SALIENCE IN MUSICAL SCENE ANALYSIS: PSYCHOACOUSTIC EXPERIMENTS AND MODELS

Von der Fakultät für Medizin und Gesundheitswissenschaften der Carl von Ossietzky Universität Oldenburg zur Erlangung des Grades und Titels eines

Doktor der Naturwissenschaften Dr. rer. nat.

eingereichte Dissertation

von Herrn Michel Bürgel

geboren am 18.06.1988 in Dormagen (Deutschland)

Gutachter Prof. Dr. Kai Siedenburg

Weitere Gutachter

Prof. Dr.-Ing. Steven van de Par

Dr. Trevor Agus

Tag der Disputation: 10. Oktober 2024

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to everyone who supported me throughout my PhD journey, contributing to our collective understanding of human perception.

First and foremost, I extend my deepest thanks to my supervisor, Kai Siedenburg. His exceptional scientific knowledge and mentoring skills have been invaluable to my research. Beyond his professional expertise, his kindness fostered a spirit of community and friendship, making this journey truly memorable. I am also grateful for his constructive feedback and the numerous opportunities and projects he involved me in, which greatly enriched my PhD experience.

I am also indebted to my colleagues and friends from the Music Perception and Processing Lab. Our scientific and informal exchanges were incredibly enlightening, and our conference trips and social gatherings were filled with joy.

I am thankful to Robin Hake, Simon Jacobsen & Iris Mencke for providing valuable feedback during the thesis review process, and to Steven van de Par and Trevor Agus for serving as members of the thesis committee.

Special thanks to the Freigeist Fellowship of the Volkswagen Foundation and the SFB project of the Deutsche Forschungsgemeinschaft for funding my research projects.

Thanks to all the researchers who provided thought-provoking feedback and critically reviewed my work, significantly enhancing the quality of my research.

Additionally, I am thankful to my longtime friends outside the academic sphere for their welcome distractions and motivation during times of need.

Above all, I want to thank my wonderful partner, Szascha, for her unwavering support, encouragement, and for always making everything better.

Oldenburg, 22.07.2024
Michel Bürgel

ABSTRACT

The human auditory system excels at disentangling complex auditory scenes with overlapping audio signals. A prime example are musical mixtures, where various instruments and singing voices create a rich acoustic scene. Within these mixtures, some sounds are perceived as more prominent, with the singing voice being especially salient. Understanding what makes certain sounds attract auditory attention is crucial for comprehending the human auditory system in complex scenes. This dissertation investigates the origins of salience in musical mixtures from a psychoacoustical perspective through four studies comprising ten experiments.

The first study explores the detection of individual target instruments and vocals in 2-second multi-track mixtures of popular music. Results reveal that prior information about the target sound enhances detection accuracy, especially for bass instruments, which are otherwise obscured by other elements in the mixture. In contrast, vocals show no effect of prior information and are detected with the highest accuracies, implying a unique ‘vocal salience.’ Aligning target instruments with vocals in sound level and spectral filtering enhances their overall prominence but does not counteract vocal salience.

The second study continues to analyze the origins of vocal salience using the same detection paradigm. The focus is set on investigating the influence of the main melody, phonetic cues, and frequency micro-modulations (FMM) in singing voices. Findings indicate that having instruments play a song’s main melody enhances their prominence in the mixture, but neither main melody nor phonetic cues are sufficient to enable vocal salience. However, FMM emerges as a significant factor, correlating with enhanced detection and salience of vocals. FMM adds acoustic cues about pitch continuity, irregularities, and emotional prosody, facilitating prioritized processing.

The third study examines the recognition of stationary short vocal and instrumental tones presented either in isolation or with a spatially separated piano interferer. Contrary to expectations, the results show no general effect of enhanced vocal sound recognition and no effect of FMM on recognition. The lack of FMM effect likely stems from the short stimuli duration and the use of stationary tones. However, most vocal sounds demonstrate robustness to interference, suggesting a degree of intrinsic salience, while vowel /u/ sounds lack this robustness. This implies that being a vocal sound or a singing voice does not inherently produce salience. An acoustic analysis indicates that multiple acoustic features such as spectral similarities interact to enable vocal salience.

The final study investigates whether the auditory system has a perceptual bias towards specific frequency regions that may explain the bass’s lack of attention attraction. Using pseudo-randomized pure tone melodies in spectrally separated frequency bands, the results reveal enhanced salience for the lowest and highest frequency bands, suggesting that the auditory system focuses on the edges of the frequency spectrum within an auditory scene. This implies that the bass's lack of salience results from complex spectro-temporal in naturalistic instrument sounds.

Overall, the studies highlight the significant interplay between bottom-up and top-down processes in complex auditory scenes. Prior information helps search rich acoustic scenes, emphasizing top-down processing in auditory scene analysis. The ability to hear non-cued instruments in musical mixtures implies some degree of holistic perception, with information about both local (the focus of attention) and global stream organizations (the whole mixture). Findings of salience at spectral edge frequencies suggest that inferior detection of bass instruments results from musical structures with distinct lead and accompaniment roles, as well as the interplay of spectral patterns produced by complex tones. Vocal salience suggests that inherent acoustic properties of the human singing voice make it particularly prominent in auditory scenes, with FMM contributing significantly to this salience by adding acoustic cues for emotional prosody processing. Additionally, the lack of some vocal sounds to generate such salience combined with the persistence of vocal salience across various conditions, and the enhancement of instrument prominence when vocal attributes are transferred indicate that multiple acoustic factors, beyond just being vocal sounds, contribute to enable salience in musical scenes.

ZUSAMMENFASSUNG

Das menschliche auditive System zeichnet sich durch die Fähigkeit aus, komplexe auditiven Szenen mit überlappenden Audiosignalen zu entschlüsseln und einzelne Klänge zu fokussieren. Ein Paradebeispiel dafür sind musikalische Mixturen, in denen verschiedene Instrumente und Gesangsstimmen eine reiche Klanglandschaft schaffen. Innerhalb dieser Mischungen werden einige Klänge als prominenter wahrgenommen, wobei die Gesangsstimme besonders hervorsticht. Zu verstehen, warum bestimmte Klänge die auditive Aufmerksamkeit auf sich ziehen, ist entscheidend für das Verständnis des menschlichen auditiven Systems in komplexen Szenen. Diese Dissertation untersucht die Ursprünge der Salienz in musikalischen Mixturen aus psychoakustischer Perspektive durch vier Studien mit insgesamt zehn Experimenten.

Die erste Studie untersucht die Detektion (Hörbarkeit) einzelner Zielinstrumente und Gesangsstimmen in 2-sekündigen Mixturen populärer Musik. Die Ergebnisse zeigen, dass vorherige Informationen über das Zielgeräusch die Detektionsgenauigkeit verbessern. Dies zeigt sich insbesondere bei Bassinstrumenten, die sonst von anderen Elementen der Mixtur überdeckt werden. Im Gegensatz dazu zeigen Gesangsstimmen keinen Effekt durch vorherige Informationen und werden mit den höchsten Genauigkeiten erkannt, was auf eine einzigartige „vokale Salienz“ hinweist. Das Angleichen der Zielinstrumente an die Gesangsstimmen in Bezug auf Schallpegel und spektrale Filterung erhöht deren Gesamtprominenz, hebt jedoch die vokale Salienz nicht auf.

Die zweite Studie setzt die Analyse der Ursprünge der vokalen Salienz mit demselben Detektionsparadigma fort. Der Fokus liegt auf der Untersuchung des Einflusses der Hauptmelodie, phonologischer Attribute und Frequenzmikromodulation (FMM) in Gesangsstimmen. Die Ergebnisse zeigen, dass das Übertragen der Hauptmelodie von der Gesangsstimme auf Instrumente, die Prominenz der Instrumente in der Mixtur erhöht. Weiter zeigen die Ergebnisse, dass weder die Übertragung der Hauptmelodie noch phonologische Attribute ausreichen, um vokale Salienz zu ermöglichen. FMM hingegen erweist sich als bedeutender Faktor, der mit einer verbesserten Erkennung und erhöhter Salienz korreliert. Diese Modulationen fügen akustische Hinweise auf Tonkontinuität und emotionale Prosodie hinzu, was eine priorisierte Verarbeitung begünstigt.

Die dritte Studie untersucht die Erkennung stationärer, kurzer Vokal-, Gesangs- und Instrumentaltöne, die entweder isoliert oder mit einem räumlich getrennten Störsignal (einem Klavierakkord) präsentiert werden. Entgegen den Erwartungen zeigen die Ergebnisse keine allgemein verbesserte Erkennung von Vokalklängen und keinen Effekt von FMM. Der fehlende FMM-Effekt kann wahrscheinlich auf die kurze Stimulusdauer und die Verwendung stationärer Töne zurückgeführt werden. Die meisten Vokal- und Gesangsklänge zeigen jedoch eine Robustheit gegenüber Störungen, was auf ein gewisses Maß an intrinsischer Salienz hindeutet. Hierbei stechen /u/ Klängen heraus, welche diese Robustheit nicht aufweisen. Dies impliziert, dass die Eigenschaft eine menschliche Gesangstimme zu sein allein keine inhärente Salienz erzeugt. Eine akustische Analyse legt nahe, dass mehrere akustische Merkmale wie spektrale Ähnlichkeiten zwischen präsentierten Signalen interagieren, um vokalen Salienz zu ermöglichen.

Die letzte Studie untersucht in der Wiederverwendung des Detektionsparadigmas, ob das auditive System eine Wahrnehmungspräferenz für bestimmte Frequenzbereiche hat, welche die fehlende Aufmerksamkeit für Bassinstrumente erklären könnte. Hierzu werden Mixturen aus pseudo-randomisierte Melodien von Reintönen in spektral getrennten Frequenzbändern präsentiert. Die Ergebnisse zeigen eine erhöhte Salienz für die tiefsten und höchsten Frequenzmelodien, was darauf hindeutet, dass das auditive System die Ränder des Frequenzspektrums innerhalb einer auditiven Szene fokussiert. Dies impliziert, dass die fehlende Salienz der Bassinstrumente auf komplexe spektral-temporale Muster zurückzuführen ist, die durch natürliche Instrumentenklänge erzeugt werden.

Zusammengefasst heben die Studien das bedeutende Zusammenspiel von Bottom-up- und Top-down-Prozessen in komplexen auditiven Szenen hervor. Vorinformationen helfen, reichhaltige akustische Szenen zu durchsuchen, was die Bedeutung der Top-down-Verarbeitung in der auditiven Szenenanalyse unterstreicht. Die Fähigkeit, Instrumente in musikalischen Mixturen zu hören, ohne dass mit Vorinformationen auditive Aufmerksamkeit auf diese gelenkt wurde, impliziert eine gewisse ganzheitliche Wahrnehmung mit Informationen über lokale (der Fokus der Aufmerksamkeit) und globale Organisationen (die gesamte Mixtur). Die Ergebnisse zur Salienz der spektralen Randfrequenzen deuten darauf hin, dass die geringere Erkennung von Bassinstrumenten auf das Zusammenspiel von Tonkomplexen und den daraus resultierenden spektralen Mustern zurückzuführen ist. Die vokale Salienz zeigt, dass die menschliche Gesangsstimme inhärente akustische Eigenschaften besitzen kann, die sie in auditiven Szenen besonders hervorhebt. Die FMM im natürlichen Gesang tragen erheblich zu dieser Salienz bei, indem sie akustische Hinweise zur Verarbeitung emotionaler Prosodie hinzufügen und so deren Trennung in musikalischen Mixturen erleichtern. Zudem zeigt die fehlende Salienz bei einigen Vokalklängen, dass eine menschliche Vokalisation zu sein nicht ausreicht, um diese Salienz automatisch zu erzeugen. Kombiniert mit der gesteigerten Prominenz von Instrumenten bei Übertragung von Eigenschaften der Gesangsstimme impliziert dies, dass mehrere akustische Faktoren zur vokalen Salienz beitragen.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Allow yourself a moment and imagine listening to one of your favorite pieces of music – and for the purpose of the mental journey, a piece of popular music with a singing voice. Immerse yourself in this musical scene, focusing on your mental representation. You may conjure a rich arrangement of several instruments playing together, creating a mixture of sounds. Within this mixture it is likely that some sounds appear more prominent in the foreground of the musical scene, while others seem to fade into the background of this perceptual hierarchy. It is likely that the first sound that comes to your mind is the singing voice and its melody, while other instruments play an accompanying role. Yet, amidst this musical mixture, have you ever pondered about the acoustic underpinnings that shape such hierarchies? This research seeks to explore precisely that question. It specifically aims to dissect the role of lead vocals and aims to uncover the acoustic foundations that contribute to their ability to attract the listeners attention in musical mixtures – a phenomenon labeled in this dissertation as 'vocal salience'.

1.1 Auditory scene analysis

To gain a deeper understanding of the auditory mechanisms at play, it is essential to examine the intricate interplay of sounds within a given auditory scene. One renowned example frequently discussed in the literature is the 'cocktail party problem' (Cherry & Taylor, 1954). As the name implies, this scenario illustrates the challenges faced by the auditory system when multiple sound sources coexist in an environment where numerous individuals are conversing simultaneously, while a listener endeavors to focus on a specific speaker. Analogous to the music example in the introduction, the sounds in such scenarios, including speech signals and other auditory inputs, often overlap spectrally and temporally, resulting in a perplexing mixture of auditory stimuli.

Despite this complexity, the human auditory system demonstrates remarkable proficiency from infancy in disentangling such auditory mixtures into distinct mental representations of auditory streams, enabling individuals to selectively attend to specific sounds as they unfold simultaneously (Demany, 1982; McAdams & Bertoncini, 1997; Winkler et al., 2003). A well-established concept that aims to explain the mechanisms enabling such stream formation is auditory scene analysis (ASA; Bregman, 1990). Bregman described this concept as follows:

> "Auditory scene analysis (ASA) is the process by which the auditory system separates the individual sounds in natural-world situations, in which these sounds are usually interleaved and overlapped in time and their components interleaved and overlapped in frequency." (Bregman & McAdams, 1994, p. 2).

At its core, ASA operates in accordance with Gestalt psychology principles, whereby sensory elements are organized into coherent perceptual units. Applied to ASA, these principles entail that the auditory system extracts acoustic characteristics, referred to as cues, from the heard audio to form distinct perceptual auditory streams. Temporal and spectral acoustical cues are partitioned into subsets that may correspond to different sounds sources. Sounds are segregated or integrated based on distinctive or common cues. Matching or closely aligned cues are more likely to be integrated together than separated into distinct streams.

Such cues include synchrony of onsets and onsets (Bregman & Pinker, 1978), fundamental frequency (Bregman et al., 1990), spatial location (Darwin & Hukin, 1999), and amplitude modulation (Bregman et al., 1985). Regarding synchrony, the timing of onset and offsets of sounds foster an integration into a perceptual stream; for example, in a metal band, when two guitarists strike their strings simultaneously in a riff, temporal proximity of onsets and offsets influences whether the sounds blend into the perception of a single unified sound stream or whether they are perceived as two separated streams. The fundamental frequency of a sound contributes to stream formation; in a choir, the difference in fundamental frequency between the soprano, alto, tenor, and bass voices allows listeners to perceive each vocal range as a separate auditory stream, even when they sing simultaneously. Spatial location cues help listeners identify where sounds are coming from and group them accordingly; in stereo recordings of pop music, panning techniques simulate a spatial arrangement of instruments and voices so that they appear to come from various directions, thus facilitating the segregation of auditory streams. Amplitude modulation, or changes in loudness over time, can indicate whether sounds belong together; in classical music, a crescendo performed by the string section can be perceived as a single, evolving auditory stream, distinct from another string section, playing a steady unmodulated sustained note.

Such acoustical cues are both evaluated in simultaneously occurring sounds in different spectral locations (Micheyl & Oxenham, 2010), as well as sequentially by change over time, allowing listeners to continuously update their perceptual representations of the auditory scene as new information becomes available (Moore & Gockel, 2012). Sequential processing enables an additional grouping mechanism, i.e. the principle of common fate, in which cues that change in a coordinated manner tend to be perceived as belonging to the same source; in a stereo recording a sound that gradually moves from one side to the other is more likely to be perceived as the same sound source as a sound that suddenly shifts sides. Time also plays a decisive role, as continuous exposure to a sound increases the potential for stream segregation (Bregman, 1978). However, a single sudden change in a sounds acoustic cue can reset this process, as the change is interpreted as a new sound (Haywood & Roberts, 2010).

This grouping of auditory elements directly influences various aspects of perception, including the perceived number of sounds that are present, location, loudness, pitch, timbre, and their relation to another. Even subtle deviations in these cues can facilitate the segregation of sound sources. For instance, in a musical scene where multiple instruments align in onsets and pitch cues with no deviations, distinct instruments can be discerned by timbral cues alone despite playing in perfect unison (Cusack & Roberts, 2000; Kalinli & Narayanan, 2009). However, if all instruments in this example were replaced by an orchestration only using multiples of the same instrument, the timbral blend would likely lead to an auditory fusion, where a segregation of individual streams fail, even though individual instruments contribute to form this fusion (Siedenburg, Saitis, & McAdams, 2019, pp. 217-218)

Naturally, acoustic scenes often comprise a mixture of complex acoustic sources, making accurate organization challenging. This can result in the inability to separate individual speakers or a single instrument as such. An example that illustrates this complexity is a study conducted by Huron (1989) in which participants were asked to identify the number of melodies in a polyphonic music mix with up to five melodies played simultaneously. The results showed that after three melodies, the error rate increased significantly, leading to frequent confusion. With a touch of sarcasm, Huron concluded that auditory perception effectively counts “one, two, three, many” (Huron, 1989, p.19). Additionally, stream formation is not impervious to misinterpretations, altering the perceived auditory scene. One such phenomenon is the emergence of auditory illusions like implied polyphony, wherein a single instrument is perceived as multiple instruments or voices by playing an interleaved melody that abruptly shifts between low and high notes (Bregman & Campbell, 1971). This violates the principle of common fate as the note distances surpass a pitch threshold for grouping together. This phenomenon is evident in the works of composers like Bach, where intricate musical compositions evoke the perceptual experience of multiple independent voices despite originating from a single instrument (Davis, 2006; for example “Fugue from Violin Sonata No. 3 in C Major” or “Courante from Violin Partita No. 1 in B Minor”).

Auditory stream formation involves interwoven bottom-up and top-down processes in the perceptual organization (Bregman, 1990; Ciocca, 2008; Shamma & Micheyl, 2010). Bottom-up processes, often referred to as primitive or stimulus-driven processes, entail the extraction of sensory information directly from the acoustic input, analyzing the physical properties of sound signals to extract features such as pitch, timbre, and spatial location. In contrast, top-down processes, also known as knowledge-based processes, involve the use of prior knowledge, expectations, and cognitive strategies to guide perceptual organization. Operating in a goal-directed manner, these processes utilize contextual information and knowledge to interpret and make sense of auditory scenes (Snyder et al., 2012). Both bottom-up and top-down processes are not mutually exclusive and influence each other.

1.2 Musical scene analysis

ASA is widely studied for speech signals and naturalistic speaker scenarios. The investigation of musical scenes is more limited and often only utilizes elementary auditory tasks (Alain & Bernstein, 2015), whereas more ecological approaches, using realistic musical scenes and audio material, are less common. While both musical and speaker scenarios can be used to study ASA, they also present some distinct challenges. In the realm of speech, primary challenges lie in the variability of speech sounds, influenced by factors such as speaker characteristics, speaking rate, and emotional expression. Conversely, music signals encompass a broad range of different timbres including the singing voice and simultaneous melodies that may overlap in time and can be scattered throughout the entire auditory spectral range. Furthermore, the perceptual grouping and segregation of musical sounds are influenced by factors such as tonal relationships, rhythmic patterns, and timbral similarities, adding further complexity to the analysis process. However, these relationships are also used within musical mixtures to help structure the auditory scene, providing acoustic cues and thus fostering its transparency (Huron, 2001). Examples include having a global meter that synchronizes instruments within the mixture (Keller & Burnham, 2005) and distinct timbres (Iverson, 1995) that play notes in separate spectral regions (Jones et al., 1995). The structured organization of a musical scene contrasts with a cocktail party scenario with multiple speakers, where incoherent and varying speech rates and background noises may overlap independently. To reflect the unique challenges faced in the perceptual organization of musical scenes, Bregman coined the term "Musical Scene Analysis" (Bregman, 1990).

1.3 Auditory attention

As there are limits to how much information we can process at once in an auditory scene (Saults & Cowan, 2007; Cowan et al., 2013; Molloy et al., 2015), auditory stream formation can be significantly influenced by auditory attention (Elhilali et al., 2009). Attention is a multifaceted construct with varying definitions across different fields (Hommel et al., 2019). A definition by Huang and Elhilali is used in this dissertation, who described the term as follows:

> *“Attention is at the center of any study of sensory information processing in the brain. It describes mechanisms by which the brain focuses both sensory and cognitive resources on important elements in the stimulus. Intuitively, the brain has to sort through the flood of sensory information impinging on its senses at every instance and put the spotlight on a fraction of this information that is relevant to a behavioral goal.”* (N. Huang & Elhilali, 2017)

Following this definition, attention enables individuals to selectively focus on specific aspects of the auditory scene while suppressing others. This process involves the allocation of cognitive resources to enhance the processing of certain sounds, making them more prominent in conscious awareness. Attention is both based on top-down and bottom-up processes. Studies investigating ASA under the aspects of auditory attention demonstrated that unattended streams can still be tracked in auditory memory, and thus that stream segregation unfolds in a pre-attentive process that is governed by ASA principles (Sussman, Ritter, & Vaughan, 1999; Alain & Arnott, 2000; Sussman, 2005).

As such, distinguishable acoustical cues are needed to successfully form auditory streams. When successful, background streams are still monitored, with segregation and integration processes acting on unattended sounds as well, even if they are irrelevant to the performed task (Sussman & Winkler, 2001). As such, this implies that a kind of passive or automatic sound organization plays a role in auditory scene analysis. Additionally, attention can be directed to highlight and process events within one stream (Sussman, Ritter, & Vaughan, 1999; Sussman, 2006; Fritz et al., 2007). For example, listeners can selectively attend and focus on a specific instrument within a musical scene, allowing them to perceive a more detailed representation of the sound. This process of attentive tracking helps listeners parse complex auditory scenes and extract meaningful information about individual sound sources. As the scene unfolds over time, the tracking demonstrates to become more effective in focusing on a sound. In conjunction, the more difficult it is to unravel the scene, the more time it takes (Best et al., 2008). Still, tracking appears to be quite robust, even in challenging scenes where similar acoustic cues are shared between sounds (Woods & McDermott, 2015). Prior knowledge of a specific attribute, such as its intensity, spatial and spectral position, can be used to cast a spotlight onto the scene and guide attention in a goal-directed manner (Luce & Green, 1978; Mondor & Bregman, 1994; Mondor & Zatorre, 1995; Bey & McAdams, 2002). Furthermore, selectively attending to features in an auditory scene can foster segregation to such an extent, that streams that are not automatically organized into segregated streams become distinguishable, thereby modulating stream formation processes (Botte et al., 1997; Eramudugolla et al., 2005; Sussman & Steinschneider, 2009). However, as processing resources are limited, selectively guiding attention comes at a cost: the resources available for the formation of background streams are curtailed (Sussman et al., 2005; Pannese et al., 2015).

1.4 Auditory salience

In an auditory scene, certain acoustic features particularly provoke auditory attention, standing out amidst the variety of features and becoming a focal point of auditory attention. This heightened prominence is commonly referred to as salience. Kothinti and Elhilali (2023) investigated auditory salience in various audio stimuli ranging from environmental to musical scenes, describing it as follows:

> “Auditory salience is a fundamental property of a sound that allows it to grab a listener's attention regardless of their attentional state or behavioral goals.” (Kothinti & Elhilali, 2023, p. 1).

This definition highlights two important aspects of salience: firstly, its ability to draw attention to a specific scene within a scene, and secondly, its capability to involuntarily distract listeners from other sounds towards the salient sound. Taken together, these aspects demonstrate that salience plays a crucial role in the perceptual organization and shapes the interpretation of complex auditory stimuli. Salience can arise from both bottom-up and top-down processes. Bottom-up salience can occur through disturbance of acoustic regularities within an auditory scene, such as a sudden unexpected loud siren sound. Conversely, a prime example of top-down salience is the heightened salience emerging from hearing one’s own name (Moray, 1959). This phenomenon can involuntarily capture a listener’s attention, overriding attentional goals such as participating in a conversation, even if the name is mentioned by an uninvolved speaker or mistakenly heard.

Compared to the modality of vision, the understanding of what makes a sound particularly salient in a complex acoustical scene is still limited (N. Huang & Elhilali, 2017) , especially in the context of musical scenes. However, computational models of auditory salience that predict neural responses and compare these predictions with brain responses have highlighted that the cues contributing to a sound’s salience are interconnected, multidimensional, and may change over time as the scene unfolds (Kaya & Elhilali, 2014). Key bottom-up candidates identified include cues such as intensity, frequency, as well as loudness, pitch, and timbre (Kaya & Elhilali, 2017; Kaya et al., 2020). Context also plays an important role in the perception of salience, as the acoustic context preceding a sound and predictions of the acoustical scene can alter its ability to attract attention. For example, the sudden sound of a cat meowing in a scenario where cat sounds are expected, such as while petting a cat, is likely to attract less attention than hearing the same sound in a situation where such sounds do not occur, e.g. during a conversation (Liang et al., 2022). Additionally, the control of attention through focusing on sounds of interest have demonstrated that top-down cues increase the contrast between the target sound and background sounds (Alain & Woods, 1997; Petkov et al., 2004; Shamma & Fritz, 2014).

Specific sources of salience have been explored in the compositional structure of musical mixtures with multiple melodies playing at the same time, supporting the observation that music is often composed in such a way that a separation of sounds was facilitated (Huron, 2001). As the musical scene evolves over time, novel melodic lines are observed to draw attention, while repeating lines tend to perceptually decrease in salience (Taher et al., 2016). Similarly, when multiple rhythmic motifs unfold simultaneously, the more irregular rhythms capture listeners’ attention more effectively (Jones et al., 1981), especially when attention is directed towards specific motifs beforehand (Devergie et al., 2010). The employment of these principles can be found across various musical genres. For instance, in progressive house, tracks are often structured in a way that starts with a single repetitive melodic line while new melodic lines and rhythmic patterns are added progressively over time. This creates novel listening experiences and potentially reflects the pleasure derived from parsing new elements within musical mixtures.

The spectral position of melodies also contributes to their salience, affecting pitch and rhythm perception in contrasting ways. On the one hand, in polyphonic music, when two independent melodies play simultaneously, the melody with the highest pitch trajectory is reported to attract more attention than the lower melody – a phenomenon known as the high voice superiority effect (Fujioka et al., 2005). This effect is assumed to arise from perceptual suppression of harmonic structures, where harmonic parts of the lower melody are suppressed by those of the higher melody (Trainor et al., 2014). Supporting this assumption, the effect is known to be present in infants (Marie & Trainor, 2013), indicating that the effect is based on fundamental principles of auditory perception rather than by learned musical schemata that prioritize higher voices. This is further highlighted by the observation that the effect occurs in musicians regardless of the instrument they trained on. However, the effect is more pronounced in musicians who play instruments in the soprano range compared to those who play in the bass range (Marie et al., 2012). On the other hand, the onsets in the lower pitch trajectory provide particularly salient cues for beat or rhythmic perception (Hove et al., 2014). Taken together, these two effects demonstrate that spectral and temporal perception is shaped by sometimes opposing cues. These effects align with the musical principle that the lead melody is usually played by higher-pitched instruments, which particularly attract attention towards the melody, while the rhythmic foundation is typically played by bass instruments.

Another cue that contributes to salience in musical scenes is timbre. Timbre is a complex auditory cue often described as a multidimensional attribute that enables the discrimination of sound sources, even when they match in other perceptual cues such as loudness and pitch (Siedenburg & McAdams, 2017a). McAdams referred to timbre as a “structuring force” (McAdams, 2019), as timbre contains important spectral and temporal cues for stream segregation (Cusack & Roberts, 2000; Kalinli & Narayanan, 2009). However, studies regarding the salience of timbre are mostly focused on environmental sounds (N. Huang & Elhilali, 2017; Kaya et al., 2020). One study that stands out with its focus on timbral salience in musical mixture was by Chon & McAdams (Chon & McAdams, 2012) in which salience of timbre was investigated under the hypothesis that each instrument’s timbre has a unique degree of salience. They suggested that this instrument-dependent salience varies with context: for example, the salience of a French horn might be less pronounced in a full orchestral setting composed of different instrument groups such as brass, strings, percussions and woodwinds. In contrast the French horn is significantly more prominent in a smaller brass ensemble, where the contrast within the more limited pool of different timbre is greater. In their experiment, a salience map was created using 15 orchestral instruments. Instruments were clustered based on their ability to attract listeners’ attention by tapping to a beat created by the instrument. However, only modest correlations were found between salience measures and subjective judgments of the instruments' ability to blend with other simultaneously played instruments. Additionally, no significant salience effects emerged when timbres were played in a polyphonic musical piece composed of two to three simultaneous melodies. They concluded that the salience of the highest voice likely overshadowed the timbre effect. Although a consistent effect of timbre salience was not found within the study, other auditory experiments have demonstrated that a non-instrumental timbre proved to be particularly salient among musical signals. This will be addressed in the following chapter.

1.5 Vocal salience

Among the diverse array of timbres, the human voice emerges as a particularly prominent candidate for enhanced auditory salience. It has long been assumed that vowel sounds are perceived in preference to other sounds even in infancy (Pisoni, 1979; Vouloumanos & Werker, 2007; Krentz & Corina, 2008; Gervain & Geffen, 2019). Neurophysiological experiments have demonstrated that isolated vocal sounds elicit enhanced cortical “voice-specific” responses when presented alongside non-vocal environmental sounds (Belin et al., 2000; Belin et al., 2002), and other musical instrument sounds (Levy et al., 2001; Gunji et al., 2003). Specific neural populations have been identified that respond selectively to music featuring singing voices, but not to instrumental music mixtures (S. V. Norman-Haignere et al., 2022). The recognition of vocal sounds also showed unique advantages over other musical sounds; isolated vocal sounds consistently yield faster reaction times and higher accuracies compared to isolated instrumental sounds (Agus et al., 2012). Subsequent experiments have further highlighted the recognition advantages of vocal sounds, showing that they are identifiable from shorter sound snippets compared to other musical instruments (Suied et al., 2014; Isnard et al., 2019). Comparative analyses of vocal and instrumental melodies have demonstrated that vocal melodies are more accurately recognized than instrumental melodies (Weiss et al., 2012), even when sung without lyrics (Weiss et al., 2021).

