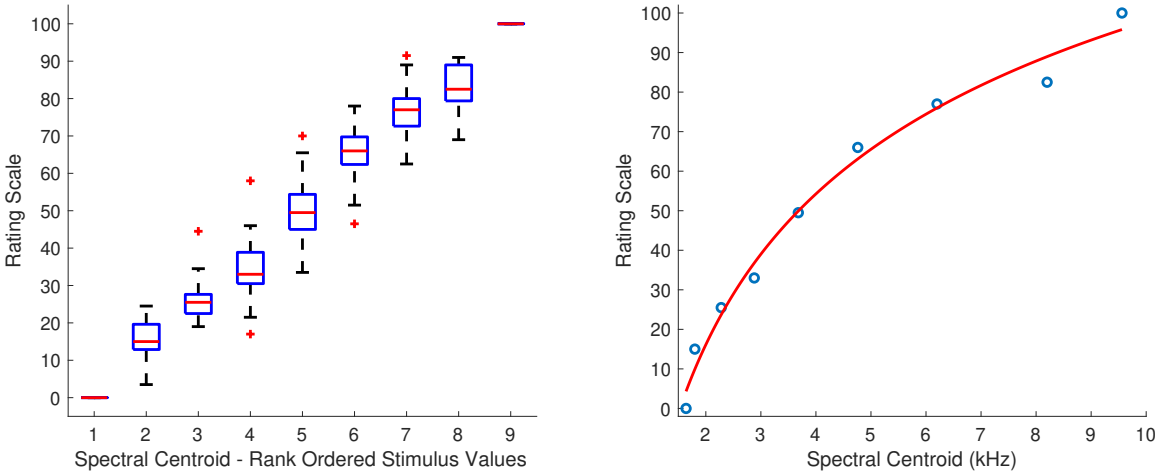


**Fig. 4.2** Boxplots and shape of the fitting function for spectral centroid. Whiskers extend to 2.7 SD.

For the sound sets with centroids at 1640 and 7800 Hz the fitting function was a quadratic polynomial with the following coefficients for the two sound sets respectively:

$$f_{1640}(x) = (-34.61 \cdot 10^{-5})x^2 + (41.59 \cdot 10^{-2})x - 22.36, \qquad R^2 = 0.99 \tag{4.2}$$

$$f_{7800}(x) = (-1.371 \cdot 10^{-5})x^2 + (8.84 \cdot 10^{-2})x - 15.04, \qquad R^2 = 0.98 \tag{4.3}$$

For the sound set with centroid at 5600 Hz the fitting function was a power function of the following form:

$$f_{5600}(x) = 10.02 \cdot x^{0.3917} - 75.15, \qquad R^2 = 0.99 \tag{4.4}$$

The fact that the fitting functions are of different forms, and with considerably different coefficients, reflects the unequal spacing of spectral spread between the sound sets of different spectral centroids.

### Spectral Skewness

Figure 4.4 shows the combined median ratings and interquartile ranges calculated over all stimulus sequences, and the shape of the fitting function. The sequences that had the lowest reliability scores ranging from less than poor to questionable were the first ($\alpha = 0.58$) and third ($\alpha = 0.46$) with centroid at 1640 Hz, as well as the third one ($\alpha = 0.68$) with centroid at 7800 Hz. For the rest of the sequences the reliability ranged from acceptable to good ($0.72 \leq \alpha \leq 0.80$). From the boxplots in the left panel of Figure 4.4 it can be seen that although the median ratings increase monotonically for increasing spectral skewness, the stimuli seem to be grouped in three different clusters: 2 to 4, 5 to 6, and 7 to 9. This

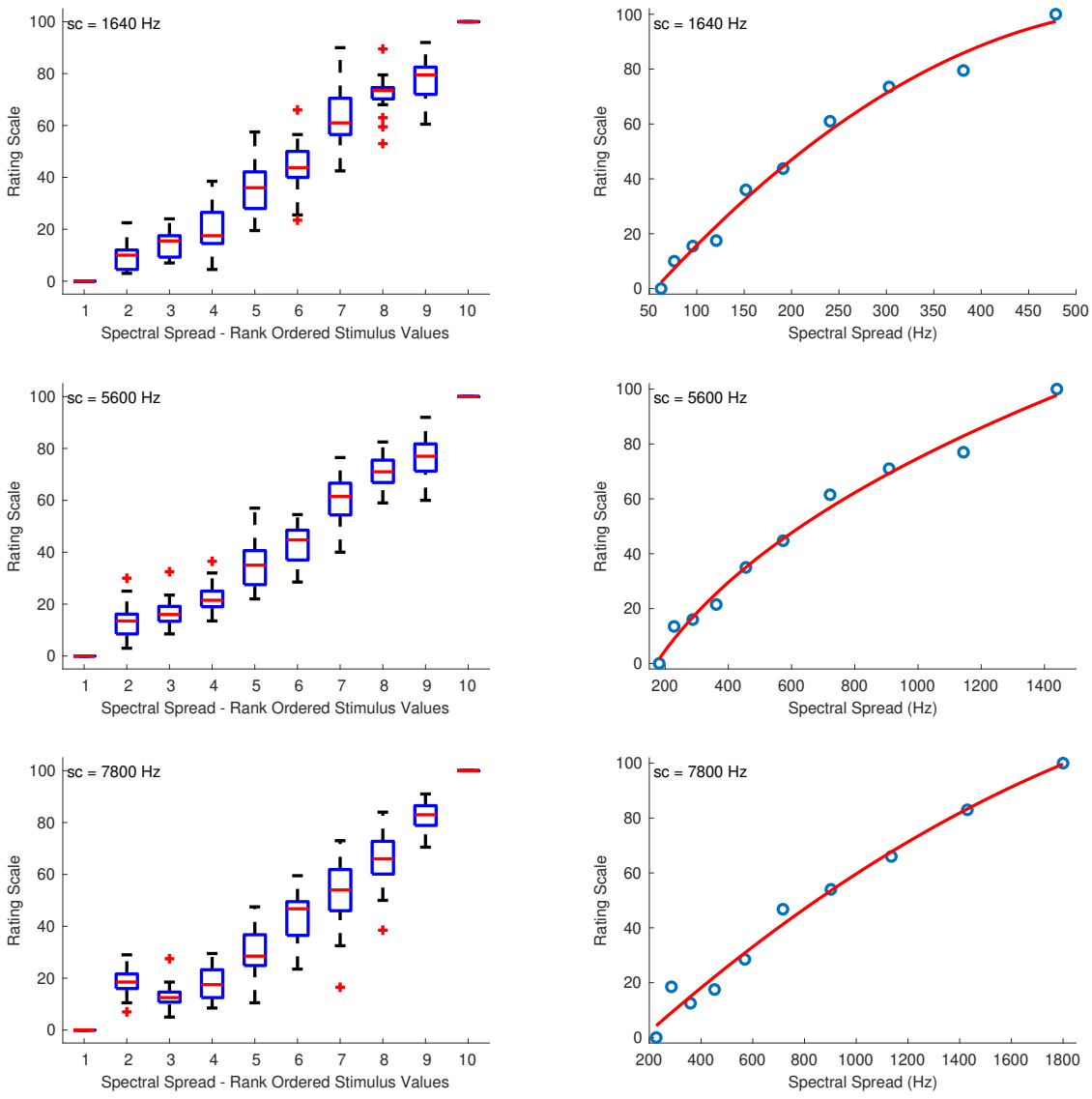


**Fig. 4.3** Boxplots and shape of the fitting functions for spectral spread. sc: spectral centroid. Whiskers extend to 2.7 SD.

trend of the data was best captured with a fifth order polynomial which had the following coefficients for each sound set with centroids at 1640, 5600, and 7800 Hz respectively:

$$f_{1640}(x) = 91.21x^5 - 24.86x^4 - 83.3x^3 + 24.59x^2 + 61.63x + 37.37, \quad R^2 = 0.99 \quad (4.5)$$

$$f_{5600}(x) = 99.72x^5 - 9.505x^4 - 88.99x^3 + 15.49x^2 + 55.97x + 34.24, \quad R^2 = 0.99 \quad (4.6)$$

$$f_{7800}(x) = 95.34x^5 - 17.78x^4 - 96.95x^3 + 19.02x^2 + 66.96x + 38.97, \quad R^2 = 0.99 \quad (4.7)$$

### Odd-to-Even Ratio

Figure 4.5 shows the combined median ratings and interquartile ranges calculated over all stimulus sequences, and the shape of the fitting function. From the boxplots it can be seen that the median ratings increased monotonically with increasing odd-to-even ratio. The reliability of interval estimations ranged from poor to excellent ($0.53 \leq \alpha \leq 0.89$). The first sequence had a questionable reliability score with $\alpha$'s at 0.64, 0.63, and 0.66, for $f_0$'s at 120, 300, and 720 Hz, respectively. The second sequence had $\alpha$'s at 0.53, 0.77, and 0.56, and the third sequence had $\alpha$'s at 0.89, 0.56 and 0.67, for $f_0$'s at 120, 300 and 720 Hz, respectively.

Because of the large range of odd-to-even ratios, the values were first log10-transformed. All the stimulus sets were best fit with a power function and were therefore transposed to strictly positive before fitting the function according to: $x' = log_{10}(x) + 10$, where x is the original stimulus value. The constant value of +10 ensured that the scale includes ratio values greater than $log_{10}(10^{-10})$. The following coefficients were used for each sound set with $f_0$'s at 120, 300, and 720 Hz respectively:

$$f_{120}(x') = (-2.904 \cdot 10^{10})x'^{-8.63} - 109.6, \quad R^2 = 0.98 \quad (4.8)$$

$$f_{300}(x') = (-2.061 \cdot 10^{10})x'^{-8.499} + 107.7, \quad R^2 = 0.99 \quad (4.9)$$

$$f_{720}(x') = (-9.95 \cdot 10^{10})x'^{-9.196} + 108, \quad R^2 = 0.98 \quad (4.10)$$

### Spectral Deviation

Figure 4.6 shows the combined median ratings and interquartile ranges calculated over all stimulus sequences, and the shape of the fitting function plotted on the actual stimulus values of spectral deviation. With the exception of the stimulus set at the $f_0$ of 720 Hz, in which the third stimulus was ranked higher than its preceding and succeeding stimulus values, the rest of the median values in all sound sets increased monotonically with increasing spectral deviation. The highest reliability scores were good and were observed

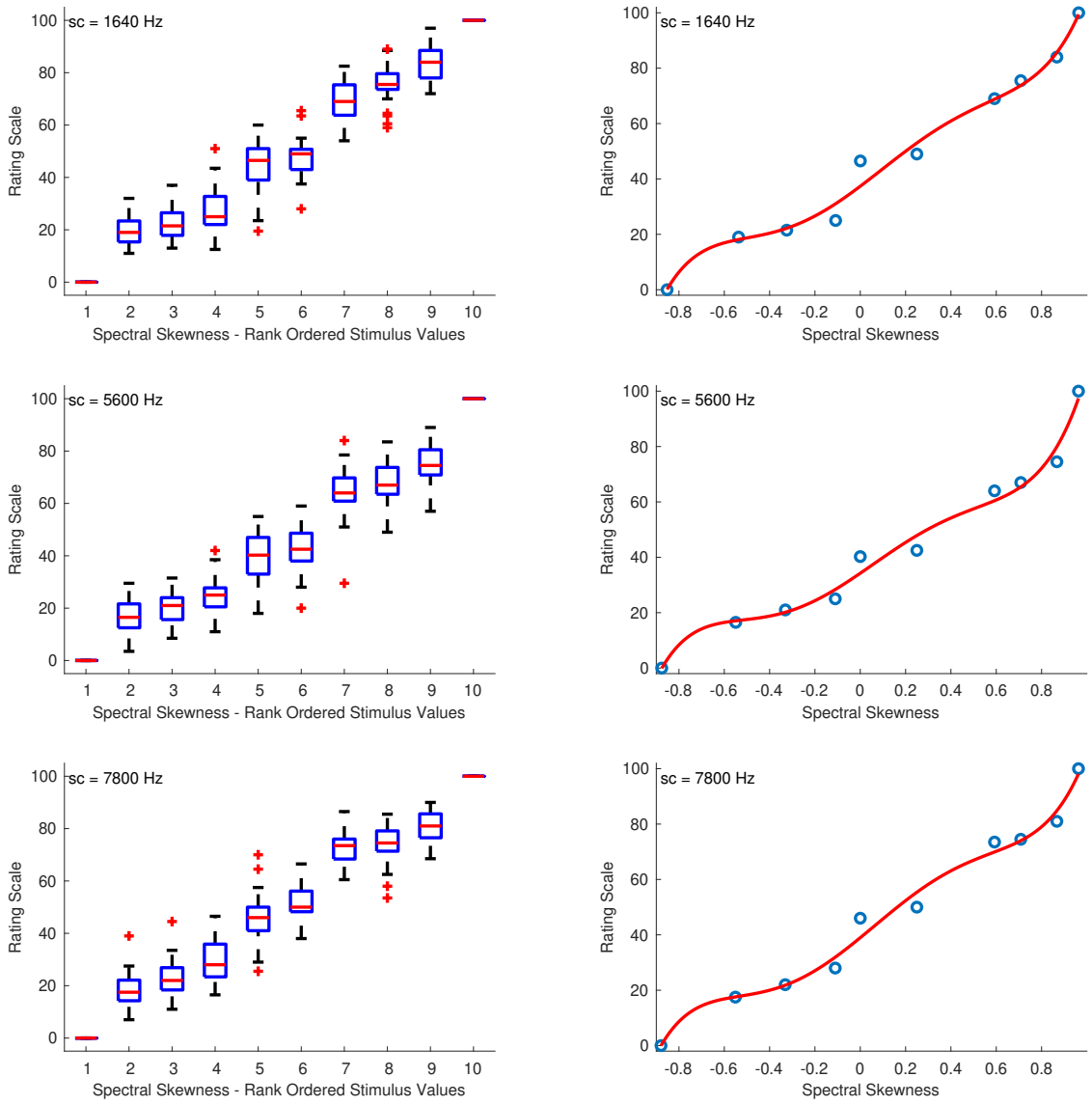


**Fig. 4.4** Boxplots and shape of the fitting functions for spectral skewness. sc: spectral centroid. Whiskers extend to 2.7 SD.

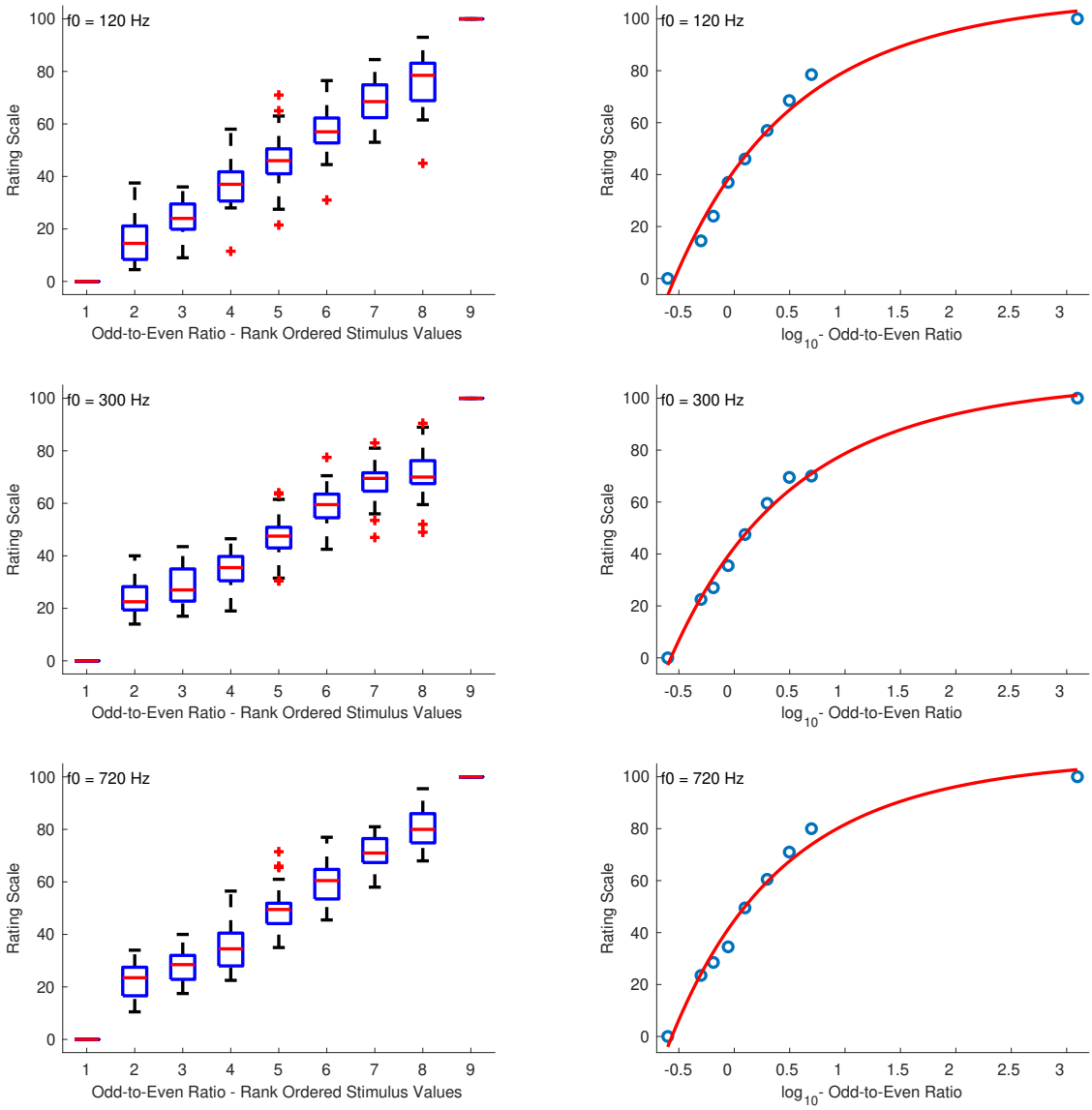


**Fig. 4.5** Boxplots and shape of the fitting functions for odd-to-even ratio. Whiskers extend to 2.7 SD.

for the second sequence in which the intermediate stimuli were spaced closer to the middle stimulus value than the edges and had $\alpha$'s at 0.83, 0.80, and 0.86 for $f_0$'s at 120, 300, and 720 Hz respectively. The reliability scores for the first and third sequences were 0.72, 0.70, 0.76, and 0.56, 0.76, 0.08, for $f_0$'s at 120, 300, and 720 Hz, respectively. For all the sound sets, the best fitting function was a power function and, as in the previous case, the values were transposed to strictly positive according to: $x' = x + 1$, where x is the original stimulus value and 1 is an arbitrary constant. The following coefficients were used for each sound set with $f_0$'s at 120, 300, and 720 Hz respectively:

$$f_{120}(x') = (14.22 \cdot 10^{-3})x'^{156.1} + 10.7, \qquad R^2 = 0.98 \tag{4.11}$$

$$f_{300}(x') = (47.76 \cdot 10^{-3})x'^{134.6} + 9.563, \qquad R^2 = 0.98 \tag{4.12}$$

$$f_{720}(x') = (55.78 \cdot 10^{-4})x'^{173.3} + 9.844, \qquad R^2 = 0.98 \tag{4.13}$$