1.6 Musical sophistication

The interplay of musical sophistication and associations of how it influences the perception of musical scenes and auditory salience deserves to be mentioned. Several studies have reported that musical sophistication and training are positively associated with enhanced performance in various aspects of scene analysis, such as pitch, rhythm, and timbre discrimination (Micheyl et al., 2006; Kannyo & DeLong, 2011; S. M. K. Madsen et al., 2017). Additionally, higher levels of musical sophistication are associated with better performance in detecting melodies and instruments within mixtures (Marozeau et al., 2010; Slater & Marozeau, 2016; Siedenburg et al., 2020; Hake et al., 2023). Furthermore, studies suggest an association between musical sophistication and auditory salience as musicians exhibit enhanced cortical responses when hearing isolated sounds produced by their trained instruments (Pantev et al., 2001; Shahin, Roberts, & Trainor, 2008; Strait et al., 2012). However, whether these enhanced responses translate into heightened salience when those instruments are part of a musical mixture remains unclear. Contrary, other studies have found no significant group differences between instrumentalists and singers, nor timbre-specific salience (Kannyo & DeLong, 2011; Martins et al., 2022) or differences in timbre discrimination between musicians and non-musicians (Allen & Oxenham, 2014; Siedenburg & McAdams, 2017b; Bigoni F. & Dahl S., 2018).

It is important to note that the relationship between auditory perception and musical sophistication is a topic of ongoing debate with contradictory conclusions. One key aspect of this discussion is the nature vs. nurture question. On the one hand, some researchers argue that predispositions in the brain determine auditory abilities and the extent to which musical training can enhance those abilities (Drayna et al., 2001; Mosing et al., 2014; Zuk et al., 2023; Schneider et al., 2023). On the other hand, others argue that extensive musical training is the driving factor shaping the brain's plasticity, enabling enhanced auditory abilities (Zatorre & McGill, 2005; Bidelman et al., 2011; Herholz & Zatorre, 2012; Bayanova et al., 2024). Additionally, the transfer of beneficial effects to other auditory domains like speech recognition and even to realms outside of auditory perception is highly discussed (for a review, see Benz et al., 2015; Schellenberg & Lima, 2024). Some studies demonstrate far-transfer effects for speech perception (Dubinsky et al., 2019; Puschmann et al., 2019; Zendel et al., 2019) and general cognitive abilities (Talamini et al., 2017; Sala & Gobet, 2020; Rodriguez-Gomez & Talero-Gutiérrez, 2022), while others show a lack of such effects (Ruggles et al., 2014; S. M. K. Madsen et al., 2017; S. M. K. Madsen et al., 2019).

1.7 Aims & relevance of this dissertation

To better understand which and why sounds stand out in an auditory scene is the primary goal of this dissertation. Understanding the acoustic underpinnings that contribute to salience within musical scenes is crucial for several reasons beyond the realm of musical scene analysis.

Firstly, dissecting the mechanisms of vocal salience can provide deeper insights into the fundamental principles of auditory perception. By examining how the human auditory system prioritizes certain sounds over others, we can better understand the cognitive and neural processes involved in ASA even outside of musical scenes. This knowledge not only enriches the field of psychoacoustics but also enhances our understanding of general auditory processing, including how we navigate natural auditory environments.

Secondly, the study of vocal salience has practical implications for various applied domains. In the field of music production and sound engineering, insights into how tracks are highlighted can be incorporated into mixing and mastering techniques, ultimately improving the listening experience. For instance, understanding the acoustic features that enhance vocal prominence can guide sound engineers in emphasizing these elements during the production process, ensuring that vocals stand out clearly in a mix.

Furthermore, this research has implications for the development of auditory prostheses and hearing aids. By identifying the key acoustic features that contribute to vocal salience, we can improve the design of these devices to better support users in isolating and focusing on speech in noisy environments. This is particularly relevant for individuals with hearing impairments, who often struggle to distinguish speech from background noise.

Lastly, the findings of this study can have broader implications for cognitive science and psychology. Understanding how and why certain sounds capture our attention can inform theories of attention and perception, shedding light on the interplay between sensory input and cognitive processing.

1.8 Dissertation structure

The dissertation focuses on exploring salience within musical mixtures, particularly emphasizing the role of singing voices. Chapters 2 to 5 present studies conducted by the author, each commencing with an introduction and concluding with a synopsis that link the studies of this dissertation.

In Chapter 2, the initial study focuses on the detection of instruments and singing voices in musical mixtures extracted from pop music excerpts. Participants are either cued towards a to be detected target instrument before listening to a mixture or hear the mixture before the target cue. Results reveal a unique vocal salience for singing voices that is unmatched by other instruments while at the same time demonstrating an inferior detection of bass instruments. Whether these findings emit from differences in spectral masking or sound level between the tested sounds is also explored, with both the bass inferiority and vocal salience persisting throughout a change in the studied acoustic properties.

Chapter 3 continues the exploration of features that enable vocal salience in a study that utilized the same detection paradigm as in Chapter 2. Here the focus is set on the effects of musical structure, phonological cues, and frequency micro-modulations (FMM) present in singing voices due to the imperfect intonation of human singing. While both musical structure and phonological cues failed to reveal effects regarding vocal salience, a correlation between the frequency modulation depth and salience effect is revealed. Further, it is shown that removing naturally occurring FMM within singing voices eliminates vocal salience.

Building on the effect of FMM, the final study in Chapter 4 presents a study on the recognition of short naturalistic instrumental and vocal sounds. Utilized sounds are extracted from two different music databases and are presented in isolation or accompanied by a piano while the impact of preserving or eliminating the FMM of the instrument or vocal sounds is observed. Contrary to expectations, no effect of FMM is observed, nor is there evidence of a general superiority in vocal recognition. Instead, differences between sung vowels are prominent, with some vowels performing better than instrument sounds, while others are outperformed by instrument sounds. Acoustical analysis implicates that spectral similarities impact the recognition, underpinning that singing voices do not automatically trigger a facilitated recognition, but that the unique role of singing voices is the result of multiple contributing factors.

The study presented in Chapter 5 returns to investigate the perceptual hierarchies observed in musical mixtures, particularly the inferior detection of bass instruments observed in Chapter 2 and facilitated salience of high voices (Fujioka et al., 2005; Marie et al., 2012; Marie & Trainor, 2013). The study examines whether perceptual biases towards distinct frequency regions can explain such salience effects. Stimuli consist of multiple randomized pure tone melodies that are presented in spectrally distinct frequency bands. Results reveal salience for both the lowest and highest melodies, thus a perceptual bias towards the edges of the musical mixture. This suggests that bass inferiority is not solely based on biases in the auditory system but rather on the interplay of complex sounds in musical scenes.

Chapter 6 concludes the dissertation by summarizing all studies presented in the preceding chapters, contextualizing them, and discussing their findings in relation to existing literature. The chapter closes with an outlook on future research.

2. LISTENING IN THE MIX: LEAD VOCALS ROBUSTLY ATTRACT AUDITORY ATTENTION IN POPULAR MUSIC

2.1 Introduction

To investigate the perceptual organization of musical mixtures and the trajectory of attention, we conducted a study incorporating excerpts of pop music. This methodology was chosen to provide a more ecologically valid approach, closely mirroring everyday listening behavior. The task of the study involved detecting individual instruments and vocals in 2-second excerpts of pop music and varying whether information about instrument or vocal sound was given before or after the mixture was presented. This study set the stage for further exploration of auditory attention and salience in musical scenes.

2.2 Study 1

This chapter has been published as: Bürgel, M., Picinali, L., & Siedenburg, K. (2021). Listening in the Mix: Lead Vocals Robustly Attract Auditory Attention in Popular Music. *Frontiers in Psychology, 12, 769663. https://doi.org/10.3389/fpsyg.2021.769663.* The content of this chapter is identical to the manuscript.

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, carried out the experiments, analyzed the data and wrote the final paper. Lorenzo Picinali provided the stimuli and revised the manuscript. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

______________________________ 20.07.2024

(name) Date

Supervisor

2.2.1 Abstract

Listeners can attend to and track instruments or singing voices in complex musical mixtures, even though the acoustical energy of sounds from individual instruments may overlap in time and frequency. In popular music, lead vocals are often accompanied by sound mixtures from a variety of instruments such as drums, bass, keyboards, and guitars. However, little is known about how the perceptual organization of such musical scenes is affected by selective attention, and which acoustic features play the most important role. To investigate these questions, we explored the role of auditory attention in a realistic musical scenario. We conducted three online experiments in which participants detected single cued instruments or voices in multi-track musical mixtures. Stimuli consisted of two-second multi-track excerpts of popular music. In one condition, the target cue preceded the mixture, allowing listeners to selectively attend to the target. In another condition, the target was presented after the mixture, requiring a more “global” mode of listening. Performance differences between these two conditions were interpreted as effects of selective attention. In Experiment 1, results showed that detection performance was generally dependent on the target’s instrument category, but listeners were more accurate when the target was presented prior to the mixture rather than the opposite. Lead vocals appeared to be nearly unaffected by this change in presentation order and achieved the highest accuracy compared with the other instruments, which suggested a particular salience of vocal signals in musical mixtures. In Experiment 2, filtering was used to avoid potential spectral masking of target sounds. Although detection accuracy increased for all instruments, a similar pattern of results was observed regarding the instrument-specific differences between presentation orders. In Experiment 3, adjusting the sound level differences between the targets reduced the effect of presentation order, but did not affect the differences between instruments. While both acoustic manipulations facilitated the detection of targets, vocal signals remained particularly salient, which suggests that the manipulated features did not contribute to vocal salience. These findings demonstrate that lead vocals serve as robust attractor points of auditory attention regardless of the manipulation of low-level acoustical cues.

2.2.2 Introduction

In everyday life, our sense of hearing is exposed to complex acoustical scenes that need to be analyzed and interpreted. The ability to segregate an acoustic scene into a mental representation of individual streams is known as auditory scene analysis (ASA; Bregman & McAdams, 1994). A prime example of this is listening to music with multiple instruments playing at once. Human listeners can focus and track a single instrument remarkably well, even though the acoustic signal is a potentially ambiguous clutter of diverse instrument signals.

Two interwoven analytical processes are used in ASA: endogenous top-down and exogenous bottom-up processes. Endogenous processes are based on cortical functions such as expectations, learned patterns and volition. Exogenous processes are driven by pre-attentive processes based on the temporal and spectral properties of a sound, from which auditory attributes such as duration, pitch, or timbre are computed, and which are pivotal for grouping auditory information into separate sound events. Timbre, often simply described as “texture” or “tone color” (Helmholtz, 1885), is a multidimensional attribute (Siedenburg & McAdams, 2017a) that enables the discrimination of sound sources (e.g., sounds from a keyboard vs. a guitar), even though they may match in other acoustic cues such as loudness and pitch.

A well-established approach to the study of ASA and auditory attention is the use elementary auditory tasks, such as the presentation of sequential or simultaneous streams of tones (for a review see, Alain & Bernstein, 2015). Bey and McAdams (2002) investigated the influence of selective attention in ASA using two-tone sequences, one of which was interleaved with distractor tones. The semitone spacing between the distractor tones and the target sequence was varied from 0 to 24 semitones, thereby varying the strength of exogenous cues that allow for bottom-up stream segregation. Participants had to judge whether the sequences were different or identical and had to ignore the distractors. To vary the dependency on selective attention, in one condition the stream with distractor tones was presented first, followed by the melody without distractors; in a second condition selective attention was facilitated by presenting the melodies without distractors first, thus providing a pattern that could be compared with the following mixture. The results showed that participants achieved higher recognition rates when the melodies without distractors were presented first, thus being able to selectively attend to the target melody.

Another more ecological approach uses polyphonic music to study ASA. In polyphonic music, multiple relatively independent melodies (also referred to as voices) are played or sung simultaneously. Behavioral studies showed that when listening to polyphonic music a superior perception of timing and meter is found in the lower voices (Hove et al., 2014), whereas tonal and melodic perception is facilitated in the highest voice (Crawley et al., 2002). Accordingly, the so-called high-voice superiority effect states that the voice with the highest pitch trajectory is most salient in polyphonic mixtures (Fujioka et al., 2005). It has been shown that this effect is present in infants (Marie et al., 2013) and that it can be enhanced by musical training (Marie et al., 2012). Using a model of peripheral auditory processing, results by Trainor et al. (2014) suggest that the origin of high-voice superiority may be based on physiological factors such as cochlear filtering and masking patterns.

Another factor that has been shown to affect musical scene perception and the specific trajectory of auditory attention is related to the repetitiveness of musical voices. Taher and McAdams (2016) found that when a repetitive and non-repetitive voice are playing simultaneously, attention is drawn to the non-repetitive voice. Barrett et al. (2021) investigated whether the coherent timings between instruments in a piece of music facilitate stream segregation. The authors either slowed down one instrument or recomposed an instrumental line so that it no longer matched with the other lines. The results suggested that, when instruments are temporally coherent, attention is not directed to a particular instrument, and therefore instruments are integrated into one percept. For incoherent musical lines, attention was drawn towards one instrument while the other instrument was ignored. A study by Disbergen et al. (2018) focused on the effect of timbre dissimilarity for distinguishing between two melodic voices in polyphonic music. Although no clear effect for a modification of timbral dissimilarity could be observed, the results implied a trend that a reduction of timbral dissimilarity and thus a reduction of acoustical cues leads to a deterioration of stream segregation, further suggesting that a minimum of exogenous cues is necessary to track and separate single streams. In Siedenburg et al. (2020), listeners had to hear out instruments and melodies of varying sound level masked by a simultaneously playing instrument. It was found that participants were able to exploit dips in the masker signal, allowing them to hear the target instrument at lower levels than with a masker that did not contain these dips.

Several of the aforementioned studies used (simplified or stylized) excerpts of Western classical instrumental music. In Western popular music, the lead melody and thus the centerpiece of a song is sung by a human voice (lead vocals), which is accompanied by a variety of instruments and, at times, background vocals. Recent studies have shown that the voice occupies a unique role among other sound sources (e.g., Belin et al., 2000; Levy, 2001; Agus et al., 2012; Suied et al., 2014; Isnard et al., 2019). In a neurophysiological study, Belin et al., (2000) examined the response to speech, vocal non-speech sounds and non-vocal environmental sounds. The data implied not only that cortical activity to vocal speech and non-speech sounds was higher than to non-vocal environmental sounds but also that specific regions in the human cortex responded more strongly to vocal sounds, suggesting a specialized processing of speech sounds. Levy et al., (2001) measured neurophysiological data from participants in an oddball task in which single instruments and singing voice were presented sequentially. A piano sound was used as a target, while other sounds were used as distractors. The results showed a stronger response to the presentation of the human voice, termed the “voice-specific response”. The authors hypothesized that this response represented a gating mechanism in which the auditory system allocates the input to be processed phonologically. In Agus et al. (2012), accuracy and reaction times were investigated in a sound classification task. Single notes were played by instruments, sung by voices, or played by interpolations between instruments and voices (i.e., chimeras). Accuracy for voices was higher and reaction times were faster than for all other target categories, indicating an advantage in processing voices. Studies by Suied et al. (2014) and Isnard et al. (2019) focused on the recognition of timbre in short glimpses of recorded sounds that differed only in timbre. Again, singing voices stood out by achieving recognition above chance level with a sound duration of only 4 ms, while all other instrument categories required 8 ms durations.

In the present study, we aim to investigate auditory attention in an instrument and singing voice detection task inspired by everyday music listening of popular music. To study how the detection of different instruments is modulated by auditory attention, we vary the presentation order of mixture and target cue. In one order of presentation, a cue from a target vocals or instrument is presented first, followed by the mixture, such that the cue can be used to search the mixture for the target. In the reverse presentation order, the mixture is presented first followed by the target cue. Based on experiments such as Bey and McAdams (2002), we expect that the order in which a cue is presented first facilitates detection of the target. Motivated by the distinct role of singing voices that has been reported in the literature, we investigated whether the lead vocals in popular music would play a special role in auditory scene analysis and selective listening. Based on this assumption, we hypothesize that lead vocals achieve distinctly higher accuracies in both presentation orders.

2.2.3 General Methods

For our experiment, we used short excerpts of popular music in which either a cued target instrument or target vocal was present or absent in a mixture of multiple instruments (see Figure 2.1B). To test the effects of selective auditory attention, we interchanged the presentation order of the cue and mixture (Bey and McAdams, 2002). This yielded two different listening scenarios: one requiring selective listening, and the other requiring a rather global mode of listening. When the target was presented prior to the mixture, selective attention could be used to detect the target in the mixture. When the target was presented after the mixture, listeners had to be aware of possibly all components of the mixture and hence listen more globally to the excerpts. In that case, attention could be affected by exogenous factors, for instance the salience of individual sounds in the musical scene. We conducted three experiments aimed to study the role of attention in the processing of popular music mixtures and whether acoustic modifications of the excerpts would manipulate the detection of instruments or vocals. For the first experiment, we left the excerpts unmodified and investigated the detection accuracy in the complex musical scene and how it was affected by the presentation order and different instruments. In the second experiment, we aimed to suppress energetic masking of the target by means of bandpass/bandstop-filtering. To control the influence of instrument dependent sound levels, we equalized the sound levels ratios between the different targets in the third experiment. A schematic overview of the experiments is shown in Figure 2.1C. The same general methods were applied in all three experiments. Specific modifications of the methods are described in detail in the respective experiments (see Sec. 3, 4, 5).

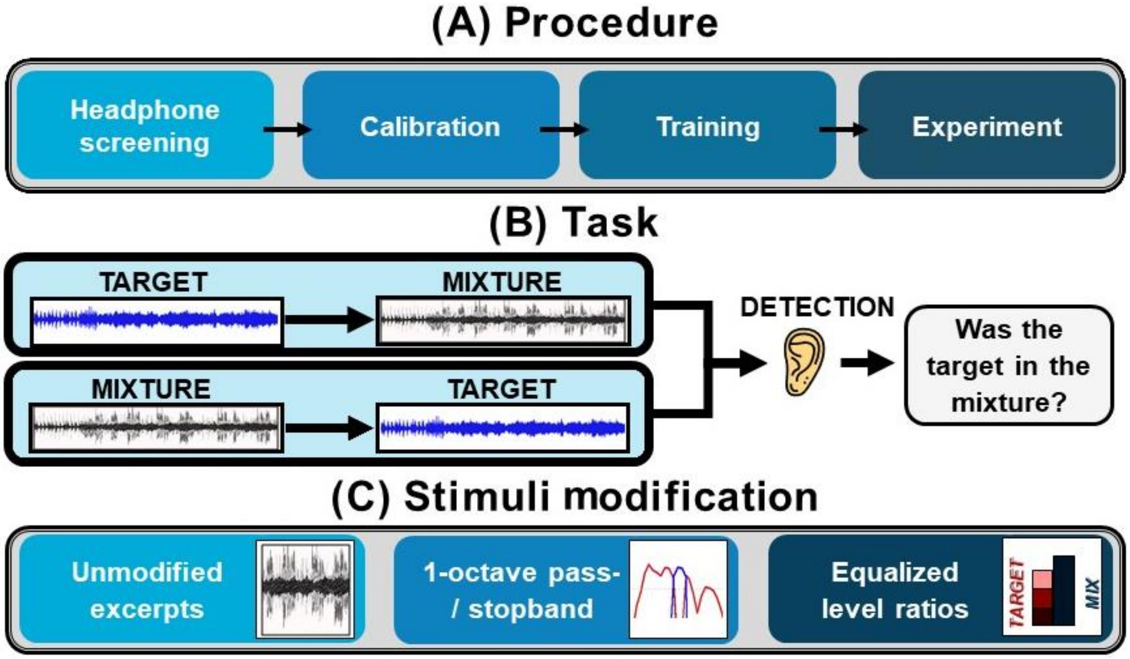

Figure 2.1: ***Schematic overview of the experiments.***

(A) Procedure: The experiment started with a headphone screening task, followed by a subjective sound level calibration, a training section where participants were familiarized with the instrument detection task and finally the main experimental section. (B) Task: An instrument detection task was used in the experiments: Participants either took part in an experiment where the targets were preceding the mixtures or where the mixtures were preceding the targets. (C) Stimuli modification: In the first experiment, excerpts unmodified from their original state were used. In the second experiment, the target instruments were filtered in an octave band to create a spectral region in which the target instrument could pass without being spectrally masked. In the third experiment, the individual sound level differences between the diverse instruments were adjusted to one of three possible level ratios.

Participants

All participants were students recruited via an online call for participation at the e-learning platform of the University of Oldenburg. General information about the experiment and exclusions criteria were given. The criteria included the use of headphones, a stable internet connection and self-reported normal hearing. Participants could start the online experiment at any time via a link that was provided in a personalized email. Participation was compensated monetarily. We acquired information about the participants musical training using five questions: Number of instruments played, hours practiced during the period of greatest musical interest, years of lessons in music theory, years of lessons for an instrument, self-designation (non-musician, amateur musician, professional musician).

Stimuli & Task

An illustration of the stimuli extraction is shown in Figure 2.2. Stimuli were generated using a Matlab script (MathWorks Inc., Natick, MA, USA) that extracted two-second excerpts from a multitrack music database. The database was created by Tency Music and is used within the Musiclarity web-app (Eastgate et al., 2016). It consists of sound alike reproductions of well-known popular music with English lyrics and individual audio files for each instrument. The Instruments in the database were coarsely categorized as: Backing Vocals, Bass, Drums, Guitars, Lead Vocals, Piano, Percussion, Strings, Synthesizer, Winds. For each excerpt, one to-be attended instrument was chosen (target). Other instruments in the excerpt that were not from the same category as the target served as maskers (mixture). Instruments from the same category that were not used as a target were excluded from the mixture. When lead vocals were assigned as the target, all backing vocals were also excluded. Songs were drawn pseudo-randomly, with the same song chosen as infrequent as possible. To investigate which instruments were audible at any given time, the sound level of each instrument was analyzed using a 500 ms sliding window. In each window, the root-mean-squared (RMS) sound level was calculated. Windows were qualified as potential candidates for the excerpt extraction if one instrument in the target category and six to nine additional instruments had sound levels above -20 dB relative to the instrument’s maximum sound level across the full song. A previously unused 2000 ms time slice containing four qualified adjacent 500 ms windows was randomly drawn. Three monophonic signals were compiled from each 2-second excerpts: 1) a signal only containing the target, 2) a signal containing a mixture of five to eight instruments from non-target categories plus the target. 3) A signal containing a mixture of six to nine instruments without the target. For mixtures, the full number of instruments was used, which were also present in the original excerpt of the song. A logarithmic fade-in and fade-out with a duration of 200 ms was applied to the beginning and end of all extracted signals. For half of the trials, the mixture signals were arranged to contain the target signals, and for the other half the mixture did not contain the target signal. From these signal combinations, two stimuli with a duration of 4500 ms were created using different presentation orders for the target and mixture signal. In the “Target-Mixture” condition, the target signal was followed by a 500 ms pause and the mixture signal; in the “Mixture-Target” condition, the presentation order was reversed. For the use on the online platform, the stimuli were converted from WAV format to MP3 with a bit rate of 320 kbit/s. Example stimuli are provided on our website: https://uol.de/en/musik-wahrnehmung/sound-examples/listening-in-the-mix

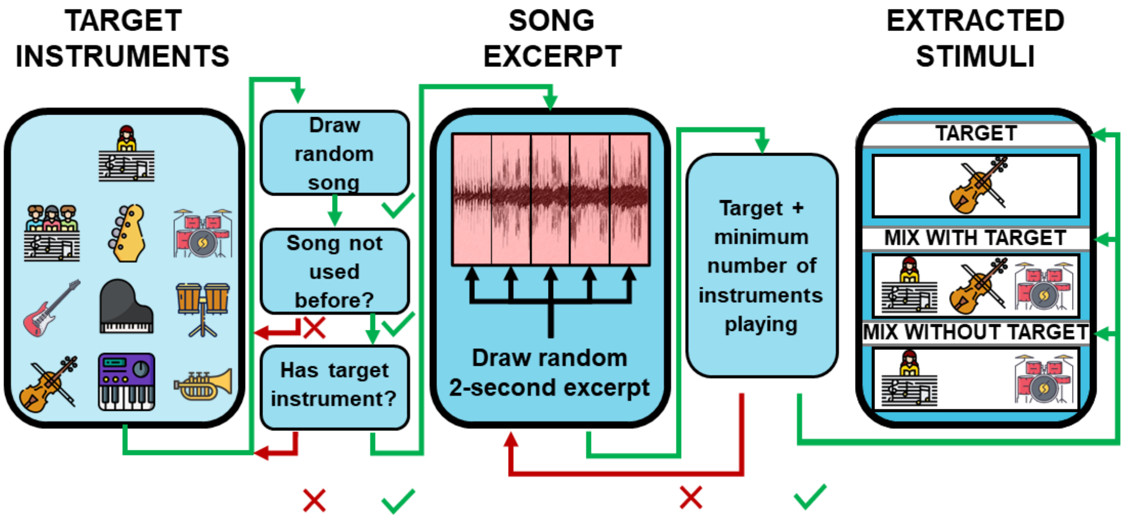

Figure 2.2: ***Stimuli extraction***

Short excerpts from a multitrack database containing reproductions of popular music were used as stimuli. The schematic shows the workflow of the stimulus construction. For details, see the text.

Procedure

The experiments were approved by the ethics board of the University of Oldenburg and carried out online via the web platform www.testable.org. Participants were divided into one of two groups. For group one all stimuli had the presentation order "Target-Mixture", whereas for group two the presentation order was reversed (see Figure 2.1B). The same excerpts were used for both groups, thus the only differences were in the order of presentation. Each experiment was further divided into four consecutive segments (see Figure 2.1A).

At the beginning of the experiment, participants had to fill in a form regarding personal data (see 2 A). To get an indication of whether participants were using headphones, a headphone screening task was performed at the beginning of the experiments. For Experiment 1 headphone screening was based on Woods et al. (2017), employing a sequential presentation of three pure tones, where one of the tones was quieter. The tones were phase shifted on the left side by 180° degrees and therefore appeared attenuated when listening over loudspeakers but not attenuated when headphones were worn. Therefore, a matching volume judgment should only be achieved by wearing headphones. Listeners had to detect the quiet tone and passed the test if five out of six detections were correct. For Experiments 2 and 3, the headphone screening was based on Milne et al. (2020), which provides a higher selectivity for headphone users than the headphone screening used in Experiment 1. Here, a sequence of three white noise signals were presented, where one of the noise signals was phase shifted by 180 degrees in a narrow frequency band at around 600 Hz on the left headphone channel. When headphones were worn, the phase shift was perceived as a narrow tone embedded in the broadband noise. Listeners had to detect the tone and passed the test if five out of six detections were correct. Participants who failed the headphone screening were removed from the data analysis.

After the headphone screening, three song excerpts were presented aiming to provide an impression of the dynamic range of the stimuli. During the presentation, participants were instructed to adjust the sound to a comfortable level. This was followed by a training phase, where participants were familiarized with the detection task. Participants listened to stimuli akin to those used in the main experiment and were asked whether the target was present or absent in the mixture. For each category, one stimulus with and without target was presented. To help participants understand the task and to make them more sensitized for the acoustic scene, feedback was given after each answer. This was followed by the main experiment where the same procedure was used but no feedback was given. Stimuli presented in the training segment were not reused in the experiment segment. All stimuli were presented in a random sequence that intermixed all conditions (except for the between-subjects factor of presentation order). The number of stimuli, the conditions and the target categories differed from experiment to experiment and are therefore described in the sections on the individual experiments below.

Data analysis

Following the methodology recommended by the American Statistical Association (Wasserstein et al., 2019), we refrain from the assignment of binary labels of significance or non-significance depending on an immutable probability threshold. We provide mean detection accuracies, followed by a square bracket containing the 95% confidence intervals computed by means of bootstrapping and round brackets containing the decrease or increase through a change in presentation order.

A generalized binominal mixed-effect model (West, 2014) was used for the statistical analysis. All mixed-effects analyses were computed with the software R (R Core Team, 2014) using the packages lme4 (Bates et al., 2015) which was also used to estimate marginal means and confidence intervals. Our model included random intercepts for each participant and item (i.e., stimulus). All binary categorical predictors were sum-coded. The correlation coefficients of the model are given as standardized coefficients (β), followed by 95% confidence intervals in square brackets and probability (p). To summarize the main effects and interactions, results are presented in the form of an ANOVA table, derived from the GLME models via the *anova* function from the car package (Fox et al., 2019). A detailed view of the behavioral results, models and statistic evaluations for each experiment are presented in the supplementary material (see Table. 1-6).

Method validation

Since the experiment was conducted online, and therefore did not undergo the strict controls of a laboratory experiment, we compared results for using calibrated laboratory equipment and consumer devices. In order to achieve this, a pilot experiment that was very similar to Experiment 1 was completed by the members of the Oldenburg research lab. In one condition, participants used their own computer and headphones. In another condition, they used calibrated audio equipment, and the presentation order of these two conditions was counterbalanced across participants. The calibrated equipment consisted of a laptop, RME Babyface soundcard, and Sennheiser HD650 headphones. The long-term sound level was set to 75 dB SPL (A), measured with Norsonic Nor140 sound-level meter using music-shaped noise as the excitation signal. Results showed very similar data for both types of equipment (for details see supplementary Figure 2.1), which did not indicate any systematic problem in conducting the present study via online experiments.

2.2.4 Experiment 1 – Unmodified excerpts

The first experiment was our starting point to investigate selective auditory attention in musical scenes. We left the excerpts in their original state (as described in 2 B). As target categories besides the lead vocals, we chose four instrument categories that had shown rather diverse results in a pilot experiment.

Participants

A total of 84 participants with a mean age of 25.1 years (SD = 4.5, range = 19-44) were tested in the experiment. A total of 25 out of 42 participants passed the headphone screening for the Target-Mixture condition and 22 out of 42 for the Mixture-Target condition (age = 25.3, SD = 5, range = 19-44). Only participants passing the headphone screening were included in further analysis. 11 participants in the Target-Mixture condition and 10 participants in the Mixture-Target condition described themselves as either amateur or professional musicians.

Stimuli & Procedure

For the first experiment, the following five target categories were selected: lead vocals, bass, synthesizer, piano, drums. Headphone screening was based on Woods et al. (2017). In the training phase of the main experiment, one excerpt with a target and one excerpt without a target were presented for each of the five target categories, summing up to 10 stimuli in total. In the experimental phase, 150 stimuli were presented, divided into 30 stimuli for each of the five target categories. The average duration of the experiment was 25 minutes.

Results & Discussion

Figure 2.3 displays the average results of the first experiment for each instrument and presentation order (for numerical values, see supplementary Table 1&2). Detection accuracy differed depending on the target category and order of presentation which was also evident in our model (Instrument: $\chi^2 = 97.881$, $p < 0.001$, Order: $\chi^2 = 38.878$, $p < 0.001$.). Averaged across target categories, the Target-Mixture condition yielded the highest accuracy of 84% [70% - 97%], which deteriorated in the Mixture-Target condition to 72% [55% - 88%] (-12%). This decline was strongest for the bass category in which the mean accuracy dropped by -19% from the Target-Mixture to the Mixture-Target condition. A nearly identical decrease was found for the synthesizers (-11%), piano (-

8%) and drums (-11%). The lead vocals had the best performance overall and were least affected by a change in the presentation order (-2%). This resulted in an interaction effect between the instrument factor and the presentation order ($\chi^2 = 13.059$, $p = 0.011$).

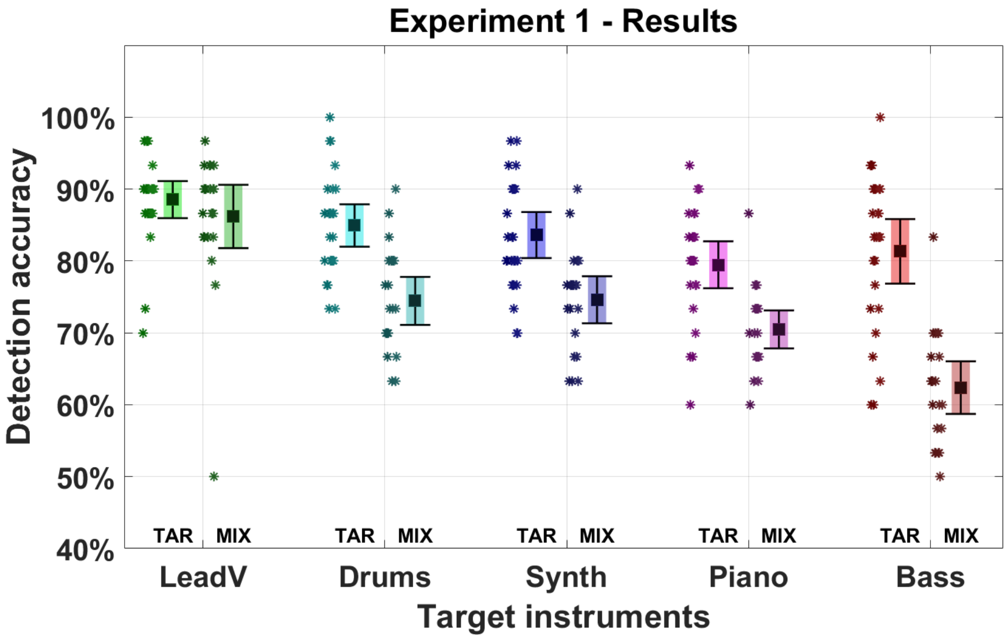

Figure 2.3: *Detection accuracy in experiment 1*

Five instrument categories were used as targets (lead vocals, drums, synthesizer, piano, bass). The Square marks the mean detection accuracy for a given instrument category. Error bars indicate 95% confidence intervals. Asterisks represent the average accuracy of an individual participant for the given instrument category. "TAR" denotes the presentation order "Target-Mixture" where the target instrument cue was presented followed by a mixture. "MIX" denotes the presentation order "Mixture-Target" where a mixture was presented followed by the target instrument cue.

All instruments except the lead vocals showed degraded detection accuracy when listeners were required to listen to the musical scenes without a cue. While the degradation of detection accuracy in a global listening scenario was to be expected (Bey & McAdams, 2002; Janata et al., 2002; Richards, 2004), the specific attentional bias towards lead vocals is, to our knowledge, a novel finding. We will refer to this unique characteristic as "lead vocal salience" in the following. This finding is in line with the unique role of singing voices documented in previous experiments, where voices were processed faster and more accurately in comparison to other musical instruments (e.g. Agus et al., 2012; Suied et al., 2014; Isnard et al., 2019) and were shown to have a unique cortical voice-specific-response indicating a specialized processing for human voices (e.g. Levy, 2001).

The bass was found to be most strongly affected by a change in presentation order, having a medial detection accuracy in the Target-Mixture condition that, however, decreased almost twofold compared to the other instruments. One explanation for this could be tied to the spectral characteristics of the bass. The bass mostly occurs in a rather narrow band in the low frequencies, whereas other instruments cover a wider frequency range. When a cue is given, attention may be focused selectively towards that frequency band, and thus narrow signals like the bass can be reliably perceived. Another explanation could be derived from the high-voice superiority effect that has been observed in polyphonic music. The effect describes a pre-attentive attentional bias (Trainor, 2014), which, in the presence of multiple voices, draws attention towards the highest voices. In the current experiment, bass signals naturally correspond to low voices, and hence high-voice superiority may come into play.

It is to be noted, that our analysis revealed no systematic differences between participants who declared themselves as musician and those who did not. This held true across all three experiments, even though in previous studies musicians showed improved results in ASA tasks (e.g., Başkent et al., 2018; S. M. K. Madsen et al., 2019; Siedenburg et al., 2020). The most likely reason to explain this may be that we did not specifically control for an equal number of musicians and non-musicians in a large sample; thus, the proportion of participants considered musicians were only a fraction of the total participants, and therefore the sample size may be too small for an adequate statistical comparison. We further analyzed how performance was affected by possible fatigue over the course of the experiment. Considering performance over the duration of the experiment averaged across subjects suggested that the difference between performance at the beginning and end of the experiment was negligible (for details, see supplementary Figure 2).

To further evaluate the acoustic origins of the lead vocal salience, we analyzed the music database in terms of spectral features and sound levels features. For each song and target category, we evaluated the broadband sound level as well as the sound level on an ERB-scale between all instruments and voices in a category and all other instruments and voices. We used a sliding window of 500 ms moving over the duration of a song and discarded all windows in which the sound level was less than 20 dB below the maximum sound level of the instruments, voices, or mixtures. The results of the time windows were then averaged for each song and are displayed in Figure 2.4.

The spectral analysis revealed a frequency region from 0.5 kHz to 4 kHz where the difference between the lead vocals and remaining mixtures had a positive level ratio (up to 2.5 dB), meaning that the lead vocals exhibited higher levels than the sum of accompaniment instruments and were therefore released from energetic masking in those spectral regions. While the lead vocals had a relative sound level of more than 0 dB in such a broad spectral region, only the bass and drums showed similar levels in either low or high frequencies. The other instruments did not have such a differentiated spectral range and their level was substantially below the level of the lead vocals. This was also evident in the broadband level analysis, in which the lead vocals had a significantly higher level than the other instruments. Accordingly, two acoustically-based explanations for the superior detection accuracy of the lead vocals could be a) less susceptibility to masking by other instruments or b) higher loudness levels of lead vocals. To scrutinize these two hypotheses, we conducted a second experiment where the vocals and the instruments were released from masking in the same frequency band, and a third experiment equalizing the sound level differences between lead vocals and instruments.

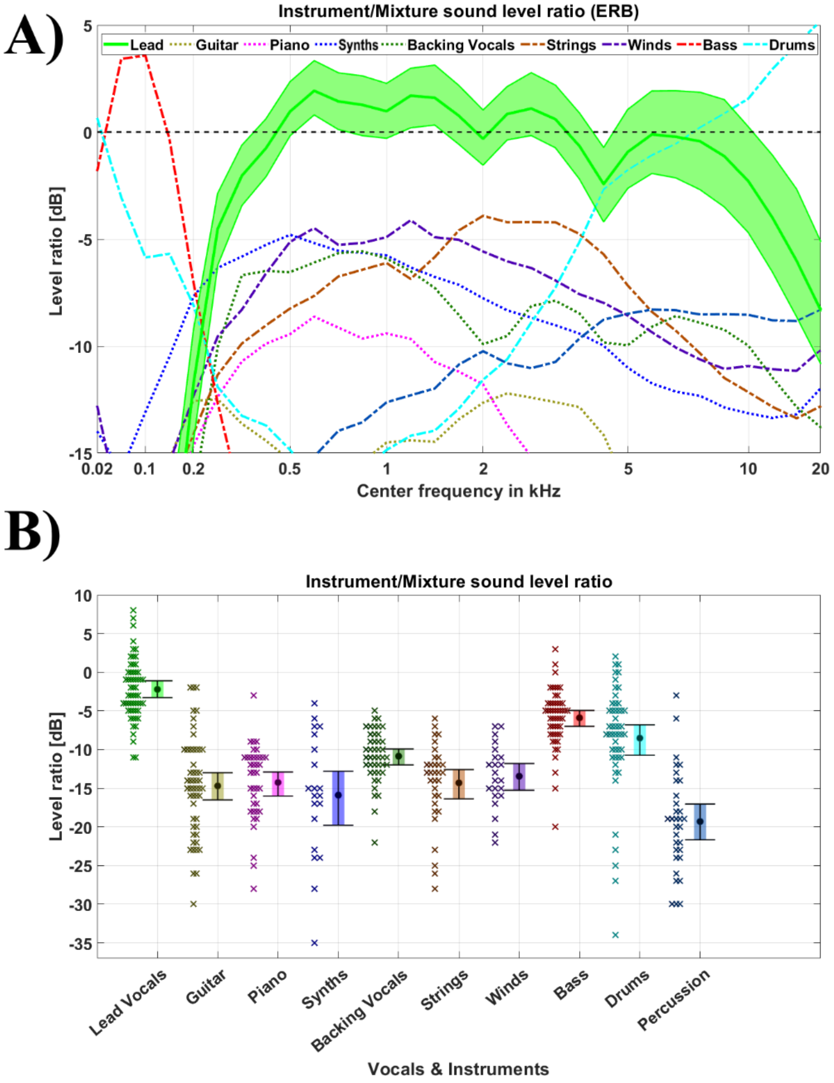

Figure 2.4: ***Database feature analysis***

We analysed the average sound level in ERB-bands (A) and broadband sound level (B) between each voice or instrument and the remaining mixture for each song. (A) Each coloured line represents the average sound level for the given centre frequency. The filled area represents the 95% confidence intervals for the lead vocals. (B) The circle marks the mean detection accuracy for a given instrument category. Error bars indicate 95% confidence intervals. Crosses represent the average level of an individual song for the given instrument category.

2.2.5 Experiment 2 – Spectral unmasking equalization

To investigate whether the observed lead vocal saliency was due to spectral masking, we here examined the spectral regions where vocals tended to be unmasked and applied the same unmasking to different target instruments. For this purpose, we analyzed the database for spectral regions in which the lead vocals exhibited particularly high sound levels. A broad spectral region from about 0.5 to 5 kHz was found. To provide equal masking and unmasking for all vocals and instruments, we used octave bands adjacent to the center of this region (1-2 kHz and 2-4 kHz) and designed filters to pass signals only into one of the two bands (bandpass) or to suppress signals only into this range (bandstop). To compensate for level-dependent differences, the sound levels of all target instruments were adjusted identically. Only instruments with relevant intensity in the selected frequency bands were considered as targets for the experiment. Therefore, lead vocals, guitars and piano were used as target categories. To avoid listeners focusing only on the octave bands, a randomly drawn accompaniment instrument was passed through the octave band for one third of trials, whereas the target category sound was attenuated in the octave band.

Participants

A total of 49 participants with a mean age of 25.6 years (SD = 4.2, range: 20-39) were tested in the experiment. A total of 20 out of 25 participants passed the headphone screening for the Target-Mixture condition and 20 out of 24 or the Mixture-Target condition (age = 23.5, SD = 2.9, range: 20-29). Only participants passing the headphone screening were included in the analysis. Among these, 12 participants in the Target-Mixture condition and 4 participants in the Mixture-Target condition described themselves as either amateur or professional musicians.

Stimuli & Procedure

In two out of three excerpts, the target was filtered through a passband either from 1 to 2 kHz or from 2 to 4 kHz, while the mixture was filtered through a bandstop in the same octave band. Excerpts filtered in this way are referred to as “TBP” in the following. To prevent participants to focus on only one of the two octave bands, in one third of the excerpts, a randomly drawn accompanying instrument was filtered through a passband of either 1 to 2 kHz or 2 to 4 kHz, while the other accompaniment instruments and the target was filtered through a bandstop in the same octave band. Excerpts filtered this way are referred to as “TBS” in further analysis. Bandpass and bandstop filters were designed and applied using the corresponding Matlab functions *bandpass* and *bandstop* (Signal Processing Toolbox Release 8.3, MathWorks Inc., Natick, MA, USA). The filtered target signal was used both during the presentation of the cue and when it was presented in the mix. The signal components in the stopband were attenuated to -80 dB FS (decibels relative to full scale). Sound levels ratios between targets and mixtures were adjusted for all targets to -10 dB. In a final step, the average sound level of each stimulus was normalized to -15 dBFS. As target categories, lead vocals, guitar and piano were chosen.

The headphone screening test was based on Milne et al. (2020). In the training phase of the main experiment, one stimulus with and one without a target were presented for each of the three target categories, each of the two octave bands and one additional stimulus for each target category and octave band where the target was filtered by a bandstop and an accompaniment instrument was filtered by a bandpass, summing up to 18 stimuli in total. In the experimental phase of the main experiment, 180 stimuli were presented, divided into groups of 60 for each of the three target categories and further subdivided into 20 stimuli for each octave band where the target was filtered by a bandpass plus 10 for each octave band where the target was filtered by a bandstop. The average duration of the experiment was 35 minutes.

Results & Discussion

Results are displayed in Figure 2.5 (for details, see supplementary Table 3&4). Detection accuracy was affected by the filter type (TBP = target is filtered with a bandpass, TBS = target is filtered with bandstop), presentation order and instrument type. While we used two different adjacent octave bands to filter the signals (1 – 2 kHz, 2 – 4 kHz), results for both frequency bands showed nearly identical results with no systematic differences (differences for all conditions between both octave bands: $\text{Difference\_}_{\text{MEAN}}$ = 2.5%, $\text{Difference\_}_{\text{MIN}}$ = 1.5%, $\text{Difference\_}_{\text{MAX}}$ = 3.5%). This finding was underpinned by the GLME model, which revealed no effect for the usage of different octave bands (Octave: χ2 = 0.002, p = 0.963).

As in Experiment 1, the detection accuracy was better in the Target-Mixture condition and best when the target signal was filtered by a bandpass. The influence of both the order and the filter was reflected in our model (Order: χ2 = 3.547, p = 0.06, Filter: χ2 = 18.657, p < 0.001). For the Target-Mixture TBP condition an average accuracy of 96% [95% - 97%] was observed compared to the 85% [82% - 89%] (-11%) in the Mixture-Target condition. For the Target-Mixture TBS condition an average accuracy of 85% [82% - 88%] was achieved compared to the Mixture-Target condition 76% [72% - 80%] (-9%).

Lead vocals performed best with an accuracy of 96% [93% - 99%] and showed the smallest decrease by changing the order (TBP: -1%, TBS: -2%) or removing the isolation by changing the filtering (Target-Mixture: -4%, Mixture-Target: -5%). This was followed by the guitar with an accuracy of 84% [80% - 86%], which in contrast to the vocals and pianos, achieved higher accuracies in the Target-Mixture TBS than in the Mixture-Target TBP condition and almost as well in the Mixture-Target TBP and the Mixture-Target TBS conditions (difference by order TBP: -17%, TBS: -7%. Difference by filter Target-Mixture: -11%, Mixture-Target: -1%). The piano with an accuracy of 79% [75% - 83%], showed a similar pattern as for the lead vocals and was generally better when it was isolated than when the isolation was lifted (difference by order TBP: -15%, TBS: -15%. Difference by filter Target-Mixture: -18%, Mixture-Target: -18%). This dependence on instruments was also corroborated by our model (Instrument: χ2 = 42.177, p < 0.001).

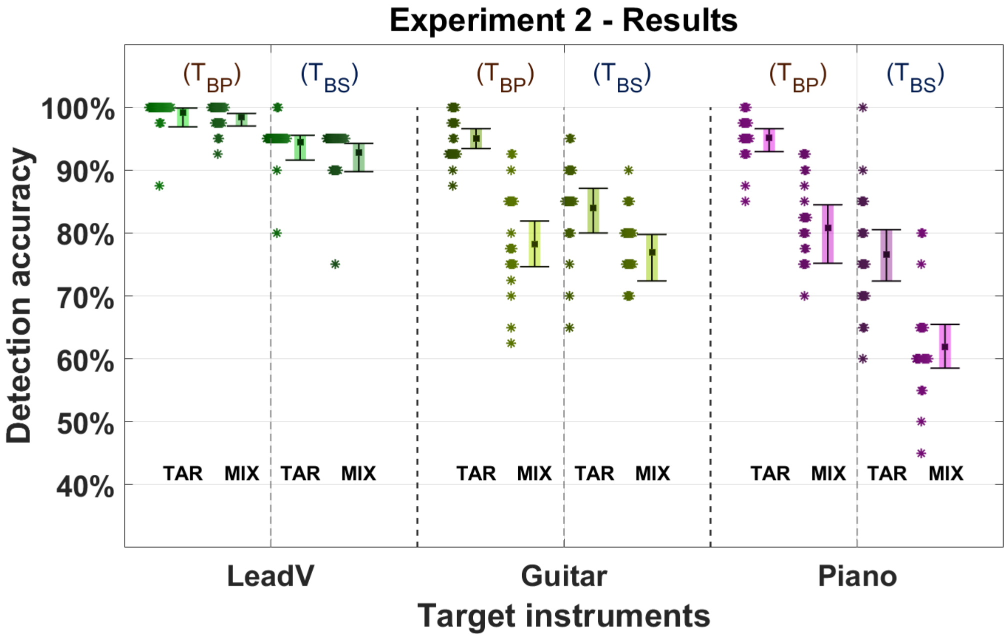

Figure 2.5: ***Detection accuracy in experiment 2***

Three instrument categories were used as targets (lead vocals, guitar, piano). Either a bandpass or bandstop was applied to the filter and the mixture. The target instrument filter type is listed in the upper area of the figure with T_{BP} indicating a bandpass was used and T_{BS} indicating a bandstop was used. The Square marks the mean detection accuracy for a given instrument category. Error bars indicate 95% confidence intervals. Asterisks represent the average accuracy of an individual participant (n = 40) for the given instrument category. "TAR" denotes the presentation order "Target-Mixture" where the target cue was presented followed by a mixture. "MIX" denotes the presentation order "Mixture-Target" where a mixture was presented followed by the target cue.

Compared to the unmodified stimuli in the first experiment, applying a bandpass filter to the target improved the detection of instruments for both presentation orders by up to 16%. Specifically, this improvement raised the accuracies in the Target-Mixture condition to 99% (Exp.1: 88%) for the lead vocals, 95% for the guitar and 95% for the piano (Exp.1: 79%). This indicates that whereas the frequency content of the instrument signals was narrowed down to an octave band and isolated, the additional selective attention in the Target-Mixture condition may have acted as searchlight, allowing for the detection of the target with an improved accuracy. However, whereas the overall accuracy was generally higher compared to the first experiment, the gaps between the accuracy in the Target-Mixture and the Mixture-Target conditions were larger than before for all instruments except the lead vocals. This gap was smallest and almost non-existent for the lead vocals (Exp.1: -2%, Exp.2 TBS: -2%), and enhanced for the guitar (in comparison to the average of non-bass instruments in Exp.1: -10%, Exp.2 TBS: -

17%) and the piano (Exp.1: -8%, Exp.2 TBS: -15%). An instrument specific deterioration was underpinned by our model, which revealed a notable smaller contribution of the order alone (Order: $\chi 2 = 3.5474$, $p = 0.060$) and a much stronger contribution for the interaction between instruments and presentation order (Interaction: $\chi 2 = 8.3447$, $p < 0.015$).

Relative to Exp. 1, the increased effect of presentation order in Exp. 2 could be interpreted as related to the narrowband nature of the target signals, as it was already discussed for the bass in the first experiment. In a global mode of listening, listeners are required to distribute attention across the whole musical scene, which may make it easier to miss narrowband signals in a mixture of wideband signals, or not to perceive them as individual signals. In contrast to the bass in Experiment 1, instruments in Experiment 2 occurred in frequency ranges in which the human hearing is particularly sensitive, which in turn still led to a generally high detection accuracy. Here, the lead vocals also showed advantages over other instruments, which suggests that other characteristics of the lead vocals can be detected within the narrow band, leading to better detection accuracy.

Detection accuracies additionally dropped in all target categories and for both presentation orders when the passband-filter was applied to an accompaniment instrument rather than the target. Again, the lead vocals were by far the least affected target category, showing that the lead vocal salience remains prominent even when the voice is suppressed in frequency regions where it is usually mixed louder than the mix. The general deterioration for all instruments and orders could be the by-product of a strategy in which participants listened primarily to the octave bands in which two-thirds of the targets appeared. Another reason for this pattern of results could be that the target did no longer occur in single octave band and thus targets were again subject to masking. A further possibility would be that the passband used here covers a particularly sensitive frequency region of human hearing, so that sound events in this area may be particularly salient.

Taken together, the spectral filtering guaranteed that the target stood out from the mixture, hence resulting in high detection accuracies in the Target-Mixture condition. In the Mixture-Target condition, accuracies were distinctly lower. As in the first experiment, the lead voice showed by far the smallest difference between the different orders of presentation. When the isolation of the target was removed, all instruments showed a deterioration of detection accuracies. Again, the lead vocals achieved the highest accuracy compared to the other instruments, yet with a smaller deterioration across presentation orders. Thus, an explanation of the lead vocal salience does not seem to be due to less susceptibility to masking in frequency regions in which the vocals are mixed with higher levels than the sum of the accompanying instruments.

2.2.6 Experiment 3 – Sound Level Equalization

Motivated by the relatively high sound levels of the lead vocals, here we aimed to manipulate the level ratios of targets relative to the accompaniment to investigate whether this manipulation would affect detection performance and the observed differences between the presentation orders. As target categories, we selected the bass and lead vocals categories, because both were shown to be the conditions with lowest and highest performance in the first experiment, respectively. Since both target categories differ greatly in their spectral components, with bass being present mainly in the low frequencies and lead vocals in the mid and high frequencies, listeners could adopt a strategy where they would only listen to one of the distinct spectral regions. To avoid this, we added an additional experimental condition that contained instruments from all other target categories.

Participants

A total of 55 participants with a mean age of 24.4 years (SD = 5.1, range = 18-33) were tested in the experiment. A total of 20 out of 27 participants passed the headphone screening for the Target-Mixture condition and 20 out of 28 for the Mixture-Target condition (age = 24, SD = 3.5, range: 19-33). 9 participants in the Target-Mixture condition and 7 participants in the Mixture-Target condition described themselves as either amateur or professional musicians. Only participants passing the headphone screening were included in the analysis.

Stimuli & Procedure

The target categories lead vocals, bass, and individual instruments from the categories drums, guitar, piano, synthesizer, strings, and winds were chosen as targets for the instrument conditions. Excerpts with targets from lead vocals, bass, and the mixed category appeared equally often. The sound level ratio between the targets and mixtures was set to one of three possible levels where the broadband level of the target was either 5 dB, 10 dB or 15 dB below the level of the mixture (referred here as -5 dB, -10 dB and -15 dB condition). To accomplish this, the 2-second instrument and mixture signal was separately analysed using a 100 ms sliding window. For every window, the A-weighted sound level was computed using the *weightingFilter* function in Matlab (Audio Toolbox Version 2.1, MathWorks Inc., Natick, MA, USA) followed by a sound level estimation via RMS calculation. We normalized the average sound levels of each stimulus to -15 dBFS.

The headphone screening test was based on Milne et al. (2020). In the training phase of the main experiment, one excerpt with and without a target were presented for each of the three target categories and for each of the three sound level ratio conditions, summing up to 18 stimuli in total. In the experimental phase of the main experiment, 180 stimuli were presented, divided into 60 stimuli for each of the three target categories and further subdivided into 20 stimuli for each sound level conditions. The average duration of the experiment was 35 minutes.

Results & Discussion

Results for the third experiment are shown in Figure 2.6 (for details see supplementary Table 5&6). Changes in the sound level ratio, presentation order and target category affected the detection accuracy. The best performing condition was the lead vocal target category with a level ratio of -5 dB with an averaged accuracy of 99% [98% - 100%] in the Target-Mixture condition and 100% [100% - 100%] in the Mixture-Target order. Lowest detection accuracy was achieved by the lowest level ratio of -15 dB in the bass category ranging from 60% [57% - 64%] in the Target-Mixture condition to 58% [53% - 63%] in the Mixture-Target order. Within the same presentation order, the Target-Mixture condition achieved a generally higher accuracy, whereas the mean accuracy of all categories in the Mixture-Target condition deteriorated from 83% [82% - 84%] to 78% [75% - 80%] (-5%).

Averaged across all sound level ratios, the lead vocals showed the highest detection accuracy and was most unaffected by a change in presentation order showing a slightly better accuracy of 97% [96% - 98%] in the Target-Mixture condition compared to the Mixture-Target condition with an accuracy of 95% [93% - 96%] (-2%). In the category of multiple instruments, the detection accuracy deteriorated from 84% [81% - 86%] in the Target-Mixture condition to 79% [76% - 82%] (-5%) in the Mixture-Target condition. The bass achieved the overall lowest accuracy dropping from 71% [68% - 74%] in the Target-Mixture condition to 63% [60% - 66%] (-8%) in the Mixture-Target condition. These results are similar to the results of Experiment 1, where the lead vocals performed best while the bass performed worst.

In summary, contrary to our assumptions using higher sound levels and equalizing the levels had not resulted in a cancellation of the order influence, as it was still present in most conditions but not present for the lead vocals. This reasoning was further supported by the statistical model, which showed relevant interaction between the instruments and presentation orders for the third experiment (Interaction: $\chi2 = 1.257$, $p < 0.001$). However, if we draw a comparison between Exp. 1 and Exp. 3, the effect of presentation order was reduced considerably (Bass: Exp.1 = -19%, Exp.3 = -10%. Other instruments: Exp.1 = -12%, Exp.3 = -5%).

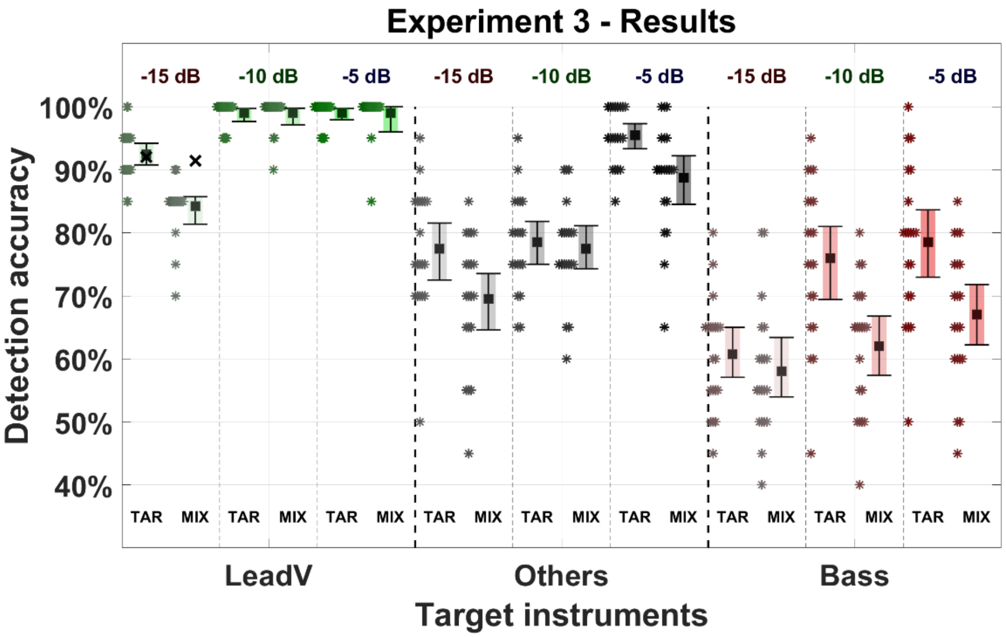

Figure 2.6: ***Detection accuracy in experiment 3***

Three instrument categories were used as targets (lead vocals, bass, others = drums, guitar, piano, strings, synthesizer, winds). The sound level ratio between the target and mixture was adjusted to either -5 dB, -10 dB, -15 dB and is listed in the upper area of the figure, decreasing from right to left. The Square marks the mean detection accuracy for a given instrument category. Error bars indicate 95% confidence intervals. Asterisks represent data from individual participants for the given instrument category. "TAR" denotes the presentation order "Target-Mixture" where the target cue was presented followed by a mixture. "MIX" denotes the presentation order "Mixture-Target" where a mixture was presented followed by the target cue. The green cross above the lead voice in the -15 dB condition marks the averaged detection accuracy when all stimuli that were consistently answered incorrectly were excluded (for details see results).

Similar to the first and second experiment, the lead vocals stood out and achieved the highest accuracy. A decline in accuracy of 7% in the Target-Mixture and a considerably larger decline of 15% in the Mixture-Target conditions could only be observed in the lowest sound level condition. An observation of the individual sound levels shows a clear difference between both presentation orders in the lowest level condition: -0% (-5 dB), -0% (-10 dB), -8% (-15 dB). Yet, a closer look at individual stimuli revealed that this decrease was based on a few distinct stimuli that achieved low detection accuracies (for a detailed view see supplementary Figure 3&4). In the Target-Mixture condition, 17 out of 20 stimuli exceeded 90% detection accuracies, while one stimulus was close to chance level at 56%, whereas two stimuli were almost collectively answered incorrectly, achieving an accuracy of only 15%. This agreement was even stronger in the Mixture-Target condition where 15 out of 20 stimuli achieved 100% accuracy, one stimulus achieved 95% accuracy, and the last four stimuli achieved an accuracy of less than

16%. When we excluded all stimuli that were consistently answered incorrectly (detection accuracy of 0%), the results remained identical for all conditions except for the lead vocals in the lowest level ratio. Here, accuracy in the Target-Mixture condition remained at 92% (+0%) and in the Mixture-Target condition from 86% to 91% (+5%), almost closing the gap between the two presentation orders that arose in the -15 dB condition (from an order effect of 6% to 3%), although we only conservatively screened out stimuli that were consistently answered incorrectly by all participants (for a detailed view, see supplementary Table 5). For these reasons, a generalization of the results of the lead vocals at the lowest level ratios seems questionable, because accuracies here seem to be mainly driven by a few stimuli rather than the systematic change in level ratio.

The target category “others” was most affected by a level decrease, declining by 18% in both presentation orders. Differences between presentation orders in this target category varied at different levels: -7% (-5 dB), -1% (-10 dB), -8% (-15 dB). At the -10 dB condition revealed an ambiguous result, where the difference between the two presentation orders is only marginal.   Considering all seven remaining conditions, which show clear effects, we interpret the present pattern of results as indication that the adjustment of the sound level ratios did not eliminate the order effect for the instruments and did not cause any robust order effect for the lead vocals.

The bass was slightly less affected by a decrease in level, achieving 18% in the Target-Mixture and 9% in the Mixture-Target conditions. With decreasing level, a consistently deteriorating detection of the mixture-target condition could be observed: -12% (-5 dB), -14% (-10 dB), -3% (-15 dB).