### Spectral Slope

Figure 4.7 shows the combined median ratings and interquartile ranges calculated over all stimulus sequences, and the shape of the fitting function plotted on the actual stimulus values of spectral slope. In all sound sets, the median values increased monotonically with increasing spectral slope. As in the previous feature set of spectral deviation, the highest reliability scores were observed for the third sequence of this sound set in which the intermediate stimuli were spaced closer to the middle stimulus value than the edges. For that sequence, the reliability scores were overall marginally excellent with $\alpha$'s at 0.92, 0.89, and 0.89 for $f_0$'s at 120, 300, and 720 Hz respectively. The first and second sequences had overall lower reliability with $\alpha$'s at 0.85, 0.71, 0.75 and 0.64, 0.71, 0.64 for $f_0$'s at 120, 300, and 720 Hz, respectively. The second sequence in which the intermediate stimulus values ranged from negative to zero spectral slope seems to have had an effect on the reliability of stimulus ratings, which were overall questionable. The fitting function was again a power function, and, as in the previous cases, the values had to be rescaled to strictly positive before applying the fitting function according to: $x' = x + 24$, where $x$ is the original stimulus value. The rescaling constant of +24 allowed values above –24 dB/octave to be included in the scale. The coefficients of the power function for sound sets with $f_0$'s at 120, 300, and 720 Hz respectively, are the following:

$$f_{120}(x') = 0.8718x'^{1.393} + 0.7432, \qquad R^2 = 1 \tag{4.14}$$

$$f_{300}(x') = 1.144x'^{1.303} + 2.16, \qquad R^2 = 1 \tag{4.15}$$

$$f_{720}(x') = 1.02x'^{1.336} + 2.46, \qquad R^2 = 0.99 \tag{4.16}$$

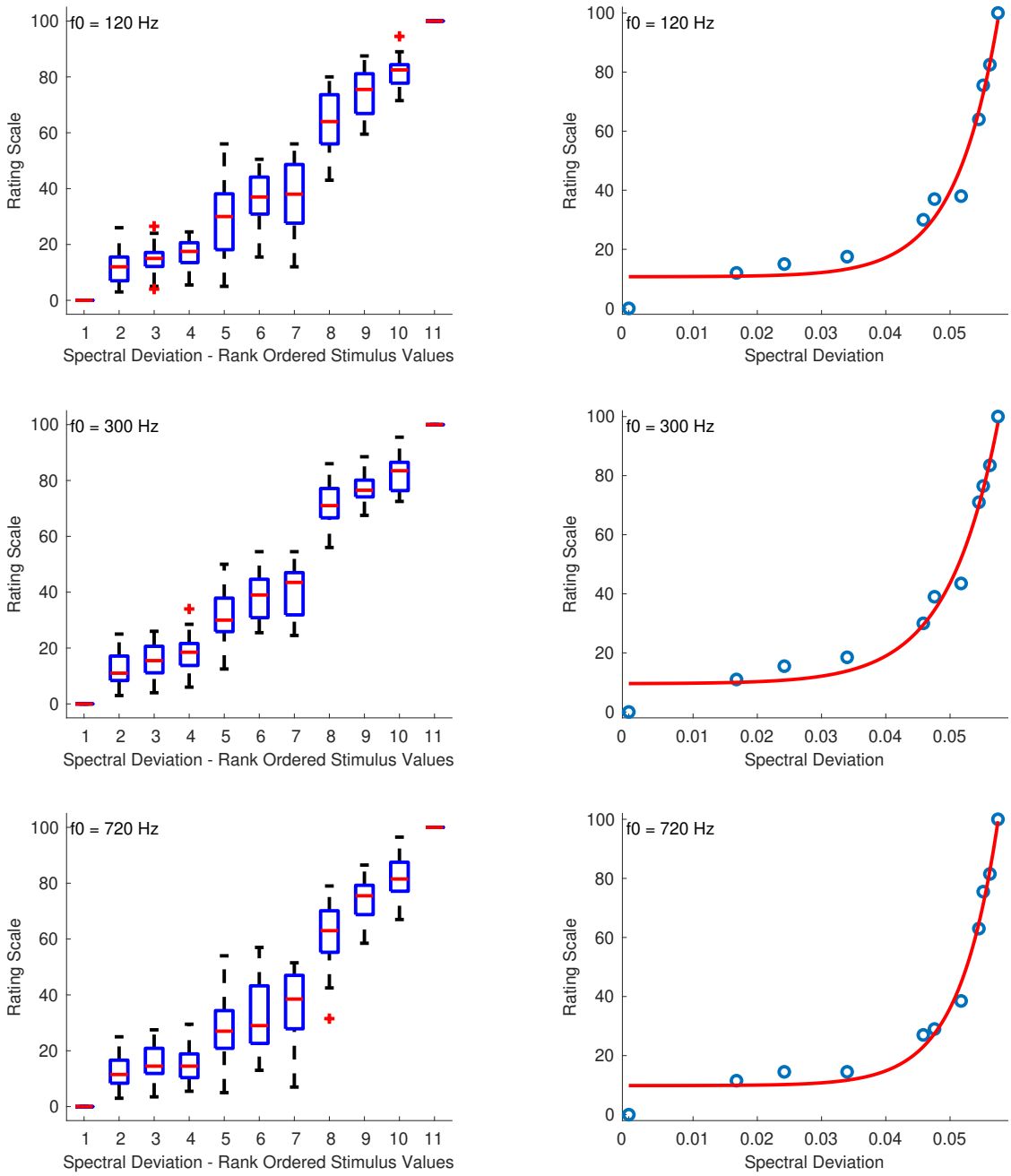


**Fig. 4.6** Boxplots and shape of the fitting functions for spectral deviation. Whiskers extend to 2.7 SD.

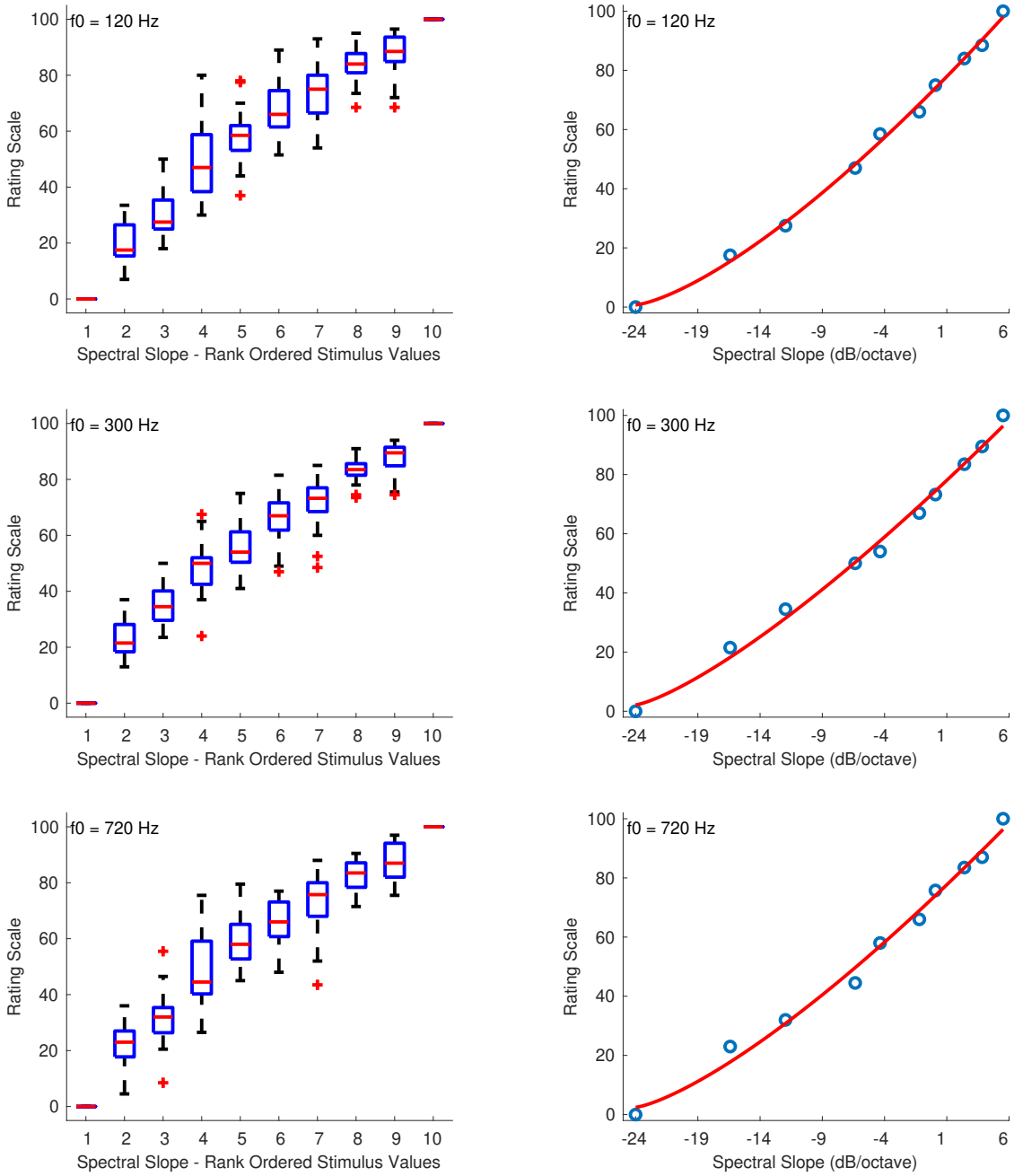


**Fig. 4.7** Boxplots and shape of the fitting functions for spectral slope. Whiskers extend to 2.7 SD.

### 4.2.3 Discussion

With the exception of the two anomalous points in one of the sound sets of spectral spread and spectral deviation, the rest of the median values of the interval estimations increased monotonically with increasing stimulus values, which indicates that the experiment was successful, and that the listeners were able to estimate intervals of the tested audio features. However, it is well-known that one of the biases of interval scales usually results from the initial selection and the limited pool of stimuli used for the estimations. This bias was controlled for by presenting subsets of an initial pool of stimuli covering a wide range of each feature with different spacing in each trial. Furthermore, any hysteresis effects were taken into account by presenting the stimulus sequences in both ascending and descending directions. Although we tried to control for the aforementioned biases, another important source of bias that was not possible to account for was the centering tendency, which afflicts all rating scales (Stevens, 1975). As became evident from most of the plots of the fitting functions, listeners tended to use the more central positions of the rating scale and avoid the extremes.

The forms of the derived fitting functions were the same within each audio feature when tested at different ranges albeit with different coefficients which indicates that the listeners' perceptions of these features depend on the spectral location at which each feature is presented in terms either of fundamental frequency or of spectral centroid. One exception to this was the interval estimations of spectral spread where the derived fitting functions within the whole tested range were not of the same kind, which reflects the unequal spacing of spread values between the stimuli constructed at different spectral centroids.

The large variability of the reliability scores measured according to *Cronbach's alpha* for each set of stimuli within the same feature and for a particular sequence indicates that the spacing of the stimuli had a big effect on the internal consistency of the listeners. The lowest alphas were observed for the stimuli of spectral skewness with centroid at 1640 Hz, and for the odd-to-even ratio. In addition, the overall lowest reliabilities were mainly observed for the sequences in which the stimuli were not (approximately) equidistantly spaced, and when the second and next-to-last stimuli were placed closer to the edges rather than to the middle stimulus value of the sequence. We hypothesize that this could be because listeners might were using the middle stimulus of the sequence as a reference (standard stimulus) for their interval estimations and as mentioned in the Introduction section as well, judgments tend to be more accurate for stimuli placed closer to the standard rather than far away from it (Gescheider & Hughson, 1991). Another factor that might have played a role in the observed variance of interval estimations, the anomaly points, and for some cases in the relatively low reliability scores, could be that for some features, the audible differences between the stimuli in the combined set were marginally supraliminal (albeit clearly supraliminal within each sequence). However, this was a direct consequence of the narrow perceivable range of some features (e.g., odd-to-even ratio) and the constraints imposed from the synthesis procedure for constructing the stimuli (e.g., narrow permissible

range of spectral skewness due to the Skew-normal distribution).

In conclusion, the largest biases of the derived interval scales resulted from the centering tendency of the listeners and in some cases from the marginally supraliminal spacing of stimulus values. Despite these biases, the experiment should be considered as an exploratory step, which confirmed the ability of the listeners to estimate intervals of the tested audio features. It also allowed us to proceed to the construction of ratio scales presented in the next section.

## 4.3 Experiment 2: Equisection Scaling

The aim of this experiment was the derivation of ratio scales of audio features provided that listeners are able to estimate intervals, which was confirmed from the results of Experiment 1. In this experiment, listeners had control over the stimulus values and were asked to equisect a continuum of a particular audio feature. Each equisection was performed using the progressive solution (Gescheider, 1997) according to which listeners progressively partition the continuum formed by the stimuli into a number of equal-sounding intervals. The equality of sensory intervals implies that the intervals themselves have ratio properties (Marks & Gescheider, 2002) and thus, the results of this experiment led to ratio scale measurements.

### 4.3.1 Method

#### Participants

Twenty participants, 6 female and 14 male, with a median age of 25 years (range: 18–41) were recruited from the Schulich School of Music, McGill University. All of them were self-reported amateur or professional musicians with formal training in various disciplines such as performance, composition, music theory, and sound engineering. Participants were compensated for their time. One participant reported perfect pitch, and another one synesthesia.

#### Stimuli and Presentation

All the stimulus sets were constructed with the procedures described in Subsection 4.2.1 and at the same $f_0$'s and spectral centroids. In order to create a continuum within a range of a particular feature, several stimuli were constructed with multiple imperceptible successive differences. The total number of sounds used for each stimulus set and the ranges of feature values for a particular set are indicated in Table 4.1. The extreme feature values of each stimulus set were the same as those used in Experiment 1. Figure 4.1 shows for a particular stimulus set, the spectral envelopes of the anchor and mid-point stimuli used in the rating scales of the present experiment.

### Procedure

Other than the experimental task, the general procedure was the same as the one used in Experiment 1. In a first step, listeners divided the continuum of an audio feature into two equal-sounding intervals, by triggering each stimulus with a cursor along a horizontal bar that contained the stimuli, and by placing a marker over the stimulus-bar. Each section was then bisected in the next step. In total there were three bisections: the first one was made between the stimuli of the total range, and the other two between the lower and upper bisected ranges. The order of presentation of the upper and lower half bisections was randomized. In a final step, listeners were presented with all their bisections and were instructed to make further fine adjustments so that all four intervals they had created in the previous steps sounded equal. The order of presentation of features and the $f_0$ or spectral centroid of each feature was randomized. As in the previous experiment, the stimuli were presented in both ascending and descending conditions at separate trials: i.e., the stimulus bar would start either from the lowest (ascending condition) or the highest (descending condition) stimulus value of the feature set. The experiment took approximately 60 minutes to complete.

## 4.3.2 Results

The ratio scales were constructed by fitting a function to the median values of each equisection computed on both the ascending and descending presentations of the stimulus sets. As in the previous experiment, whenever the best fitting function was a power function, it was necessary to first transpose the stimulus values to strictly positive before fitting the function. The criteria used for choosing the form of the function were again monotonicity and maximum explained variance, but also good continuation outside the tested range (i.e., no oscillations outside the tested range). After identifying the form of the function, the *zero point* of the scale was determined either empirically by extrapolating the function to a point that marks the lower limit of perception for a particular audio feature, or wherever applicable, according to the physical stimulus value in which case "zero" has a physical meaning (e.g., zero skewness). Finally, the *units* of the psychophysical scales were defined by assigning specific numerals to the points of the equisection scale. *Cronbach's alpha* was again used to evaluate the reliability of the derived scales across listeners. In all cases, the reliability was overall excellent, with the scales of spectral centroid, and spectral skewness with centroid at 7800 Hz having the highest reliability ($\alpha = 0.96$). The lowest reliability was observed for the equisections of spectral spread with centroid at 7800 Hz ($\alpha = 0.89$), and odd-to-even ratio with $f_0$ at 720 Hz ($\alpha = 0.78$).

### Spectral Centroid

Figure 4.8 on the left shows the fitted power function on top of the median ratings and interquartile ranges. At that point, the ordinate was assigned arbitrary units which rep-

resent equal spectral centroid-distances as perceived by listeners. Figure 4.8 on the right shows the extrapolated function for centroids in the range of 20 Hz to 20 kHz. The location of the zero point on the ordinate was assigned the value of 20 Hz, which marks the lowest limit of pitch perception, and finally, the units of the scale were derived by assigning the numeral 10 to the 1 kHz spectral centroid. The coefficients of the final fitting equation after the unit assignment are:

$$f(x) = -34.61x^{-0.1621} + 21.2985, \qquad R^2 = 1 \tag{4.17}$$

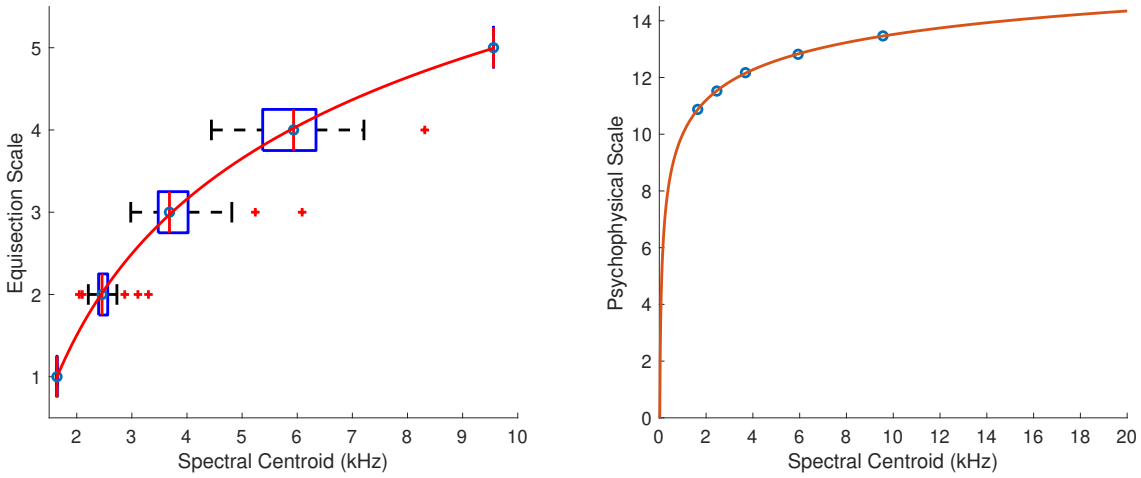


**Fig. 4.8** Equisection and psychophysical scales of spectral centroid. On the left: boxplots and fitting function on the median ratings. Whiskers extend to 2.7 SD. On the right: psychophysical scale and extrapolated fitting function.

### Spectral Spread

Figure 4.9 shows the fitted power functions on top of the median ratings and interquartile ranges for the sound sets of spectral spread with centroids at 1640 ($R^2 = 1$), 5600 ($R^2 = 1$), and 7800 Hz ($R^2 = 1$), respectively. In order to find a single psychophysical function of spectral spread covering the entire range independently of the centroid used in each sound set, the three functions for the overlapping tested ranges needed to be combined. To this end, Torgerson's (1958) method was used according to which the scale values in the lower and upper ranges are converted into scale units of the middle range, resulting in a single function. The conversion was performed for the overlapping portions of spectral spread's range by linearly regressing both the lower ($R^2 = 1$) and upper ranges ($R^2 = 1$) over the mid-range. The conversion equations for the lower and upper ranges were $f_l(x) = 0.7345x - 1.159$, and

$f_u(x) = 1.086x + 0.3256$ respectively, and their respective plots are shown in the top panel of Figure 4.10. The final fitting power function ($R^2 = 1$) covering the entire tested range is shown on the left of the bottom panel of Figure 4.10 where the vertical distances on the graph represent spectral-spread distances as perceived by the listeners.

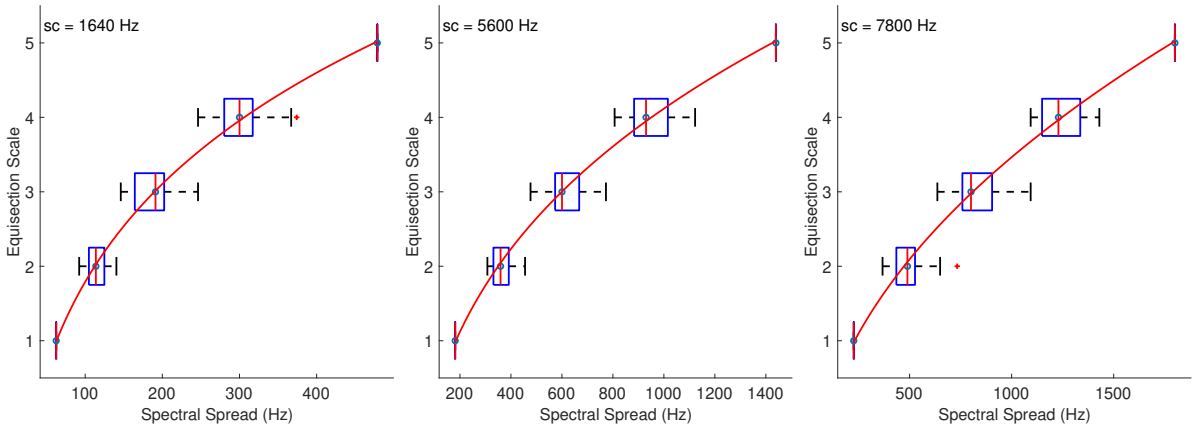


**Fig. 4.9** Equisection scales of spectral spread. Boxplots and fitting function on the median ratings. Whiskers extend to 2.7 SD.

The bottom panel of Figure 4.10 on the right, shows the extrapolated power function in the range of 0 Hz to 10 kHz[2]. The zero point on the ordinate was assigned the value of 0 Hz because in this case, a spectral spread of 0 Hz has a physical meaning indicating the presence of just a single component in the spectrum. The units of the final scale were derived after assigning the numeral 10 to the spread of 10 kHz. The final fitting equation covering the entire tested range and after the unit assignment is:

$$f(x) = 0.6134x^{0.3031}, \qquad R^2 = 1 \tag{4.18}$$

### Spectral Skeweness

Figure 4.11 shows the fitted third-order polynomial functions on top of the median ratings and interquartile ranges for the sound sets of spectral skewness with centroids at 1640, 5600, and 7800 Hz. The skewness values were the same in all sound sets, so the aim was not to derive a single function independent of the centroid used, as in the previous case, but to derive psychophysical functions of spectral skewness centered at different locations in the spectrum. Figure 4.11 on the right shows the extrapolated functions in the range of –6 to 6 spectral skewness. As in the previous case, a value of 0 skewness has a physical meaning indicating a gaussian spectral distribution and therefore, the zero point on the

[2]The power function requires that all data points be strictly positive, and therefore the actual value used was $2^{-52}$ Hz instead of 0 Hz.

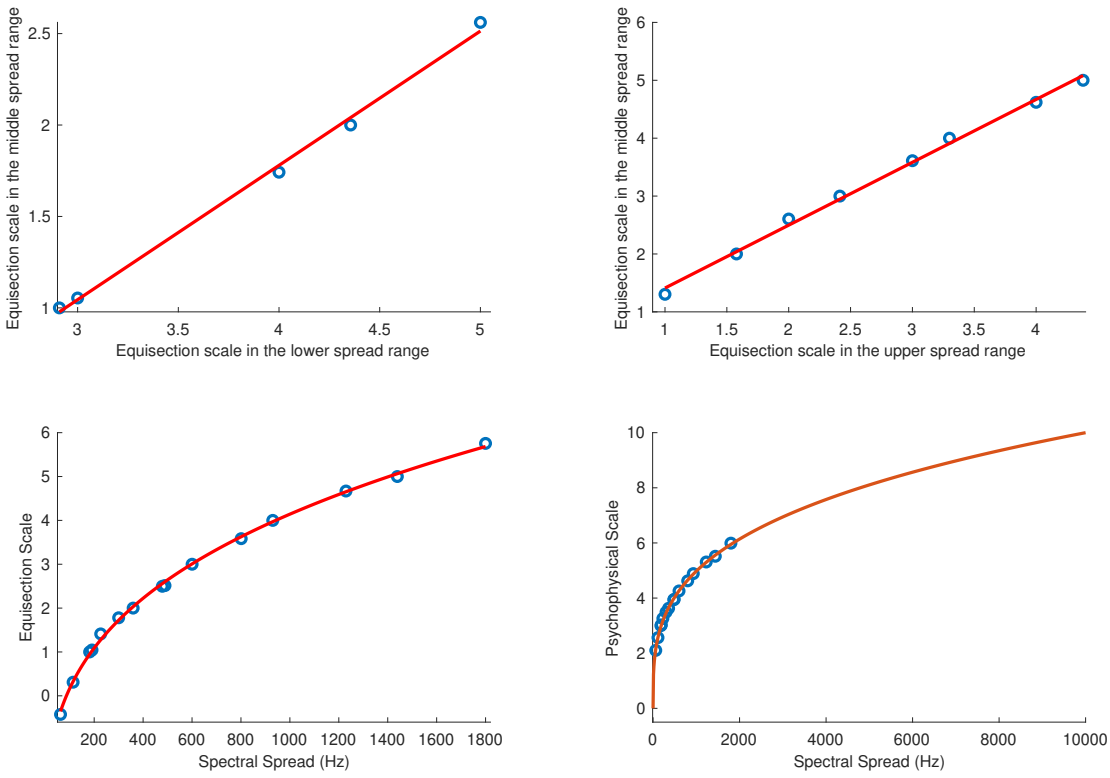


**Fig. 4.10** Combined equisection and psychophysical scales of spectral. Top panel: equisection scales as a function of middle-spread range; Bottom panel: unified equisection scale (on the left) and psychophysical scale (on the right).

ordinate was assigned the value of 0 skewness. Finally, the units of the scale were derived after assigning the numeral 0.1 to the skewness of 1. The fitting equations after the unit assignment for skewness at centroids of 1640, 5600 and 7800 Hz are:

$$f_{1640(x)} = 0.03334x^3 + 0.005587x^2 + 0.06107x, \qquad R^2 = 0.99 \tag{4.19}$$

$$f_{5600(x)} = 0.04435x^3 + 0.02014x^2 + 0.03551x, \qquad R^2 = 1 \tag{4.20}$$

$$f_{7800(x)} = 0.02951x^3 + 0.01567x^2 + 0.05482x, \qquad R^2 = 1 \tag{4.21}$$

### Odd-to-Even Ratio

The best fitting function for the odd-to-even ratio was a power function. Because of the large range of stimulus values ($x$), these were first log10-transformed and transposed to strictly positive before fitting the power function according to: $x' = log_{10}(x) + 10$. Figure 4.12 (left panel) shows the fitted functions on top of the median ratings and interquartile ranges for the sound sets with $f_0$'s at 120, 300, and 720 Hz. The abscissa holds the log-transformed values of the odd-to-even ratio, and the arbitrary units on the ordinate represent equal odd-to-even ratio-distances as perceived by the listeners. Because the perception of this feature depends on fundamental frequency (Kazazis et al., in preparation), the aim was not to find a single function independent of fundamental frequency but to derive more precise psychophysical functions of odd-to-even ratio at the different $f_0$'s tested.

The right panel of Figure 4.12 shows the extrapolated functions in the range of –0.9 to 3 of the $log_{10}$-transformed odd-to-even ratios. The zero point of the scale was defined at the odd-to-even ratio of 1, and since the values were log-transformed this point corresponds to the zeros on the abscissas in the right panel of Figure 4.12 . The units of the ordinate were derived after assigning the numeral 2 to the odd-to-even ratio of 2. The fitting equations after the unit assignment for the stimuli at $f_0$'s of 120, 300 and 720 Hz are respectively:

$$f_{120}(x') = -1.49 \cdot 10^{11} x'^{-10.29} + 7.60, \qquad R^2 = 1 \tag{4.22}$$

$$f_{300}(x') = -1.798 \cdot 10^{11} x'^{-10.38} + 7.55, \qquad R^2 = 1 \tag{4.23}$$

$$f_{720}(x') = -1.808 \cdot 10^{11} x'^{-10.38} + 7.55, \qquad R^2 = 1 \tag{4.24}$$

### Spectral Deviation

The best fitting function for the spectral deviation stimuli was a power function, and as in the previous case the stimulus values ($x$) were transposed to strictly positive before fitting the function according to $x' = x + 1$. Although it would have been possible to perform the fitting after adding a small constant just to the stimulus of zero spectral deviation, the numerical accuracy of the algorithm, and thus the fit, was found to be poorer

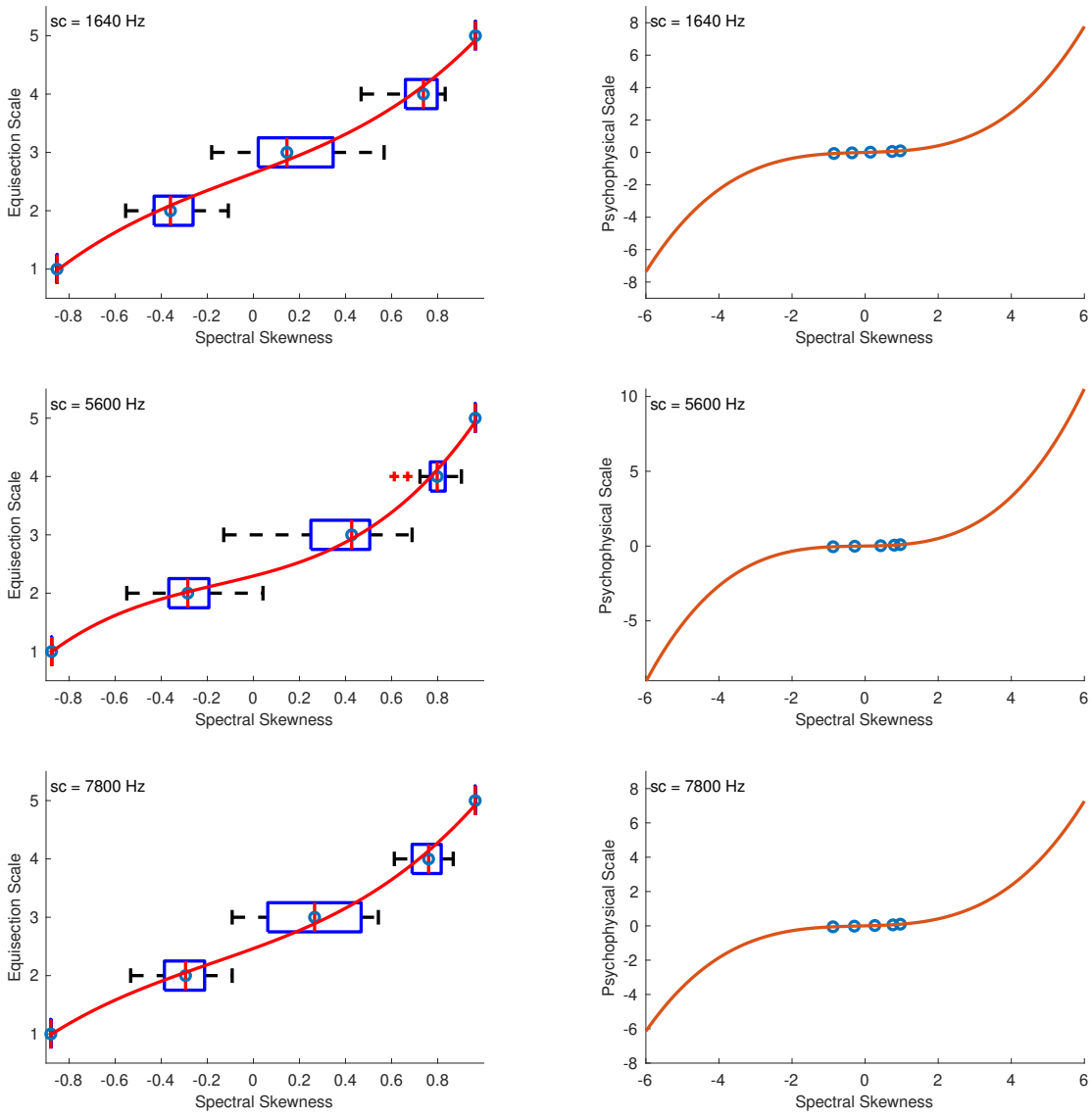


**Fig. 4.11** Equisection and psychophysical scales of spectral skewness. Left panel: boxplots and fitting function on the median ratings. Whiskers extend to 2.7 SD; Right panel: psychophysical scale and extrapolated fitting function.

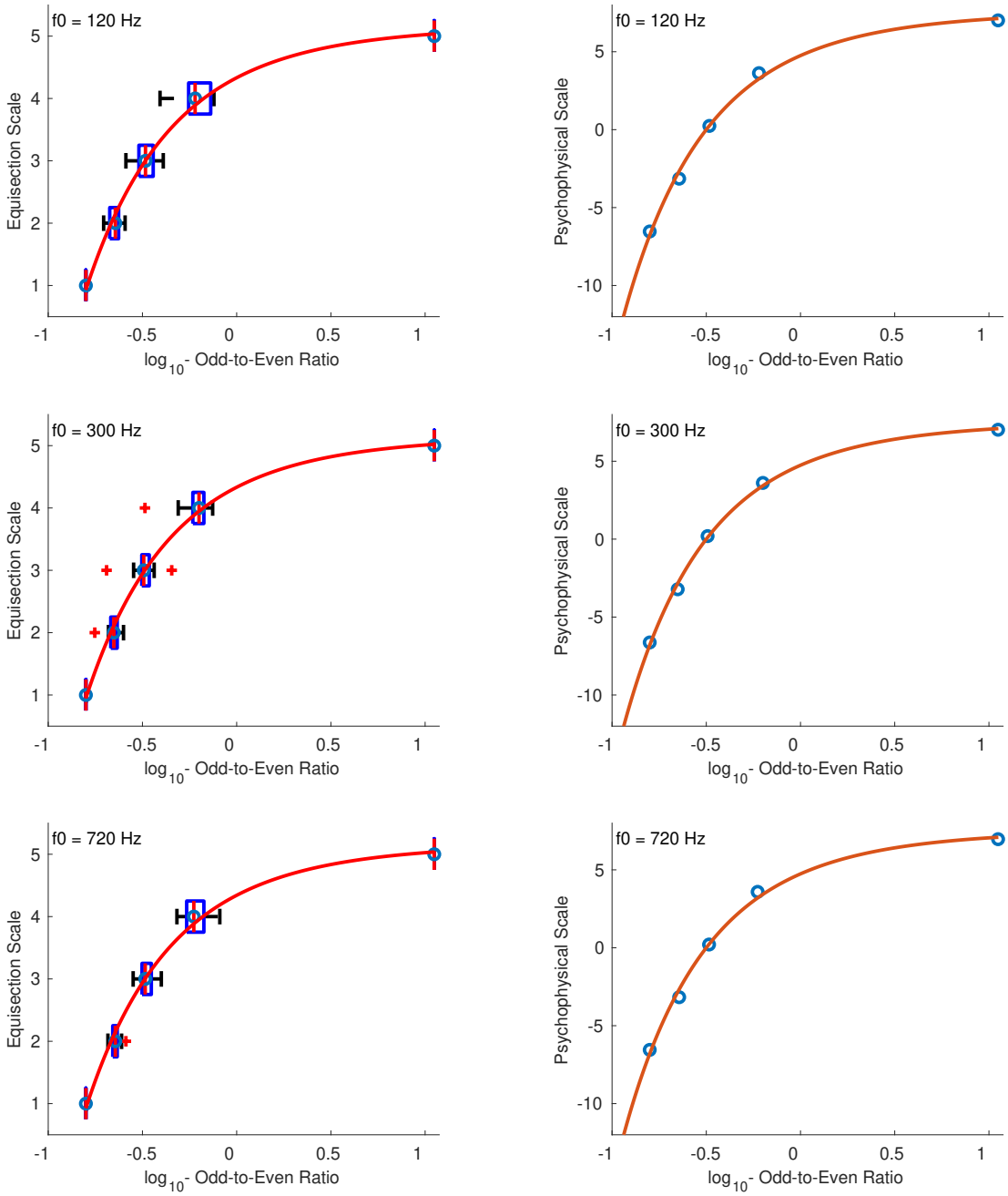


**Fig. 4.12** Equisection and psychophysical scales of odd-to-even ratio. Left panel: Boxplots and fitting function on the median ratings. Whiskers extend to 2.7 SD; Right panel: psychophysical scale and extrapolated fitting function on $log_{10}$-transformed stimulus values.

when compared to shifting all the values by a larger constant, possibly due to round-off errors. Figure 4.13 (left panel) shows the fitted functions on top of the median ratings and interquartile ranges for the stimuli with $f_0$'s at 120, 300, and 720 Hz where the abscissas hold the actual stimulus values.

The right panel of Figure 4.13 shows the fitted functions in the extrapolated range of 0 to 0.07 spectral deviation after the unit assignment. The zero point of the scale was naturally assigned the value of 0, which maps to zero spectral deviation, and finally, the units on the ordinate were derived after assigning the numeral 1 to the 0.01 spectral deviation. The final fitting equations for the stimuli at 120, 300, and 720 Hz are, respectively:

$$f_{120}(x') = 5.302x'^{17.37}\text{–}5.30, \qquad R^2 = 1 \tag{4.25}$$

$$f_{300}(x') = 4.267x'^{21.16}\text{–}4.27, \qquad R^2 = 1 \tag{4.26}$$

$$f_{720}(x') = 4.504x'^{20.15}\text{–}4.50, \qquad R^2 = 1 \tag{4.27}$$

### Spectral Slope

The best-fitting function for the stimuli of spectral slope was again a power function and the stimulus values ($x$) were rescaled to strictly positive before applying the fitting function according to: $x' = x + 24$, where x is the original stimulus value. The rescaling constant of +24 allowed values above –24 dB/octave to be included in the scale. Figure 4.14 (left panel) shows the fitted functions for stimuli at 120, 300, and 720 Hz, on top of the median ratings and interquartile ranges for the actual stimulus values of spectral slope. The right panel of Figure 4.14 shows the psychophysical scale in the range of –24 dB/octave to +6 dB/octave slope. The zero point of the scale was naturally assigned the value of 0 which corresponds to zero spectral slope, and the units of the scale were derived after assigning the numeral 1 to the value of +1 spectral slope. The final fitting equations for stimuli at 120, 300, and 720 Hz, after the unit assignments are, respectively:

$$f_{120}(x') = 9.673 \cdot 10^{-2}x'^{1.586}\text{–}14.95, \qquad R^2 = 0.99 \tag{4.28}$$

$$f_{300}(x') = 21.06 \cdot 10^{-2}x'^{1.385}\text{–}17.19, \qquad R^2 = 1 \tag{4.29}$$

$$f_{720}(x') = 16.46 \cdot 10^{-2}x'^{1.448}\text{–}16.42, \qquad R^2 = 1 \tag{4.30}$$

### 4.3.3 Discussion

For almost all features, the reliability of the derived scales within the tested range was excellent as indicated by *Cronbach's alpha*. The equisection of the physical continuum for each feature was performed on stimuli presented in both ascending and descending directions, which controlled for any hysteresis effects on the derived scales. With the

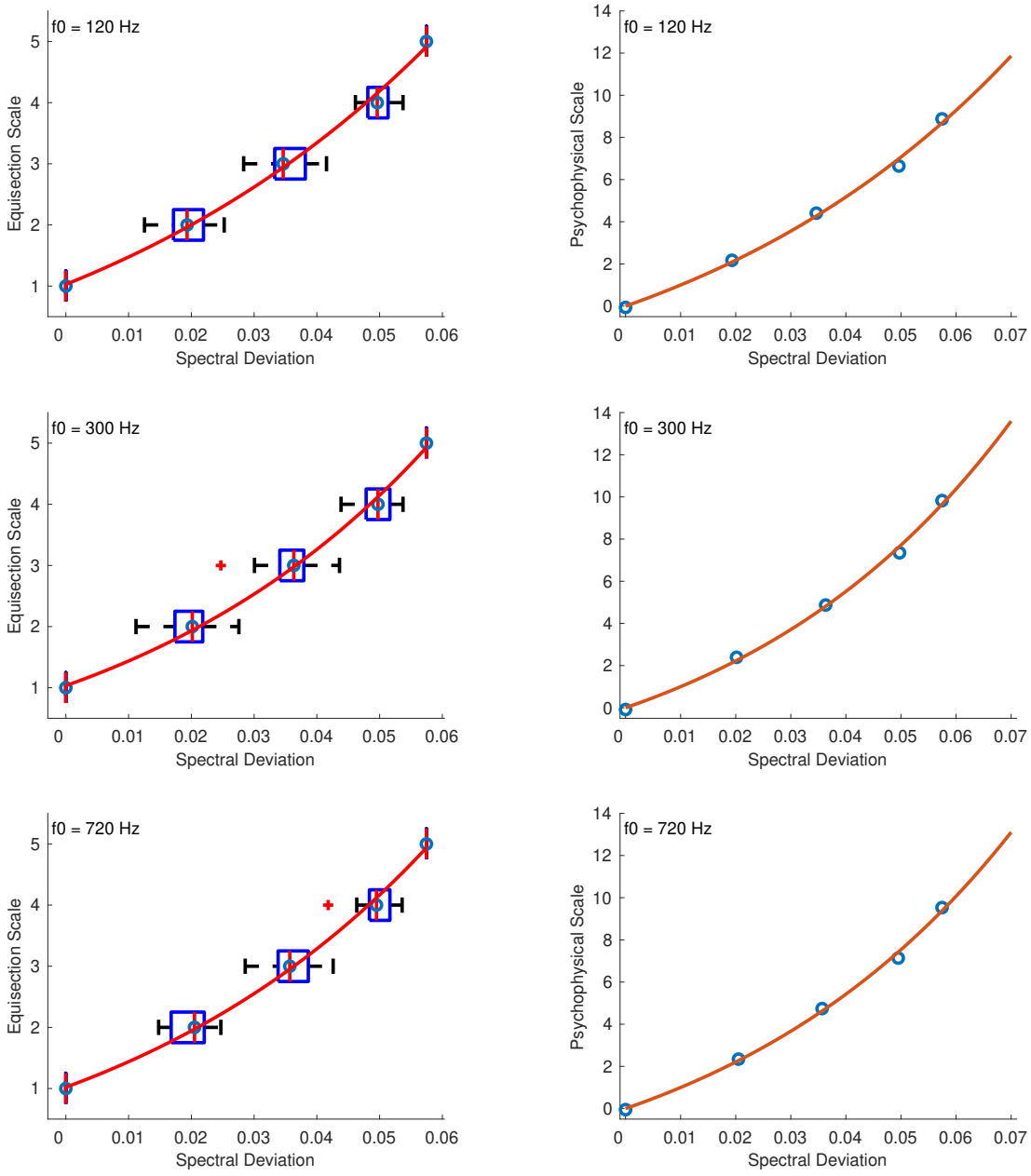


**Fig. 4.13** Equisection and psychophysical scales of spectral deviation. Left panel: boxplots and fitting function on the median ratings. Whiskers extend to 2.7 SD; Right panel: psychophysical scale and extrapolated fitting function.