In summary, by varying the target-to-accompaniment level ratio, we here observed effects of presentation order at different level ratios for a mixed category of instruments and for the bass instrument but no notable effect for the lead vocals. This once more confirmed the inherent salience of lead vocals in musical mixtures, which seems to be stable across sound levels.

2.2.7 General Discussion

In this study, we aimed to investigate auditory scene analysis for musical instruments and singing voices and its modulation by selective auditory attention. Excerpts of popular music were presented in an instrument and singing voice detection task. Participants listened to a two-second excerpt either globally with a mixture preceding a target cue or selectively with a target cue preceding the mixture. We hypothesized that listeners’ performance would be facilitated when a target cue is presented prior to the presentation of the mixture. In addition, we suspected a detection advantage of lead vocals relative to other instruments.

In line with our assumptions regarding the presentation order and previous studies (e.g., Bey & McAdams, 2002; Janata et al., 2002), detection performance was best when the target cue was presented before the mixture, highlighting the role of endogenous top-down processing to direct selective auditory attention. Accuracy worsened when listeners were presented the target after the mixture. This was the case for all target categories, apart from the lead vocals, which not only achieved the best detection accuracies among all target categories, but also showed no (or clearly much smaller) decreases of detection accuracies across the two orders of presentation. Although we initially assumed a higher detection accuracy for the lead vocals, the latter finding exceeded our expectations about vocal salience in musical mixtures.

In a second and third experiment, we investigated how manipulations of acoustical features would affect lead vocal salience by eliminating differences between the target categories in relative sound level or release from spectral masking. However, contrary to our hypothesis, even when targets were completely unmasked from the mixture, or when the same sound levels were applied, lead vocals retained a unique role and robustly achieved the highest detection accuracies results across all manipulations, with a clear advantage over all other instruments. These findings support a unique role of the lead vocals in musical scene perception. More generally, this pattern of results is consistent with previous work in which singing voices have been shown to be perceptually privileged compared to other musical instruments by yielding faster processing (Agus et al., 2012) and more precise recognition rates (e.g., Suied et al., 2014; Isnard et al., 2019) as well as a stronger cortical representation (e.g. Levy, 2001) compared to other instruments. Our results demonstrate that auditory attention is drawn to the lead vocals in a mix, which complements knowledge about pre-attentive perceptual biases in musical scene analysis such as the high-voices superiority effect (Trainor et al., 2014).

From a music production point of view, it may be argued that the facilitated detection of lead vocals could be a result of acoustic cues that arise from common tools such as compression and notch filtering, which allow the vocals to “come through” and be perceived as the most prominent sound “in front of” the mixture. The results of Exp. 2 and 3 render this hypothesis unlikely, however. Despite complete unmasking of target categories in Exp. 2 and drastic changes of level in Exp. 3, the lead vocals remained the only target category that did not show an order effect and hence may be interpreted as the only target category with specific auditory salience.

In several recent studies, Weiss and colleagues provided evidence for a memory advantage of vocal melodies compared to melodies played by non-vocal musical instruments. Analyses of the recognition ratings for old and new melodies revealed that listeners more confidently and correctly recognized vocal compared to instrumental melodies (Weiss et al., 2012). It was further shown that the presentation of vocal melodies, as well as previously encountered melodies, was accompanied by an increase in pupil dilation (Weiss et al., 2016), often indirectly interpreted as indicator of raised engagement and recruitment of attentional resources. Our results directly highlight that those vocal melodies appear to act as a type of robust attentional attractors in musical mixtures, hence providing converging evidence for a privileged role of voices in auditory scene analysis.

The lead vocal salience observed here could be due to a human specialization to process speech sounds. Therefore, lead vocals may have benefited from their speech features. Previous studies have demonstrated that phonological sounds, such as words and pseudo-words, are easier to detect than non-phonological complex sounds (Signoret et al., 2011). Therefore, an idea worth exploring is whether the advantage of lead vocals is still present when the vocal melody is sung with non-phonological speech, sung by humming or played by an instrument. Another speech-like aspect that could make vocals more salient is the semantic content of the lyrics. Trying to grasp the meaning behind the lyrics could therefore draw attention to the vocals. Although test participants were not English native speakers, in Germany, it is common to listen to songs with English lyrics.

Another origin for the lead vocal salience could lie on a compositional level. In the used excerpts of popular music, lead voices certainly acted as the melodic center of the songs. The resulting melodic salience is known to dominate the perception of a musical scene (Ragert et al., 2014). A question which would be interesting to examine is whether the vocal salience found here would also be found if the main melody were played by another instrument and whether in this case the instrument would show enhanced auditory salience.

2.2.8 Conclusion

We used short excerpts of popular music in a detection task to investigate the influence of selective auditory attention in the perception of instruments and singing voices. Participants were either directed to a cued target vocal or instrument in a musical scene or had to listen globally in the scene before the cued target was presented. As expected, in the presentation order where no cue was given before the mixture and thus no additional support for endogenous top-down processing was provided, detection accuracy deteriorated. Whereas all instruments were affected by a change in the presentation order, the lead vocals were robustly detected and achieved the best detection accuracies among all target categories. To control for potential spectral and level effects, we filtered the target signals so that they were unmasked in a particular frequency band and eliminated sound levels differences between the targets. This facilitated instrument detection for the presentation order where the target was presented first, but not for the order where the mixture was presented first. These results indicate that the observed lead vocal salience is not based on acoustic cues in frequency region where the lead vocals are mixed at higher levels than the sum of accompanying instruments. It was further found that higher sound levels resulted in more similar scores across the presentations orders, but there remained clear order effect for all instruments except for the lead vocals, suggesting that the higher level-ratios of vocals are not the origin of the lead vocal salience. This confirms previous studies on vocal significance in auditory scene analysis. Further research is needed to assess whether these features are based on its unique vocal qualities, semantic aspects of the vocal signal, or on the role of the center melody in musical mixtures.

2.2.9 References

Agus, Trevor R.; Suied, Clara; Thorpe, Simon J.; Pressnitzer, Daniel (2012): Fast recognition of musical sounds based on timbre. In *The Journal of the Acoustical Society of America* 131 (5), pp. 4124–4133. DOI: 10.1121/1.3701865.

Alain, Claude; Bernstein, Lori J. (2015): Auditory Scene Analysis. In *Music Perception* 33 (1), pp. 70–82. DOI: 10.1525/mp.2015.33.1.70.

Barrett, Karen Chan; Ashley, Richard; Strait, Dana L.; Skoe, Erika; Limb, Charles J.; Kraus, Nina (2021): Multi-Voiced Music Bypasses Attentional Limitations in the Brain. In *Frontiers in neuroscience* 15, p. 588914. DOI: 10.3389/fnins.2021.588914.

Başkent, Deniz, Christina D. Fuller, John J. Galvin III, Like Schepel, Etienne Gaudrain, and Rolien H. Free. "Musician effect on perception of spectro-temporally degraded speech, vocal emotion, and music in young adolescents*." The Journal of the Acoustical Society of America 143, no. 5 (2018): EL311-EL316.* DOI: 10.1121/1.5034489

Bates, Douglas; Mächler, Martin; Bolker, Ben; Walker, Steve (2015): Fitting Linear Mixed-Effects Models Using lme4. In *J. Stat. Soft.* 67 (1). DOI: 10.18637/jss.v067.i01.

Belin, P.; Zatorre, R. J.; Lafaille, P.; Ahad, P.; Pike, B. (2000): Voice-selective areas in human auditory cortex. In: *Nature 403 (6767),* S. 309–312. DOI: 10.1038/35002078.

Bey, Caroline; McAdams, Stephen (2002): Schema-based processing in auditory scene analysis. In *Perception & psychophysics* 64 (5), pp. 844–854. DOI: 10.3758/bf03194750.

Bregman, Albert S.; McAdams, Stephen (1994): Auditory Scene Analysis: The Perceptual Organization of Sound. In *The Journal of the Acoustical Society of America* 95 (2), pp. 1177–1178. DOI: 10.1121/1.408434.

Crawley, Edward J.; Acker-Mills, Barbara E.; Pastore, Richard E.; Weil, Shawn (2002): Change detection in multi-voice music: The role of musical structure, musical training, and task demands. In *Journal of Experimental Psychology: Human Perception and Performance* 28 (2), pp. 367–378. DOI: 10.1037/0096-1523.28.2.367.

Disbergen, Niels R.; Valente, Giancarlo; Formisano, Elia; Zatorre, Robert J. (2018): Assessing Top-Down and Bottom-Up Contributions to Auditory Stream Segregation and Integration With Polyphonic Music. In *Frontiers in neuroscience* 12, p. 121. DOI: 10.3389/fnins.2018.00121.

Eastgate, Richard, Lorenzo Picinali, Harshada Patel, and Mirabelle D'Cruz. "3d games for tuning and learning about hearing aids." *The Hearing Journal 69, no. 4 (2016): 30-32.* DOI: 10.1109/VR.2018.8446298

Fox J, Weisberg S (2019): An R Companion to Applied Regression, *3rd edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html*

Helmholtz, H. (1885/1954): On the Sensations of Tone as a Physiological Basis for the Theory of Music. *New York, NY: Dover. trans. by A. J. Ellis of 4th German Edn. 1877 republ. 1954 Edn*.

Hove, Michael J.; Marie, Céline; Bruce, Ian C.; Trainor, Laurel J. (2014): Superior time perception for lower musical pitch explains why bass-ranged instruments lay down musical rhythms. In *Proceedings of the National Academy of Sciences of the United States of America* 111 (28), pp. 10383–10388. DOI: 10.1073/pnas.1402039111.

Isnard, Vincent; Chastres, Véronique; Viaud-Delmon, Isabelle; Suied, Clara (2019): The time course of auditory recognition measured with rapid sequences of short natural sounds. In *Scientific reports* 9 (1), p. 8005. DOI: 10.1038/s41598-019-43126-5.

Janata, Petr; Tillmann, Barbara; Bharucha, Jamshed J. (2002): Listening to polyphonic music recruits domain-general attention and working memory circuits. In *Cognitive, affective & behavioral neuroscience* 2 (2), pp. 121–140. DOI: 10.3758/CABN.2.2.121.

Levy, D. A.; Granot, R.; Bentin, S. (2001): Processing specificity for human voice stimuli: electrophysiological evidence. In *Neuroreport* 12 (12), pp. 2653–2657. DOI: 10.1097/00001756-200108280-00013.

Marie, Céline; Fujioka, Takako; Herrington, Leland; Trainor, Laurel J. (2012): The high-voice superiority effect in polyphonic music is influenced by experience: A comparison of musicians who play soprano-range compared with bass-range instruments. In *Psychomusicology: Music, Mind, and Brain* 22 (2), pp. 97–104. DOI: 10.1037/a0030858.

Marie, Céline; Trainor, Laurel J. (2013): Development of simultaneous pitch encoding: infants show a high voice superiority effect. In *Cerebral cortex (New York, N.Y. : 1991)* 23 (3), pp. 660–669. DOI: 10.1093/cercor/bhs050.

Madsen, Sara M. K.; Marschall, Marton; Dau, Torsten; Oxenham, Andrew J. (2019): Speech perception is similar for musicians and non-musicians across a wide range of conditions. In: *Scientific reports 9 (1)*, S. 10404. DOI: 10.1038/s41598-019-46728-1.

Milne, Alice E.; Bianco, Roberta; Poole, Katarina C.; Zhao, Sijia; Oxenham, Andrew J.; Billig, Alexander J.; Chait, Maria (2020): An online headphone screening test based on dichotic pitch. In *Behavior research methods*. DOI: 10.3758/s13428-020-01514-0.

Ragert, Marie; Fairhurst, Merle T.; Keller, Peter E. (2014): Segregation and integration of auditory streams when listening to multi-part music. In *PloS one* 9 (1), e84085. DOI: 10.1371/journal.pone.0084085.

Richards, Virginia M.; Neff, Donna L. (2004): Cuing effects for informational masking. In *The Journal of the Acoustical Society of America* 115 (1), pp. 289–300. DOI: 10.1121/1.1631942.

Siedenburg, Kai; McAdams, Stephen (2017): Four Distinctions for the Auditory "Wastebasket" of Timbre. In: *Frontiers in psychology 8*, S. 1747. DOI: 10.3389/fpsyg.2017.01747.

Siedenburg, Kai; Röttges, Saskia; Wagener, Kirsten C.; Hohmann, Volker (2020): Can You Hear Out the Melody? Testing Musical Scene Perception in Young Normal-Hearing and Older Hearing-Impaired Listeners. In *Trends in hearing* 24, 2331216520945826. DOI: 10.1177/2331216520945826.

Signoret, Carine; Gaudrain, Etienne; Tillmann, Barbara; Grimault, Nicolas; Perrin, Fabien (2011): Facilitated auditory detection for speech sounds. In *Frontiers in psychology* 2, p. 176. DOI: 10.3389/fpsyg.2011.00176.

Suied, Clara; Agus, Trevor R.; Thorpe, Simon J.; Mesgarani, Nima; Pressnitzer, Daniel (2014): Auditory gist: recognition of very short sounds from timbre cues. In *The Journal of the Acoustical Society of America* 135 (3), pp. 1380–1391. DOI: 10.1121/1.4863659.

Taher, Cecilia; Rusch, René; McAdams, Stephen (2016): Effects of Repetition on Attention in Two-Part Counterpoint. In *Music Perception* 33 (3), pp. 306–318. DOI: 10.1525/mp.2016.33.3.306.

Trainor, Laurel J.; Marie, Céline; Bruce, Ian C.; Bidelman, Gavin M. (2014): Explaining the high voice superiority effect in polyphonic music: evidence from cortical evoked potentials and peripheral auditory models. In *Hearing research* 308, pp. 60–70. DOI: 10.1016/j.heares.2013.07.014.

Wasserstein, Ronald L.; Schirm, Allen L.; Lazar, Nicole A. (2019): Moving to a World Beyond “p < 0.05”. In *The American Statistician* 73 (sup1), pp. 1–19. DOI: 10.1080/00031305.2019.1583913.

Weiss, Michael W.; Trehub, Sandra E.; Schellenberg, E. Glenn (2012): Something in the way she sings: enhanced memory for vocal melodies. In: *Psychological science 23 (10),* S. 1074–1078. DOI: 10.1177/0956797612442552.

Weiss, Michael W.; Trehub, Sandra E.; Schellenberg, E. Glenn; Habashi, Peter (2016): Pupils dilate for vocal or familiar music. In: *Journal of experimental psychology. Human perception and performance 42* (8), S. 1061–1065. DOI: 10.1037/xhp0000226.

West, Brady T.; Welch, Kathleen B.; Galecki, Andrzej T. (2014): Linear Mixed Models: Chapman and Hall/CRC.

Woods, Kevin J. P.; Siegel, Max H.; Traer, James; McDermott, Josh H. (2017): Headphone screening to facilitate web-based auditory experiments. In *Attention, perception & psychophysics* 79 (7), pp. 2064–2072. DOI: 10.3758/s13414-017-1361-2.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

MB and KS designed the study. LP provided the stimuli. MB collected and analyzed the data. MB wrote a first draft of the manuscript. KS and LP revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by a Freigeist Fellowship of the Volkswagen Foundation to K.S.

Acknowledgments

We thank Tency Music for allowing us to use excerpts from the musiclarity database for our experiments. We also express our gratitude to the reviewers whose constructive feedback improved the paper.

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.769663/full#supplementary-material

2.3 Synopsis

The result of the study illustrates the significant influence of top-down processing on the perception of auditory scenes and the varying salience of different target categories. Prior information about a target sound enabled listeners to focus their attention on it, emphasizing sounds that would otherwise be lost in the mixture; this effect was particularly pronounced for bass instruments. Furthermore, a distinct vocal salience emerged, with vocals being recognized at the same accuracy regardless of whether prior information was provided. Investigations into whether spectral masking or level differences caused this saliency ruled out these factors, indicating that vocal salience is not based on these acoustic features.

3. SALIENCE OF FREQUENCY MICRO-MODULATIONS IN POPULAR MUSIC

3.1 Introduction

To further investigate the fundamentals of vocal perception, additional experiments were conducted using the same recognition paradigm. The emphasis here focused on how the main melody of a song, phonetic cues, and frequency micro-modulations of vocals contribute to recognizability. In addition, stimuli were extracted from a different music database to investigate whether the saliency effect persists across different databases that may differ in their titling technique.

3.2 Study 2

This chapter has been published as: Bürgel, M., & Siedenburg, K. (2023). Salience of Frequency Micro-modulations in Popular Music. Music Perception, 41(1), 1–14. https://doi.org/10.1525/mp.2023.41.1.1: The content of this chapter is identical to the manuscript.

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, carried out the experiments, analyzed the data and wrote the manuscript. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

____________________	20.07.2024
(name) Supervisor	Date

3.2.1 Abstract

Singing voices attract auditory attention in music unlike other sounds. In a previous study, we investigated the salience of instruments and vocals using a detection task, in which cued target sounds were to be detected in musical mixtures. The presentation order of cue and mixture signals influenced the detection of all targets except the lead vocals, indicating that listeners focus on voices regardless of whether these are cued or not, highlighting a unique vocal salience in music mixtures. The aim of the present online study was to investigate the extent to which phonological cues, musical features of the main melody, or frequency micro-modulation (FMM) inherent in singing voices contribute to this vocal salience. FMM was either eliminated by using an autotune effect (Experiment 1) or transferred to other instruments (Experiment 2). Detection accuracy was influenced by presentation order for all instrumental targets and the autotuned vocals, but not for the unmodified vocals, suggesting that neither the phonological cues that could provide a facilitated processing of speech-like sounds nor the musical features of the main melody are sufficient to drive vocal salience. Transferring FMM from vocals to instruments or autotuned vocals reduced the magnitude of the order effect considerably. These findings suggest that FMM is an important acoustical feature contributing to vocal salience in musical mixtures.

3.2.2 Introduction

Who has not experienced it: While listening to music, the ear seamlessly picks up a catchy vocal melody from a musical mix. A melody emerges in the mind of the listeners, seemingly independent from the musical background that it was embedded in. Notwithstanding the ease of auditory processing, multi-instrumental music confronts listeners with complex acoustic scenes, in which instruments and voices overlap in both time and frequency. Despite the potential complexity of musical scenes, the auditory system analyzes and groups musical mixtures into representations of individual streams. This ability to organize sounds into perceptual streams is referred to as auditory scene analysis (ASA; Bregman, 1994). This framework assumes that ASA is determined by primitive (bottom-up) and schema-driven (top-down) processing. The latter is thought to incorporate processes of scene parsing based on attention, memory, and knowledge. Selective attention in ASA has been studied using an interleaved melody recognition paradigm with simple melodies (Bey & McAdams, 2002), which has listeners detect a target sound in a mixture. The target can be presented before or after the mixture and the resulting difference in detection accuracy is assumed to be due to processes of selective attention. In a previous study (Bürgel et al., 2021), we found that all sound categories except the lead vocals showed effects of selective attention. Because accuracy was particularly high and independent of selective attention for vocals, we dubbed this pattern of results *vocal salience*. Here, we wished to further explore the basis of vocal salience in popular music. Generally, this approach extends previous research by using mixtures of popular music as highly realistic and representative stimuli for ASA research.

Auditory attention, such as the reflex-like focusing on a loud sound or deliberate listening to an instrument in a mixture, modulates the cognitive representation of the acoustic scene by allocating processing resources to distinct elements of a scene (e.g., Shamma, Mounya, Christophe, 2010; Sussman, 2017). Studies of auditory attention in musical scenes found that the voice occupies a unique role among other sound sources, enabling the voice to stand out from other instruments in a mixture: When human listeners are asked to recognize isolated voices and instruments, responses to voices occur faster and with higher accuracy (Agus et al., 2012). Moreover, voice sounds require a shorter time of exposure for recognition compared to other musical instrument sounds (Suied et al., 2014; Isnard et al., 2019). When comparing vocal melodies and instrumental melodies, previously presented vocal melodies are more precisely recognized compared to instrumental melodies (Weiss, Trehub, Schellenberg, 2012). Neurophysiological experiments underpin this unique role of the vocals, showing an enhanced cortical response when vocal signals are presented in isolation among speech and non-vocal environmental sounds (Belin et al., 2000; Belin, Zatorre, & Ahad, 2002), and among other instruments (Levy, Granot, Bentin, 2001; Gunji et al., 2003). Further, when presented in a musical mixture, specific neural populations were found that respond distinctively to music with singing voices but not to instrumental music (S. V. Norman-Haignere et al., 2022).

This facilitated processing of vocals also plays out in multi-instrumental musical mixtures. Previously, we investigated the detection of cued target instruments and voices in short excerpts of popular music mixtures (Bürgel, Picinali, & Siedenburg, 2021). The cue consisted of an isolated instrument or voice and was either presented before or after the mixture. Notably, all target signals except the lead vocals showed a clear surplus of detection accuracy when the target cue was presented before the mixture, highlighting the intrinsic salience of the vocals that attracts the listeners attention regardless of the presentation of a cue. This salience persisted and was unmatched by other instruments, even when the instruments and vocals were matched in sound level or were spectrally filtered to pass through the mixture unmasked.

The question arises as to which features of vocal signals contribute to their unique role among natural sounds. Here, we considered three candidate features. First, it may seem reasonable to suggest that the unique salience of vocals could arise from the phonological information they contain. Language specific processing may potentially activate increased attentional resources (Signoret et al., 2012). Second, another feature contributing to the unique presence of the vocals could be their favorable musical role in the multi-instrument mixtures. In Western popular music, the lead vocals contribute the main melody of a song and thus are composed to possess a prominent role with respect to the accompanying instruments and background vocals. When listening to music hierarchically structured into main melody and accompaniment, previous studies have shown that attention is drawn towards the main melody (Ragert, Fairhurst & Keller, 2014).

Third, a more acoustically based candidate feature may be related to frequency micro-modulation (FMM). Here, we understand FMM as non-stationary frequency changes in acoustic signals, usually less than one semitone, which are not perceived as irregular or as intonation errors. In singing, FMM tends to be caused by imperfect control of intonation caused by vocal-motor control adjustments of the human voice (Hutchins, Larrouy-Maestri, & Peretz, 2014) and is present even in highly trained singers (e.g., Sundberg, Prame, Iwarsson, 1996; Mori et al., 2004; Hutchins & Campbell, 2009). Even though pitch detection for vocals seems to be less precise than for musical instruments (Hutchins, Roquet, Peretz, 2012; Sundberg, Lã, Himonides, 2013; Gao & Oxenham, 2022), FMM influences the perception of intonation (Larrouy-Maestri & Pfordresher 2018), is known to facilitate the prominence of vowel sounds (McAdams 1989; Marin &

McAdams, 1991), and evokes cortical responses that can be traced by neurophysiological measurements (Wang, Tan & Martin, 2013). Experiments with speech signals indicate that both the exaggeration and reduction of the modulations result in decreased speech intelligibility (Miller, Schlauch, Watson, 2010); frequency modulations naturally inherent in speech signals were associated with highest speech intelligibility scores.

The purpose of the present study is to further investigate the unique ability of the vocals to be the focal point of auditory attention in musical scenes (vocal salience), which was found in our previous experiments (Bürgel, Picinali & Siedenburg, 2021). More precisely, we investigate how these three candidate features contribute to vocal salience. We analyze the role of FMM as well as phonological cues in natural singing voices, either by eliminating the modulations in the vocals (Experiment 1) or by transferring the modulations to instruments (Experiment 2). We further examine how having instruments play the vocal melody affects their salience in the mixture (Experiment 1 & Experiment 2). We use the same experimental paradigm as in our prior experiments (Bürgel, Picinali & Siedenburg, 2021): participants are asked to detect a cued target signal (vocal or instrument) embedded in a mixture of multiple instruments. Because detection accuracy is influenced not only by the salience of the target but also by factors such as sound level or spectral masking (Bürgel et al, 2021; Siedenburg et al., 2020), we test the effect of the presentation order of target cue and mixture to isolate how the detection of the target signal is modulated by auditory attention. For one half of the participants, the target cue is presented first and followed by the mixture, allowing the cue to be used to "search" the mixture for the target. This order is used to measure detection accuracy in a facilitated listening situation where participants have prior knowledge of the target. For the other half of participants, the presentation order is reversed, with the mixture presented first, so that the detection of targets strongly depends on the salience of the target in the mixture. A comparison between both presentation orders allows us to quantify the influence of the effect of selective attention through the surplus of the accuracy in the target-mixture condition compared to the mixture-target condition.

For the conditions where FMM is eliminated from the vocal signals, we speculate on two possible outcomes: Either the facilitated detection for singing voices remains intact because it is driven by phonological cues that encourage a facilitated processing of speech-like sounds, and that are retained throughout the pitch quantization. Alternatively, detection of singing voices degrades, because vocal salience is a result of the human sensitivity towards FMM. Considering the role of the melodic material, we speculate that in trials in which instruments replace the vocals and play the main melody detection accuracy is clearly facilitated. For transferring the melody and FMM of the lead vocals to instruments, we expect that the presence of FMM that are uncommon for the instruments introduces a cue that results in an increase of detection accuracy compared to conditions presenting the main melody without FMM. If the FMM is driving the facilitated detection of vocals, this transfer of FMM may decrease or even eliminate the effect of presentation order.

3.2.3 Method

Participants

All participants were recruited via an online call for participation on the e-learning platform of the University of Oldenburg. The call included a briefing, a link to the online experiments, and inclusion criteria such as the use of headphones, a stable internet connection, and self-reported normal hearing. Participants could take part in the experiment online at any time during a one-month time window. Participants who took part in Experiment 1 were not permitted to take part in Experiment 2. A total of 69 participants (age: $\bar{x}$ = 25.1, std = 3.5) took part in Experiment 1 and 70 participants (age: $\bar{x}$ = 24.7, std = 3.5) in Experiment 2.

In Experiment 1, the overall scores of individual listeners were distributed bimodally, with three participants exhibiting drastically worse results (< 60% correct responses) compared to most other listeners, indicating that they did not actively participate in the experiment and were therefore discarded from the analysis. A histogram with overall accuracies of included and excluded participants is part of to the supplementary material (see Individual results). The same was true for two participants in Experiment 2 (< 60% correct responses). The results of 67 participants (age: $\bar{x}$ = 25.1, std = 3.2) in Experiment 1 and 67 participants (age: $\bar{x}$ = 24.7, std = 3.4) in Experiment 2 were analyzed. In both experiments, participants were randomly assigned to one of two groups that determined the order in which the target cue and mixture were presented: 33 participants (age: $\bar{x}$ = 24.8, std = 3.3) in Experiment 1 and 33 participants (age: $\bar{x}$ = 24.8, std = 3.3) in Experiment 2 were assigned to the order in which the target was presented before the mixture. For the reverse order, 34 (age: $\bar{x}$ = 25.3, std = 3) participants in Experiment 1 and 35 participants (age: $\bar{x}$ = 24.7, std = 3.8) in Experiment 2 were assigned. We acquired information on the participants' musical abilities using a subset of the Gold-MSI (Müllensiefen et al., 2014) consisting of nine questions on music perception abilities and seven questions on musical training.

Stimuli and Task

Stimuli were generated in MATLAB (MathWorks Inc., Natick, MA, USA) by extracting two-second excerpts of a single target instrument or vocals and a mixture of multiple instruments and vocals from a multitrack music database ("MedleyDB", https://medleydb.weebly.com/), see Figure 3.1A for a schematic. The database consisted of 127 royalty-free songs covering a wide range of popular music genres, with individual audio files for each instrument and vocals. The majority of the songs had English lyrics. Instruments and vocals were mixed so that the overall mix adhered to the conventions of popular music. We coarsely categorized the Instruments and vocals in the database as: Backing Vocals, Bass, Drums, Guitars, Lead Vocals, Piano, Percussion, Strings, Synthesizer, Winds. For each excerpt, a to-be attended instrument or vocal was chosen (target). Remaining instruments or voices in the excerpts that did not belong to the same category as the target functioned as maskers (mixture). Instruments or voices in the excerpt that belonged to the same category as the target were not included in the excerpt. In the case where the lead vocals were assigned as the target, all backing vocals were also excluded. Guitar, synthesizer, and winds were selected as instrument targets and the category lead vocals was selected as vocal targets. For guitar, synthesizer and wind targets that were adapted to the main melody, excerpts of the lead vocals were used as the basis.

To examine song excerpts for potential stimuli, we computed an instrument and vocal activity analysis for each song, indicating which instrument or vocals were likely audible in a given time frame. The activity analysis was created by calculating the sound level of each instrument and vocal in each song using a 500 ms sliding window. In each window, the root-mean-square value (RMS) of the sound level was calculated. For each instrument or vocal, the instrument or vocal was considered active in a time window, when the sound level in the window was above -20 dB relative to the maximum sound level of the entire song of the respective instrument or vocal. To further control the complexity of musical scenes in our stimuli, we removed all time windows from the activity map in which fewer than five and more than nine instruments or vocals were active. For each target category, we drew a 2000 ms excerpt with four adjacent, previously unused 500 ms time windows in which the target category and up to seven other vocals or instrument categories were considered active. Time slices were drawn from pseudo-randomly selected songs, with a preference to use the same song as infrequently as possible. In this way a total of 30 excerpts for each instrument target and 150 excerpts for vocal targets were drawn. The excerpts for the vocal target were then subdivided to be used either as vocal target, pitch-quantized vocal targets (autotune), or instrument targets playing the main melody. Furthermore, the excerpts contained sung English words, which could foster a potential facilitated processing of phonological features.

120 vocal tracks were pitch quantized using the pitch correction software *Melodyne* (Melodyne Version 5, Celemony Software). The corresponding manipulation of FMM is illustrated in Figure 3.1B by two exemplary excerpts and in the supplementary Figure 2. Quantization was set to both match pitch to a tempered scale and to eliminate all FMM, resulting in a robotic voice quality typical of the autotune effect. Thirty vocal tracks were modified in this way and were used as targets for the “autotune” category. The pitch of the remaining 90 quantized vocal tracks was used as a basis for the instrument main melody targets by having the melodies being played by three different MIDI-based instruments that corresponded to a guitar, synthesizer, or wind sound, thus creating 30 tracks for each of the three instruments. MIDI notes were programmed manually to accurately match the vocals in pitch, on- and offset times. For Experiment 2, the original frequency trajectory of the unquantized vocal tracks was reapplied to the autotuned vocals und instrument main melody targets by using the *Auto-Tune Pro* Plugin (Auto-Tune Pro, Antares).