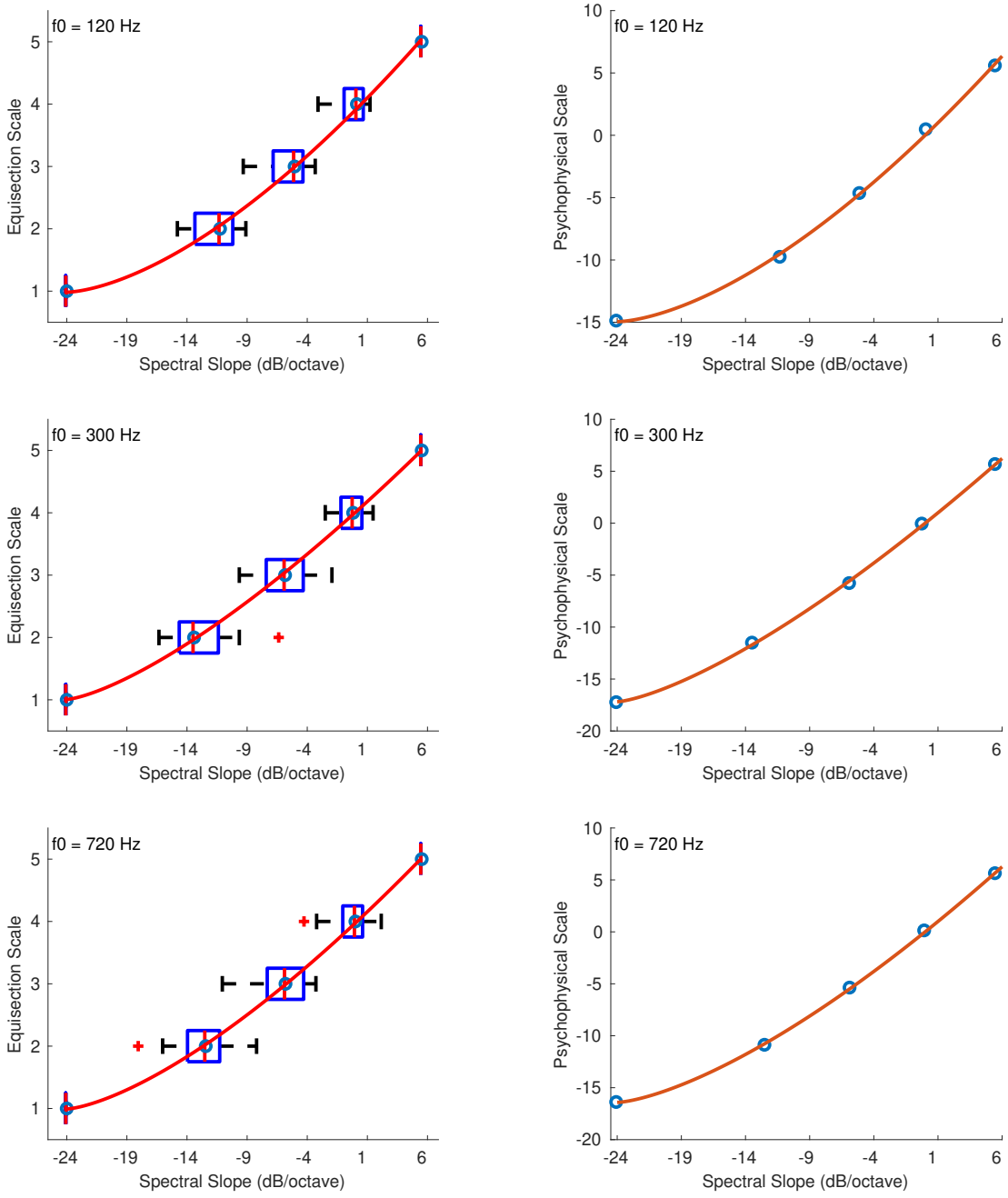


**Fig. 4.14** Equisection and psychophysical scales of spectral slope. Left panel: boxplots and fitting function on the median ratings. Whiskers extend to 2.7 SD; Right panel: psychophysical scale and extrapolated fitting function.

exception of spectral skewness, for which the best fitting function on the median (averaged) ratings was a third-order polynomial, for the rest of descriptors the best fitting functions were all power functions albeit exhibiting significantly different shapes, which indicates that each descriptor is perceived on a different psychophysical scale. At this point, it should also be pointed out that Torgerson's (1958) method, which was used for deriving a single psychophysical scale of spectral spread, would not have been applicable if listeners were not internally consistent. The linearity of the functions used to convert the overlapping scale values of the upper and lower ranges into the values of the middle range indicates that the estimated equisection points were in fact a function of spectral spread and not of the presented range.

With the exception of spectral centroid, for which the zero point of the scale was derived by extrapolating the fitting function, the rest of the scales were assigned a zero point that has a physical meaning and maps naturally to the physical value of the stimulus. For spectral centroid, which is related to the perception of auditory brightness (see, for instance, Schubert & Wolfe, 2006), the zero point was assigned empirically at 20 Hz, which marks the lower limit of pitch perception. Although it can be argued that the centroid will in general have values above 20 Hz, there can still be cases in which the centroid is evaluated on spectra with minimal or zero spectral spread. With respect to such cases, in which the spectral centroid would match the stimulus fundamental frequency, and after taking into account Schubert and Wolfe's (2006) conclusion that brightness is dependent upon $f_0$ to the extent that increasing $f_0$ also increases spectral centroid, it was concluded that frequencies as low as 20 Hz should not be excluded from the psychophysical ratio scale of spectral centroid. The numerical ranges of the scales corresponding to the minimum and maximum physical values of the audio features are all comparable in terms of magnitude, because the unit assignment, albeit arbitrary, was performed in such a way as to facilitate comparisons between different audio features when these are extracted from a given stimulus.

In previous experiments on ordinal scaling (Chapter 2) it was shown that the perception of some audio features depends on the fundamental frequency or its spectral centroid, and therefore, with the exception of the scales of spectral centroid and spectral spread, the psychophysical scales for the rest of the features were derived separately for each fundamental frequency or centroid tested. A scale for a particular $f_0$ or spectral centroid that falls in between the tested range of this study, can be derived by using a weighted interpolation scheme between the coefficients of the fitting functions in log-frequency. Of course, that could not have been possible if the fitting functions that were used to derive the ratio scales of a particular feature and at each $f_0$ or centroid were not of the same form.

## 4.4 Conclusion

The aim of the present study was to test listeners' abilities to estimate intervals of audio features, and the construction of perceptual ratio scales, when each of the presented fea-

tures was independently controlled through specifically designed synthesis algorithms. The experimental design used in both experiments controlled for the biases of order effects, in which listeners' judgments of a particular stimulus often depend on 1) the preceding stimuli, 2) hysteresis effects in which judgments are biased from the ascending or descending presentations of stimulus values, and wherever possible, 3) for range effects, which occur when listeners' judgments are performed on a limited range of stimulus values. However, for the stimulus sets of spectral centroid, spread, and especially skewness, the range effects were the hardest to control for due to the constraints imposed by the synthesis algorithms such as: the choice of using nine harmonics at a fixed $f_0$ for keeping the spectral spread and skewness fixed in the centroid stimulus sets; bandwidth restrictions due to fixed $f_0$, centroid and skewness for the spectral spread stimulus sets; or, the narrow permissible range of skewness in the Skew-normal distribution, which was employed for keeping the centroid and spread fixed in the skewness stimulus sets.

In the first experiment, listeners made estimations based on successive differences between stimuli of a given audio feature, and thus this experiment provided interval scale measurements. There was a large variability in the reliability of the ratings measured according to *Cronbach's alpha*, which was dependent on the presented spacing of stimuli. The largest biases of the derived interval scales resulted from the centering tendency of the listeners, and for some features, from the marginally supraliminal stimuli used in the combined stimulus set. Despite these biases, the experiment is to be considered successful because in general, the median values of the interval estimations increased monotonically with increasing stimulus value, and thus confirmed the ability of listeners to estimate intervals between stimuli of a given audio feature.

The method of equisectional scaling, which was employed in Experiment 2, leads to equal sensory intervals that have built-in ratio properties and thus, the results of that experiment provided ratio scale measurements. The interval scaling in Experiment 1 was a prerequisite for proceeding to the construction of ratio scales, which was the ultimate goal of this study, because without any prior evidence that listeners were actually capable of estimating intervals of audio features, the results of the second experiment would have been subject to the uncertainty of whether they were visually bisecting or quartering the displayed range of stimulus values, instead of performing estimations according to prescribed psychophysical ratios (i.e., halving and quartering) of auditory stimuli. As evidenced by *Cronbach's alpha*, the reliabilities of the derived psychophysical scales were overall excellent. With the exception of spectral centroid, where the zero point was derived by extrapolating the fitting function, the rest of the zero points of the derived psychophysical scales were mapped naturally to the stimulus physical values, and the units, albeit arbitrary, were assigned as to facilitate in a listeners' mind any comparisons across the values of different features when these are extracted from a single stimulus. Due to constraints imposed in the synthesis process for independently controlling each tested feature value and constructing perceptually uncorrelated stimuli, the extreme values of the psychophysical scales were

derived by extrapolating the fitting functions. Nevertheless, although the extrapolation was well behaved (in terms of monotonicity) further experiments are needed to verify the presented scales on the extrapolated regions.

The results of the two experiments are not directly comparable because in the first experiment, the listeners' task was to estimate intervals between successive stimuli, whereas in the second experiment the task was to equisect a given range of each features' continuum, which after the zero point assignment on the fitting function, led to ratio scales with internally consistent judgments. In addition, in the first experiment, the reliability scores for some stimulus sequences of a particular feature were considerably lower than the overall excellent reliability observed in the ratio scaling experiment. However, in both experiments, the form of the fitting functions on the interval and ratio estimations of each feature were of the same kind (i.e., power functions), with the exception of spectral spread and skewness. For spectral spread, the fitting function for the stimulus sequences with 1640 and 7800 Hz spectral centroid were quadratic polynomials, whereas for the intermediate values of spectral spread with 5600 Hz centroid the fitting function was a power function which was also used on the results of the ratio scaling experiment. For spectral skewness, although in the first experiment the best fitting function was a fifth-order polynomial whereas in the second experiment was a third-order polynomial, the resultant shapes of both functions highlight the asymmetrical judgments between negative and positive skewness, which were also pointed out in Chapter 2.

If interval measurements are needed, then these can be derived from the ratio scales, because ratio scales subsume the interval scales. In most cases, in order to test over a wide range of each features' values, separate scales were derived for each of the fundamental frequencies or spectral centroids used within each stimulus set. If a psychophysical scale for a feature at an intermediate fundamental frequency or centroid is needed, it can be derived by interpolating the coefficients of the fitting functions derived from the present study. However, for sounds having fundamental frequencies or centroids that are located below or above the tested range of this study, the derived scales should be used with caution, as their validity remains questionable before conducting any further experiments.

The construction of psychophysical scales based on univariate stimuli, allowed for the establishment of cause and effect relations between audio features and perceptual dimensions, contrary to past research that has relied on multivariate stimuli and has only examined the correlations between the two. Finally, the psychophysical scaling of audio features presented in this study is a prerequisite and essential step before one starts to study timbre as a phenomenon that emerges from a combination of audio features and explore its attributes through perceptual dominance hierarchies of those features.

# References


Almeida, A., Schubert, E., Smith, J., & Wolfe, J. (2017). Brightness scaling of periodic tones. *Attention, Perception & Psychophysics*, *79:1892*.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, *32*, 159–188.

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

Farbood, M. M., & Price, K. C. (2017). The contribution of timbre attributes to musical tension. *Journal of the Acoustical Society of America*, *141*, 471–482.

Genesis S. A. (2009). History and description of loudness models. (s.l.: Loudness Toolbox for Matlab)

George, D., & Mallery, P. (2003). *Spss for windows step by step: A simple guide and reference* (4th ed.). Boston: Allyn & Bacon.

Gescheider, G. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Gescheider, G., & Hughson, B. A. (1991). Stimulus context and absolute magnitude estimation: A study of individual differences. *Perception & Psychophysics*, *50:45*.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, *63*, 1493–1500.

Hellman, R. P., & Zwislocki, J. (1961). Some factors affecting the estimation of loudness. *Journal of the Acoustical Society of America*, *33*, 4687–694.

ISO 389-8. (2004). *Acoustics – Reference Zero for the Calibration of Audiometric Equipment – Part 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones.* (International Organization for Standardization, Geneva, Switzerland)

ISO/IEC. (2002). *MPEG-7: Information Technology – Multimedia Content Description Interface - Part 4: Audio.* (ISO/IEC FDIS 15938–4:2002)

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, *94*, 2595–2603.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caracterisation du timbre des sons complexes. 2: Analyses acoustiques et quantification psychophysique. [Characteri-

zation of the timbre of complex sounds. 2: Acoustic analysis and psychophysical quantification]. *Journal de Physique*, *4*, 625–628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzén & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (Vol. 1989, pp. 43–53). Amsterdam: Excerpta Medica.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, *62*, 426–1439.

Laurier, C., Lartillot, O., Eerola, T., & Toivianen, P. (2009). Exploring relationships between audio features and emotion in music. In *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009)* (pp. 260–264). Jyväskylä, Finland.

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology: Perception and motivation; learning and cognition* (pp. 3–74). Oxford, England: John Wiley & Sons Inc.

Marks, L. E., & Gescheider, G. (2002). Psychophysical scaling. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Perception and motivation; learning and cognition* (pp. 91–138). Hoboken, NJ, US: John Wiley & Sons Inc.

Martin, F. N., & Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *Journal of the American Academy of Audiology*, *11*, 64–66.

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

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

Montgomery, H. (1975). Direct estimation: Effect of methodological factors on scale type. *Scandinavian Journal of Psychology*, *16*, 19–29.

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

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

Poulton, E. C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, *69*, 1–19.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Schubert, E., & Wolfe, J. (2006). Does timbral brightness scale with frequency and spectral centroid? *Acta Acoustica united with Acoustica*, *92*, 820–825.

Siedenburg, K., & Müllensiefen, D. (2017). Modeling timbre similarity of short music clips. *Frontiers in Psychology*, *8:639*.

Smith, B. K. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. In *Proceedings of the meeting of the Society for Music Perception and Cognition.* Berkeley, CA: University of California, Berkeley.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, New Series*, *103*, 677–680.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review*, *78*, 426–450.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects.* Reprint, New York, NY: Routledge, 2017.

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*, 329–353.

Torgerson, W. S. (1958). *Theory and methods of scaling* (3rd ed.). Oxford, England: Wiley and Sons, Inc.

# Chapter 5

# Conclusion

The aim of this thesis was to investigate whether, how, and to what extent listeners perceive magnitude differences along timbre-related audio descriptors. To this end, several experiments were conducted to test whether audio features can be perceived on ordinal and interval scales, and to finally construct psychophysical ratio scales of each descriptor. Throughout these experiments, the following spectral and temporal audio descriptors were tested: spectral centroid, spread and skewness; harmonic odd-to-even ratio, spectral deviation, and spectral slope; attack and decay time, temporal centroid with fixed attack and decay times, and inharmonicity. The results of those experiments indicated that all the spectral (Chapter 2) and temporal (Chapter 3) audio features that were tested can be perceived on ordinal scales. Furthermore, the results of the experiments described in Chapter 4 indicated that listeners can perceive intervals and ratios between spectral feature values and therefore, enabled the construction of perceptual ratio scales for each feature tested, which is the most informative type of scale.

An important and determinant factor of the experimental methodology, was the construction of synthesized stimuli by controlling each feature independently of the rest, as well as determining empirically, after numerous informal listening tests, the appropriate ranges and intermediate feature values that would be presented in the experiments, with respect to each feature's peculiarities and the imposed limitations by the synthesis algorithms used to construct each feature's stimulus set. Another issue, which would not have been resolved without using the appropriate synthesis techniques and making a proper selection of stimulus values, is that throughout all the experiments, listeners had to discover the attribute of study (i.e., which audio descriptor was being tested) themselves by exploring the presented ranges and intermediate stimulus values of a particular audio feature, and without receiving any verbal indication from the experimenters. That is because contrary to common perceptual attributes such as pitch or loudness, which musically sophisticated participants would have been familiar with, presenting a consistent description and explanation of audio features to participants prior to the experiment, would require the use of the mathematical formulations that were used to define each audio feature, and the use of

technical terminology with which participants may not have been familiar with. Instead, the experimental methods allowed participants to develop a mental representation of each audio feature by identifying the covariance and invariances of audio feature values within a stimulus set.

## 5.1 Summary of Methods and Results

Chapter 2 perceptually validated spectral audio features through an ordinal scaling experiment. The results indicate that listeners were overall able to sort the stimuli of a particular feature set when presented with an appropriate spacing of feature (magnitude) values. However, there were cases in which most listeners could not discriminate magnitude differences between some feature values of a stimulus set (most notably for spectral skewness) and thus, these stimuli were not ordered correctly. An inspection of the respective auditory excitation patterns of these stimuli revealed that these patterns were very similar despite the equidistant spacing of feature values either on a linear or logarithmic scale, and led to almost identical percepts, which explains listeners' confusions in the ordering task. As a result of this analysis, a practical outcome of this experiment was also the identification of clearly discriminable feature values among the stimulus sets. Discriminability was required before conducting the interval and ratio scaling experiments described in Chapter 4, because the differences between stimulus values in such experiments (i.e., global psychophysical experiments) need to be supraliminal.