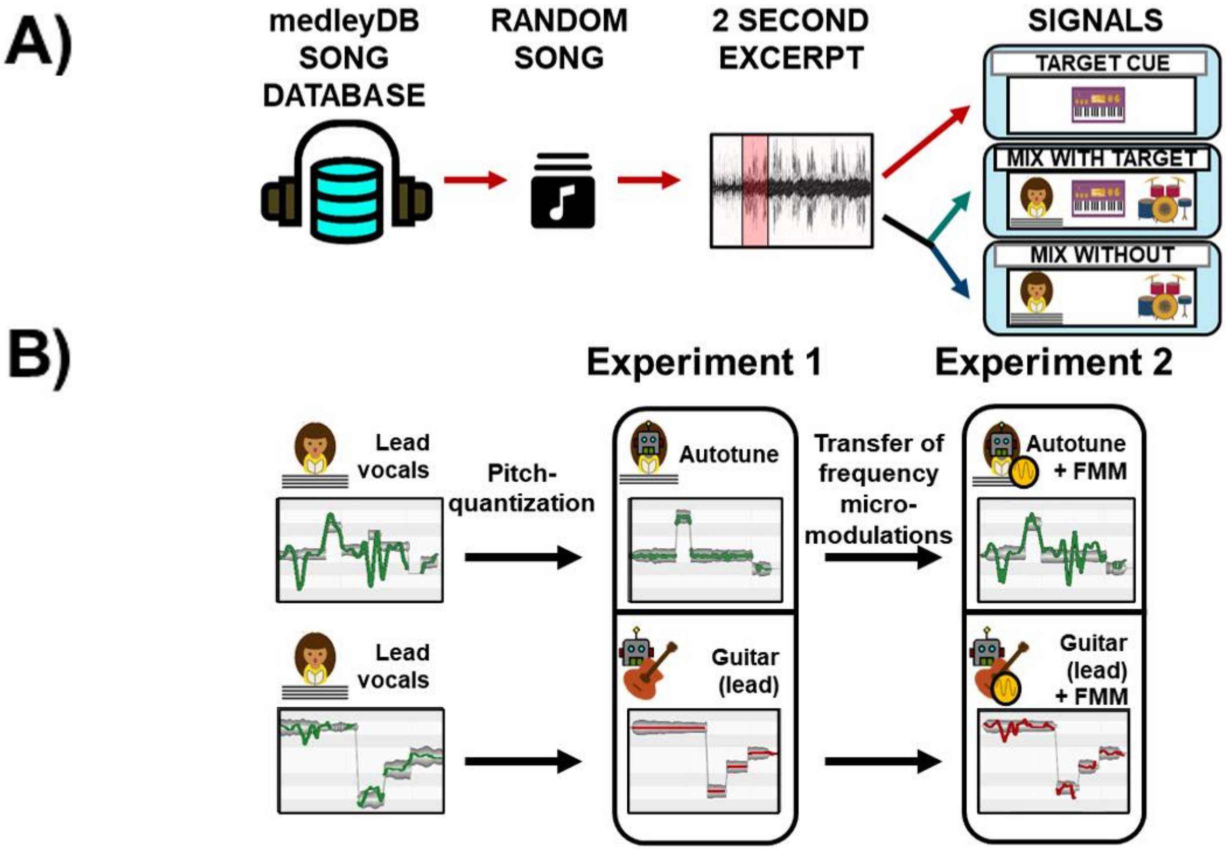

Figure 3.1: Schematic overview of the methods.

(A) Stimulus extraction: Short excerpts from the open source "medleyDB" multitrack databasewere used. Songs were drawn randomly without replacement. From each excerpt, two signals were extracted: One signal containing only the target signal, another signal either containing the mixture with or without the target signal. See the text for details. (B) Vocal manipulation: Lead vocal excerpts were pitch-quantized to create autotune or instrument main melody targets (lead) in Experiment 1. The original frequency trajectory of the unquantized vocal tracks was reapplied to the autotune und instrument main melody targets in Experiment 2. The gray waveform is representing the amplitude of the excerpt over time. Within the waveform, colored lines indicate the frequency trajectory. The light and dark gray shades indicate divisions of a chromatic scale in semitone steps.

Three monophonic signals were compiled from each 2-second excerpt: 1) a signal containing only the target, 2) a signal containing a mixture of five to eight instruments or vocals from non-target categories plus the target, 3) a signal containing a mixture of six to nine instruments or vocals without the target. For mixtures, the full number of instruments that were also present in the original excerpt of the song were used. A logarithmic fade-in and fade-out with a duration of 200 ms was applied to the beginning and end of all extracted signals. The sound level ratio between the target and the mixture was adjusted to -10 dB (cf., Bürgel, Picinali, & Siedenburg, 2021). For half of the trials, the mixture signals were arranged to contain the target signals; for the other half, the mixture did not contain the target signal. To prevent the presence of the three MIDI instruments from serving as a cue of the target, the lead vocals of the mix were replaced by one of the MIDI instruments using the same sound level as the vocals in one-third of the excerpts where an accompanying instrument was the target. Stimuli were created using the isolated target signal and a mixture signal in which the target was either present or absent. A 500 ms pause was inserted between the two signals, resulting in a total stimulus duration of 4500 ms. By interchanging the presentation order of target and mixture signal, two order conditions were created: In the "Target-Mixture" condition, the target signal was followed by a pause and the mixture signal; in the "Mixture-Target" condition, the presentation order was reversed. For use on the online platform, stimuli were converted from WAV format to MP3 at a bit rate of 320 kbit/s. Example stimuli and sound samples are provided on the website: https://uol.de/en/musik-wahrnehmung/sound-examples/akrs

Procedure

The experiments were approved by the ethics committee of the University of Oldenburg and conducted online via the web platform www.testable.org. Experiment 1 and Experiment 2 were identical in design, used the same song excerpts, and differed only in the absence (Experiment 1) or presence (Experiment 2) of FMM in the autotune and main melody instrument targets. Participants were automatically assigned to one of two groups, determining the presentation order of target cue and mixture. In the first group all stimuli appeared in the "Target-Mixture" presentation order, whereas for the second group the order was reversed to the "Mixture-Target" order. Each experiment used the same excerpts and was structured into five consecutive segments.

In the first segment, participants had to complete a headphone screening task based on Milne et al (2020). Here, a sequence of three white noise signals were presented, with one of the noise signals being phase shifted by 180 degrees in a narrow frequency band at around 600 Hz on the left headphone channel. When headphones were worn, the phase shift is perceived as a narrow tone embedded in the broadband noise. The task began with an instruction and a presentation of the noise signal, a 600 Hz tone in isolation and three mixtures of the tone in noise. Listeners had to detect the tone and passed the test if five out of six responses were correct. Participants who did not pass the headphone screening were returned to the instruction panel and reminded that they must pass the headphone screening before they were allowed to continue.

After the headphone screening, three song excerpts were presented to provide an impression of the dynamic range of the stimuli. During the presentation, participants were instructed to adjust the sound to a comfortable level. This was followed by a training phase, to familiarize participants with the detection task. Participants were presented with stimuli that were very similar but different from those used in the main experiment and were asked whether the target was present or absent in the mixture. Participants were allowed as much time as they needed to respond to the questions. To help participants understand the task, feedback was given after each answer. One stimulus with target and one without target in the mixture among the target categories lead vocals, autotune, guitar (accompaniment), synth (accompaniment) and winds (accompaniment) were presented. After the ten stimuli, participants had the option to repeat the training section or to continue with the main experiment.

During the main experiment, the same procedure as in the training was used, but no feedback was given. In this section, a total of 240 stimuli were presented in random order, corresponding to 30 stimuli for each of the eight target categories.

The final section of the experiment consisted of a questionnaire regarding personal data, questions from the Gold-MSI and a debriefing that presented the achieved average detection accuracy. On average participants took 41 minutes to complete the experiments.

Behavioral Analysis

Detection accuracy was determined directly from participants' responses. Following recommendations by the American Statistical Association (Wasserstein et al. 2019), we avoid assigning binary labels of "significance" to empirical results but instead provide confidence intervals of estimates where possible. Accuracies are always structured as a pair, with the first indicating the result of the target-mixture condition and the second indicating the result of the mixture-target condition. We provide mean detection accuracies followed by round brackets containing the decrease or increase through a change in presentation order.

Generalized binominal mixed-effect models (GLME; West et al., 2014) were used for statistical analyses. All mixed-effects analyses were computed in MATLAB using the *glme* function in the *Statistics and Machine Learning Toolbox* (Statistics and Machine Learning Toolbox Release 8.7, MathWorks Inc., Natick, MA, USA). Our model included random intercepts for each participant and item (i.e., stimulus). All binary categorical predictors were sum-coded. To summarize the main effects and interactions, results are presented in the form of an ANOVA table, with fixed effects coefficients provided as statistical parameter (F) and probability (p), derived from the GLME models via MATLAB's *anova* function. A detailed view of the behavioral results, models and statistic evaluations are presented in the supplementary material (see supplementary Tables 1-4).

Frequency micro-modulation analysis

To measure the difference in FMM between the original vocal and its pitch-quantized counterparts, we evaluated the range of FMM in short time windows for unmodified vocal excerpts and pitch-quantized vocals and instruments (see supplementary Figure 3). We used a sliding window of 10 ms over the duration of the excerpt and extracted f0 via the MATLAB function *pitch* (Audio Toolbox Release 3.7, MathWorks Inc., Natick, MA, USA). Given that the extraction contained artifacts such as irregular fluctuations, which occurred especially in the offsets and onsets of the vocals, additional artifact suppression was applied to the extracted f0s. The artifact rejection was based on a threshold for tonal components in the time window (harmonic ratio) as provided in the *pitch* function, excluding samples below a harmonic ratio of 75%. Additionally, a threshold for maximum f0 distance within a 100 ms sliding time window with 50% overlap was applied, excluding frequencies with a distance greater than one octave relative to the median pitch within the time window. For each excerpt and signal, the FMM range was obtained within a 100 ms sliding window by evaluating the difference in cents between the highest and lowest note. As a final step, the median across the windows was evaluated for each excerpt and signal. This excluded the relative rare time windows that contained tonal transitions. Results are presented in Figure 3.2 and the supplementary material (see supplementary Table 5).

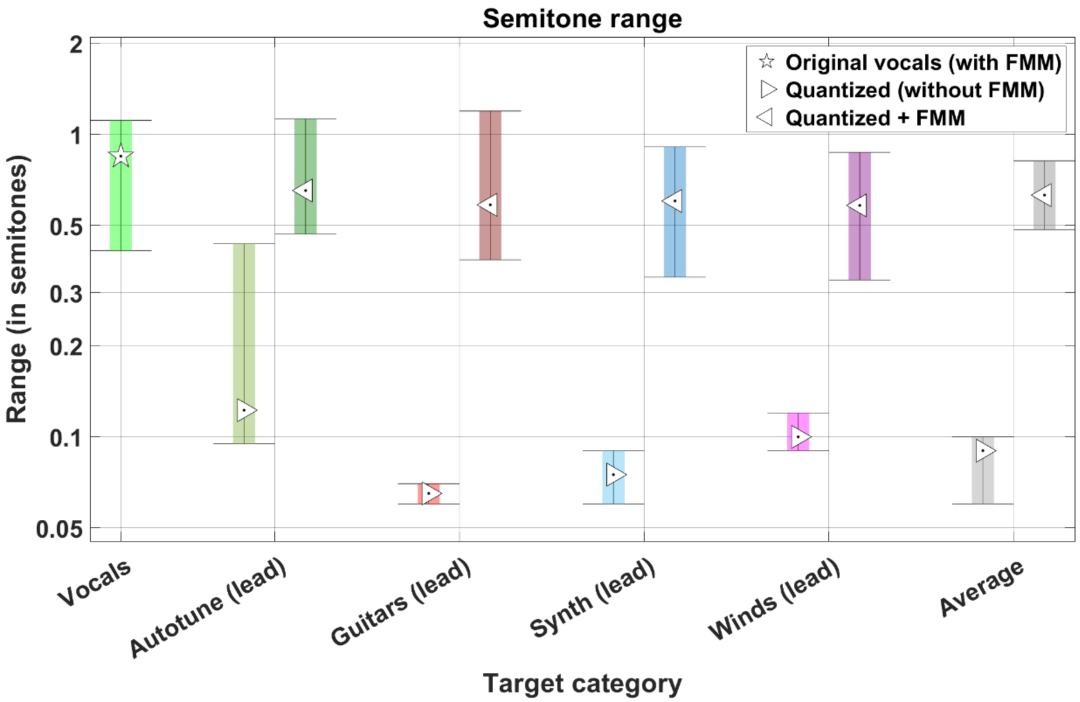

Figure 3.2: Frequency modulation analysis

To quantify the change in frequency micro-modulations between the original lead vocals, their pitch-quantized counterparts and their pitch-quantized counterparts with added frequency micro-modulation (FMM), the extracted f0 trajectories were transformed to cents and the f0 range in 100 ms time windows was evaluated. The median of the range was computed across all stimuli (30 excerpts) in each target category. "Quantized" refers to the FMM range in the autotune or melody instruments without FMM as used in Experiment 1. "Quantized + FMM" refers to the FMM range in the autotune or melody instruments with FMM as used in Experiment 2.

The pitch-quantized vocal alteration showed the smallest FMM range of 0.09 semitones, whereas the FMM range of 0.51 semitones for unquantized vocals and of 0.63 semitones for the quantized alteration with FMM were considerably higher. Autotune excerpts generated directly from pitch-quantized voices showed a higher range than the excerpts generated by MIDI instruments. An additional analysis of the distance analysis between estimated f0 to perfect tempered scale tone is included in the supplementary material.

3.2.4 Results

Experiment 1 – Pitch-quantized targets

Detection accuracies of Experiment 1 are displayed in Figure 3.3 (for numerical values, see supplementary Table 1). A GLME included presentation order and target categories as fixed effects (see supplementary Table 2). Accuracy varied by presentation order and target category: averaged across target categories, the Target-Mixture condition yielded a higher accuracy of 88% compared to the reverse Mixture-Target condition 80% (-8%). A decline of the accuracy between the two orders was present in almost every target category but differed in size. These effects were reflected by the GLME model, with pronounced effects for the presentation order (F = 9.78, p=0.002), the targets (F = 15.10, p < 0.001) and the interaction between the order and targets (F =5.93, p < 0.001). For readability, the following results are presented in pairs, with the first detection rate indicating the accuracy for the Target-Mixture order, and the subsequent detection rate indicating the accuracy for the Mixture-Target order. When examining the target categories, the best performing category was lead vocals with an accuracy of 99% and a minuscule decrease to 97% (-2%). The quantized voice had an accuracy of 96% but showed a decline to 87% (-9%). Targets in which the original lead vocals were replaced with an instrument showed the following accuracies: guitar from 90% to 80% (-10%), synths from 93% to 81% (-12%) and winds from 86% to 78% (-8%). Targets containing the instrumental excerpts taken from the original mixtures reached the following accuracies: guitar from 78% to 69% (-9%), the synths from 79% to 64% (-15%) and the winds from 88% to 78% (-10%).

Inspecting differences between instrument categories part of the accompaniment and those playing the main melody (guitar, synths, winds), the main melody instruments yielded clearly higher accuracies. However, the average accuracy of all main melody targets decreased considerably between presentation orders, from 89% to 79% (-10%). A similar decrease was observed for the accompanying categories with a decline from 82% to 71% (-11%). Differences between the two instrument types were analyzed using a GLME that included presentation order and instrument types as fixed effects (see supplementary Table 3). The model reflected the differences between accompaniment and main melody targets (F = 6.953, p = 0.009) and the influence of the presentation order (F = 151.96, p < 0.001). The presentation order affected each instrument in a similar way as indicated by the lack of an interaction effect between order and instrument type (F = 0.912, p = 0.340). The winds category behaved differently compared to the other instruments as it showed no benefit when playing the main melody, but rather a minor increase when playing in the accompaniment in the Target-Mixture order (+2%) and a decrease of the same quantity in the Mixture-Target order (-2%).

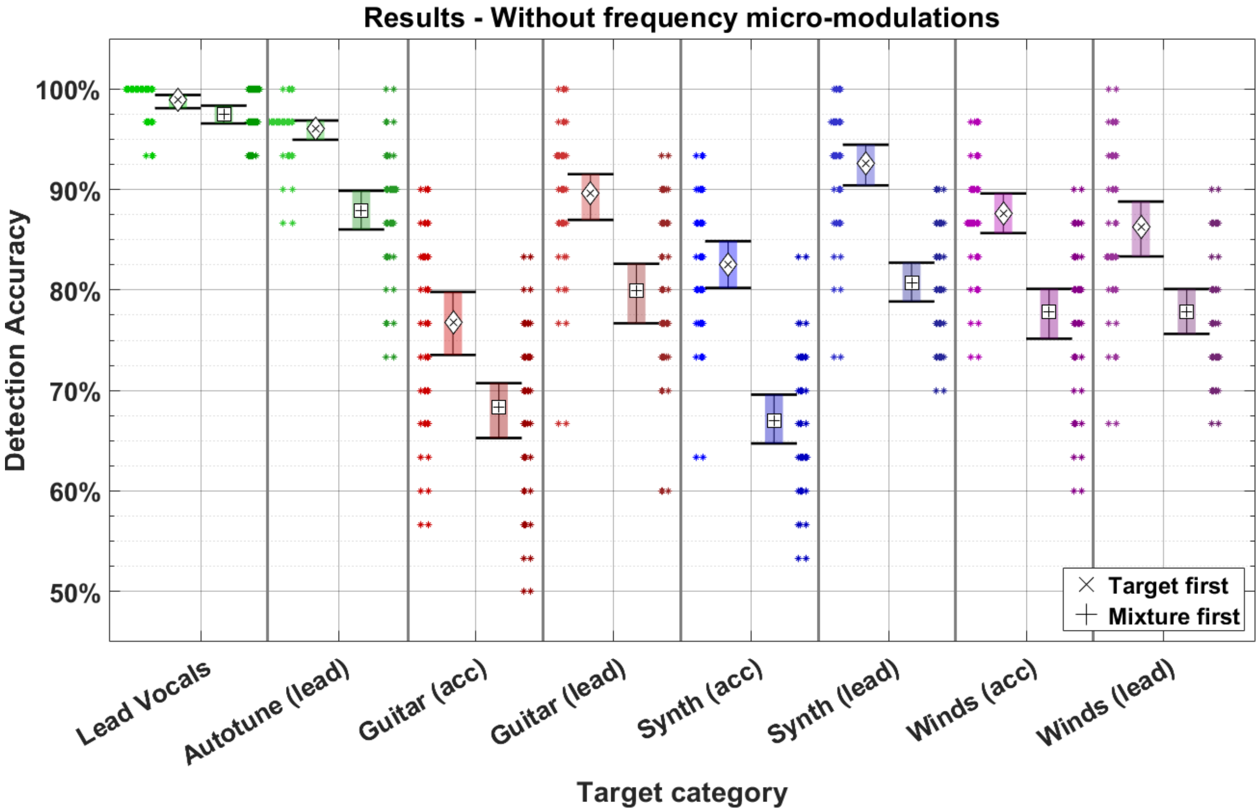

Figure 3.3: Detection accuracies for Experiment 1

Six instruments and two vocal categories were used as targets. Each instrument category was used twice either using the instrument track which was present in the excerpt (acc) or replacing the lead vocals in the excerpt by MIDI instruments using the same melody as the vocals (lead). The "x" marks the mean detection accuracy for a given target category in the presentation order "Target-Mixture". The "+" marks the mean detection accuracy for a given target category in the presentation order "Mixture-Target". Error bars indicate 95% confidence intervals. Asterisks left and right to the average of a category present average accuracies of individual participants for the given condition.

In summary, there was an effect of presentation order for all targets except the original vocals. Targets were detected considerably better when the isolated target was presented first, followed by the mixture. This was also evident when target instruments that otherwise played in the accompaniment replaced the vocals in the main melody. In contrast to the original vocals with FMM, the pitch-corrected vocals without FMM showed a clear effect of presentation order. This raised the question whether transferring FMM from vocals to instrumental signals could increase their salience. Thus, we repeated the experiment with a slight modification of the targets: we transferred the

FMM of the original vocals to the respective pitch quantized vocal and main melody instrument targets.

Experiment 2 – Targets with frequency micro-modulations

The average detection accuracies of the second experiment are displayed in Figure 3.4 (for numerical values, see supplementary Table 1). A GLME included presentation order and target categories as fixed effects (see supplementary Table 2). Accuracy differed depending on the target category and order of presentation which was also evident in our model (Order: F = 0.414, p = 0.03, Target: F = 11.054, p < 0.001, Interaction: F = 3.486, p = 0.001). Similar to Experiment 1, when inspecting the difference of presentation orders by averaging over target categories, the Target-Mixture condition held a higher accuracy of 90% than the Mixture-Target condition with an accuracy of 82% (-9%). When looking into the target categories, targets maintaining the original frequency trajectory of the vocals (lead vocals, autotune and main melody instruments) revealed a clearly smaller decrease between both presentation orders than the accompanying instrument categories. This result was most pronounced in the lead vocals, which performed best with an accuracy of 98% and a decrease to 95% (-3%).

Inspecting the differences between the instrument categories playing an accompanying role and those replacing the lead vocals, the main melody instruments yielded higher accuracies. Average accuracies of all main melody targets decreased across presentation orders from 91% to 87% (-4%). A larger decrease was shown for the targets part of the accompaniment with a decline from 83% to 70% (-13%). Differences between the two instrument types were analyzed using a GLME that included presentation order and musical material (accompaniment vs. main melody) as fixed effects (see supplementary Table 3). Our model reflected the differences between accompaniment and main melody targets (F = 6.953, p = 0.003), the influence of the presentation order (F = 93.70, p < 0.001), and in contrast to Experiment 1, that the presentation order affected the accompaniment and instrument targets differently by revealing an effect of the interaction between order and instrument type (F = 47.166, p < 0.001).

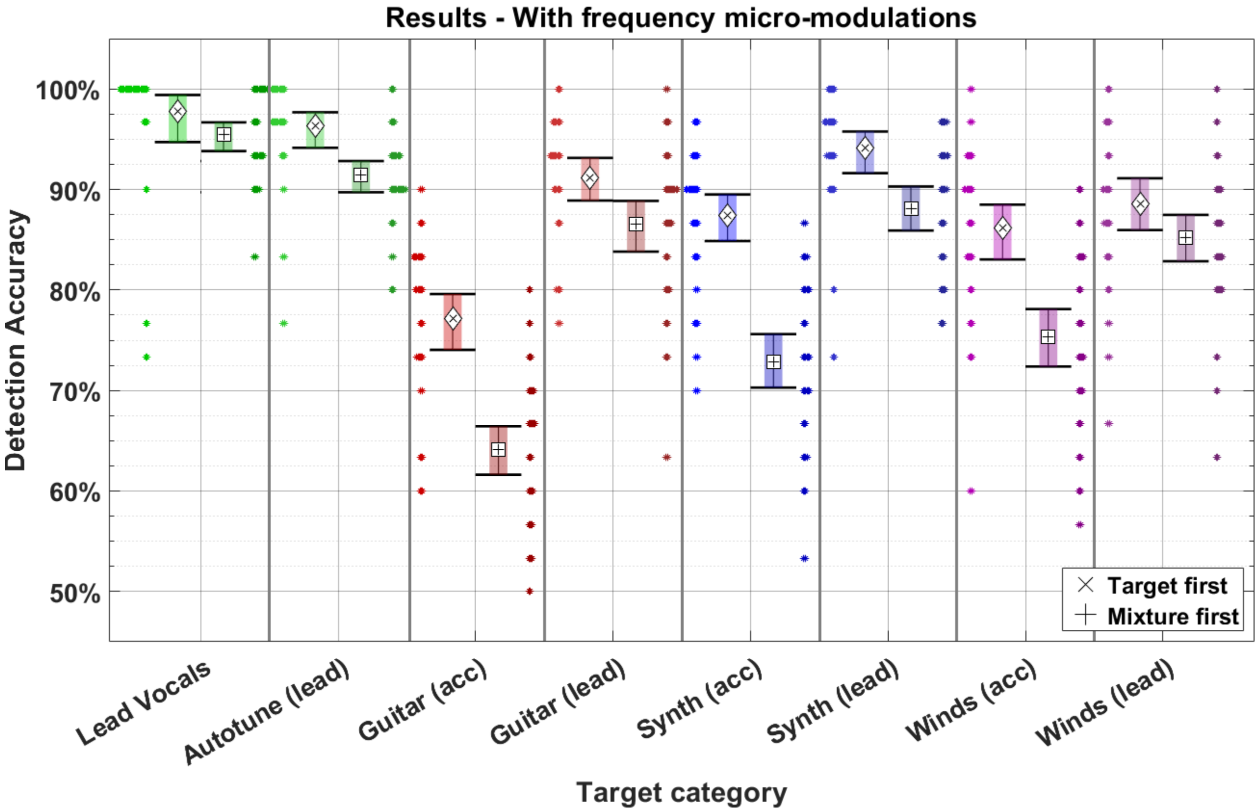

Figure 3.4: Detection accuracies for Experiment 2

Six instrument and two vocal categories were used as targets. Each instrument category was used twice either using the instrument track which was present in the excerpt (acc) or replacing the lead vocals in the excerpt by MIDI instruments using the same melody and frequency trajectory as the vocals (lead). Graphical conventions otherwise identical to Fig. 3.3.

Musical experience

Musical experience was analyzed in a questionnaire using a subset of the Gold-MSI. Nine questions regarding perceptual abilities and seven questions regarding musical training were included in the questionnaire. Scores between 1 and 7 could be obtained for each question. For Experiment 1, participants reached a score of 43.4 in the perceptual abilities subscale and a score of 22.4 in the musical training subscale. The correlation based on perception abilities for the Target-Mixture order was $R^2=0.001$ ($p = 0.89$) and for the Mixture-Target at $R^2=0.05$ ($p = 0.22$). Similar results were shown for the set regarding musical training, with a correlation for the Target-Mixture order of $R^2=0.008$ ($p = 0.23$) and for the Mixture-Target order $R^2=0.054$ ($p = 0.2$). Regarding Experiment 2, participants reached an average score of 43.1 in the perceptual abilities' subscale and an average score of 19.4 in the musical training subscale. As in Experiment 1, no notable correlations were found between the individual musical experience scores and detection accuracies. The correlation based on perception abilities for the Target-Mixture order was R^2=0.003 (p = 0.80) and for the Mixture-Target at R^2=0.07 (p = 0.28). Similar results were shown for the set regarding musical training, with a correlation for the Target-Mixture order of R^2=0.025 (p = 0.23) and for the Mixture-Target order R^2=0.112 (p = 0.10). Because we did not specifically recruit separate groups of participants with diverse degrees of musical experience, the lack of an effect of musical experience observed here was not surprising and consistent with previous research (Bürgel et al., 2021).

Comparison of both Experiments

The stimuli between Experiment 1 and Experiment 2 differed only in the exclusion of FMM (Experiment 1) and the inclusion of FMM (Experiment 2) for the autotune vocals and target instruments playing the main melody. The average detection accuracy across all instruments between the two presentation orders revealed a slightly better performance in Experiment 2 with a miniscule difference of two percentage points between experiments in both presentation orders. Stimuli that remained consistent across experiments showed differences in accuracy from zero to four percentage points. Yet overall performance was similar, with an average difference between the vocals and accompanying instruments of less than one percentage point. A direct comparison of detection accuracies in both experiments for the autotune and main melody instruments is shown in Figure 3.5A. There were negligible differences in the Target-Mixture condition by about one percentage point. However, in the Mixture-Target condition, the autotune and melody instruments in Experiment 2 showed an enhanced detection of six percentage points compared to Experiment 1. To statistically evaluate the differences between both experiments, a GLME was utilized that included presentation order, musical role, and the different experiments as fixed effects. The model corroborated the influence of FMM (see supplementary Table 4) by indicating no interaction between presentation order and musical role when averaged across both experiments (F = 0.624, p = 0.430), but a three-way interaction between presentation order, musical role, and experiment (F = 11.227, p < 0.001). This underlines that the presence of FMM in Experiment 2 boosted performance in the otherwise difficult Mixture-Target condition of the main melody targets (see Fig. 5A). In addition, a strong correlation of R^2 = 0.9 was found between the FMM range and the order effect expressed as difference in detection accuracy of both presentation orders (see Fig 5B). Taken together, this further suggests that FMM enriches the vocals by an important factor for creating auditory salience in musical scenes.

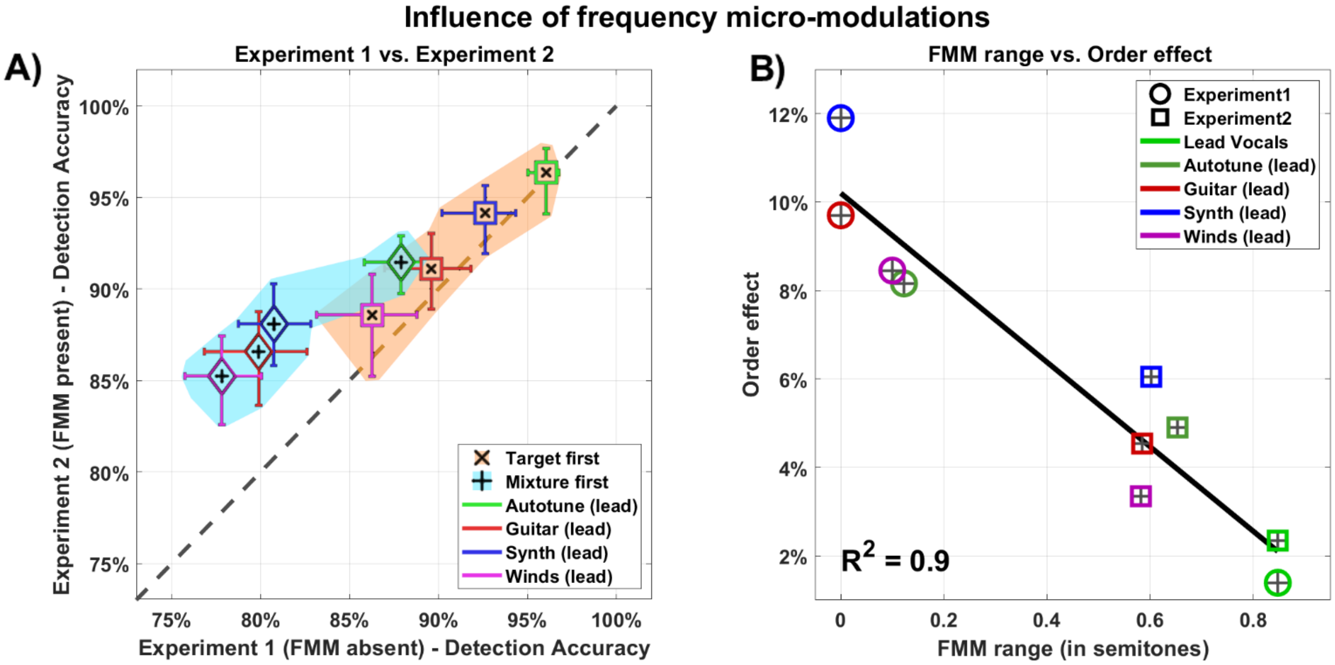

Figure 3.5: Influence of frequency micro-modulations

(A) Detection accuracy in selected conditions from Experiment 1 (x-axis) and Experiment 2 (y-axis): Two-dimensional error bars indicate 95% confidence intervals. Note that for the presented target categories, average accuracies in the “Mixture-first” conditions were significantly higher in Experiment 2 (with FMM) compared to Experiment 1 (without FMM), whereas this was not the case for “Target-first” conditions. (B) Correlation of frequency micro-modulation range and order effect: The FMM range is represented by the median range of each lead-melody target from Experiment 1 and Experiment 2. The order effect is quantified for each lead-melody target as the difference between the average detection accuracy of the “Target-first” and “Mixture-first” conditions in the respective experiment and condition.