A similar approach was followed in Chapter 3, which presented an experiment in which the listeners' task was to rank order stimuli that varied according to temporal audio features including inharmonicity. The results underpinned the importance of amplitude envelope features in the discrimination between different attack times and decay times, and the importance of spectral envelope features for discriminating between different inharmonicity levels. In addition, the results provided further evidence for the asymmetry found in the auditory system when processing the attack and decay times of stimuli constructed with a slow attack and a fast decay, versus stimuli constructed with a fast attack and a slow decay. The synthesis process used for constructing the stimulus sets of temporal centroids enabled us to disentangle attack time (or decay time) from temporal centroid by keeping the attack (or decay) time fixed and controlling the temporal centroid by shaping the amplitude envelope during the attack segment (or decay). Although there were many confusions in ordering short attack and decay times, listeners performed well in ordering temporal centroids even at very short attack and decay times. A meta-analysis of six timbre spaces was therefore conducted to test the explanatory power of attack time versus the attack temporal centroid (ATC) along a perceptual dimension derived from multidimensional scaling (MDS). The results indicate that ATC is a robust feature for explaining dissimilarity ratings along a spectrotemporal dimension, which has overall greater explanatory power than attack time itself.

Chapter 4 presented the interval and ratio scaling experiments of spectral audio features. There were two main factors that needed to be taken into account before concluding on which experimental method would be the most appropriate for the purpose of each experiment. The first factor was that most participants (excluding sound engineers or listeners familiar with psychoacoustic experiments) rarely experience most of the tested features in isolation and thus, as previously mentioned, they were not familiar with the attributes of study. The second factor is that some features are essentially perceived by detecting loudness differences between individual spectral components and therefore have a narrow usable range (e.g., harmonic odd-to-even ratio). This factor has important consequences for the pool of stimuli that experimenters have at their disposal when conducting magnitude estimation experiments, for which the presentation of supraliminal stimuli is a requirement. After conducting several pilot experiments that took these factors into account, it was concluded that partition scaling methods were the most appropriate for deriving interval and ratio scale measurements of spectral audio features. Moreover, the experimental design used in each of the two experiments allowed us to control for the biases of *order*, *hysteresis*, and *range* effects.

The listeners' task in the first experiment was the estimation of the relative differences between successive levels of a particular audio feature, and thus this experiment provided interval scale measurements. The results of this experiment indicate that listeners were overall able to perceive intervals of spectral audio features. As previously mentioned, for this particular task, it was imperative that the presented stimuli were *supraliminal*, which were therefore chosen based on the results of the ordinal scaling experiment described in the second chapter of this thesis. Contrary to the first experiment, which was based on interval estimation and in which listeners had no control over the stimulus values, in the second experiment, listeners had total control over the stimulus values and were asked to partition a continuum into a number of equal-sounding intervals. The equality of sensory intervals implies that the intervals themselves have ratio properties and thus, the results of that experiment led to ratio scale measurements that enabled the construction of psychophysical ratio scales of spectral audio features.

At this point, the reader should be reminded that a main difference between interval and ratio scales is that the former are constructed using an arbitrary point across all the possible stimulus values as a reference, whereas the latter are constructed using the point of absolute zero as a reference. According to Stevens (1975), the absolute zero point in psychophysical scaling is only an abstraction because if literally interpreted it would indicate the total absence of the stimulus rather than a perceptual threshold above which sensation levels can be measured. However, most audio features do exhibit an absolute zero: zero spectral spread indicates the presence of just a single frequency component in the spectrum, which has the same value with the spectral centroid; zero skewness indicates a symmetrical distribution around the spectral centroid; zero spectral slope indicates a flat spectrum; zero spectral deviation indicates that each spectral component has the same

amplitude with the average of itself and its two neighbor components, leading to a smooth (averaged) amplitude distribution of harmonic components rather than to a "jagged" amplitude distribution; a zero point mapped to an odd-to-even ratio of one, indicates that the odd frequency components have the same energy with the even components and therefore, the result of subtracting the energies of the two would be zero. The only exception is spectral centroid, which is related to the perception of timbral brightness, and for which the zero point was assigned empirically at 20 Hz, which marks the lower limit of pitch perception.

## 5.2 Contributions to knowledge

To the best of our knowledge, there have been no previous attempts in psychophysical scaling of timbre-related audio features other than perhaps the preliminary results of Almeida, Schubert, Smith, and Wolfe (2017) who attempted to derive a ratio scale of auditory brightness as a function of spectral centroid. However, the authors only provide an equation that approximates the spectral centroid ratio needed to double the perceived brightness, and which is only valid for centroids in the range of 500 to 1160 Hz. It should also be noted that the spectral centroid of those stimuli was adjusted by varying the spectral slope, which although in that particular case was linearly dependent on centroid, also co-varied strongly with spectral spread and skewness (see also Caclin, McAdams, Smith, & Winsberg, 2005 for a similar approach used for adjusting the spectral centroid). Therefore, it can be concluded that the presented equation of perceived brightness should relate not only to changes in spectral centroid but also to changes in both spectral spread and skewness. In the present thesis, a significant amount of effort was invested in constructing synthesized stimuli by controlling each feature independently of the rest, which therefore allowed the establishment of cause and effect relations between each of the tested audio features and perceptual dimensions through a preplanned sequence of psychophysical scaling experiments.

Throughout this thesis, the sequence of conducted experiments was determined according to the perceptual hierarchy of scales, starting from the least informative ordinal scale and progressing to the most informative ratio scale. However, a reader might question the necessity of the experiments conducted prior to the experiment which provided ratio scale measurements, and to whom we may respond by posing the following rhetorical questions: i) Would the results of the interval estimation experiment and the derived interval scales be trustworthy if the assumption of ordinal scalability was violated? Or in other words, could we confidently attribute all the observed variances in listeners' ratings to imprecise interval estimations without presuming ordinal scalability of the presented features? ii) Would the production of equal sensory intervals and thus, the ratio scale measurements be trustworthy, if listeners had not been able to estimate intervals in the first place? And how should the results be interpreted if we had observed a minimum (or in some cases even zero) variance between listeners' equisections? Would that imply that listeners could

estimate ratios with extreme precision, or should the results be regarded as disappointing and attributed to a bad experimental design, given that a certain amount of variance was to be expected from the results of the interval estimation task?

The results of the ordinal scaling experiment on spectral audio features (Chapter 2) outlined trajectories of spectral audio features that causally correspond to listeners' perceptions. In addition, the ordinal scalability of all the tested features suggests that people can perceive contours in other audio features as well, besides spectral centroid and loudness that have been previously examined from McDermott, Lehr, and Oxenham (2008). These authors reported that contours in those features are nearly as useful as pitch contours for recognizing familiar melodies that are normally conveyed via pitch. More recently, Siedenburg and Müllensiefen (2017) empirically demonstrated that timbre-related audio features play a major role in listeners' similarity judgments of short music clips with a maximum duration of 800 ms. Thus, the concept of timbral contours (i.e., contours extracted from timbre-related audio descriptors) could lead to better audio processing strategies for music information retrieval (MIR) tasks such as music recommendation or genre classification systems that would better reflect listeners' notions of genre and musical similarity. Another potential application field for timbral contours would be the development of computational music theory tools for the analysis of contemporary music, especially for musical works where pitch itself plays a minor role, as well as contours in other features that dominate listeners' perceptions of the musical material (see, for instance, Noble & McAdams, 2020).

The results of the ordinal scaling experiment on temporal audio features (Chapter 3) suggest that the ordinal scalability of attack and decay time does not solely depend on those features alone, but also on the shape of the amplitude envelope which is implicitly encoded in the temporal centroid computed over the attack (i.e., the attack temporal centroid: ATC) or decay time. Although the slope of attack time (Peeters, Giordano, Susini, Misdariis, & McAdams, 2011) also captures some aspects of the amplitude envelope during the attack time, it has only been used as a variable for refining predictive models of perceptual attack time (Collins, 2006; Gordon, 1987), and its perceptual relevance has not been directly investigated. In addition, its computation as offered in the Timbre Toolbox (Peeters et al., 2011) and the MIRtoolbox (Lartillot & Toiviainen, 2007) give inaccurate predictions to empirical results derived from listening experiments (Nymoen, Danielsen, & London, 2017). A qualitative interpretation of the results related to the ordinal scaling of temporal centroids leads to the conclusion that differences between short temporal centroids perceptually manifest as spectral, whereas differences between longer temporal centroids manifest as temporal. The above conclusion combined with the results of the meta-analysis on previously reported timbre spaces, indicates that the temporal centroid and attack time collapse onto a single spectrotemporal perceptual dimension associated with ATC, which in the previous studies had been considered to be strictly temporal and was associated only with attack time. This finding has practical implications in blend prediction models that rely on timbre spaces and which take into account not only the spectral but also the tem-

poral aspect of the constituent sounds (Kendall & Carterette, 1993; Sandell, 1995; Tardieu & McAdams, 2012). In addition, given that ATC is a robust descriptor for explaining dissimilarity ratings along a spectrotemporal perceptual dimension, it could also be used as an additional constraint in sound morphing strategies, which would therefore lead to perceptually smoother transitions between the morphed amplitude envelopes of the source and target sounds.

Some researchers have suggested that the timbre dimension associated with spectral centroid may only exist on an ordinal scale (McAdams & Siedenburg, 2019). However, the results of the interval and ratio scaling experiments (Chapter 4) demonstrate that listeners can also perceive intervals as well as ratios between spectral descriptor values, which allows for the construction of psychophysical ratio scales. The assigned units on the derived psychophysical ratio scales are all of comparable orders of magnitude, in order to ease quantitative comparisons between perceptual intervals of different descriptors. The derived scales along with their respective units designate a *perceptual coordinate system of audio features*, in which sounds can be grouped and ordered (on an approximately uniform grid formed by the units of the scales) according to their perceived sound qualities that relate to each descriptor. This approach expands the concept of timbre spaces, which are derived from MDS algorithms, by limiting the possibility of having *metameric matches* between perceptually different sounds located within a perceptual feature space. The perceptual coordinate system of descriptor values could potentially act as a "control surface" for music technology applications that include computer aided orchestration, perceptually motivated sound effects and synthesis algorithms, and constrained approaches to sound morphing (Chapter 1).

As mentioned in Chapter 1, audio descriptors have been widely used as predictor variables in statistical regression models for interpreting and predicting listeners' responses on a variety of tasks that relate to timbre. However, the physical values of these predictors may lead to false-positive interpretations about their perceptual significance on a particular task. The derived scales allow timbre researchers to use perceptually informed values of spectral descriptors as predictors in their statistical models that may lead to more sustainable conclusions and accurate interpretations in terms of perception. In addition, a practical consequence of using perceptually informed descriptor values is that descriptors which may be physically correlated may become uncorrelated when measured on psychophysical scales, which offers many advantages in the statistical analyses. However, the opposite is also true: descriptors that may be physically uncorrelated may become correlated when measured on perceptual scales. In that unfortunate scenario, the researchers at least have in their disposal the units of the psychophysical scales that can be used as indicators for evaluating the (individual) magnitude contributions of each descriptor in terms of perceptual feature-intervals between stimuli.

## 5.3 Limitations and future directions

Unlike controlled experimental conditions that use synthesized stimuli such as the ones in the presented experiments and in which audio descriptors remained constant over the duration of the stimuli, in most applications descriptor values vary over successive analysis frames (especially during the attack time). Although time series of descriptor values have been used in sound morphing (Appendix 2), computer orchestration environments (Carpentier & Bresson, 2010), and for evaluating the quality of synthesized sounds that imitate natural instrument sounds (Kendall, Carterette, & Hajda, 1999), their effect on the perception of timbral contours remains unclear. In order to derive timbral contours extracted from sounds presented either in isolation or sequentially, further research is needed for determining perceptually relevant temporal windows over which descriptors are computed, or the appropriate summary statistics of descriptor values (e.g., mean, median, interquartile ranges) when computed over several successive analysis frames.

This thesis has only investigated interval and ratio perception of spectral descriptors. A similar experimental paradigm could be used for investigating interval and ratio perception of temporal descriptors. However, such undertaking would first require having a reliable estimation method for predicting perceptual attack time (PAT), which is a complicated phenomenon that relates both to the spectral envelope and amplitude envelope characteristics of a particular sound, but not necessarily to attack time itself. Current models of PAT do not generalize to all kinds of sounds and are only useful for a subset of stimuli with similar spectrotemporal properties (for a recent overview of available models see Bechtold & Senn, 2018; London et al., 2019). Nonetheless, experiments on that direction could start by investigating the interval and ratio properties of temporal centroid, which can be partly disentangled from the total duration of a stimulus when a procedure such as the one described in Chapter 3 is used.

For some descriptors, especially for the spectral descriptors centroid, spread, and skewness, the synthesis procedure used for constructing the stimuli imposed limitations on the control of range effects. In other words, the scales for those descriptors were derived from stimuli that covered only a portion of each descriptor's total usable range and by extrapolating the fitting function outside the tested range. Therefore, more experiments are needed to validate the derived scales outside the tested range. In addition, the psychophysical scales were derived according to listeners' estimations on harmonic stimuli. Even though this procedure has its own advantages, it would be instructive to test in the future whether the scales derived from harmonic stimuli can generalize to sounds that exhibit noisy spectra.

The perceptual coordinate system of descriptor values, with its axes derived from the psychophysical scales of each descriptor, does not yet constitute a timbre space. The reason for that is because, although most of the tested descriptors are physically independent, the independence in the descriptor space may not hold along orthogonal perceptual dimensions, which is the basic premise of MDS algorithms that are used to derive the axes of a timbre space. This thesis has provided evidence that individual descriptors can be perceived

on perceptual scales when the rest remain constant, but it did not test the extent to which each descriptor is independently perceived when multiple descriptors covary. Indeed, several studies have pointed out the perceptual interference among audio features (Caclin et al., 2005; McAdams, Beauchamp, & Meneguzzi, 1999; McDermott et al., 2008) and pose the question of whether the source of these dependencies is from learned associations (or covariation) between timbre, pitch, and dynamics found in mechanical instruments, music in general, or in speech (McAdams, Tse, & Wang, 2016; McDermott et al., 2008) . Along those lines, further experiments are needed to test the relative perceptual importance of each descriptor when their values covary, and the results of this thesis constitute an essential first step toward that direction.

In addition, it would be interesting to conduct experiments using the experimental paradigm of *functional measurement* (Anderson, 1970) for identifying conditions under which psychological dimensions result from a combination of descriptors. For instance, in Chapter 3 a spectrotemporal dimension was identified through a single descriptor constructed by a combination of attack time with temporal centroid. In general, such dimensions are most likely to arise from a combination of physically interrelated descriptors such as spectral slope, odd-to-even ratio, spectral deviation, and harmonic tristimulus values, that are possible to individually control (up to a certain extent), and which all contribute to a description of the global spectral envelope. However, in most studies, the psychological dimensions that may result from a combination of descriptors are usually interpreted through correlational analyses between listeners' ratings and a set of descriptors derived after data reduction techniques (e.g., PCA), which are agnostic with respect to the relative contributions of individual descriptors and their potential interactions (McAdams, Roussarie, Chaigne, & Giordano, 2010). The experimental paradigm of functional measurement in conjunction with the perceptual scales derived from the present thesis would allow timbre researchers to abandon such agnostic approaches, which only pertain to a particular stimulus set, in favor of perceptually informed metrics that could generalize along a broader set of sounds generated either from mechanical or electroacoustic sound sources.

## 5.4 Concluding remarks

Although indeed, *Timbre is a Many-Splendored Thing* (Thoret, Goodchild, & McAdams, 2018), I chose to follow a relatively unexplored timbre path full of cobwebs, which fortunately led to the establishment of psychophysical correspondences between perception and several timbre-related audio descriptors. It is also true that if any of my initial hypotheses tested throughout the various stages of this research were violated (i.e., ordinal, interval, and ratio scalability of audio descriptors), I would have been thesis-less by now, a possibility that was also brought up by one of the committee members who commented on my initial thesis proposal. Fortunately, the facts show otherwise. The findings of this thesis advance the current knowledge on timbre perception both by establishing cause-and-effect

relations between audio descriptors and perceptual dimensions, and by expanding previous research in which the acoustical interpretations of perceptual dimensions were made solely under the prism of correlational analyses. The psychophysical correspondences between perception and audio descriptors reported in this thesis will hopefully serve as a basis for future research, which may attempt to study timbre as a phenomenon that emerges from a combination of audio features and explores its psychophysical attributes through perceptual dominance hierarchies of those features.

# Appendix A

# A Performance Evaluation of the Timbre Toolbox and the MIRtoolbox on Calibrated Test Sounds

This appendix is based on the following conference paper:

Kazazis, S., Esterer, N., Depalle, P. and McAdams, S. (2017). A performance evaluation of the Timbre Toolbox and the MIRtoolbox on calibrated test sounds. In *Proceedings of the 2017 International Symposium on Musical Acoustics (ISMA2017)*, Montreal, Canada, June 18–22, 2017

**Abstract** We evaluate the accuracy of the Timbre Toolbox (v.1.2) and the MIRtoolbox (v.1.6.1) on audio descriptors that are putatively related to timbre. First, we report and fix major bugs found in the current version of the Timbre Toolbox, which have gone previously unnoticed in publications that used this toolbox as an analysis tool. Then, we construct sound sets that exhibit specific spectral and temporal characteristics in relation to the descriptors being tested. The evaluation is performed by comparing the theoretical (real) values of the sound sets to the estimations of the toolboxes.

## A.1 Introduction

The Timbre Toolbox [1] and the MIRtoolbox [2] are two of the most popular MATLAB [3] toolboxes that are used for audio feature extraction within the music information retrieval (MIR) community. They have been recently evaluated according to the number of presented features, the user interface and computational efficiency [4], but there have not been performance evaluations of the accuracy of the extracted features. The aim of this paper is: (1) to detect and summarize the bugs in the current version of the Timbre Toolbox and (2)

to evaluate the robustness of audio descriptors these toolboxes have in common and that are putatively related to timbre. For this purpose, we synthesized various sound sets using additive synthesis, calculated the theoretical (real) values of each descriptor tested, and compared these values with the estimations of the toolboxes. Section A.2 summarizes the bugs found in the current publically available version of the Timbre Toolbox (v.1.2). Section A.3 describes the construction of the sound sets used for evaluating the performance of the MIRtoolbox (v.1.6.1) and a beta version of the Timbre Toolbox, which fixes the reported bugs. Section A.4 presents the results of the evaluation and Section A.5 summarizes our findings.

## A.2 Points of consideration and bug fixing in the Timbre Toolbox

In this section, we report the bugs found in the current version of the Timbre Toolbox and some issues related to user interaction. The Timbre Toolbox incorporates the following sound models of the time-domain signal for extracting audio descriptors: the temporal energy envelope; the short-term Fourier transform (STFT) on a linear amplitude scale (STFTmag) and a squared amplitude scale (STFTpow); the output of an auditory model based on the concept of the Equivalent Rectangular Bandwidth, which is either calculated using recursive gammatone filters (ERBgam), or their finite impulse response approximation using the fast Fourier transform (ERBfft); and a sinusoidal harmonic model [1].

In some cases, especially when the amplitude of the lower frequencies is lower than the upper ones, the harmonic representation using the default amplitude threshold for detecting harmonics will not analyze even strictly harmonic sounds. Furthermore, the default analysis limit of 20 harmonics could also be problematic for analyzing low-frequency sounds having spectral energy that increases with harmonic number. However, this scenario is very unlikely to occur in natural sounds, but it is still possible with synthetic sounds used in psychoacoustic experiments (e.g., [5]) or in electroacoustic music. Another conceptual bug is the estimation of inharmonicity: according to Eq. A.1, which is presented in [1], a signal with a fundamental frequency of 100 Hz and a partial at 150 Hz will be less inharmonic than a signal with the same fundamental and a partial at 190 Hz even though the partial of the second signal is only detuned by 10 Hz below the next harmonic.

$$\text{inharm} = \frac{2}{f_0} \frac{\sum_{h=1}^{\text{H}} (f_h - h f_0) a_h^2}{\sum_{h=1}^{\text{H}} a_h^2} \tag{A.1}$$

In v.1.2, the end user only had access to summary statistics, and as such it was not possible to evaluate the time-varying patterns of audio descriptors. Furthermore, the export format of the results was a text file. This did not facilitate further processing of the results especially in the case of a batch analysis where the output consists of several text files. Also, MATLAB ran out of memory when the Timbre Toolbox processed long audio files.

According to Peeters et al. [1], the window that should be used for the harmonic analysis is a Blackman window. However, in the toolbox's implementation, the window is a boxcar (i.e., no window weighting at all), but we also noticed that the removal of the window's energy contribution to the input sound was implemented incorrectly. Furthermore, some calculations on audio descriptors returned the results in normalized frequency (including the spectral centroid) without warning the user and led to misinterpretations (e.g., [6]).

Although the actual sampling rate is read directly from the file, in some sound models it was not actually used: the parameters related to the FFT analysis were specified according to a fixed sampling rate of 44.1 kHz no matter the actual sampling rate of the input file. Finally, in most of the employed sound models, the computations of spectral spread, skewness, kurtosis and spectral slope were implemented incorrectly.

The analysis results presented in this paper are based on a beta version of the Timbre Toolbox that fixes and takes into consideration all of the above-mentioned points except the calculation of inharmonicity and the threshold settings used in the harmonic representation.

## A.3 Construction of the test sound sets

The sounds were constructed using additive synthesis, which allows for a direct computation of the audio descriptors. Each sound set was designed to exhibit specific sound qualities that are directly related to the descriptors being tested. In this way, we are able to systematically test the performance of the toolboxes by tracking the circumstances under which certain audio descriptors are poorly calculated. All sounds were synthesized at 44.1 kHz with 16-bit resolution and peak amplitude of 6 dB relative to full scale (dBFS). To avoid spectral spread induced by an abrupt onset and offset when performing the FFT on these synthetic sounds, we applied a 10-ms raised inverse cosine ramp to all sounds except the ones used to test the attack time and the attack and decrease slopes. Durations were fixed at 600 ms and all sounds contained harmonics up to (but not including) the Nyquist frequency.

We used the following fundamental frequencies for all the sound sets except those related to the temporal energy envelope: C#1 (34.65 Hz), D2 (73.42 Hz), D#3 (155.56 Hz), C4 (261.63 Hz), E4 (329.63 Hz), F5 (698.46 Hz), A5 (880 Hz), F#6 (1479.98 Hz), G7 (3135.96 Hz) and B7 (3951.07 Hz). The C4 was slightly detuned from 261.63 Hz to 258 Hz in order to match exactly the frequency of an FFT bin and to test whether the estimations would be improved; for a sampling frequency of 44.1 kHz and an FFT size of 1024 samples (default setting of the Timbre Toolbox) the bins are spaced 43 Hz apart. We used such a wide frequency range because as the fundamental frequency increases and approaches the Nyquist limit, the number of "significant" FFT bins decreases, which may affect the accuracy of the results, especially in the presence of noise.

### A.3.1 Attack Time, Attack Slope and Decrease Slope

The Timbre Toolbox uses the "weakest effort method" for estimating the attack time and the attack and decrease slopes [1], whereas the MIRtoolbox uses a similar method based on Gaussian curves [2]. In these adaptive methods, the threshold energy level that the signal must surpass is not fixed, but is determined as a proportion of the maximum of the signal's energy envelope. The attack and decrease slopes are then estimated as the average temporal slope during the start and end times of the attack portion. An 'effort' is defined as the time it takes for the signal to go from one threshold value to the next. It is therefore logical to assume that if the signal varies rapidly and non-linearly during the attack time, the true attack time values may be poorly estimated.

For testing the accuracy of this method, we constructed nine attack envelopes for each of ten logarithmically spaced attack times ranging from 1 to 300 ms. The shape of the envelopes was determined by:

$$y(t) = mt^{\mathrm{b}} \tag{A.2}$$

where $m$ controls the slope of the attack time and b is a curvature constant which was assigned the following values: 3, 2.5, 2 and 1.5 for an exponential shape; 1 for a linear shape; and 0.67, 0.5, 0.4 and 0.33 for a logarithmic shape. The attack envelopes were then applied to a flat harmonic spectrum with a fundamental frequency of 258 Hz and a total duration of 600 ms. A similar procedure was used for testing the estimations of decrease slope.

### A.3.2 Spectral Centroid

For this sound set, we used a flat spectrum with octave-spaced harmonics and included in the above-mentioned set a lower fundamental of C0 (16.35 Hz). In order to systematically test the accuracy of spectral centroid estimation, we iteratively removed just one harmonic from the initial spectrum up to the last one for every fundamental. This way, the sounds generated from the last fundamental just contain a single frequency component, because there is only one harmonic present due to the Nyquist limit, and therefore the spectral centroid ideally should match the value of the fundamental frequency estimation.

### A.3.3 Spectral Spread, Skewness, Kurtosis and Roll-off

For testing the estimations of spectral spread, skewness, kurtosis, and roll-off, we designed a sound set in which the sounds vary by fundamental frequency and according to spectral slopes. By precisely controlling the spectral slopes, we directly alter in a predictable way the higher statistical moments of the spectrum and the frequency below which 95% of the signal energy is contained. In our analysis, we took into account the fact that the MIRtoolbox uses a default value of 85%. For every fundamental, we constructed a spectrum that contained

both odd and even harmonics with a $1/n^2$ power decrease, where $n$ denotes the harmonic number. Then in nine steps we altered linearly the energy distribution of the harmonics until we reached a flat spectrum. The same procedure was repeated by starting from a flat spectrum and reaching in nine steps a positive slope of the harmonics which had an $n^2$ power increase.

### A.3.4 Harmonic Spectral Deviation and Spectral Irregularity

Spectral deviation (in the Timbre Toolbox) and spectral irregularity (in the MIRtoolbox) are the same descriptors but are computed slight differently with respect to a scaling factor. MIRtoolbox offers two estimation methods based on Jensen [7] and Krimphoff et al. [8]. Here, we only tested the estimation based on Krimphoff's method (Eq. A.3.), which is the only option available in the Timbre Toolbox. For every fundamental, we started from a flat spectrum that only contained the fundamental with even harmonics, and we gradually increased the level of the odd ones until we reach a flat spectrum in ten steps.

$$\mathrm{dev} = \sum_{h=2}^{\text{H-1}} \left| a_h - \frac{a_{h-1} + a_h + a_{h+1}}{3} \right| \tag{A.3}$$

### A.3.5 Spectral Flatness

To evaluate the accuracy of the estimations of spectral flatness, we applied a Gaussian spectral window centered at the middle harmonic to a flat spectrum that contained both odd and even harmonics, and progressively altered its standard deviation in ten steps so that the last window resulted in an extremely peaky spectrum. For altering the width of the window we used the following coefficients, which are proportional to the reciprocal of the standard deviation: 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7 and 8. This process was done for the whole range of fundamentals.

### A.3.6 Inharmonicity

This sound set is similar to A.3.2, but here we used inharmonic spectra. The inharmonic components were kept fixed in the whole sound set, and were spaced according to an inharmonicity coefficient that controlled the amount of deviation from each harmonic, which varied linearly from 0 to 0.5 with respect to the harmonic number. Inharmonicity was increased by gradually increasing the amplitude of the inharmonic components instead of increasing their deviation from the harmonics. The inharmonic components were initially attenuated with a $1/n^2$ envelope to reach a flat spectrum in ten steps $\mu$ by gradually increasing linearly their energy distribution.

## A.4 Results

We evaluate toolbox performance by analyzing the sound sets with each toolbox and calculating the normalized root mean squared (RMS) error between their output and the theoretical values. The theoretical values were calculated using either the power or magnitude scale depending on the input representation being tested. MIRtoolbox's default input representation using 'mirspectrum' is based on a STFT with a Hamming window and a half overlapping frame length of 50 ms, which is similar to the 'STFTmag' representation used in the Timbre Toolbox. For the Timbre Toolbox, we tested all the available input representations because there is no default option. For analyzing the sounds, we used the default settings of each toolbox, and the summary statistics from the frame-by-frame analysis were derived using the median values.

### A.4.1 Temporal Energy Descriptors

MIRtoolbox uses two estimation methods for calculating the attack and decrease slopes: 'Diff', which computes the slope as a ratio between the magnitude difference at the beginning and end of the attack period and the corresponding time difference; and 'Gauss', which is similar to Peeters' method [1]. Table A.1 shows the results of the error analysis. The observed general trend for both toolboxes was that short attack times (about less than 40 ms) were significantly overestimated, whereas longer attack times were mainly underestimated. The Timbre Toolbox also systematically estimated the exponential attacks as being longer than the logarithmic attacks.

| **Descriptors** | **Timbre Toolbox** | **MIRtoolbox** (Diff / Gauss) |
|---|---|---|
| Attack Time | 24.40 | 21.57 |
| Attack Slope | 36.85 | 36.15 / 36.82 |
| Decrease Slope | 37.31 | 37.53 / 37.36 |

**Table A.1** RMS error (%) of temporal energy descriptors.

### A.4.2 Spectral and Harmonic Descriptors

Although we tested the accuracy of extracted descriptors on all sound sets, due to space limitations, the evaluation results presented in Table A.2 are based only on the designated sets for each descriptor, which were presented in the previous section. Also, we only report the most accurate results (i.e., the minimum RMS error) among the Timbre Toolbox's different input representations. In the following, we present a qualitative inspection of the errors with respect to the sound sets.

*Centroid*: MIRtoolbox always overestimates slightly the centroids, whereas the Timbre Toolbox returns accurate results for fundamentals of 65.4 Hz and above.

*Higher-order moments of the spectrum and roll-off*: the MIRtoolbox was numerically unstable returning 'Not a Number' (NaN) in the estimation of spectral centroid for the sets with fundamentals of 34.65 Hz and 73.42 Hz. Table A.2 summarizes the results after removing the sounds for which MIRtoolbox returned NaNs. Timbre Toolbox's STFTpow representation provides overall the most accurate estimations even when all sounds were included in the analysis, in which case it produced a 1.37% RMS error for spectral roll-off.

*Spectral Flatness*: MIRtoolbox again returned Not a Number for some of the sounds with fundamentals of 34.65 Hz and 73.42 Hz, and although this sound set was not designed to test the estimation of spectral irregularity, MIRtoolbox did not provide any results for the estimation of this descriptor and exited with an error message. We also noted that in both toolboxes, as the fundamental frequency increases and spectral spread decreases, the estimations of spectral spread become more erroneous, although limited within a small margin. Although the spectral centroid and higher moments were estimated quite accurately in both toolboxes, the estimation of spectral flatness was inaccurate.

*Spectral irregularity* (or *deviation*): Harmonic spectral deviation is only available in the harmonic representation of the Timbre Toolbox. However, we were not able to run the analysis on the whole sound set using the default amplitude threshold setting for harmonic detection: as the fundamental frequency increases, the settings should be lowered, otherwise the sound will not be further analyzed (in the beta version tested here, the user gets warned whenever this situation occurs). The MIRtoolbox also proved to be erroneous for the estimation of this descriptor. However, both toolboxes returned quite accurate results for the spectral centroid and higher-order moments for this sound set.

*Inharmonicity*: The estimations of inharmonicity could not be quantitatively evaluated due to the current behavior of the Timbre Toolbox, as mentioned previously, and the unavailability of the precise equation used by MIRtoolbox. Qualitatively, and given the way this sound set was constructed (section A.3.6), we expect the estimation of inharmonicity to increase for the subsets of each fundamental. Fig. A.1 shows the estimations of the MIRtoolbox, which seem to be more plausible after the fifth set of fundamentals (i.e., from F5 up to B7, section A.3).

## A.5 Conclusions

Before evaluating the accuracy of the toolboxes, we reported and fixed in a beta version the major bugs, configuration, and presentation issues that were encountered in the current version of the Timbre Toolbox (v. 1.2). Our evaluation on synthetic test sounds shows that for spectral descriptors, the Timbre Toolbox performs more accurately and on some sound sets outperforms the MIRtoolbox, with the short-term Fourier transform power representation (STFTpow) being overall the most robust. The estimations of spectral centroid and

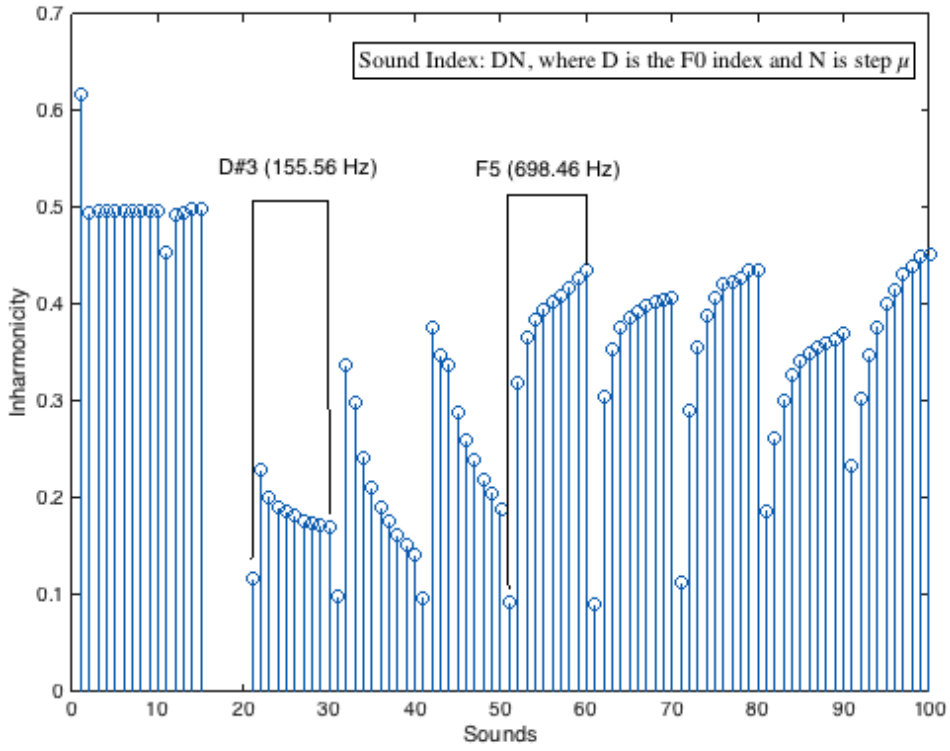


**Fig. A.1** Inharmonicity estimation in the MIRtoolbox. The horizontal axis indicates the sound index $DN$, and the vertical axis the relative deviation of the partials from purely harmonic frequencies. The missing values correspond to NaN.

higher order moments of the spectrum were quite accurate with small errors except the estimation of spectral flatness, which both toolboxes estimated erroneously. The Timbre Toolbox failed to analyze some sounds using the harmonic representation with the default settings even though all sounds were strictly harmonic. In the beta version tested here, if this situation occurs, the estimation of fundamental frequency is automatically set to zero, which affects the calculation of all descriptors related to this representation. However, the user receives a warning message in order to alter the default settings appropriately. The MIRtoolbox's estimations of spectral centroid on some sounds, and spectral irregularity and inharmonicity on a specific sound set proved to be numerically unstable returning NaN, or exiting with error messages without providing any results. For descriptors that are based on the estimation of the temporal energy envelope, both toolboxes perform almost equally but poorly. We noticed that in this case the errors depend both on the attack or decay times and on the shape of the slopes. The test sound sets are available at:https://www.mcgill.ca/mpcl/resources-0/supplementary-ma-terials

| **Descriptors** | **Timbre Toolbox** | **MIRtoolbox** |
|---|---|---|
| Centroid | 01.21 (STFTpow) | 03.56 |
| Spread | 00.00 (STFTpow) | 02.95 |
| Skewness | 02.06 (STFTmag) | 03.82 |
| Kurtosis | 04.31 (STFTmag) | 06.87 |
| Roll-off | 00.00 (STFTpow) | 01.57 |
| Flatness | 34.87 (ERBgam) | 51.82 |
| Irregularity | N/A | 31.36 |

**Table A.2** RMS error (%) of spectral energy descriptors. In the Timbre Toolbox, spectral irregularity could not be evaluated after the fifth set of fundamentals (section A.3).

# References

[1] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals," *J. Acoust. Soc. Am.,* 130, pp. 2902-2916, 2011.

[2] O. Lartillot, *MIRtoolbox 1.6.1 User's Manual,* Technical report, Aalborg University, Denmark. December 2014.

[3] http://www.mathworks.com

[4] D. Moffat, D. Ronan, and J. D. Reiss, "An Evaluation of Audio Feature Extraction Toolboxes," in *Proc. DAFx,* Trondheim, Norway, 2015.

[5] A. Caclin, S. McAdams, B.K. Smith, and S. Winsberg, "Acoustic Correlates of Timbre Space Dimensions: A Confirmatory Study Using Synthetic Tones," *J. Acoust. Soc. Am.,* 118 (1), pp. 471-482, 2005.

[6] C. Douglas, *Perceived Affect of Musical Instrument Sounds,* Master's thesis, McGill University, Montreal, Canada. June 2015.

[7] K. Jensen, *Timbre Models of Musical Sound: From the model of one sound to the model of one instrument*, PhD thesis, University of Copenhagen. 1999.

[8] J. Krimphoff, S. McAdams, and S. Winsberg, "Caractérisation du timbre des sons complexes. II : Analyses acoustiques et quantification psychophysique," *Journal de Physique,* 4(C5), 625-628, 1994.

# Appendix B

# Sound Morphing by Audio Descriptors and parameter interpolation

This appendix is based on the following conference paper:

Kazazis, S., Depalle, P. and McAdams, S. (2016). Sound morphing by audio descriptors and parameter interpolation. In *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx16)*, Brno, Czech Republic, September 5–9, 2016

**Abstract** We present a strategy for static morphing that relies on the sophisticated interpolation of the parameters of the signal model and the independent control of high-level audio features. The source and target signals are decomposed into deterministic, quasi-determini-stic and stochastic parts, and are processed separately according to sinusoidal modeling and spectral envelope estimation. We gain further intuitive control over the morphing process by altering the interpolated spectrum according to target values of audio descriptors through an optimization process. The proposed approach leads to convincing morphing results in the case of sustained or percussive, harmonic and inharmonic sounds of possibly different durations.

## B.1 Introduction and related work

Sound morphing plays an important role in many areas including sound design for compositional applications and video games, speech manipulation, and in generating stimuli with specific and controllable acoustic parameters that are used in psychoacoustic experiments [1, 2]. Despite the extensive literature on this topic, there is no consensus on a single definition of audio morphing, and an extensive discussion on different viewpoints can be

found in [3]. In this paper we present a strategy for stationary morphing, as opposed to *dynamic morphing*, in which a source sound gets continuously transformed over time into a target sound. We consider *static morphing* as a process that hybridizes a source sound with target sounds, or target audio features, through the independent manipulation of acoustic parameters.

Additive synthesis is one of the most flexible techniques, and as such many morphing strategies rely on interpolating the parameters of a sinusoidal model [4, 5, 6, 7, 8]. Tellman et al. [4] first pair the partials of the two sounds by comparing their frequency ratios to the fundamental frequency, and afterwards they interpolate their frequency and amplitude values. They also time-scale the two sounds to morph between their tremolo and vibrato rates based on assumptions that usually do not hold in the case of most natural sounds. Osaka [5] first performs dynamic time warping (DTW), and then he finds partials' correspondences by dynamic programming. The residual is modeled with short partials and is morphed according to stochastic parameter interpolation with hypothesized distributions. Fitz et al. [6] estimate the parameters of the "Bandwidth Enhanced Model" [9] by reassigned spectrograms, and use morphing envelopes to control the evolution of the frequency, amplitude, bandwidth and noisiness of the morph. Haken et al. [7] use a similar technique to morph in real time between pre-analyzed sounds that are placed in a three-dimensional timbre control space. Boccardi and Drioli [8] use Gaussian Mixture Models (GMM) to morph only the partials' magnitudes, which are derived from Spectral Modeling Synthesis (SMS) [10]. According to Boccardi and Drioli, since the morphing is based only on magnitude transformations, the source and target signals should belong to the same instrument family.

Other morphing strategies rely on interpolating the parameters of a source-filter model. Slaney et al. [11] construct a multidimensional space that encodes spectral shape and fundamental frequency on orthogonal axes. Spectral shape is derived through Mel-Frequency Cepstral Coefficients (MFCC) and fundamental frequency by the residual spectrogram. The optimum temporal match between the source and target sounds is found using DTW based on MFCC distances. The smooth and pitch spectrograms are interpolated separately. Ezzat et al. [12] argue that interpolating the spectral envelopes by simple cross-fading, as in [11], does not account for proper formant shifting. They describe a method for finding correspondences between spectral envelopes so as to encode the formant shifting that occurs from a source to a target sound. The morphing is based on interpolating the warped versions of the two spectral envelopes, and morphing between the residuals is left for future work.