3.2.5 Discussion

In the present study, we analyzed the acoustical and musical underpinnings of the lead vocals, which contribute to their role as an elevated point of auditory attention in musical mixtures (vocal salience). We investigated the influence of frequency micro-modulation (FMM) of the lead vocals and the role of the main melody in hearing out individual instruments from a mix. Specifically, participants were asked to detect cued vocals and instruments in 2-second excerpts of western popular music. To investigate the influence of attentional cues on the detection of the target, the presentation order of cue and mixture was swapped between participants, whereby the comparison between both orders revealed to which degree detection was modulated by attention (order effect). To analyze the role of the main melody for contribution to the vocal salience, instrument targets were either used in their role as part of the musical accompaniment or they were used as a replacement for the lead vocals, that is, they played the melody of the vocals. We added a vocal target category with pitch-quantized lead vocals, eliminating FMM inherent in the vocals (Experiment 1). Additionally, we repeated a modified version of the experiment in which we transferred the FMM of the lead vocals to the pitch-quantized vocals and the instruments replacing the vocals (Experiment 2).

Order Effect and Vocal Salience

Consistent with classic studies (e.g., Bey & McAdams 2002), the presentation order of the cue played a key role in our results. When the cue preceded the mixture, listeners were able use this information to direct selective attention towards the cued signal. This resulted in higher detection rates compared to when the cue was presented subsequent to the mixture. Consistent with our previous experiments (Bürgel, Picinali, & Siedenburg, 2021) and our hypothesis, this effect was evident in all target categories except the lead vocals, which showed only a slight decrease of accuracy when the cue was presented after the mixture. This finding highlights a unique vocal salience that enables the vocals to attract the listeners attention, even when listening blindly into a musical scene. The present study used a different database of music excerpts compared to our previous work (Bürgel, Picinali, & Siedenburg, 2021). The consistency of our findings across different music databases supports our general hypothesis that vocal salience in mixtures of popular music is not the result of a specific mixing strategy in music production, but rather an effect inherent in vocal signals. Previous studies have established a perceptually privileged role of the voice through the presentation of isolated voices and instruments (e.g., Levy, Granot, Bentin, 2001; Gunji et al., 2003, Agus et al., 2012). Our present results extend this line of research by demonstrating that this effect is also present in musical mixtures.

Effect of main melody

When the guitar and synthesizer replaced the vocals as the main melody of a song, overall detection accuracy improved, supporting our hypothesis and previous studies (Ragert, Fairhurst, Keller, 2014) of a stronger perceptual salience of melody instruments over accompaniment instruments. Surprisingly, wind instruments showed no improvement whatsoever. One likely reason for this contrasting effect relates to the specific musical role of the different categories of instruments as part of the accompaniment. Whereas the guitar and synthesizer mostly played chord-based progressions in our excerpts, the winds played accompanying melodies. Consequently, the transition to the melody of the lead vocal might be a rather small change for the wind instruments, but a more drastic change of musical material for guitars and synthesizers. Nonetheless, it is important to note that the differences between the two presentation orders were still present and almost unaffected for instruments playing the vocal melodies. This implies that instruments playing the main melody are generally easier to detect, but playing the main melody does not automatically guarantee salience in a musical mixture (i.e., does not automatically attract auditory attention without a cue signal). It should be kept in mind that we used a consistent MIDI instrument for each of the individual instrument categories, which potentially may have added detection cues, although such cues would have been identical for both presentation orders. Whereas the timbre of accompanying instruments could vary between excerpts (because we used the original instruments within a song), the timbre of the three instruments playing the vocal melody did not vary. Even though we attempted to balance this aspect of experimental design by interspersing excerpts in which the vocals were replaced by instruments while the target was an accompaniment instrument, we cannot rule out that participants became accustomed to the timbre of the MIDI instruments over the duration of the experiment and implicitly memorized specific timbral properties of the MIDI instruments (Agus, Thorpe, Pressnitzer, 2010; Siedenburg & Müllensiefen, 2019; Siedenburg & McAdams, 2018).

Effect of frequency micro-modulations

The pitch-quantized vocal category showed degradation in the Mixture-Target order, while also performing somewhat worse compared to the lead vocals in the Target-Mixture order. This suggests that excessively pitch-corrected voices do not capture listeners’ attention to the same extent as more naturalistic singing voices and therefore are more likely to fuse with elements of the accompaniment in musical mixtures. This pattern of results further refutes the assumption that phonological cues are the basis of vocal salience, because pitch quantization did not affect the phonological content of the vocals. One reason for the loss of attentional cues in the quantized vocals appears to be the lack of FMM, which was reduced compared to the original vocals. An acoustical analysis corroborated this interpretation by revealing a greater range of FMM for the unquantized vocals and instruments compared to their quantized counterparts that strongly correlated with the strength of the order effect. This finding is consistent with previous studies that have shown specific facilitated processing of speech with naturalistic frequency modulations, which is more intelligible compared to speech without, with decreased or exaggerated modulations (e.g., Wingfield et al., 1984, Miller, Schlauch, Watson, 2010). Furthermore, FMM has been shown to facilitate the detection of concurrently presented vowel sounds (McAdams, 1989, Marin & McAdams 1990). Our findings extend the literature in this regard by demonstrating that the salience of vocals in musical mixtures strongly relies on frequency modulations that are present in naturalistic singing voices, helping the vocals to stand out from the mixture and attract listeners' attention.

The influence of FMM was further corroborated in our second experiment. We repeated the experiment using the same excerpts while adding the FMM of the original vocal excerpts to the instruments substituting the main melodies and quantized vocals. For the signals with artificially added FMM, our results showed a considerably reduced difference between the presentation orders in comparison to the first experiment. Interestingly, when the cue was presented before the mixture, the targets achieved very similar results across both experiments. This contradicted our hypothesis because the additional FMM did not increase overall detection but only seemed to increase the detection in the Mixture-Target order. Thus, the modulations appeared to increase the salience of the target when no prior cue was provided, drawing the attention towards the target in a similar way as seen in the lead vocals.

Curiously, even the pitch-quantized vocals with micro-modulations showed small differences between the orders of presentation, although the differences to the original vocals were supposed to be eliminated by the transfer of FMM. This result implies that although the micro-modulations make a strong contribution to vocal salience, it seems that the full salience effect may emerge from the conjunction of multiple features of the vocals. One of the features might be the pitch offset of the unaltered vocals that was eliminated by quantizing the pitch to a tempered scale. These intonation deviations occur even in professional singers (Sundberg, Prame & Iwarsson, Sundberg et al., 1996, pp. 291–306; Mori et al., 2004; Hutchins & Campbell, 2009) and are an inevitable consequence of imperfect motor controls of the voice (Hutchins, Larrouy-Maestri & Peretz, 2014.). Even though these deviations were unlikely to be perceived as intonation errors (Hutchins, Roquet & Peretz, 2012), it is possible that these deviations yield auditory grouping cues that let the vocals stand out of the mixture. Furthermore, singers intentionally create such deviations to add expressivity to the sound (Sundberg, Lã, Himonides, 2013) and therefore may add an important cue to the unaltered vocals, that is lost in pitch-quantization.

More speculatively, the pitch quantization and re-introduction of pitch variation may have also altered timbral features of the vocals. Timbre is a multidimensional attribute (Siedenburg, Saitis, & McAdams, 2019) that enables the discrimination and identification of sound sources (e.g., sounds from a keyboard vs. a guitar), even though they may match in other acoustic cues such as loudness and pitch. Previous studies focusing on the recognition of instruments and voices showed that the human singing voice has an advantage over other instruments supposedly based on timbre alone (Agus et al., 2012; Suied et al., 2014; Isnard et al., 2019). Voice specific cortical areas remain selective to timbre of naturalistic vocal sounds even when vocal and non-vocal sounds where matched in acoustic cues (Bélizaire et al., 2007). Further, the facilitated recognition and cortical selective was observed only for natural vocals and was absent when "chimeras", i.e., interpolations between instruments and vocals were presented (Agus et al., 2012; Agus et. al., 2017). Even though we think that in the present experiments timbre changes were subtle, if noticeable at all, this interpretation would suggest that vocal salience could be a result of the joint contribution of timbre and pitch cues in auditory scene analysis. A distortion of such joint features due to the autotuning and f0-modulation could have hindered voice-specific processing to occur, thus hindering the full salience effect to arise for our modified vocals.

In summary, in line with previous experiments, the detectability of all non-vocal instruments was affected by a change in the presentation order, whereas lead vocals were detected with similarly high accuracies in both presentation orders. This effect corroborates a unique vocal salience that automatically attracts listeners' attention. Instruments replacing vocals showed better detection accuracies compared to instruments playing as part of the musical accompaniment, but still exhibited reduced accuracy when the mixture preceded the target. Even for pitch-quantized vocals, this dependency on presentation order was evident, implying that phonological features that engage a facilitated processing of speech sounds are not sufficient to drive vocal salience. The difference between the presentation orders decreased considerably when the FMM originally present in the vocals were transferred to the instruments and pitch-quantized vocals. Overall, this also implies that excessive pitch correction may strip vocals of a unique acoustical feature that helps turning the human voice into a focal point of musical scenes.

3.2.6 References

Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. Scientific Reports, 7(1), 11526. https://doi.org/10.1038/s41598-017-11684-1

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. The Journal of the Acoustical Society of America, 131(5), 4124–4133. https://doi.org/10.1121/1.3701865

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. Neuron, 66(4), 610–618. https://doi.org/10.1016/j.neuron.2010.04.014

Belin, P, Zatorre, R. J, Lafaille, P., Ahad, P, & Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature, 403(6767), 309–312. https://doi.org/10.1038/35002078

Belin, P, Zatorre, R. J , & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. Cognitive Brain Research, 13(1), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

Bélizaire, G., Fillion-Bilodeau, S., Chartrand, J.-P., Bertrand-Gauvin, C., & Belin, P (2007). Cerebral response to 'voiceness': A functional magnetic resonance imaging study. Neuroreport, 18(1), 29–33. https://doi.org/10.1097/WNR.0b013e3280122718

Bey, C., & McAdams, S. (2002). Schema-based processing in auditory scene analysis. Perception & Psychophysics, 64(5), 844–854. https://doi.org/10.3758/bf03194750.

Bregman, A. S., & McAdams, S. (1994). Auditory Scene Analysis: The Perceptual Organization of Sound. The Journal of the Acoustical Society of America, 95(2), 1177–1178. https://doi.org/10.1121/1.408434

Bürgel, M., Picinali, L., & Siedenburg, K. (2021). Listening in the Mix: Lead Vocals Robustly Attract Auditory Attention in Popular Music. Frontiers in Psychology, 12, 769663. https://doi.org/10.3389/fpsyg.2021.769663

Gao, Z., & Oxenham, A. J. (2022). Voice disadvantage effects in absolute and relative pitch judgments. The Journal of the Acoustical Society of America, 151(4), 2414. https://doi.org/10.1121/10.0010123

Gunji, A., Koyama, S., Ishii, R., Levy, D., Okamoto, H., Kakigi, R., & Pantev, C. (2003). Magnetoencephalographic study of the cortical activity elicited by human voice. Neuroscience Letters, 348(1), 13–16. https://doi.org/10.1016/s0304-3940(03)00640-2

Hutchins, S., & Campbell, D. (2009). Estimating the time to reach a target frequency in singing. Annals of the New York Academy of Sciences, 1169, 116–120. https://doi.org/10.1111/j.1749-6632.2009.04856.x

Hutchins, S., Roquet, C., & Peretz, I. (2012). The Vocal Generosity Effect: How Bad Can Your Singing Be? Music Perception, 30(2), 147–159. https://doi.org/10.1525/mp.2012.30.2.147

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. Attention, Perception & Psychophysics, 76(8), 2522–2530. https://doi.org/10.3758/s13414-014-0732-1

Isnard, V., Chastres, V., Viaud-Delmon, I., & Suied, C. (2019). The time course of auditory recognition measured with rapid sequences of short natural sounds. Scientific Reports, 9(1), 8005. https://doi.org/10.1038/s41598-019-43126-5

Larrouy-Maestri, P., & Pfordresher, P. Q. (2018). Pitch perception in music: Do scoops matter? Journal of Experimental Psychology. Human Perception and Performance, 44(10), 1523–1541. https://doi.org/10.1037/xhp0000550

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. Neuroreport, 12(12), 2653–2657. https://doi.org/10.1097/00001756-200108280-00013

Marin, C. M., & McAdams, S. (1991). Segregation of concurrent sounds. Ii: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. The Journal of the Acoustical Society of America, 89(1), 341–351. https://doi.org/10.1121/1.400469

McAdams, S. (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. The Journal of the Acoustical Society of America, 86(6), 2148–2159. https://doi.org/10.1121/1.398475

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. The Journal of the Acoustical Society of America, 128(1), 435–443. https://doi.org/10.1121/1.3397384

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. Behavior Research Methods, 53(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Mori, H., Odagiri W., Hideki., Honda, K. (2004). Transitional Characteristics of Fundamental Frequency in Singing. n Internal Congress on Acoustics (ICA), pages 499–500

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. PloS One, 9(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. Current Biology : CB, 32(7), 1470-1484.e12. https://doi.org/10.1016/j.cub.2022.01.069

Ragert, M., Fairhurst, M. T., & Keller, P. E. (2014). Segregation and integration of auditory streams when listening to multi-part music. PloS One, 9(1), e84085. https://doi.org/10.1371/journal.pone.0084085

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. Speech Communication, 46(3-4), 405–417. https://doi.org/10.1016/j.specom.2005.01.010

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. Trends in Neurosciences, 34(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Signoret, C., Gaudrain, E., Tillmann, B., Grimault, N., & Perrin, F. (2011). Facilitated auditory detection for speech sounds. Frontiers in Psychology, 2, 176. https://doi.org/10.3389/fpsyg.2011.00176

Siedenburg, K., & McAdams, S. (2018). Short-term Recognition of Timbre Sequences. Music Perception, 36(1), 24–39. https://doi.org/10.1525/mp.2018.36.1.24

Siedenburg, K., Saitis, C., & McAdams, S. (2019). The Present, Past, and Future of Timbre Research. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), Springer Handbook of Auditory Research. Timbre: Acoustics, Perception, and Cognition (Vol. 69, pp. 1–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_1

Siedenburg, K., & Müllensiefen, D. (2019). Memory for Timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), Springer Handbook of Auditory Research. Timbre: Acoustics, Perception, and Cognition (Vol. 69, pp. 87–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_4

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., & Pressnitzer, D. (2014). Auditory gist: Recognition of very short sounds from timbre cues. The Journal of the Acoustical Society of America, 135(3), 1380–1391. https://doi.org/10.1121/1.4863659

Sundberg, J., Prame, E., & Iwarsson, J. (1996). Replicability and Accuracy of Pitch Patterns in Professional Singers. In P. J. Davis, & N. H. Fletcher (Eds.), Vocal Fold Physiology : Controlling Complexity and Chaos (pp. 291-306). Singular Publishing Group, Inc.. Vocal fold physiology series

Sundberg, J., Lã, F. M. B., & Himonides, E. (2013). Intonation and expressivity: A single case study of classical western singing. Journal of Voice : Official Journal of the Voice Foundation, 27(3), 391.e1-8. https://doi.org/10.1016/j.jvoice.2012.11.009

Sussman, E. S. (2017). Auditory Scene Analysis: An Attention Perspective. Journal of Speech, Language, and Hearing Research: JSLHR, 60(10), 2989–3000. https://doi.org/10.1044/2017_JSLHR-H-17-0041

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond “ p < 0.05”. The American Statistician, 73(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way she sings: Enhanced memory for vocal melodies. Psychological Science, 23(10), 1074–1078. https://doi.org/10.1177/0956797612442552

West, Brady T.; Welch, Kathleen B.; Galecki, Andrzej T. (2014). Linear Mixed Models. CRC Press.

Wingfield, A., Lombardi, L., & Sokol, S. (1984). Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation. Journal of Speech and Hearing Research, 27(1), 128–134. https://doi.org/10.1044/jshr.2701.128

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

MB and KS designed the study. MB collected and analyzed the data. MB wrote a first draft of the manuscript. KS revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by a Freigeist Fellowship of the Volkswagen Foundation to K.S.

Acknowledgments

We thank research assistant Ghifar Aldebs for helping create the stimuli for our experiments.

3.3 Synopsis

The results underscore the effects observed in Study 1, highlighting the influence of top-down processing on the perception of auditory scenes, the varying salience of different target categories, and a unique vocal salience. While neither the main melody nor phonetic cues were found to drive vocal salience, frequency micro-modulations stood out as a promising candidate, as a correlation between FMM intensity and salience was observed. This finding raises questions about whether FMM also contribute to other reported perceptual benefits of vocals.

4. IMPACT OF INTERFERENCE ON VOCAL AND INSTRUMENT RECOGNITION

4.1 Introduction

Motivated by the effect of frequency micro-modulations (FMM) and vocal salience in the detection of voices in musical mixtures, this study investigated whether similar patterns can be found in other areas of perception. The third study focused on the recognition of short sounds, where it has been demonstrated that vocals are recognized faster and more accurately than other instrument sounds. The aim was to investigate the influence of FMM on these recognition advantages and to determine whether the recognition of sounds accompanied by interfering sounds demonstrates parallels to salience in musical mixtures, that would manifest with vocals showing a distinct robustness to interference.

4.2 Study 3

This chapter is currently in press at *The Journal of the Acoustical Society of America* as: Bürgel. M. & Siedenburg, K. (2024) Impact of Interference on Vocal and Instrument Recognition. The content of this chapter is identical to the manuscript.

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, carried out the experiments, analyzed the data and wrote the final paper. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

____________________________________ 20.07.2024

(name) Date

Supervisor

4.2.1 Abstract

Voices arguably occupy a superior role in auditory processing. Specifically, studies have reported that singing voices are processed faster and more accurately and possess greater salience in musical scenes compared to instrumental sounds. However, the underlying acoustic features of this superiority and the generality of these effects remain unclear. This study investigates the impact of frequency micro-modulations (FMM) and the influence of interfering sounds on sound recognition. Thirty young participants, half with musical training, engage in three sound recognition experiments featuring short vocal and instrumental sounds in a go/no-go task. Accuracy and reaction times are measured for sounds from recorded samples and excerpts of popular music. Each sound is presented in separate versions with and without FMM, in isolation or accompanied by a piano. Recognition varies across sound categories, but no general vocal superiority emerges and no effects of FMM. When presented together with interfering sounds, all sounds exhibit degradation in recognition. However, whereas /a/ sounds stand out by showing a distinct robustness to interference (i.e., less degradation of recognition), /u/ sounds lack this robustness. Acoustical analysis implies that recognition differences can be explained by spectral similarities. Together, these results challenge the notion of general vocal superiority in auditory perception.

4.2.2 Introduction

The human auditory system possesses remarkable abilities to detect and distinguish sounds, even in complex acoustic scenes where various sounds occur simultaneously. This complexity is exemplified in musical scenes featuring multiple instruments and voices. Despite the simultaneity of sounds, the auditory system excels at identifying and selectively focusing on individual instruments and vocals within a musical scene. This ability is achieved through the process of auditory scene analysis (ASA, Bregman, 1990), wherein sounds are separated and organized into mental representations of distinct auditory streams. Acoustic features of sounds play a crucial role in this process, providing cues to organize the auditory input into meaningful components. In the context of music, these cues encompass loudness, pitch, and timbre. Timbre, often simply described as “texture” or “tone color” (Helmholtz, 1877), is a multidimensional feature (Siedenburg et al., 2019) that enables the discrimination of sound sources (e.g., sounds from a singing voice vs. a cello), even when other acoustic features match.

Neurophysiological experiments have demonstrated enhanced cortical “voice-specific” responses when isolated vocal sounds are presented alongside non-vocal environmental sounds (Belin et al., 2000; Belin et al., 2002), as well as other musical instrument sounds (Levy, et al., 2001; Gunji et al., 2003). Moreover, specific neural populations have been identified that respond selectively to music featuring singing voices but not to instrumental music mixtures (Norman-Haignere et al., 2022). This facilitated processing of vocal sounds extends to multi-instrumental musical scenes, where vocal sounds exhibit a unique salience that attracts listeners' attention unlike other sounds (Bürgel et al., 2021). In a comparative analysis of vocal and instrumental melodies, vocal melodies were shown to be more accurately recognized than instrumental melodies (Weiss et al., 2012) even when the melodies are sung without lyrics (Weiss et al., 2021). We here refer to the faster and more precise recognition of vocal sounds as ‘vocal superiority’.

Agus and colleagues investigated the ability to recognize instrumental and vocal sounds in a multi-experiment study (Agus et al., 2012). The study used sounds excerpts with a duration of 250 ms extracted from a database of isolated musical sounds. Sounds were controlled in level and pitch with the aim of isolating timbre as the distinctive factor.

Participants were tasked with recognizing sounds of a target timbre in a sequence of diverse sounds, responding actively when detecting a target. As recognition was anticipated to be highly accurate, response times were measured to provide insights even in circumstances of perfect or indifferent recognition accuracy. Participants showed near-perfect accuracy and fast reaction times for all targets. Nevertheless, vocal superiority emerged via consistently faster reaction times and higher accuracy compared to instrumental sounds. Subsequent experiments further underscored recognition advantages for vocal sounds, revealing that vocals are recognizable from shorter sound snippets compared to other musical instruments (Suied et al., 2014; Isnard et al., 2019).

The specific acoustical features responsible for triggering vocal superiority are still unknown. Agus et al. (2012) argued that the full spectro-temporal envelope of sounds must be involved in vocal recognition, that is, neither solely spectral nor solely temporal features suffice for robust vocal recognition. This argument appears to be inconsistent with other studies on recognition of vocal sounds from short sound snippets below 8 ms duration (Suied et al., 2014), so short that reliable temporal cues are likely to be eradicated. Spectral envelope cues of vocal sounds such as formants have also been shown to be highly informative of instrument identity and the natural basis of vowel recognition (Reuter et al., 2018). Using automatic instrument classification on a large set of sound samples, Siedenburg et al. (2021) observed that spectral envelope features alone sufficed to accurately discriminate vocal sounds from other harmonic musical instrument sounds. Thus, whether spectral envelope features alone contribute substantially to perception tasks based on fast recognition of vocal and musical instrument sounds remains to be determined.

Our previous experiments on auditory attention in musical scenes highlighted yet another candidate feature: frequency micro-modulations (FMM) present in singing voices (Bürgel & Siedenburg 2023). In the context of this study, FMM refers to non-stationary frequency changes in pitched sounds, usually smaller than one semitone (Larrouy-Maestri & Pfordresher 2018). In singing, FMM arises from imperfect control of intonation caused by vocal-motor control adjustments (Hutchins et al., 2014), that persist even in highly trained singers (e.g., Sundberg et al., 1996; Hutchins & Campbell, 2009), but can also be utilized intentionally as a form of expressive intonation (Sundberg, et al., 2013). Although pitch perception for vocals appears to be less precise than for musical instruments (“vocal generosity effect” - Hutchins et al., 2012; Sundberg et al., 2013; Gao & Oxenham, 2022), the expressivity of FMM still provides perceptible additional musical information (Larrouy-Maestri & Pfordresher, 2018) and plays a role in enhancing the prominence of vowel sounds (McAdams 1989; Marin & McAdams, 1991). Hence, this line of work raises the question whether FMM, enhancing the prominence of singing voices in musical scenes, also plays a role in the fast and precise recognition of vocal sounds.

The aim of this study is to critically revisit the phenomenon of vocal superiority in order to highlight various acoustical factors that affect the recognition of vocal and instrumental sounds. We aim to investigate recognition of vocals under consideration of the specific vowel type within simplified musical scenes and explore the influence of FMM and other spectral features in a regression model. Furthermore, we investigate whether recognition is dependent on the used audio material or remain consistent across different stimulus sets and vocal sounds. Overall, this may help to further disentangle the roles of acoustical features and perceptual categorization processes in the perception of voice sounds.

The experimental design emulates that of Agus et al. (2012): in a recognition go/no-go task participants are presented both vocal and instrumental sounds and are instructed to respond to one type of sound while ignoring the other (Agus et al., 2012). Participants are instructed to respond as quickly as possible when hearing sounds of a target category while ignoring sounds of the non-target category (distractors). We measure response times and recognition accuracy. All sounds are aligned in duration, sound level and controlled in pitch. The targets consist of either instrumental sounds (wind or string instruments) or vocal sounds (sung vowels or singing voices). To investigate the effect of FMM on sound recognition, here each sound is presented in both an unmodified version and a version with eliminated FMM. Throughout the experiment the target category alternates between blocks to gather category-dependent responses. Additionally, in one block for both vocal and instrument targets, sounds are accompanied by a spatially separated piano accompaniment forming a minor or major triad with the target. This design aims to assess recognition abilities when the target is embedded in a simplified musical scene.

We conduct three experiments as illustrated in Figure 4.1: In Experiment 1A, sounds are extracted from a sample database, encompassing sung vowels /a/ and /u/ in both alto and soprano registers as vocal sounds, along with bassoon, trumpet, cello, and violin sounds as instrumental sounds. The results indicate frequent confusion between specific vocal and wind sounds, and the absence of an effect for FMM. To further investigate this issue, two additional experiments are conducted. Experiment 1B replicates Experiment 1A but excludes wind instruments. In Experiment 2, sounds are extracted from a popular music database, featuring female and male singing voices as vocal sounds, alongside string and wind instruments as instrument sounds. This experiment aims to compare professionally manufactured samples of isolated sounds with relatively small FMM against excerpts of naturalistic popular music with relatively large FMM.

Expanding upon previous studies, we hypothesize a discernible effect of vocal superiority, anticipating that vocal sounds are recognized faster and more accurately than instrument sounds. Furthermore, we expect that the human auditory system is specifically sensitive to the FMM occurring in singing voices, yielding pronounced vocal superiority for singing voices with FMM. Notably, our previous experiments (Bürgel & Siedenburg, 2023) demonstrated a high correlation between a larger frequency range of FMM and sound salience. Therefore, we expect the influence of FMM to be even more pronounced in the experiment that utilizes pop music excerpts. Given the reported robust detection of vocal sounds in complex musical scenes, demonstrating a partial immunity to interference from other sounds present within such scenes, we speculate that the introduction of an accompanying piano interferer has a comparatively small impact on the recognition of vocals.

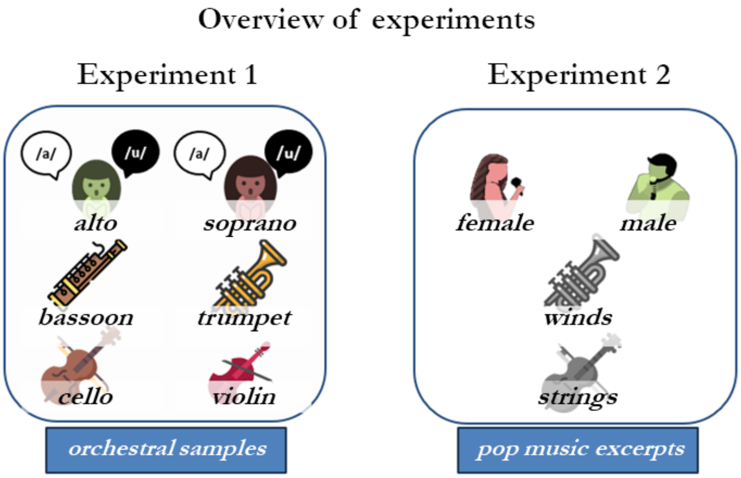

Figure 4.1: Overview of experiments

Illustration of the sound categories used in the experiments. In Experiment 1A and 1B, sounds were extracted from a sample database of isolated sounds, including sung vowels /a/ and /u/ from both alto (green) and soprano (red) registers, along with bassoon, trumpet, cello, and violin sounds. For Experiment 2, sounds were extracted from a popular music database, featuring female and male singing voices, along strings and wind instruments sounds.

4.2.3 General Methods

Participants

All participants were recruited via calls for participation on the online learning platform of the University of Oldenburg. Separate calls for subjects with and without musical training were posted to ensure a diverse range of musical abilities in our sample. The inclusion criterion for musically trained participants was a minimum of four years of musical training on at least one instrument. All participants were required to self-report normal hearing as a prerequisite for participation. Information on the participants' musical abilities was acquired using a subset of the Gold-MSI (Müllensiefen et al., 2014) consisting of nine questions on music perception abilities and seven questions on musical training. The number of participants is listed in the respective experiment sections.

Stimuli and Task

Stimuli were generated in MATLAB (MathWorks Inc., Natick, MA, USA) by using modified excerpts from two distinct databases. For Experiment 1, a sample database of orchestral instruments was utilized (Vienna Symphonic Library VI Series, https://www.vsl.co.at/en, see supplementary material Table 10). All sounds in the VSL database had a uniform duration, but featured variations of attack and decay length (see supplementary material S6). In Experiment 2 excerpts from a popular music multitrack database were employed ("MedleyDB", https://medleydb.weebly.com/). A schematic illustration of the stimuli extraction is shown in Figure 4.2A.

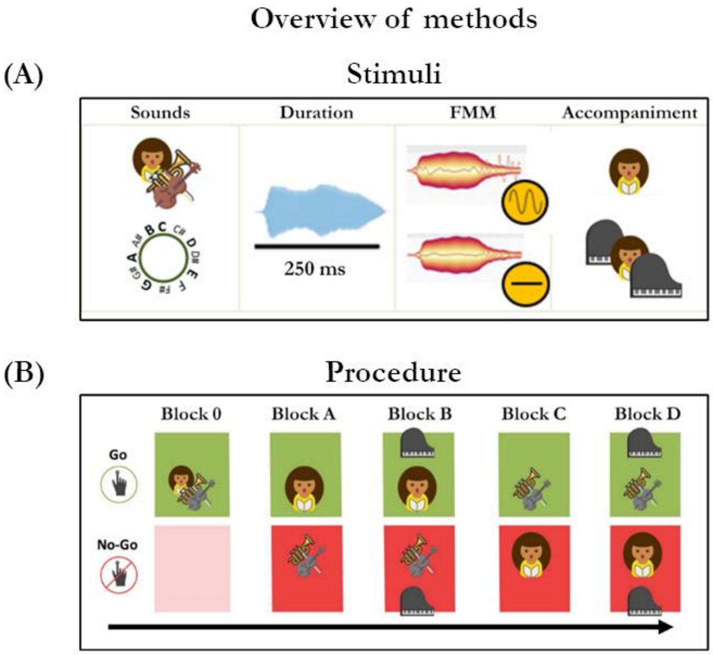

Figure 4.2: Overview of methods

(A) Stimuli encompassed vocal sounds, as well as wind and string instruments. Each sound category was represented by twelve distinct sounds, covering fundamental frequencies across a one-octave range. Targets were aligned to a 250 ms duration and normalized in sound level. Two distinct variations were created for each target: one with inherent frequency micro-modulation (FMM) and another where FMM was eliminated. Half of the time the target sound was embedded in two piano sounds, creating a major or minor triad with the target. (B) The experimental procedure comprised five blocks. The initial block entailed a detection task, where participants were instructed to respond to all sounds regardless of their timbre. Subsequent blocks were randomized in order and encompassed a go/no-go recognition task, wherein participants were instructed to solely respond to either instrumental or vocal sounds, while ignoring the other sounds. Two blocks featured isolated sounds, whereas in the remaining two blocks, sounds were presented simultaneously with a piano accompaniment.