Other authors claim to control synthesis parameters or to mor-ph according to perceptual dimensions by using high-level audio features. Hoffman and Cook [13] propose a general framework for feature-based synthesis according to an optimization scheme that maps synthesis parameters to target feature values. The results are very preliminary: the source sound consists of stationary sinusoids and noise that is spectrally shaped through

MFCCs; the target features are limited to spectral centroid, spectral roll-off and fundamental frequency histograms. Park et al. [14] treat single features as modulation signals that are applied to a source sound. According to their proposed scheme, different features cannot be controlled independently and thus the combination of multiple target features leads to unpredictable results. Mintz [15] uses linear constrained optimization on audio descriptors to control the parameters of an additive-plus-noise synthesizer. Williams and Brookes [16] morph using SMS according to verbal attributes that correlate with audio descriptors and in [17] employ a similar technique to morph between prerecorded sounds and sounds captured in real time. Hikichi and Osaka [18] adjust the parameters of a physical model using the spectral centroid as a reference to morph between piano and guitar sounds, and Primavera et al. [19] focus on the importance of decay time when morphing between percussive sounds of the same family. Coleman and Bonada [20] derive analytic relations for the spectral centroid and standard deviation to control adaptive effects for resampling and band-pass equalization. Caetano and Rodet in [21] investigate spectral envelope representations, which lead to linearly varying values of audio descriptors when linearly interpolated according to a morphing factor, and in [22] use optimization techniques based on genetic algorithms to obtain morphed spectral envelopes that approximate target audio descriptor values.

Other approaches rely strictly on the time domain [23] or on time-frequency representations [24, 25]. Röbel [23] models the signals as dynamical systems using neural networks and morphs by interpolating their corresponding attractors. According to the author, the attractors of the two sounds should be topologically equivalent for achieving a convincing morphing. Ahmad et al. [24] propose a scheme for morphing between transient and non-stationary signals using the discrete wavelet transform (DWT) along with singular value decomposition (SVD) for interpolating the wavelet coefficients. Olivero et al. [25] propose a sound morphing technique without making any presumptions about the nature of the signal or its underlying model. The technique relies on the interpolation of Gabor masks and its penalty-based version is shown to encompass typical cross-synthesis strategies used in computer music applications. Furthermore, the interpretation of one of the strategies in terms of Bregman divergences allows them to include constraints that force morphing intermediates to exhibit a predesigned temporal sequence of centroids. This approach works well only as long as there is overlapping energy between the sounds and in our opinion, certain presumptions about the nature of the signal are necessary for choosing an appropriate morphing strategy.

Table B.1 shows a brief comparison between the above-presented methods that are applicable to static morphing and the current approach. In Section B.2 we present an overview of our proposed approach. Section B.3 describes in detail the morphing process based on parameter interpolation, and Section B.4 presents the optimization scheme used for morphing based on higher-level audio features. In Section B.5 we present our concluding remarks and future improvements of our method.

| **Author(s) and papers** | **Sound model & morphing strategy** | **Partial matching** | **High-level audio features** |
|---|---|---|---|
| Osaka [5] | Sinusoidal modeling. The residual is modeled with short partials according to hypothesized distributions. | Yes | No |
| Tellman et al. [4] | Sinusoidal modeling. No treatment of the residual. | Yes | No |
| Haken et al. [7] | Noise-enhanced sinusoidal modeling. | No | Amplitude and fundamental frequency |
| Boccardi and Drioli [8] | GMM applied to SMS. No treatment of the residual. | No | No |
| Caetano and Rodet [22] | Spectral envelopes for the deterministic and stochastic parts. | No | Spectral audio descriptors |
| Röbel [23] | Dynamical systems. | Not applicable | No |
| Ahmad et al. [24] | DWT with SVD. | Not applicable | No |
| Olivero et al. [25] | Gabor transform with constrained Gabor masks. | Not applicable | Arithmetic, harmonic and geometric centroids |
| Kazazis et al. [present document] | Sinusoidal modeling and spectral envelopes. | Yes | Spectral and harmonic audio descriptors |

**Table B.1** *A brief comparison of methods for static morphing.*

## B.2 A hybrid approach to sound morphing

The morphing scheme presented here requires a source sound, to which we apply timbral transformations according to a morphing factor "$\alpha$" ($0 \leq \alpha \leq 1$), and a target. A value of $\alpha = 0$ corresponds to the source sound and a value of $\alpha = 1$ corresponds to the target sound. The target could consist only of specific audio descriptor values that are obtained according to a morphing factor $\alpha_{\mathrm{d}}$ and applied to the source sound, or it could be a different sound from which we extract the audio descriptors that we want to morph accordingly, but we also interpolate between the spectrotemporal fine structures of the two according to a morphing factor $\alpha_{\mathrm{p}}$. Depending on their spectral content, the source and target sounds can be decomposed into three parts as in [5]: a deterministic part, which is related to harmonic and inharmonic qualities; a quasi-deterministic part, which is more related to transients and spectrotemporal irregularities; and a stochastic part, which is related to noise color. The deterministic and quasi-deterministic parts are estimated through sinusoidal modeling from which we obtain the time-varying frequencies, amplitudes and phases of the partials. The stochastic parts are derived by subtracting the deterministic and quasi-deterministic parts from the original signals [10] and are modeled by estimating their spectral envelopes.

In the next step, we compute the time-varying audio descriptors on each of the three parts and for each analysis frame. Audio descriptors that are applicable to the current approach are presented in detail in [26]. For the purposes of this study we have experimented with: spectral centroid and higher order statistical moments of the spectrum including the standard deviation (referred to as spectral spread), spectral skewness, and spectral kurtosis; spectral decrease; and spectral deviation, which is only computed on the deterministic part of the signal. Descriptors that are applicable exclusively to harmonic (or slightly inharmonic) signals, such as tristimulus values and the odd-to-even harmonic ratio, are also applicable. Natural sounds, however, rarely exhibit such well-defined properties, and thus such descriptors would be more suitable in the case of synthetic or simplified natural sounds. Once we calculate the descriptors of the source and target sounds we can compute intermediate values according to the morphing factor $\alpha_d$, and we interpolate the model parameters of the deterministic, quasi-deterministic and stochastic parts separately. The intermediate values of audio descriptors are applied to the parameter-interpolated signals using the optimization scheme described in Section B.4.

We chose to model differently the stochastic part, on the one hand, and the deterministic and quasi-deterministic parts, on the other hand, because not all sounds exhibit a strong formant structure. As such, spectral envelopes would be a poor estimation of the signal, unless they are estimated by the tracked partials, as in [10, 27, 28]. On the other hand, it is well known that if the signal is stochastic-only, sinusoidal modeling usually leads to artifacts and so a morphing scheme based exclusively on this model would degrade the sound quality. The separation into deterministic and quasi-deterministic parts is necessary for improving the estimation of partial-to-partial correspondences, as we discuss in Section B.3.1.1. In the following we assume that the source and target sounds are equalized in

loudness, have the same fundamental frequencies, and can be of different durations.

## B.3 Parameter interpolation

In this section we describe the interpolation schemes based on the parameters of the sinusoidal model and the parameters that model the spectral envelopes of the residuals.

### B.3.1 Deterministic and quasi-deterministic parts

The following scheme is used for both the harmonic and quasi-harmonic parts. Before interpolating the parameters of the sinusoidal model, it is necessary to find partial-to-partial correspondences between the source and target sounds.

#### Estimating partial-to-partial correspondences

The deterministic part consists of partials that are long in duration, with respect to the total duration of the analyzed sound, whereas the quasi-deterministic part consists of shorter partials that are generally unstable in frequency (short chirps), have lower amplitude values, and surround the harmonic or inharmonic partials of the deterministic part. Such partials may also occur as artifacts of the sinusoidal analysis algorithm, especially in cases where the sinusoids are of low amplitude and the tracking algorithm fails to perform a reliable peak-to-peak matching.

A one-to-one correspondence between the partials of the sour-ce and target sounds is very unlikely to occur unless we limit the number of tracked partials to the most prominent ones with respect to their durations and amplitude thresholds. However, there are cases in which even if there is a limit to the number of tracked partials, the assumption of a one-to-one correspondence as described in [21] could be problematic. For example, when morphing from a sound that has odd and even harmonics to a sound that has only odd ones, we would ideally interpolate only the frequency and amplitude values of the odd harmonics of the two sounds to avoid the artifacts that would result from interpolating the odd with both the odd and even harmonics of the two sounds.

For finding correspondences between the partials of the source and target sounds, we use a k-nearest neighbors classifier (k-NN) based on Euclidean frequency proximity, and under the condition that the vector that is to be classified must have the same or a smaller number of partials. Obviously, the k-NN classifier does not return a one-to-one, but rather a many-to-one, mapping, so we choose the closest neighbor in frequency, and we treat the rest of the neighbors as unmatched partials. The unmatched partials retain their original frequencies but are initialized with zero amplitude levels, which gradually increase according to the morphing factor. After experimenting with different sounds, we concluded that such treatment does not lead to perceptual stream segregation, but rather to a seamless partial

fade-in effect that facilitates the morphing between inharmonic sounds or between sounds that consist of unequal numbers of partials (see Fig. B.1).

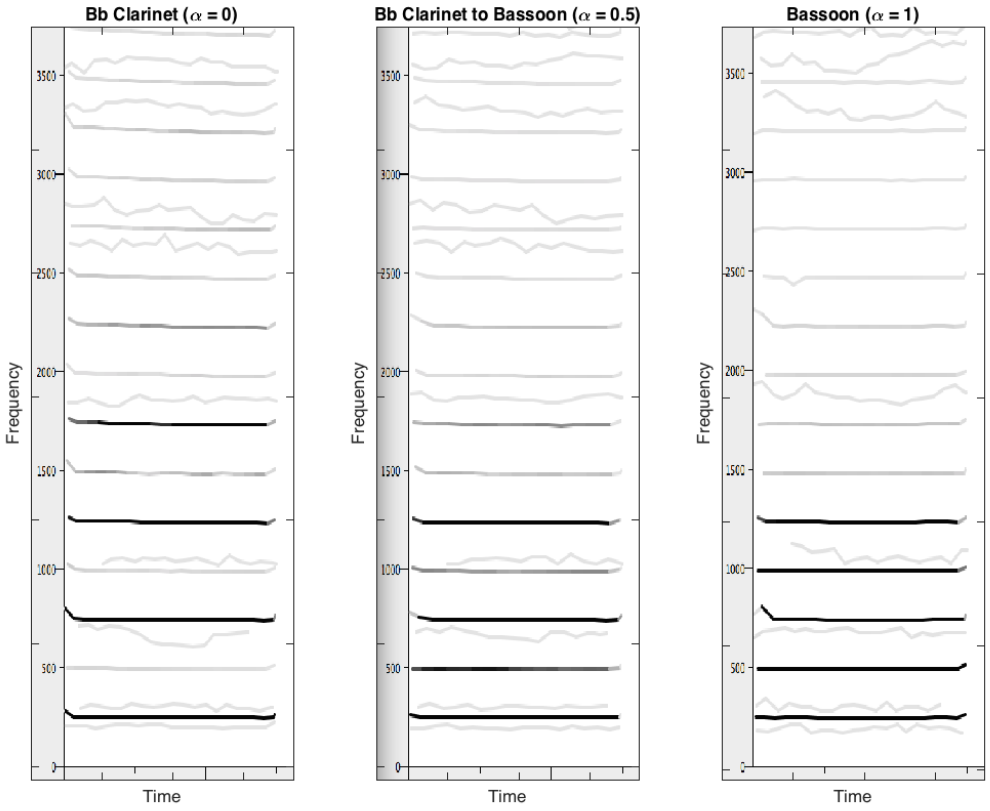


**Fig. B.1** *Partial-to-partial correspondences and parameter interpolation of the deterministic part. Morphing from a clarinet sound to a bassoon with* $\alpha_p = 0.5$. *Gray-level values correspond to the partials' amplitude values.*

### Interpolation of partials' breakpoint values

We represent each partial according to its start and end times, and with time breakpoints that are set according to its frequency and amplitude variations. If the source and target sounds have a different number of breakpoints, we simply interpolate the breakpoint values of the shorter one in order to match the number of breakpoints of the longer one. This representation enables us to interpolate the parameters at the level of events, which offers greater control over the morphing process as opposed to parameter interpolation between time frames. If the partials of the source and target sounds differ in duration, we are able

to achieve intermediate durations by interpolating the breakpoint values of each partial according to the morphing factor. Interpolating between the start and end times of the partials also allows us to morph their onset asynchrony. We use the following expressions for calculating the interpolated values of partials' frequencies and amplitudes, respectively:

$$f(\alpha_\mathrm{p}) = \alpha_\mathrm{p} f_\mathrm{s} + (1 - \alpha_\mathrm{p}) f_\mathrm{t} \tag{B.1}$$

$$\log_{10}(g(\alpha_\mathrm{p})) = \alpha_\mathrm{p} \log_{10}(g_\mathrm{s}) + (1 - \alpha_\mathrm{p}) \log_{10}(g_\mathrm{t}) \tag{B.2}$$

where the subscripts "s", "t" denote the source and target, respectively, and $\alpha_\mathrm{p}$ is the morphing factor related to parameter interpolation. Though Fig. B.1 does not show a typical harmonic spectrum of the analyzed sounds because of the very low amplitude detection threshold ($-90\,\mathrm{dB}$) that was used in the partial-tracking algorithm, and which subsequently gave rise to auxiliary harmonic components, it clearly illustrates the estimation of partial-to-partial correspondences and the interpolation of the partials' breakpoint values.

### B.3.2 Stochastic part

For morphing the stochastic part, we first estimate for every analysis frame its spectral envelope using Linear Predictive Coding (LPC), because we assume that the modeled signal is random, which fits exactly the basic assumption of LPC. We then get a temporal sequence of spectral envelopes (one for each frame), which allows us to render a time-varying Power Spectral Density (PSD) of the stochastic part. In order to morph, we interpolate for each time between the spectral envelope of the source and the target at this corresponding time. For a high-quality interpolation of the spectral envelopes, it is necessary to convert the LPC transverse coefficients to an alternative representation, because they do not interpolate well and might lead to unstable filters. Line Spectral Frequencies (LSF), Reflection Coefficients (RC) and Log Area Ratio (LAR) have been shown to interpolate smoothly, lead to stable intermediate filters, and lead to linear variations of audio descriptors when linearly interpolated [21, 29]. We choose to interpolate the LAR coefficients (Eq. 3) as they both guarantee the filter's stability and have a physical interpretation, which could be specifically useful when trying to morph between sounds that were created by physical modeling synthesis as in [5, 2]. The filters' coefficients are interpolated according to Eq. (2).

$$\mathrm{lar}(r_\alpha) = \alpha_\mathrm{p} \mathrm{lar}(r_\mathrm{s}) + (1 - \alpha_\mathrm{p}) \mathrm{lar}(r_\mathrm{t}) \tag{B.3}$$

where lar is a vector the coefficients of which read:

$$\mathrm{lar}(r)[i] = ln\left(\frac{1 - r(i)}{1 + r(i)}\right), \quad 1 \leq i \leq n \tag{B.4}$$

and $n$ is the number of reflection coefficients $r$. The morphed residual is synthesized by filtered white noise after the inversion of the LAR coefficients to LPC coefficients.

### B.3.3 Temporal Energy Envelope

In the present approach, the temporal energy envelope is a consequence of morphing. The parts of the signal that were morphed independently are added together to form the parameter-interpolated signal and thus, the energy envelope is constructed from the time-varying amplitudes of the partials and the gains of the filter.

## B.4 Feature interpolation

The desired values of descriptors along with the interpolated spectrum form an underdetermined system because in theory there are an infinite number of sounds that have the same audio descriptor values. As previously described in Section B.2, the target may consist only of target descriptor values $D_a$, in which case the morphing is based exclusively on high-level features. Fig. B.2 shows an example of two sounds exchanging time-varying spectral centroids, where $\alpha_{\mathrm{p}} = 0$, since the source is the Timpani without any parameter-based morphing, and $\alpha_{\mathrm{d}} = 1$, because we apply to the source spectrum the spectral centroid values of the Tuba, which is the target.

For each time frame, we match the audio descriptor values obtained according to a specific $\alpha_{\mathrm{d}}$ to the interpolated spectrum by optimizing the amplitudes of the sinusoids or FFT bins of the interpolated spectrum $x_j$ under the constraints of the target values of descriptors $D_a$. More formally this can be expressed as:

$$\min_x \sum_{j=1}^{N} |x_j - g_j| \quad \text{subject to} \quad D(x) = D_a \tag{B.5}$$

where $g_j$ are the parameter-interpolated amplitude values according to $\alpha_{\mathrm{p}}$, $N$ is the total number of partials or FFT bins, and $D_a$ is the target value of $D(x)$, which can be one of the following descriptors (Eq. (6) – (11)).

$$m_1 = \sum_{j=1}^{N} f_j \cdot p_j \tag{B.6}$$

$$m_2 = (\sum_{j=1}^{N} (f_j - m_1)^2 \cdot p_j)^{1/2} \tag{B.7}$$

$$m_3 = (\sum_{j=1}^{N} (f_j - m_1)^3 \cdot p_j)/m_2^3 \tag{B.8}$$

$$m_4 = (\sum_{j=1}^{N} (f_j - m_1)^4 \cdot p_j)/m_2^4 \tag{B.9}$$

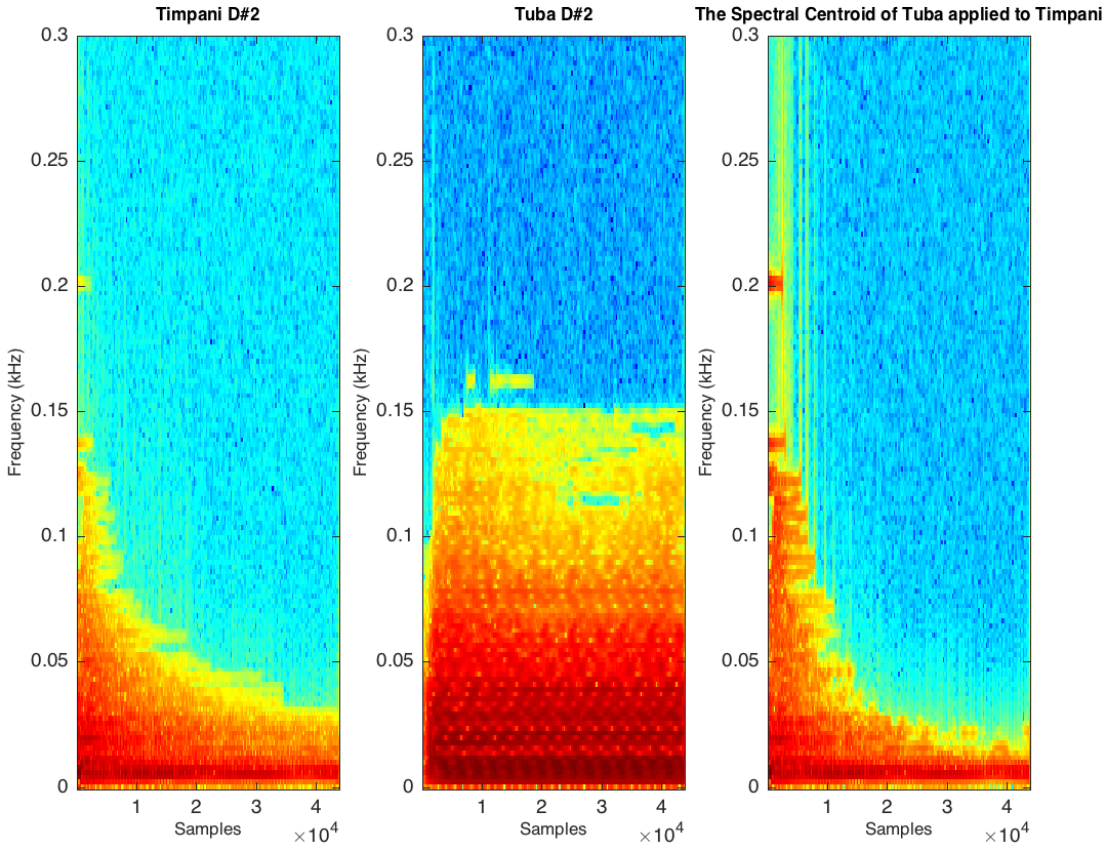


**Fig. B.2** *The spectral centroid time series of a Tuba sound applied to a Timpani (the actual values of the time series are shown in Fig. B.3.)*

$$decr = \frac{1}{\sum_{j=2}^{N} x_j} \sum_{j=2}^{N} \frac{x_j - x_1}{j-1} \tag{B.10}$$

$$dev = \frac{1}{N} \sum_{j=1}^{N} (x_j - \mathrm{SE}(f_j)) \tag{B.11}$$

where $p_j$ are the normalized values of $x_j$ [27]:

$$p_j = \frac{x_j}{\sum_{j=1}^{N} x_j} \tag{B.12}$$

*dev* denotes the harmonic spectral deviation and $\mathrm{SE}(f_j)$ is the value of the spectral envelope at frequency $f_j$, which is estimated by averaging the values of three adjacent partials; *decr*

denotes the spectral decrease; $m_1, m_2, m_3$ and $m_4$ denote the spectral centroid, spectral spread, spectral skewness and spectral kurtosis respectively. The optimization is run in Matlab using the "fmincon" function along with the "sqp" method, which are suitable for solving constrained and non-linear problems [30]. Since the audio descriptors have different ranges, it is necessary to normalize them for assessing the convergence of the algorithm. Using this optimization scheme, we are able to set different morphing factors for each descriptor independently, as long as a feasible solution among these values exists. Furthermore, the choice of the objective function (Eq. 5) forces the optimized spectrum to be as close as possible to the interpolated one by keeping its frequency content unchanged and by altering its amplitude values as little as possible. Fig. B.3 shows an example of morphing the parameter-interpolated signal according to varying morphing values of spectral centroid and spectral spread while preserving a constant value for the rest.

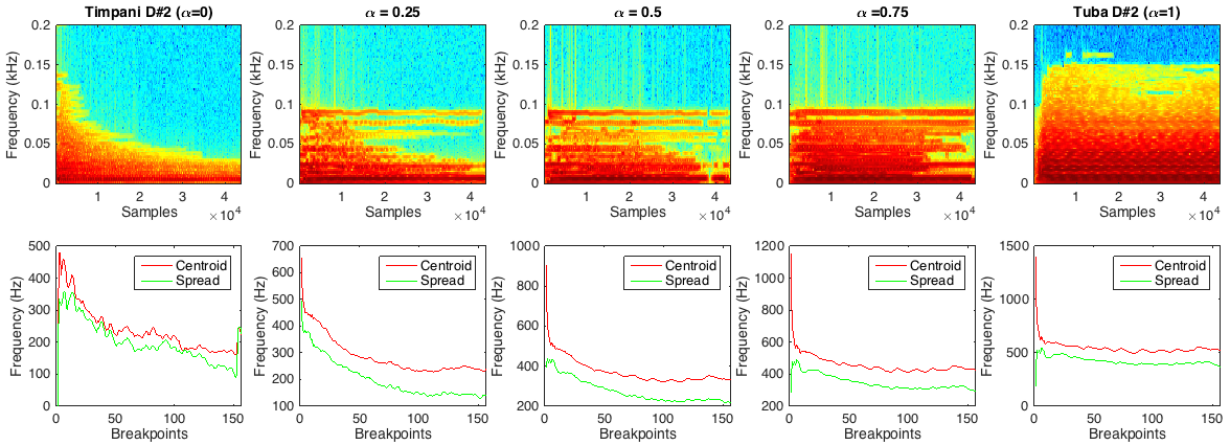


**Fig. B.3** *Morphing the parameter interpolated signal by audio descriptors. Spectral centroid and spectral spread vary according to the morphing factor $\alpha$. The rest of descriptors preserve constant target values according to their median when interpolated with $\alpha_d = 0.5$.*

Using a sinusoidal model for the deterministic and quasi-deterministic parts, the optimized values correspond directly to the parameters of additive synthesis, and the residual reaches its target values by altering the energy of the FFT bins. As in Section B.3.1.2, if the source and target sounds are of different durations, we simply interpolate the descriptor values of the shorter one in order to match them to the analysis frames of the longer one.

## B.5 Conclusions and future work

We presented a hybrid approach to sound morphing based on sinusoid-plus-noise modeling and higher-level audio features. Dividing the signal into deterministic, quasi-deterministic,

and stochastic parts and processing them separately allows for finer control of the synthesis parameters and also enables us to morph between deterministic and quasi-deterministic signals of different durations. The morphed sound is synthesized using additive synthesis for the deterministic and quasi-deterministic parts and filtered white noise for the stochastic part. The spectrum of the morphed signal is further refined according to target audio descriptor values through an optimization process. We have shown that this process allows us to control accurately and independently several audio features, provided that a feasible solution among them exists. Audio examples are available at: https://www.mcgill.ca/mpcl/resources-0/ supplementary-materials. The proposed scheme is more suitable for sustained and percussive sounds, which can either be harmonic or inharmonic, rather than textural sounds. Their residuals however, should be stationary (or pseudo-stationary) as opposed to sound texture, the residual of which is usually non-stationary and may consist of sharp transients. A refinement of our approach would be to find sophisticated ways to interpolate between different tremolo and vibrato rates while preserving the overall spectrotemporal complexity of the partials. Finally, we by no means claim that the use of high-level audio features enables a perceptually based sound morphing. Rather, it offers a more intuitive control over the morphing process, as in the case of adaptive effects [31]. Up to now only spectral centroid and log-attack time have been shown to be significantly correlated with perceptual dimensions, cf. [1, 32]. If and how such audio features collapse to single perceptual dimensions remains to be empirically determined.

## Acknowledgments

We would like to thank Philippe Esling for his advice on improving the computational efficiency of our optimization procedure, and the four anonymous reviewers for their helpful comments about improving the clarity of this document. This work was funded by Canadian Natural Sciences and Engineering Research Council grants awarded to Stephen McAdams and Philippe Depalle and a Canada Research Chair awarded to Stephen McAdams.

# References


[1] A. Caclin, S. McAdams, B.K. Smith, and S. Winsberg, "Acoustic Correlates of Timbre Space Dimensions: A Confirmatory Study Using Synthetic Tones," *J. Acoust. Soc. Am.,* 118 (1), pp. 471-482, 2005.

[2] S. Lakatos, P. Cook, and G. Scavone, "Selective attention to the parameters of a physically informed sonic model," *Acoustics Research Letters Online, J. Acoust. Soc. Am.,* March 2000.

[3] M. Caetano, *Morphing isolated quasi harmonic acoustic musical instrument sounds guided by perceptually motivated features,* PhD Thesis , IRCAM, Université Pierre et Marie Curie, Paris, France. June 2011.

[4] E. Tellman, L. Haken, and B. Holloway, "Timbre Morphing of Sounds with Unequal Numbers of Features," *J. Audio Eng. Soc.,* vol. 43, no. 9, pp 678-689, September, 1995.

[5] N. Osaka, "Timbre Interpolation of Sounds Using a Sinusoidal Model," in *Proc. ICMC,* Banff Centre for the Arts, Canada, 1995.

[6] K. Fitz, L. Haken, S. Lefvert, C. Champion, and M. O'Donnell, "Cell-Utes and Flutter-Tongued Cats: Sound Morphing Using Loris and the Reassigned Bandwidth-Enhanced Model," *Computer Music Journal,* 27 (3), 2003.

[7] L. Haken, K. Fitz, and P. Christensen, *Sound of Music: Analysis, Synthesis, and Perception,* "Beyond Traditional Sampling Synthesis: Real-Time Timbre Morphing Using Additive Synthesis," J. W. Beauchamp Ed. Springer-Verlag, Berlin, 2006.

[8] F. Boccardi, and C. Drioli, " Sound Morphing with Gaussian Mixture Models," in *Proc. DAFx,* Limerick, Ireland, 2001.

[9] K. Fitz, L. Haken, and P. Christensen, "A New Algorithm for Bandwidth Association in Bandwidth Enhanced Additive Sound Modeling," in *Proc. ICMC*, Havana, Cuba, 2001.

[10] X. Serra and J. I. Smith, "Spectral Modelling Synthesis," *Computer Music Journal,* 14 (4),1990.

[11] M. Slaney, M. Covell, and B. Lassiter, "Automatic Audio Morphing," in *Proc. ICASSP,* Atlanta, Georgia, 1996.

[12] T. Ezzat, E. Meyers, J. Glass, and T. Poggio, "Morphing Spectral Envelopes using Audio Flow," in *Proc. ICASSP,* Philadelphia, Pennsylvania, 2005.

[13] M. Hoffman,and P. Cook, "Feature-based Synthesis: Mapping from Acoustic and Perceptual Features to Synthesis Parameters" in *Proc. ICMC,* New Orleans, USA, 2006.

[14] T. Park, J. Biguenet, Z. Li, R. Conner, and S. Travis, "Feature Modulation Synthesis," in *Proc. ICMC,* Copenhagen, Denmark, 2007.

[15] D. Mintz, *Toward Timbral Synthesis: a New Method for Synthesizing Sound Based on Timbre Description Schemes,* Master's thesis, Univ. Cal, 2007.

[16] D. Williams, and T. Brookes, "Perceptually-Motivated Audio Morphing: Warmth," *AES 128th Convention,* London, UK, 2010.

[17] D.Williams, P. Randall-Page, E. R. Miranda, "Timbre morphing: near real-time hybrid synthesis in a musical installation," in *Proc. NIME,* London, UK, 2014.

[18] T. Hikichi and N. Osaka, "Sound Timbre Interpolation Based on Physical Modeling" *Acoustical Science and Technology,* 22 (2), 2001.

[19] A. Primavera, F. Piazza and J. D. Reiss, "Audio Morphing for Percussive Hybrid Sound Generation," *AES 45th Conference on Applications of Time Frequency Processing in Audio*, Helsinki, March 2012.

[20] G. Coleman, and J. Bonada, "Sound Transformation by Descriptor Using an Analytic Domain," in *Proc. DAFx,* Espoo, Finland, 2008.

[21] M. Caetano and X. Rodet, "Automatic Timbral Morphing of Musical Instrument Sounds by High-Level Descriptors," in *Proc. ICMC,* New York, USA, 2010.

[22] M. Caetano and X. Rodet, "Independent Manipulation of High-Level Spectral Envelope Shape Features for Sound Morphing by Means of Evolutionary Computation," in *Proc. DAFx,* Graz, Austria, 2010.

[23] A. Röbel, "Morphing Dynamical Sound Models," in *Proc. IEEE Workshop Neural Net Sig. Proc,* 1998.

[24] M. Ahmad, H. Hacihabiboglu, and A. M. Kondoz, "Morphing of Transient Sounds Based on Shift-Invariant Discrete Wavelet Transform and Singular Value Decomposition," in *Proc. ICASSP,* Taipei, Taiwan, 2009.

[25] A. Olivero, P. Depalle, B. Torresáni and R. Kronland-Martinet, "Sound Morphing Strategies Based on Alterations of Time-Frequency Representations by Gabor Multipliers", *AES 45th Conference, Applications of Time-Frequency Processing in Audio,* Helsinki, Finland, March 2012.

[26] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals," *J. Acoust. Soc. Am.,* 130, pp. 2902-2916, 2011.

[27] A. Röbel, and X.Rodet, "Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation," in *Proc. DAFx,* Madrid, Spain, 2005.

[28] T. Galas, and X. Rodet, "An Improved Cepstral Method for Deconvolution of Source-Filter Systems with Discrete Spectra: Application to Musical Sounds," in *Proc. ICMC,* Glasgow, Scotland, 1990.

[29] T. Islam, *Interpolation of Linear Prediction Coefficients for Speech Coding,* Master's thesis, McGill University, 2010.

[30] http://www.mathworks.com/help/optim/ug/constrained-nonlinear-optimization-algorithms.html#bsgppl4/

[31] V. Verfaille and P. Depalle, "Adaptive Effects based on STFT, using a Source-Filter model," in *Proc. DAFx*, Naples, Italy, 2004.

[32] C. W. Wun, A. Horner, and Bin. Wu, "Effect of Spectral Centroid Manipulation on Discrimination and Identification of Instrument Timbres," *J. Audio Eng. Soc.,* 62(9), 575-583, 2014.

# Master References


Almeida, A., Schubert, E., Smith, J., & Wolfe, J. (2017). Brightness scaling of periodic tones. *Attention, Perception & Psychophysics*, *79:1892*.

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review*, *77*, 154–170.

Baird, J. C., & Noma, E. (1978). *Fundamentals of scaling and psychophysics*. John Wiley & Sons, Inc.

Bechtold, T. A., & Senn, O. (2018). Articulation and dynamics influence the perceptual attack time of saxophone sounds. *Frontiers in Psychology*, *9*.

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

Caetano, M., & Rodet, X. (2010a). Automatic timbral morphing of musical instrument sounds by high-level descriptors. In *Proceedings of the 2010 International Computer Music Conference (ICMC), New York, USA*.

Caetano, M., & Rodet, X. (2010b). Independent manipulation of high-level spectral envelope shape features for sound morphing by means of evolutionary computation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx),Graz, Austria.*

Carpentier, G., & Bresson, J. (2010). Interacting with symbol, sound, and feature spaces in orchidée, a computer-aided orchestration environment. *Computer Music Journal*, *34*, 10–27.

Collins, N. (2006). Investigating computational models of perceptual attack time. In *Proceedings of the 9th International Conference on Music Perception and Cognition (ICMPC), Bologna, Italy.*

Elliott, T. M., Hamilton, L., & Theunissen, F. E. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *Journal of the Acoustical Society of America*, *133*, 389–404.

Farbood, M. M., & Price, K. C. (2017). The contribution of timbre attributes to musical tension. *Journal of the Acoustical Society of America*, *141*, 419–427.

Gescheider, G. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, J. W. (1987). The perceptual attack time of musical tones. *Journal of the Acoustical Society of America*, *82*, 88–105.

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

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, *63*, 11493–1500.

Hellman, R. P., & Zwislocki, J. (1961). Some factors affecting the estimation of loudness. *Journal of the Acoustical Society of America*, *33*, 4687–694.

Hemery, E., & Aucouturier, J.-J. (2015). One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis. *Frontiers in Computational Neuroscience*, *9*.

Hoffman, M., & Cook, P. (2006). Feature-based synthesis: Mapping from acoustic and perceptual features to synthesis parameters. In *Proceedings of the 2006 International Computer Music Conference (ICMC), New Orleans, USA.*

ISO/IEC. (2002). *MPEG-7: Information Technology – Multimedia Content Description Interface - Part 4: Audio.* (ISO/IEC FDIS 15938–4:2002)

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, *94*, 2595–2603.

Kendall, R. A., & Carterette, E. C. (1993). Identification and blend of timbres as a basis for orchestration. *Contemporary Music Review*, *9*, 51–67.

Kendall, R. A., Carterette, E. C., & Hajda, J. M. (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception*, *16*, 327–363.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caracterisation du timbre des sons complexes. 2: Analyses acoustiques et quantification psychophysique. [Characterization of the timbre of complex sounds. 2: Acoustic analysis and psychophysical quantification]. *Journal de Physique*, *4*, 625–628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzén & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (Vol. 1989, pp. 43–53). Amsterdam: Excerpta Medica.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, *62*, 426–1439.

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx),Bordeaux, France.*

Lembke, S.-A. (2014). *When timbre blends musically: perception and acoustics underlying orchestration and performance* (Unpublished doctoral dissertation). McGill University.

Lembke, S.-A., & McAdams, S. (2015). The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds. *Acta Acustica united with Acustica*, *101*, 1039–1051.

London, J., Nymoen, K., Langerød, M. T., Thompson, M. R., Code, D. L., & Danielsen, A. (2019). A comparison of methods for investigating the perceptual center of musical sounds. *Attention, Perception & Psychophysics*, *81*.

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology: Perception and motivation; Learning and cognition* (pp. 3–74). Oxford, England: John Wiley & Sons Inc.

McAdams, S. (2015). *Perception et cognition de la musique [Perception and cognition of music]*. Paris, France: J. Vrin.

McAdams, S., Beauchamp, J. W., & Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, *105*, 882–897.

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

McAdams, S., & Giordano, B. L. (2006). Generalizing timbre space data across stimulus contexts: The meta-analytic approach. *Journal of the Acoustical Society of America*, *119*, 3395.

McAdams, S., Roussarie, V., Chaigne, A., & Giordano, B. L. (2010). The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *Journal of the Acoustical Society of America*, *128*, 1401–1413.

McAdams, S., & Siedenburg, K. (2019). Perception and cognition of musical timbre. In P. J. Rentfrow & D. J. Levitin (Eds.), *Foundations in music psychology: Theory and research* (pp. 71–120). Cambridge, MA: MIT Press.

McAdams, S., Tse, A., & Wang, G. (2016). Generalizing the learning of instrument identities across pitch registers. In *Proceedings of the 14th International Conference on Music Perception and Cognition (ICMPC), San Francisco, USA.*

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

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychological Science*, *19*, 1263–1271.

McDermott, J. H., Schlemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*, 493–498.

Miller, J. R., & Carterette, E. C. (1975). Perceptual space for musical structures. *Journal of the Acoustical Society of America*, *58*, 711–720.

Noble, J., & McAdams, S. (2020). Sound mass, auditory perception, and 'post-tone' music. *Journal of New Music Research*. doi: 10.1080/09298215.2020.1749673

Nymoen, K., Danielsen, A., & London, J. (2017). Validating attack phase descriptors obtained by the Timbre Toolbox and MIRtoolbox. In *Proceedings of the 14th Sound*

*and Music Computing Conference (SMPC), Espoo, Finland.*

Olivero, A., Depalle, P., Torrésani, B., & Kronland-Martinet, R. (2012). Sound morphing strategies based on alterations of time-frequency representations by Gabor multipliers. In *Proceedings of the Audio Engineering Society Conference (AES): 45th International Conference: Applications of Time-Frequency Processing in Audio, Helsinki, Finland.*

Park, T., Biguenet, J., Li, Z., Conner, R., & Travis, S. (2007). Feature modulation synthesis. In *Proceedings of the 2007 International Computer Music Conference (ICMC), Copenhagen, Denmark.*

Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. (2011). Music in our ears: The biological bases of musical timbre perception. *PLOS Computational Biology*, *8*.

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

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing* (Vol. 1970, pp. 397–410). Leiden, Netherlands: Sijthoff.

Sandell, G. J. (1995). Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration. *Music Perception*, *13*, 209–246.

Schwarz, D., & O'Leary, S. (2015). Smooth granular sound texture synthesis by control of timbral similarity. In *Proceedings of the 12th Sound and Music Computing Conference (SMPC), Maynooth, Ireland.*

Siedenburg, K. (2016). *Perspectives on memory for musical timbre* (Unpublished doctoral dissertation). McGill University.

Siedenburg, K., Fujinaga, I., & McAdams, S. (2016). A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *Journal of New Music Research*, *45*, 27–41.

Siedenburg, K., & Müllensiefen, D. (2017). Modeling timbre similarity of short music clips. *Frontiers in Psychology*, *8:639*.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, New Series*, *103*, 677–680.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects.* Reprint, New York, NY: Routledge, 2017.

Tardieu, D., & McAdams, S. (2012). Perception of dyads of impulsive and sustained instrument sounds. *Music Perception*, *30*, 117–128.

Thoret, E., Goodchild, M., & McAdams, S. (2018). Timbre 2018: Timbre is a many-splendored thing. Montreal, QC: McGill University. Retrieved from https://www.mcgill.ca/timbre2018/files/timbre2018/timbre2018_proceedings.pdf (Last accessed 15 May 2020)

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2015). An interlanguage study of musical

timbre semantic dimensions and their acoustic correlates. *Music Perception*, *31*, 339–358.