For the instrument sounds, two wind instruments and two string instruments were selected as target categories: bassoon, trumpet, cello, and violin. For vocal sounds, the vowels /a/ and /u/ in the registers alto and soprano were selected. Each instrument or vowel was selected in the mezzo forte dynamic level and in a range of twelve semitones, ranging from A3 to G#4 (one octave), resulting in 96 sounds. An alternative version of each sound without FMM was created using the frequency modulation tool in the pitch and time correction software Melodyne (Melodyne Version 5, Celemony Software). This process involved analyzing the FMM in Melodyne using the “pitch modulation” function and subsequently removing FMM by setting all modulations to zero. A quantification of the FMM reduction is included in the supplementary material (S8). Its perceptible impact on the sound ranged from being nearly inaudible without direct comparison to sounds with FMM, particularly for some sounds in the orchestral database, to being more noticeable for the pop music extracts. For transparency we uploaded example sound files on our website (https://github.com/MichelBuergel/Data/vocalRecognition). All sounds were truncated to a duration of 250 ms, starting 5 ms before the sound level reached a threshold of -20 dB relative to its maximum level. The first 5 ms were used to create a smoothed onset, while the last 5 ms were used for as a fading offset, utilizing a 5 ms logarithmic ramp for both. Signals were converted to mono by summing up both channels and sound level was normalized in root-mean-square (rms). In total, 192 distinct target sounds were generated this way, comprising twelve sounds each with and without FMM for each of the four instruments and four vocal sounds. For conditions in which the sounds were accompanied by a piano interferer, additional piano samples (Bösendorfer grand piano) were used, spanning the pitch range from A2 to A5, to encompass all possible pitch combinations required to create a triad with the target sound (cf., Siedenburg et al., 2020).

Procedure

The experiments were approved by the ethics committee of the University of Oldenburg. All experiments shared identical designs but varied in duration and the chosen stimulus set. Figure 4.2B provides a schematic overview of the procedure. Experiment 1A utilized all 192 stimuli extracted from the VSL database. In Experiment 1B, the wind instruments were replaced with string instruments, resulting in a balanced set of instrumental and vowel sounds and a total of 96 stimuli. Experiment 2A employed 96 stimuli from the popular music database.

Each experiment started with a briefing session, during which participants were instructed about the experiment's structure, consisting of distinct blocks featuring either isolated or accompanied sounds. Additionally, participants were informed that they would need to react to and attend exclusively to vocal sounds, or sounds belonging to string or wind instruments. Subsequently, participants received specific instructions for the accompanied blocks. They were informed that an interfering piano sound would play in one ear, which they were instructed to ignore, while focusing their attention on either vocal or string and wind sounds presented in the other ear. This briefing was followed by a training section, during which participants freely listened to instrumental and vocal sounds, with or without a piano interferer. A description and icon for the presented sound were provided before and during the sound presentation. Participants had to listen to each sound category in randomized order, both with and without piano interferer, before they could proceed to the main experiment. The main experiment comprised five blocks, each containing all stimuli in a randomized order. The first block always involved an „all-go“ detection task, where participants were instructed to respond to all stimuli as fast as possible, irrespective of timbre. Subsequent blocks were “go no-go” tasks, where instrumental or vocal sounds acted as targets, and the other group had to be ignored. Sounds were either presented in isolation or accompanied by a piano dyad, forming a major or minor triad with the target sound. To attenuate effects of energetic masking , the piano signal was set to a level of -5 dB relative to the target and presented dichotically with respect to the target signal. The dichotic separation of piano and target (left or right channel), as well as the key of the triad (major or minor) and tonal position of the target in the triad (root, third, or fifth), were randomly assigned but balanced across all stimuli in one block. The number of target and distractor stimuli was also balanced. All sound levels were normalized. Instructions, using the same icons as in the trainings phase, were presented on a touch screen before each block and remained visible during each block. Participants manually continued the experiment by pressing a button on the touch screen before a block started. The stimuli presentation started after a 2000 ms pause. Stimuli were presented in a continuous stream with a 2000 ms response window and a randomized 1000 to 2000 ms inter-trail interval to prevent a rhythmic presentation. All stimuli were presented diotically, except for the dichotic condition with a piano interferer. The experiment concluded with a questionnaire gathering demographic data and a subset of sixteen questions from the Gold MSI to assess participants' musical ability.

Apparatus

The experiment was conducted in a double-walled sound booth. Participants sat in a comfortable chair and interacted with the experiment using a touch screen attached to a movable arm in front of them. Stimuli were processed through a RME Fireface UCX soundcard at a 44.1 kHz sampling rate and presented on Sennheiser HD 650 headphones. Participants’ responses were captured using a custom-made response box. Pressing a button on the box triggered a short signal burst, which was recorded simultaneously with the stimulus presentation at audio sampling rate, removing any potential time lag between stimulus presentation and response recording. Stimuli were presented at an average level of 70 dB SPL (A) as measured with a Brüel & Kjær Type 2250 light sound-level meter and a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled.

Behavioral Analysis

Participant accuracy was evaluated by analyzing the error rate, which accounted for each missed stimulus. A stimulus was deemed a miss when the response time either exceeded 2000 ms or was less than 100 ms, aiming to mitigate responses based on anticipation (Suied et al., 2010). The inverse efficiency score (IES; Townsend and Ashby, 1978), combining accuracy as error rate (ER) and reaction time (RT) as defined in Eq. (1), was used as a dependent variable. The IES has been shown to be a straightforward and useful synthesis of response times and accuracy measures, interpretable as an estimate of the average time needed to complete a correct response or simply as accuracy-corrected response time in ms (for a comprehensive review, see Vandierendonck, 2017; Mueller et al., 2020). To compute the IES, the ER and RT were averaged over every stimulus within a condition for each subject before dividing RT by 1-ER:

$$IES = \frac{RT}{1-ER} \tag{1}$$

Adhering to the recommendations of Bruyer and Brysbaert (2013), it was ensured that strong correlations between ER and RT were present, along with comparable statistical effects in both variables and the IES. A representation of recognition accuracy as ER and speed as RT is provided in the supplementary material.

Linear mixed-effect models (LME; West et al., 2014) were utilized for statistical analyses. All mixed-effects analyses were conducted in MATLAB using the *fitlme* function in the Statistics and Machine Learning Toolbox (Statistics and Machine Learning Toolbox Release 8.7, MathWorks Inc., Natick, MA, USA). The model incorporated random intercepts for each participant. In addition, IES and musical sophistication were used as numerical variables, whereas the presence of FMM, and sound category (vocal or instrumental sound) as categorical predictors. All binary categorical predictors were sum-coded. To present main effects and interactions succinctly, results are displayed in the form of an ANOVA table, with fixed effects coefficients presented as statistical parameters (F) and probability (p). These values were derived from the LME models using MATLAB's *anova* function. For a more detailed analysis, individual fixed effects coefficients are also reported as statistical parameters (t) and probability (p). For a comprehensive display of the behavioral results, models, and statistical evaluations, please refer to the supplementary material.

4.2.4 Experiment 1A

The aim of Experiment 1A was to investigate the influence of target category, FMM and accompaniment on the recognition of sounds. Two vowel sound categories were presented: vowel /a/ and /u/ in the register alto and soprano; and two instrumental sound categories: strings with cello and violin and winds with bassoon and trumpet. Each category comprised twelve sounds with notes spanning over the same one-octave range. Each sound was presented in two versions, one with naturally occurring FMM and another where FMM was eliminated. Additionally, each sound was both presented in isolation and accompanied by a piano interferer.

Participants

A total of 32 participants took part in Experiment 1A. Two participants were excluded from the analysis because they achieved notably lower accuracies in the isolated sound recognition (60% and 63%) than the average accuracy of the subjects (88%, minimum = 77%, maximum = 100%). Consequently, 30 participants (age: $\bar{x}$ = 24.1, std = 2.6) were included in the analysis. This group comprised 15 self-described musically trained participants and 15 participants with no or less than four years of musical training. There were overlaps in the musical sophistication scores of both groups: Non-musicians had scores of 32.9 (minimum: 9, maximum: 45) in the nine questions regarding musical perception and scores of 12.8 (minimum: 7, maximum: 34) in the seven questions regarding musical training. In contrast, musicians had scores of 45.7 (minimum: 38, maximum: 52) in questions related to musical perception and 32.1 (minimum: 24, maximum: 41) in questions related to musical training.

Results & Discussion

Results

In the initial "all-go" block, participants were involved in a sound detection task, where they were instructed to respond whenever they heard a sound, regardless of its timbre. The task was accomplished with perfect detection accuracy, and the average detection time was 314 ms, resulting in an IES of the same value (314 ms), with no significant effects of sound category observed in the statistical evaluation (F= 0.632, p = 0.729). The distinctions between sound categories became more pronounced during the go/no-go recognition task, for which IES scores are displayed in Figure 4.3A. Averages across sounds with and without FMM are shown because there were virtually no effects of FMM, see the statistics below.

Recognition was generally slower and less precise compared to the detection task. From a descriptive perspective, vocal sounds in the isolated presentation condition slightly outperformed instrumental sounds, yielding IES scores of 667 ms compared to instruments with a score of 692 ms. However, categories clustered into two groups, one with relatively fast and precise recognition and therefore low IES containing the /a/, strings and trumpets, and another group with slower and less precise recognition containing the /u/ and bassoon. Among vowels /a/, yielded a score of 582 ms, compared to /u/, with a score of 752 ms. Among instruments, recognition of strings and trumpet were closely similar with a score of 537 ms and a score of 553 ms, respectively, whereas the bassoon stood out with a total score of 820 ms. When embedded in a musical scene, recognition for both vowels and instruments worsened, albeit with a different impact for the categories. When comparing recognition between the isolated and accompanied presentations, voice sounds exhibited an increase of 353 ms. However, differences seen in the isolated presentation were even more pronounced among voice sounds as /a/ demonstrated robustness to the presence of the accompaniment, displaying a score increase of 64 ms, which was considerably smaller than in /u/ which showed an increase of IES by 640 ms as a result of with slower RT and an ER close to chance level. In contrast, instrument sounds exhibited an increase in IES of 413 ms. Consistent with the isolated presentation, the bassoon continued to stand out among instruments, revealing differences between the bassoon and all other instruments and having the largest score increase in 764 ms with not only a more increased RT but also an ER close to chance level. The strings and trumpet showed a deterioration similar to each other, with an increase in IES of 296 ms.

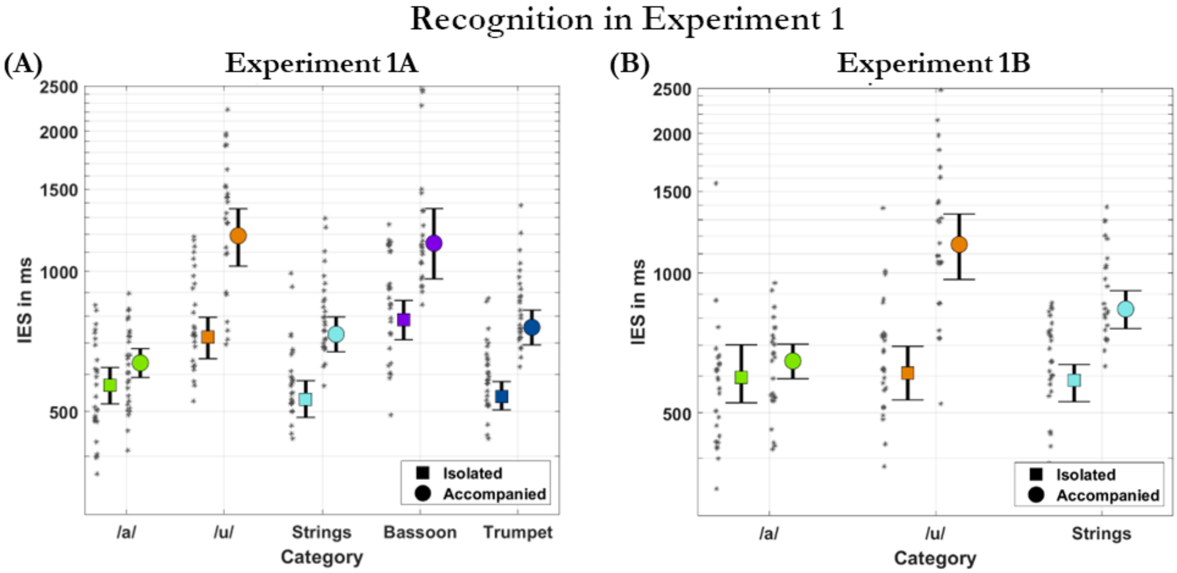

Figure 4.3: Recognition in Experiment 1

(A) *Results from Experiment 1A. Inverse efficiency scores are displayed on the y-axis, sound categories on the x-axis. Sounds were either presented in isolation (round) or accompanied by a piano (rectangles). The square marks the mean for a given sound category. Error bars indicate 95% confidence intervals. The mean of individual participant is represented by black asterisks.* ***(B)*** *Results from Experiment 1B. Graphical conventions are otherwise identical to (A).*

In terms of the statistical evaluation, the coefficient of determination (R^2) for the correlation between error rate (ER) and reaction time (RT) across all stimuli was 0.82. Contrary to our assumptions, the hypothesized vocal superiority was absent in the isolated presentation condition as sound recognition showed no differences between vocal and instrumental sounds (F = 0.014, p = 0.91). Additionally, the presence of FMM had minimal impact, yielding deviations of IES scores not larger than ±22 ms (F = 0.060, p = 0.806). Notably, no clear trend emerged for either vowels or instruments. Musical sophistication also showed no impact on the recognition of sounds, neither for musical perception scores (F = 0.326, p = 0.568), nor musical training scores (F = 1,487, p = 0.222), nor as a categorical factor by differentiating between self-identified musicians and non-musicians (F = 0.55, p = 0.477). The presentation side of the target sound in the interfered presentation also showed no effect (F = 1.153, p = 0.288). However, the presence of accompaniment showed an effect on recognition for accuracy (F = 9.311, p = 0.002). In the isolated presentation, the average IES was 680 ms which increased further in the presentation with accompaniment, resulting in a score of 1120 ms. While no general difference between the vocals and instruments in a comparison between with and without accompaniment were apparent (F = 0.040, p = 0.95), considerable effects were observable between /a/ sounds and all other categories (/u/: t = 47.450, p < 0.001, bassoon: t = 71.232, p < 0.001, and for strings & trumpet: t = 23.471, p = 0.021), highlighting a robustness to interference for /a/ sounds, unmatched by instruments or the vowel /u/.

Discussion

Experiment 1A reinforces the fundamental fact that timbre cues are sufficient to guide the recognition of sound sources. In line with Agus and colleagues and further studies (Agus et al., 2012; Moskowitz et al., 2020), no differences between sound categories emerged in the simple detection task, suggesting that sound detection is equally fast and accurate for all sounds with similar onsets. When participants were asked to recognize sound categories in a stream of sounds, both recognition times and error rates increased compared to the simple detection task. Another similarity with previous work is that recognition does not seem to be driven by a tradeoff in which faster response times lead to lower accuracy. Rather, some sounds were recognized both fast and accurately, whereas others were recognized slowly and inaccurately.

Contrary to previous work (Agus et al., 2012) and our hypothesis, a vocal superiority effect did not emerge for the isolated presentation of sounds. Instead, most of the instruments were in fact recognized slightly faster and more accurately with lower IES scores. Furthermore, differences between the vowels became apparent and revealed that the recognition of /a/ sounds were similar to the recognition of the instruments, whereas the recognition of /u/ yielded lower IES scores as a result of being both slower and less accurate. Curiously, the bassoon showed a similar pattern of performance compared to /u/ sounds. A possible explanation for this could lie in the confusion between /u/ and bassoon. This confusion seems to be a result of spectral similarity which may result from shared formant frequencies between the sounds (Reuter et al., 2018). The confusion presents characteristics akin to an informational masking effect (Tanner, 1958; Eipert et al., 2019), wherein the shared acoustical characteristics interfere with the perceptual processing of both sounds, leading to confusion and difficulty in distinguishing between the two. Consequently, this uncertainty in the recognition of /u/ and bassoon sounds may have contributed to enhanced certainty for other sounds and facilitated better recognition.

Nevertheless, this assumption does not explain the lack of vocal superiority for /a/, which is known to be spectrally distinct from /u/. Among vowels, /u/ is assigned to the group of lowest statistical format frequency as opposed to /a/, which is in the highest group (e.g., Maurer, 2016, pp 35). An argument could be made that the ambiguity of the bassoon has led to a general uncertainty in the recognition of vowels, which manifested in both vowels but especially for /u/ . Support of this assumption is the consideration that in the study by Agus, neither /u/ sounds, nor bassoon sounds were competing as target sounds, and it could be speculated that with the omission of these sounds, a vocal superiority was fostered. Coherent with this perspective, the sounds in Agus et al.'s experiment were also recognized with much higher accuracy, achieving error rates of less than two percent. This leaves room for speculation as to whether there is a general reaction time advantage only when the recognition accuracy is at ceiling level. However, it is important to note that multiple studies utilizing a comparison in the recognition of speech and instrumental sounds have also demonstrated an advantage for speech sounds (Murray et al., 2006; Moskowitz et al., 2020) and even recognition advantages for singing voices (Suied et al., 2014).

In parallel, participants' familiarity with the sound categories may have influenced recognition (Siedenburg & McAdams 2017). Although a training phase was provided to familiarize participants with the sounds, this exposure may not have been extensive enough to achieve a sufficiently consolidated mental representation of the underlying sound categories, resulting in poor classification for certain categories.

Furthermore, no effect of FMM was observed. In our previous experiment, we utilized excerpts featuring tone transitions, known to exhibit particularly perceptible FMM (Saitou et al., 2005; Hutchins & Campbell, 2009; Larrouy-Maestri & Pfordresher, 2018). This implies that the degree of FMM may have been too small to enrich sounds with perceptible information, leading to an alignment between sounds with and without FMM. Alternatively, a more straightforward explanation could be that FMMs are not a crucial feature for recognition but rather exploited for pattern matching in complex musical scenes, where they particularly help the vocals to stand out from other sounds.

Notably, a unique robustness to interference emerged for the /a/ sounds in the comparison between the presentation with and without accompaniment. As recognition worsened for all sounds, vowel /a/ stood out by achieving a considerable smaller decrease of IES in the condition with accompaniment compared to the /u/ and all other instrument sounds. This robustness was a result of a distinctly faster recognition of /a/ sounds with almost no deterioration in accuracy between the isolated and accompanied presentation which was visible in both musically untrained and trained participants. However, it should be noted that due to the high error rate of /u/, a clear interpretation of the results for this sound category is involved. If the /u/ sounds were excluded from the results, it could be assumed that the robustness to interference of the /a/ sounds might demonstrate another facet of vocal superiority that shows parallels to the robust detection of voices in musical scenes (vocal salience, Bürgel et al. 2021). However, whether this represents an actual vocal-specific recognition, or whether it is a result of spectral similarities favorable to /a/ sounds and unfavorable to /u/ sound, remains unanswered in our experiment.

Taken together, these findings prompt the question about the extent to which confusion impacted the recognition of vowel sounds and whether distinctiveness is the primary factor driving recognition scores. To investigate this, Experiment 1B was conducted with the focus of disentangling the confusion, by reducing the stimuli set to vowel and string sounds.

4.2.5 Experiment 1B

To address questions regarding potential confusion between vowels and the bassoon, Experiment 1B used the same design as Experiment 1A but omitted wind instruments from the stimuli set. To balance the number of vocal and instrumental sounds, the instrumental samples were presented twice. Additionally, each sample was presented both in isolation and accompanied by a piano dyad. All stimuli were presented in versions with FMM.

Participants

A total of 30 participants (age: $\bar{x}$ = 24.1, std = 2.6) took part in Experiment 1B. The experiment was carried out in the same session as Experiment 2, and consequently the same participants were part of both experiments. Participants were randomly assigned to start with either Experiment 1B or Experiment 2, with order counterbalanced among participants. No participants were excluded from the analysis. This resulted in 15 self-described musically trained participants and 15 participants with no or less than four years of musical training. Overlaps between both groups were observed for musical sophistication scores, with non-musicians achieving scores of 38.8 (minimum: 23, maximum: 51) in the nine questions regarding musical perception and scores of 12.4 (minimum: 7, maximum: 22) in the seven questions regarding musical training. In contrast, musicians scored 49.7 (minimum: 36, maximum: 58) in questions related to musical perception and 23.2 (minimum: 10, maximum: 42) in questions related to musical training.

Results & Discussion

Results

The performance in the recognition task is depicted in Figure 4.3B. In the isolated presentation, the average IES was 598 ms. This increased in the accompanied presentation, resulting in an IES of 878 ms. In the isolated presentation vowels and strings were closely aligned, with vowels exhibiting an IES of 603 ms and strings an IES of 588 ms. Both vowel sounds /a/ and /u/, were recognized to a similar degree, with /a/ yielding an IES of 594 ms, and /u/ an IES of 611 ms. The inclusion of musical accompaniment had a detrimental impact on the recognition of /u/ and string sounds. Vowel /a/ showed the smallest degradation among sounds with an increase of 43 ms. In contrast vowel /u/, exhibited the largest increase of 599 ms with an accuracy close to chance level. Strings exhibited an increase of 264 ms.

Error rate (RT) and reaction time (RT) shared substantial variance with an R^2 value of 0.62. Consistent with the findings in Experiment 1A, no differences in recognition emerged between vocal and instrumental sounds ($F = 1.505$, $p = 0.221$). Recognition performance deteriorated for the presence of accompaniment ($F = 11.528$, $p < 0.001$). However, in line with the robustness to interference observed in Experiment 1A, this effect was less pronounced for /a/ sounds, as underlined by effects between /a/ and /u/ ($t = 6.842$, $p < 0.001$), as well as /a/ and strings ($t = 2.745$, $p = 0.006$). Musical sophistication did not appear to affect recognition in a substantial way, as no main effects for either musical perception (F = 0.443, p = 0.506) or musical training (F = 0.031, p = 0.859) were evident. Also, in line with Experiment 1A the presentation side of the target sound in the interfered presentation showed no effect (F = 0.024, p = 0.876)

Discussion

Experiment 1B aimed to investigate vocal sound recognition in the absence of wind instruments which were frequently confused with /u/ sounds in Experiment 1A. Contrary to expectations, no recognition advantage emerged for vocals in the isolated presentation, even though only string and vocal sounds were presented, closely mirroring the conditions of the shared distractor experiment by Agus and colleagues (Experiment 2: voice-processing advantage; Agus et al., 2012). Unlike Experiment 1A, no differences were observed between vowel /a/ and /u/ in the isolated presentation. This underscores our assumption that the relatively poor recognition of /u/ in Experiment 1A was due to confusion with the bassoon, indicating that vocal sound identity itself does not guarantee superior recognition and does not make sounds unsusceptible to confusion with non-vocal sounds. This assumption is further supported by the contrasting behavior of vowels when embedded in a musical scene. As observed in Experiment 1A, /a/ sounds exhibited unique robustness to interference, with a smaller decrease in recognition performance distinct from /u/ and instrumental sounds. However, even without wind instruments, /u/ sounds yielded the worst recognition scores among all sounds.

Another interesting observation arises in the isolated presentation: an equalization between strings and voices. In Experiment 1A the strings, especially the cello, exhibited faster recognition compared to vocal sounds. However, this difference is no longer evident. It remains unclear whether this change is due to the absence of the bassoon, the reduction of target categories, or a specific behavior of the cello. Nevertheless, it underscores the necessity of contextualizing results within the stimulus set and emphasizes the importance of testing with diverse stimuli sets.

Furthermore, recognition accuracy for /u/ dropped close to chance level in the condition with accompaniment, suggesting that the piano dyad strongly interfered with the /u/ sounds. However, stimuli were deliberately designed to hinder complete energetic masking by reducing the piano’s sound level and spatially separating the target and interferer sounds, resulting in binaural presentation. An argument could be made that informational masking occurred with both sounds remaining audible but listeners’ attention shifting towards the masking sound (Pollack, 1975; Kidd et al., 2008), potentially due to the uncertainty associated with the random assignment of target and accompaniment to left and right channels. However, this would not explain why certain sound categories such as the /a/ are very robust to the presence of the accompaniment. The significantly better recognition of /u/ in the isolated performance in Experiment 1B together with the poor recognition of /u/ in Experiment 1A suggests that most likely a combination of confusion with the bassoon and interference by the piano, both of which can be attributed to informational masking, were at the source of the observed effect.

In summary, the results underline the findings from Experiment 1A and support the assumption that vocal sounds in musical scenes do not inherently evoke enhanced recognition. Instead, it appears that the recognition of vocal sounds seems to be highly influenced by the vocal sound quality itself (here, the vowel type). Thus, it would be intriguing to explore whether the recognition of other vowels or vocal sounds would be similar to the /a/ or /u/ sounds. To explore the generality of the observed effects and further assess the effect of FMM, we conducted Experiment 2, in which stimuli were generated by extracting snippets of vocals or instruments from a pop music database.

4.2.6 Experiment 2

Sounds were extracted from songs in a multitrack popular music database used in previous work, which demonstrated a correlation between FMM range and vocal salience (Bürgel & Siedenburg, 2023). The excerpts were taken from the onset of notes and thus contained pitch transitions encompassing more pronounced FMM compared to Experiment 1.

Participants
The same participants as in Experiment 1B took part.

Stimuli and procedure
The popular multi-track music database comprises 127 songs across a variety of popular music genres, each with individual audio files for instrument and vocal tracks. Given the continuous nature of the tracks and the presence of overlaid audio effects, potential sound candidates had to be manually selected to resemble the clean and unmodified samples in the VSL database. In line with the selected sound categories in the database used for Experiment 1, string and wind instruments were chosen as instrumental targets. For vocal sounds, rather than different voice registers or vowels, female and male vocal tracks were selected. To control pitch, the chosen tracks were analyzed in Melodyne, and twelve different excerpts for each target sound were extracted, covering the same pitch in a range of one octave (from A3 to G#4) as in our VSL samples. This process resulted in vocalizations that could be categorized as nine /a/ sounds, four /u/ sounds, one /o/ sounds, one /e/ sound, one hissed sound, and eight mixed sounds with multiple vowels. FMM manipulation, truncation, ramping, and normalization were performed similarly to the sounds from Experiment 1. In total, 96 different stimuli were extracted this way, including twelve sounds with and without FMM for two instrumental and two vocal sounds each. Additionally, each target sound was both presented in isolation and embedded in a musical scene with a piano dyad accompanying the target sound. The procedure was identical to Experiment 1A and B.

Results & Discussion

Results

In the detection task, all sounds were detected perfectly, with only minor differences visible for reaction time. Vocals yielded an IES of 365 ms, and instruments had an IES of

350 ms, with no significant effects present in the LME (F=0.817, p=0.44). Additionally, no difference was observed for participants who started either with Experiment 1B or Experiment 2 (F = 0.603 p = 0.616). The performance for the recognition task in Experiment 2 is displayed as averages across sounds with and without FMM in Figure 4.4A.

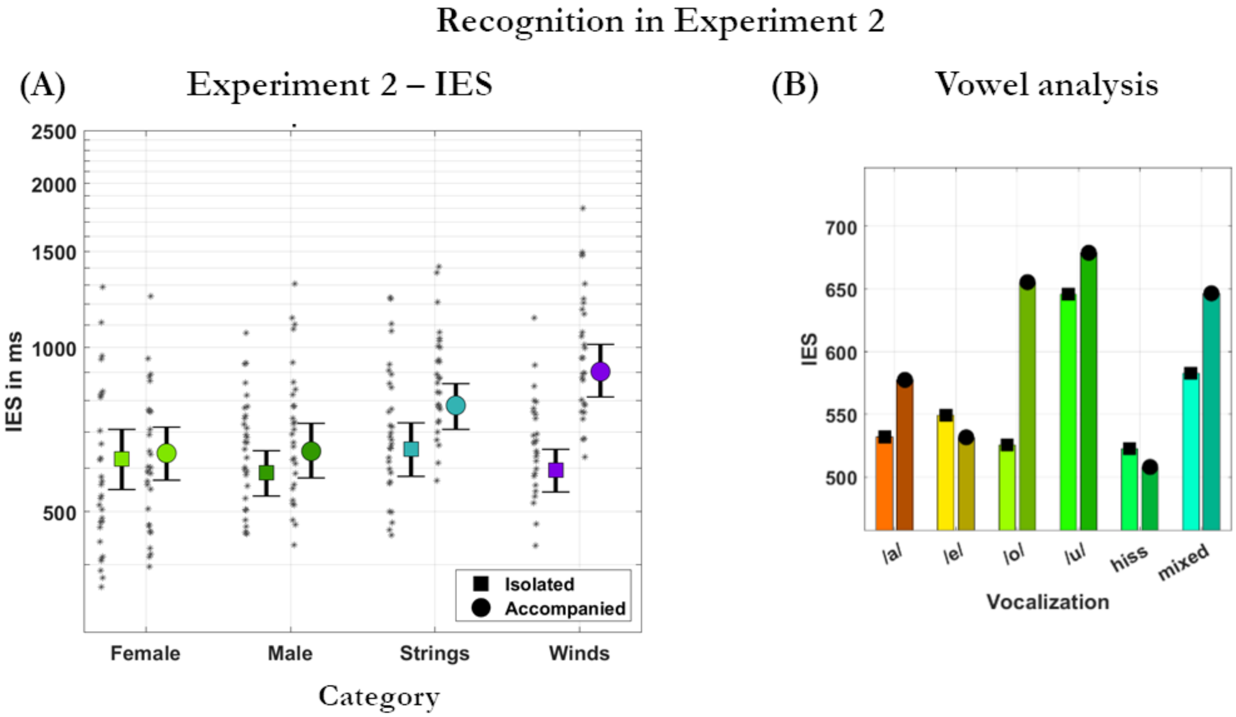

Figure 4.4: Recognition in Experiment 2

(A) IES: *Vocal sounds (female and male) and two instrument sounds (strings and winds) as targets in a go/no-go task. Graphical conventions are otherwise identical to Fig. 3.* ***(B) Vowel analysis:*** *Vocal sounds were categorized by vocalizations in nine /a/ sounds, four /u/ sounds, one /o/ sounds, one /e/ sound, one hissed sound, and eight mixed sounds with multiple vowels. The recognition for a given vocalization is represented as the average IES in isolation (rectangles) or in accompanied presentation (round).*

Overall, recognition was generally slower and less precise compared to the detection task. In the isolated presentation, the average IES was 621 ms. This increased in the presentation with accompaniment, resulting in an IES of 800 ms. The absence of FMM showed neither a positive nor negative trend, leading to average variances of IES not larger than ±24 ms. Differences between vocals and strings were closely aligned in the isolated presentation: female vocals exhibited a IES of 624 ms, male vocals of 589 ms, strings of 654 ms, and winds of 594 ms. When presented with accompaniment, recognition worsened for both vowels and strings. However, this effect was less pronounced for the vocal sounds, which exhibited an increase of 53 ms compared to an increase of 214 ms for instruments, demonstrating a robustness to interference as seen in both previous experiments for vowel /a/. The increase differed slightly between female and male vocals with no effects in the LME, with an increase of 37 ms for female vocals and an increase of 69 ms for male vocals. In contrast, strings yielded an increase of 134 ms, and winds an increase of 294 ms.

Discrimination between the vocalizations within the sounds revealed notable differences in recognition, as depicted in Figure 4.4B. The increase in IES between the isolated to accompanied presentation ranged from -18 ms to 64 ms, with the one /o/ sound standing out with an increase of 130 ms. Notably, the eight sounds with /a/ had an average IES of 531 ms in the isolated presentation, with an increase of 46 ms in the accompanied presentation. In contrast the four /u/ sounds had an IES of 646 ms in isolation, with an increase of 33 ms. Thus, /u/ sounds held higher IES, aligning with observations in Experiment 1. Contrary to those observations, no pronounced differences in the deterioration were observed between /a/ and /u/ sounds. However, it is important to note that this analysis is not balanced, as the number of stimuli within each vocalization greatly differed and the observed differences may be an artifact of individual stimuli.

Error rate (ER) and reaction time (RT) shared substantial variance with an R^2 value of 0.86. Consistent with Experiment 1, neither an effect of FMM emerged ($F = 0.776$, $p = 0.378$) nor an effect for musical sophistication in musical perception ($F= 0.939$, $p = 0.332$) or training ($F= 0.012$, $p = 0.912$) or an effect for target presentation side ($F= 0.169$, $p = 0.682$). However, an overall recognition advantage for vocal sounds was present ($F = 20.449$, $p < 0.001$), as well as an effect for accompaniment ($F = 15.813$ $p < 0.001$) and interaction effects between both ($F = 20.252$, $p < 0.001$). Differences between vocal and instrumental sounds were negligible in the isolated presentation but pronounced in the accompanied presentation with effects between vocals and strings ($t = 3.784$, $p < 0.001$), as well as vocals and winds ($t = 9.313$, $p < 0.001$), highlighting a specific robustness to interference for the vocals. As a synopsis of the experiments, Figure 4.5 displays the differences between the isolated and accompanied presentation of all three experiments.

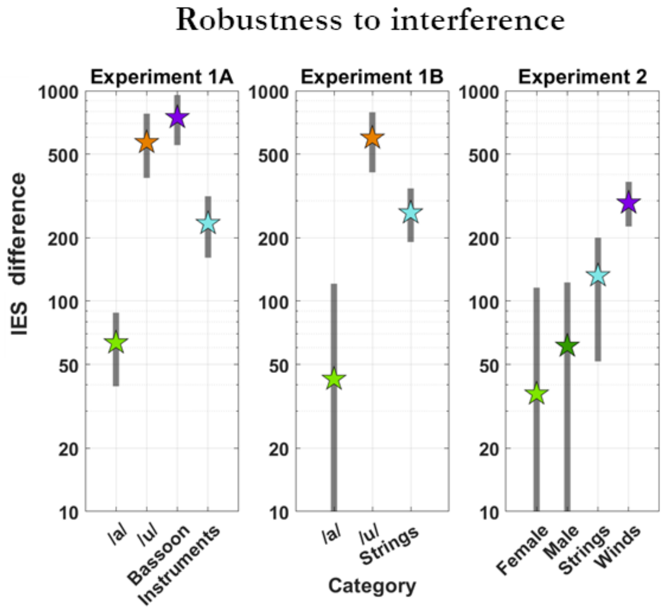

Figure 4.5: Robustness to interference

The presence of accompaniment worsened recognition for all categories. The difference in IES between the isolated and accompanied presentation is displayed for all experiments and categories. The stars mark the average difference for a given category. Error bars indicate 95% confidence intervals.

Discussion

Experiment 2 investigated the recognition of vocal and instrument sounds extracted from a pop music database, focusing on the effect of Frequency Micro-Modulation (FMM) and assessing the generalizability of recognition advantages found in previous experiments. The results mirrored those of Experiment 1. Notably, the anticipated superior recognition for vocal sounds was only present in the condition with accompanied presentation but was not present in the condition with isolated presentation. Furthermore, the presence of FMM showed no impact on sound recognition, supporting the conclusion that cues related to FMM are not exploited or do not affect the recognition of musical sounds.

Additionally, the previously observed robustness to interference for /a/ sounds persisted between the isolated and accompanied presentation for the vocal sounds of Experiment 2. Additionally, an overall less pronounced deterioration of recognition in the accompanied presentation was observed for all sound categories, even though the sounds used within a category were less homogeneous than in Experiment 1. This inhomogeneity stemmed from extracting instrumental and vocal sounds from various songs with different instruments or singers, lacking strict control for intonation dynamics and articulation. It is likely that the less controlled pop music excerpts could stand out more easily from the piano accompaniment than the orchestral samples, which were aligned with the piano in intonation dynamics and articulation. Yet, despite the inhomogeneity of sounds, a clear robustness to interference was pronounced in the experiment, indicating that diverse vocal sounds are capable of exhibiting robust recognition. Interestingly, despite the distinctions observed between /a/ and /u/ in Experiment 1, this robustness was evident for most vocalizations, this robustness was evident for most vocalizations, except for three sounds (/o/, /e/, and /hiss/) which showed either no robustness or slightly worse performance in isolated presentation. This inconsistency might be a consequence of the sound's inhomogeneity, allowing them to provide more distinct cues compared to the orchestral samples.

4.2.7 Acoustical Analysis

To explore relations between acoustic features of the sound and recognition performance, linear regression analysis was employed to predict human recognition scores using the spectral similarity between sounds and their FMM range. Spectral information of sound signals was obtained through cepstral coefficients derived of sounds' energy in filter bands with equivalent rectangular bandwidth (ERBCC). This representation served two purposes: first, to compare spectral attributes of the competing target sounds that participants were asked to recognize, and second, to assess the similarity between the target sounds and piano interferers. The ERBCC extraction process was analogous to the computation of Mel-frequency cepstral coefficients (MFCC) known for their effectiveness in computational sound classification (Monir et al., 2022), but with using an ERB-filter bank (instead of the Mel-spectrum) to better align with data of human frequency selectivity (Glasberg & Moore, 1990). The ERBCC extraction process involved extracting the long-term spectrum over the whole 250 ms duration, filtering the spectral energy of the target sounds into 64-ERB bands in the frequency range between 20 Hz to 16000 Hz, taking the logarithm, and deriving the first thirteen cepstral coefficients using discrete cosine transformation. In a last step, the first coefficient was discarded as it contains no information about the spectral shape, but only a constant level offset information and the temporal dimension was discarded by averaging each coefficient across the time windows.

In order to extract the similarity between vocal and instrumental sounds, a principal component analysis (PCA) was performed on the ERBCC data of all target signals for each target category, encompassing all twelve notes, and the same twelve notes of the piano interferer (A3 - G#4). The first two principal components (PC1 & PC2), that explained the most variation in the dataset, were used to create a two-dimensional component space. The proximity of sounds in the space indicates their spectral similarity, with sounds closer together being more similar. A measure of spectral distinctiveness was derived by calculating the Euclidean distance between a target sound and the center of the space with the rationale that more spectrally distinct sounds would occupy regions further separated from the center of the space. To represent the similarity to the piano interferer, the distance between the target and the spatial centroid of the piano sounds was computed. Additionally, to examine potential correlations between recognition accuracies and acoustic features, a confusion matrix was generated using the ERBFCC. This involved determining the similarity between sounds through a sound-by-sound correlation analysis. The resulting confusion matrix can be found in the supplementary material (S7).

To gather information regarding the FMM intensity, the FMM range as the difference between highest and lowest fundamental frequency (f0) within each sound was analyzed. To do so, the fundamental frequency was extracted using the MATLAB function *pitch* (Audio Toolbox Release 3.7) in 10 ms sliding time windows over the duration the sound. An additional artifact suppression was implemented to counteract irregular fluctuations, by applying a threshold for tonal components in the time window (harmonic ratio) as provided in the pitch function, excluding samples below a harmonic ratio of 75%. Additionally, a one octave frequency threshold around the sounds median f0 was applied to each window, to eliminate erroneous leaps and octave errors in pitch recognition. As a final step the f0 frequencies were transformed to a scale with a resolution of one cent and the distance between largest and smallest f0 was computed.

Both distance metrics and FMM range were employed as independent variables in a multiple linear regression model to predict IES. To mitigate the potential influence of adding independent variables to the regression, a bootstrap hypothesis testing was incorporated. In this process, linear regression models were generated using randomized independent variables to predict IES values in a bootstrap procedure comprising 1000 iterations. The results of the model using the true (non-randomized) similarities were then compared with the distribution of R^2 values from the bootstrap models, considering the model suitable when the adjusted R^2 of the analyzed data was greater than the 99th percentile of the bootstrap distribution.

4.2.7.1 Experiment 1

For Experiment 1, the first two principal components explained 91 percent of the variance, with PC1 accounting for 80 percent and PC2 explaining 11 percent. The two components are depicted in a component space in Figure 4.6A. Vowel /u/ and bassoon sounds are clustered throughout the component space. Along PC1, these are adjacent to /a/ sounds, whereas string and trumpet sounds create their own region on the opposite side of the space. Piano sounds appear in the center of the space and overlap on the edges mostly with /u/ and bassoon sounds. Interestingly, clusters that appeared on the opposite side of the component space, e.g., /a/ sounds, trumpet, and strings, were also the sounds with the best recognition scores. Conversely, the merged cluster of vowel /u/ sounds and instrumental bassoon sounds in the center is in concordance with assumptions about potential confusion between these categories. Further, this cluster underlines our hypothesis that /u/ and bassoon sounds had pronounced spectral similarities, which could explain their relatively poor recognition. Furthermore, the relatively small distance between those and the piano sounds implies that the piano interferer could have had a much greater influence on recognition for those sounds than for the other more distant sounds.

The analysis of FMM range is presented in Figure 4.6B. Vocal sounds had a larger range than instrument sounds, with /a/ showing a range of 49 cents and /u/ of 45 cents. However, this range was smaller than in our previous detection experiments where vocals exhibited a range of 84 cents. The elimination of FMM reduced the range to 6 cents for both sounds. Instruments initially had an overall reduced FMM range compared to vocal sounds, with a range of 28 for strings and 16 cents for winds, further minimized to 4 cents for strings and 5 cents for winds.

Results of the multiple linear regression are illustrated in Figure 4.6C. Similar to the LME analysis, a linear regression operating on the FMM range showed no considerable correlations with R^2 smaller than 0.02 in both the isolated and the accompanied presentation. Utilizing the distance between the target sounds in the component space yielded moderate correlations with R^2 values of 0.30 for the isolated presentation and 0.38 presentation with the piano interferer. When considering the distance in the component space between the target and piano sounds, the linear regression for the presentation with the interferer yielded an R^2 value of 0.18. Operating on both distances improved the model to an R^2 of 0.54. The addition of the FMM range did not enhance the model. To examine whether the presence of FMM specifically impacts the recognition of vowels, a distinct linear regression was performed, focusing solely on vowel targets. However, even in this analysis, FMM range demonstrated no significant influence, with R^2 values remaining below 0.05 in both presentation conditions. All results, except for the one only operating on the FMM range, passed the bootstrap hypothesis test, with R^2 values exceeding the 99th percentile of the bootstrap distribution. In summary, the model supports our assumptions that spectral distinctiveness of the target sounds and the similarities between the target and the piano interferer guide sound source recognition. The consistently stronger correlations in the accompanied presentation for the similarities between the stimuli suggest that these similarities are particularly impactful when the musical scene is more demanding.

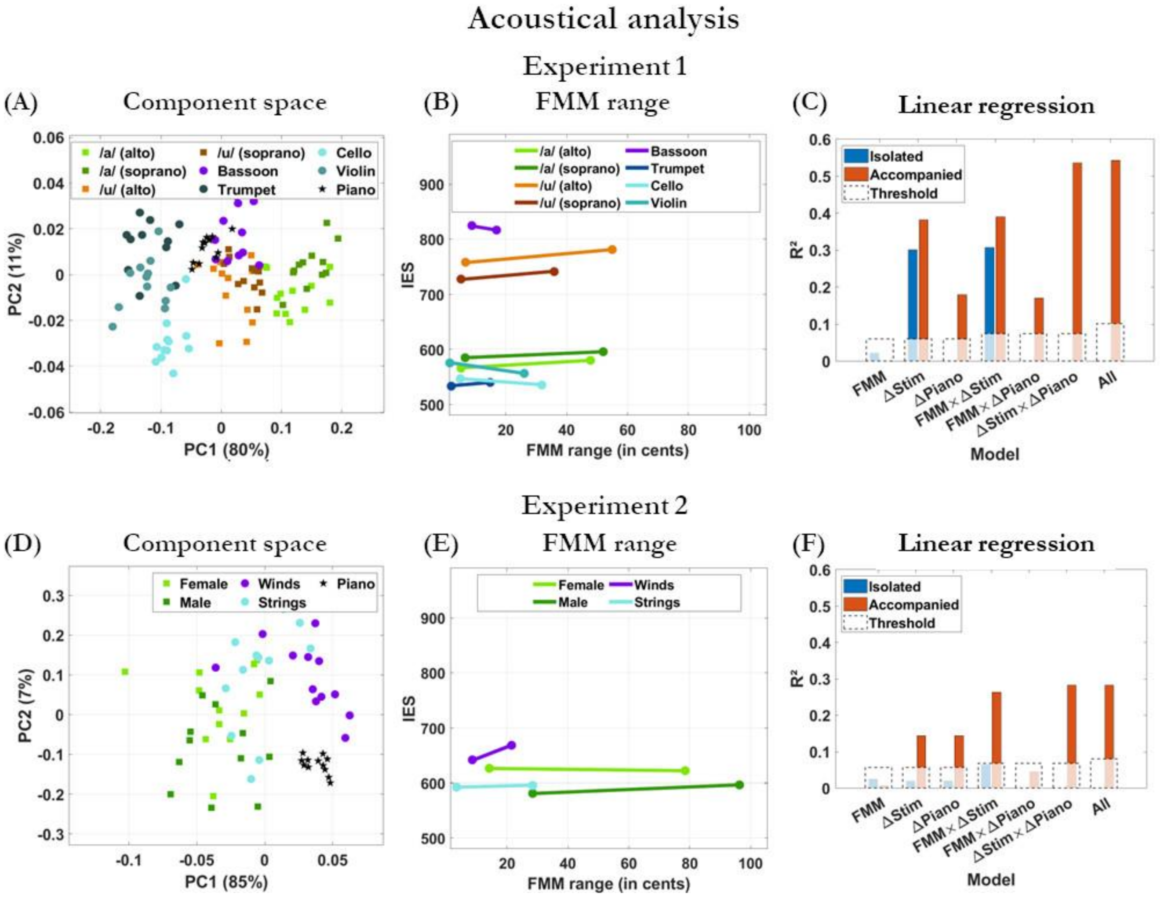

Figure 4.6: Acoustic analysis

(A) Component space: *A principal component analysis (PCA) was applied to the spectral features of all signals in Experiment 1. Sounds are presented in a component space of the first two components. Percentual values at the axis labels indicate the explained variance of the respective component. (B)* ***FMM range****: The average range of frequency micro-modulation (FMM) and average IES of the isolated presentation is displayed for each sound category in Experiment 1.* ***(C) Linear correlation:*** *A multiple linear regression was computed on features of individual sounds employed in Experiment 1 to predict IES. As features, the analysis utilized the sounds FMM range (FMM), distance between the sound and the center of a principal component space (ΔStim), distance between the sound and the mass of all piano sounds in a principal component space (ΔPiano) and combinations of all features. Transparent dotted boxes indicate the 99th percentile of a bootstrap hypothesis testing.* ***(D)****,* ***(E)****,* ***(F)*** *depicts the described analysis methods as in* ***(A)****,* ***(B)****,* ***(C)*** *but for Experiment 2.*

4.2.7.2 Experiment 2

The component space for sounds utilized in Experiment 2 is shown in Figure 4.6D. In comparison to Experiment 1, sound categories were more intermingled in space. The first two principal components explained 92 percent of the variance, with PC1 accounting for 85 percent and PC2 for 7 percent. Vocal and string sounds were mixed, while wind instruments stood out and were mostly found in a distinct quadrant. Additionally, a densely packed cluster of piano sounds was visible. The lack of separate instrument and vocal clusters can be understood as a result of the less homogeneous excerpts utilized in this experiment, further emphasized by the pronounced cluster of piano sounds originating from the VSL database. Interestingly, despite the less homogeneous sound excerpts, better recognition was observed in Experiment 2 than for the more distinct sounds in Experiment 1.

The analysis of FMM range is depicted in Figure 4.6E. Vocal sounds in Experiment 2 not only obtained a larger FMM range compared to Experiment 1 but also showed a close resemblance to the ranges found in our previous detection experiment. Female vocal sounds showed a range of 78 cents, and male sounds showed a range of 96, both being close to the salient vocal signals in our previous experiment with a range of 82 cents. After the FMM reduction, the range decreased to 15 cents for female vocals and to 29 cents for male vocals. Instrumental sounds carried smaller ranges than vocals, which for strings were reduced from 24 cents to 4 cents, and for winds reduced from 21 cents to 9 cents.

As observed in the LME, a linear regression based on the FMM range showed no substantial correlations, with R^2 approximately 0.05 in both the isolated and the accompanied presentation. Utilizing the distance between target sounds in the component space yielded weak correlations, with R^2 values of 0.24 for the isolated presentation and 0.28 for the presentation with the piano interferer. When based on the distance in the component space between the target and piano sounds, the linear regression for the presentation with the interferer yielded an R^2 value of 0.11, which did not surpass the 99th percentile threshold. Utilizing both FMM range and the distance of target sounds slightly improved the model to an R^2 of 0.26 for the isolated and 0.29 in the accompanied presentation. Operating on both distance metrics held an R^2 of 0.29. Incorporating all predictors resulted in an R^2 of 0.32. Only models that operated on the similarities within the target sounds passed the bootstrap hypothesis testing. To investigate whether the FMM might only influence vocal sounds, a separate linear regression was performed focusing exclusively on female and male target sounds. However, even in this particular analysis, the FMM range showed no considerable correlation, with R^2 values remaining below 0.05 in both presentation conditions. Taken together, these results imply that spectral similarities between the targets impacted recognition, albeit to a somewhat smaller degree than in Experiment 1, as other distinct features of the inhomogeneous sounds may have been more dominant compared to spectral similarities.

The diminished impact of the interferer sound appears to result from their spectral distinctiveness. The piano tones occupied a distinct area within the component space, lacking overlap with other sounds. This dissimilarity seems to have surpassed a critical threshold, leading to the recognition of the target sounds no longer being influenced by their similarity to the interferer. Therefore, the relatively smaller deterioration in the accompanied presentation in Experiment 2 could have been a combination of multiple differences between the target and interferer, including intonation, articulation, and spectral dissimilarity.

Despite the average FMM range of vocal sounds being comparable to our previous detection experiment, correlations between FMM range and recognition were negligible. This finding refutes our initial assumption in Experiment 1, that the absence of an FMM effect was due to insufficient FMM range. Additionally, the omission of FMM did not exhibit a consistent trend; instead, it appeared to unsystematically and marginally worsen or improve recognition. These findings reinforce our prior conclusion that cues related to FMM do not significantly affect the recognition of musical sounds.

4.2.8 General Discussion

In this study, we investigated recognition of vocal and instrumental sounds in three experiments. We tested the influence of frequency micro-modulations (FMM) and accompaniment on the recognition of vocal sounds. Sounds from multiple databases were utilized to examine the generality of effects across diverse audio material. Participants were tasked with classifying short vocal or instrumental sounds in a go/no-go task. Sounds were controlled in level and pitch and each sound was presented in versions with naturalistic FMM and with reduced FMM. Additionally, sounds were either presented in isolation or formed a harmonic triad with an accompanying but spatially separated piano interferer. The audio material of the sounds varied between experiments. To assess whether human recognition could be explained by acoustic features, a multiple linear regression employing spectral features of the sounds was utilized.

Contrary to our hypotheses and previous findings we did not observe vocal superiority (faster and more accurate recognition) across any of our three experiments. Instead, notable differences between sung vowels became apparent, with /a/ sounds outperforming /u/ sounds. Suspecting the absence of this effect due to spectral similarities between vocal and bassoon sounds (Reuter et al., 2018), as indicated by our acoustical model and observed in the behavioral data, we repeated the experiment and removed the wind instruments or used different audio material. While recognition improvement was evident, differences between the vowels persisted and the reported vocal superiority effect remained absent. An argument could be made that language differences between the singer and listener influenced the recognition of the vowels to some degree, thus potentially explaining one contributing factor to the observed confusion. However, it's essential to note that the vowels used in classical singing may differ compared to those in everyday speech. The lack of vocal superiority is only sparsely reported in the literature (Bigand et al., 2011; Ogg et al., 2017). However, a direct comparison between our study and aforementioned studies is questionable. Ogg argued that the absence of vocal superiority in speech signals found in their study may not apply to the recognition of singing voices, as the investigated scenarios were too disparate. Bigand and colleagues pointed out that the absence of superiority in their study was likely caused by a peak level normalization. Furthermore, they reported that when utilizing a root-mean-square (RMS) sound level normalization, as conducted in many sound recognition studies (i.e., Agus et al., 2012; Suied et al., 2014; Moskowitz et al., 2020) superior recognition of vocals was observed. They argued that the RMS normalization might have emphasized frequencies that facilitate a superior recognition of speech sounds. However, this suggestion is not supported by our results, as we applied root-mean-square level normalization but still observed an absence of vocal superiority, what makes our findings unique.

Contrary to our assumptions, the removal of FMM showed no influence on vocal recognition. It is assumed that FMM enriches vocal sounds with additional information about pitch continuity (Weiss & Peretz, 2019) and enhances the prominence of sung vowels compared to vowels without FMM (McAdams, 1989). Our previous study on the influence of FMM on the detection of vocals in musical scenes underpinned the importance of FMM, as the reduction of FMM led to a reduced vocal salience (Bürgel & Siedenburg 2023). Consequently, our intention was to explore the effect of FMM on sound recognition. However, in the present study no such effects were found. To investigate whether this absence was based on using an orchestral database with stationary tones and generally lower FMM ranges, we conducted an additional experiment that used sounds extracted from continuous excerpts of the same pop music database used in our previous detection experiment. Although the FMM ranges were similar to the detection experiment, the effect of the FMM on recognition remained absent. An obvious disparity between both studies is the stimulus duration: two seconds in the previous experiment and a quarter-second in the current experiment. An argument could be made that such short stimuli do not provide sufficient exposure to allow a perceivable effect of FMM to unfold. However, this reasoning seems rather unlikely, considering that established perceptual thresholds for identifying the direction of pitch modulations are as low as 20 ms (Gordon & Poeppel, 2002). Moreover, FMM perceptibility has been demonstrated for tones of considerably shorter durations, such as 80 ms (d’Alessandro and Castellengo, 1991), as well as similar durations of 220 ms (Larrouy-Maestri & Pfordresher, 2018). On the other hand, this finding aligns with studies indicating that the auditory system utilizes multiple processes with different temporal resolutions (Poeppel 2002; Santoro et al., 2014; Giroud et al., 2020). These processes encompass mechanisms specialized in extracting information such as pitch and spectral shape, within short time intervals (~30 ms), while other processes analyze longer time intervals (~200 ms) to detect changes over time. Even though the time windows identified are both shorter than the stimulus duration we employed, this observation could indicate that our stimulus duration was too brief to give rise to a significant effect of time-variant features on detection. Moreover, in the previous detection experiment, we employed musical scenes featuring a variety of instruments and vocal sounds that overlapped in time and spatial location within the scene. This stands in contrast to the relatively simple dichotic scene utilized in the current recognition experiment, where the piano and target sound were clearly separated. This less saturated scene with strict peripheral separation might have offered a simplicity that rendered FMM cues obsolete for recognition. Taken together it seems likely that the presence of FMM has no major impact on the recognition of musical sounds when other timbre cues are available.

The presence of a musical accompaniment deteriorated recognition for all sounds considerably. Unexpectedly, however, this negative effect was consistently smaller for the recognition of /a/ sounds and vocal excerpts from the pop music database. This distinct robustness to interference primarily manifested as a faster recognition compared to other sounds, with almost no deterioration in accuracy. Importantly, this effect was present across all three experiments, despite variations in the excerpts, highlighting the consistency of the effect. The observed robustness in vocal recognition may be indicative of a specialized processing mechanism for acoustic features of vocal signals during the segregation of auditory objects in a musical scene. Suggesting that, when the auditory system segregates sound into mental representations of distinct streams, vocal features could trigger a prioritized, voice-specific processing (Belin et al., 2000; Levy et al., 2001; Gunji et al., 2003; Belin et al., 2004) that in turn might facilitate a better identification of vocal sounds within the complex auditory scene, contributing to accelerated recognition. However, distinctions between the tested vowels were apparent, with /u/ sounds lacking the robustness seen in /a/ sounds, suggesting that vocal sounds do not inherently trigger facilitated recognition. When also considering the susceptibility of /u/ to be confused with instruments, our results suggest that while vocal recognition in musical scenes can be uniquely robust, it does not possess properties that make it impervious to confusion with spectrally similar sounds.

Musical sophistication showed no significant effects on sound recognition in our experiments. Musicians are often reported to have advantages in the discrimination pitch (e.g., Tervaniemi et al., 2005; Micheyl et al., 2006) or timbre (e.g., Chartrand & Belin, 2006; Kannyo & DeLong, 2011), improved resistance against informational masking (Oxenham et al., 2003) and the ability to hear out partials in tone complexes (Zendel & Alain, 2009), in chords (Fine & Moore, 1993) or even melodies in complex musical mixtures (Siedenburg et al., 2020). Furthermore, familiarity with instrumental sounds (typical for more musically experienced individuals) is known to enhance recognition (Siedenburg & McAdams, 2017). However, no effect for musical sophistication was observed in our study. It should be noted that a distinct analysis of accuracy and speed also revealed no tradeoff between them (Chartrand & Belin, 2006). Instead, most participants showed a behavior where sounds with higher accuracy tended to be detected faster. Partially in agreement with the absence of effects, studies specifically investigating low-millisecond sound recognition and discrimination have shown conflicting effects of musical sophistication, with influences either being absent (Bigoni F. & Dahl S., 2018) or present (Akça et al., 2023).

Limitations

While our study has provided valuable insights into sound recognition, it is imperative to recognize the inherent limitations that may have influenced the interpretation of our findings. One caveat of the study is that certain sounds, primarily /u/ and the bassoon, exhibited a high error rate in recognition. While reaction time measurement has proven to be a useful tool, especially when assessing supra-threshold recognition, it may not be suitable for all signals tested in this study. This is particularly critical given the limited number of stimuli used for each condition (twelve). An alternative interpretation is that the response times measured in this case may reflect a measure of confidence rather than recognition speed. This alternative view is further supported by the correlation between errors and reaction times, indicating that a low error rate was associated with a faster reaction time. Nonetheless, we deliberately included these sounds because we believe they provide a valuable insight into the influence of acoustic similarity on sound recognition. To address the issue of high error rates, one perspective could also involve extending the training phase. Previous research by Agus and colleagues (Agus et al., 2010) has demonstrated that even the recognition of seemingly meaningless noise signals can be improved above chance levels with training. One approach could be to design a training session in which a certain number of stimuli must be correctly recognized for each sound category before proceeding to the main experiment. Another attempt could involve investigating whether increased repetition of stimuli leads to a reduction in errors and how this in turn affects reaction times. In the same vein, it would also be intriguing to explore the recognition of /u/ or other vowel and instrumental sounds across different or more diverse stimuli pools containing spectrally similar or dissimilar sounds. This could shed light on how such variations influence the results, providing further insights into the hypothesized vocal superiority and the intricate interplay of spectral similarity.

Another potential source of undesired variability in our results could stem from the use of stimuli that are intended to be more ecologically valid and therefore are less controlled. While we applied a standardized method to extract stimuli based on a sound-dependent level threshold to maintain consistency, this approach may in principle have led to varying degrees of transient truncation across different signals. However, based on our own close listening of the stimuli and as highlighted by Fig. S6, we do not think that the procedure truncated onset portions severely such that onset cues remained intact (Siedenburg, Schädler, & Hülsmeier, 2019). To further mitigate this potential issue, we included diverse sound categories (e.g., different vocal registers such as alto and soprano) and utilized multiple source databases. While the consistency observed in our results suggests the absence of significant bias caused by our extraction, further investigation into the generalizability of these findings across different databases could provide valuable insights.

The methods used to investigate the effects of FMM also offer opportunities for expansion. It would be intriguing to explore whether presenting stimuli with and without FMM in separated blocks-wise presentation would yield different results. This approach could potentially impact the already challenging recognition task, by allowing participants to adopt a strategy of extracting additional FMM information for blocks with FMM. It could be argued that our methodology did not facilitate this, as the alternation of signals with and without FMM within a presentation block may not have provided a reliable strategy for exploiting FMM cues. Moreover, increasing the number of stimulus repetitions could further enhance certainty regarding the influence of FMM. Additionally, selecting stimuli with a high modulation depth could maximize the contrast between signals with and without FMM, potentially affecting recognition outcomes. Related to this, the choice of stimuli could consider the fact that FMM are notably pronounced during note transitions, so it would be especially intriguing to employ excerpts featuring transitions. Taken together, these revised methods could provide a clearer picture of the influence of FMM on the recognition of short sounds.

Conclusion

In contrast to previous studies, our work did not demonstrate a general recognition advantage for vocal sounds in the isolated presentation condition, nor did FMM influence recognition. Notably, recognition between vowels /a/ and /u/ differed considerably, which was linked to similarities with instrumental sounds. When the sounds were accompanied by a piano dyad, recognition accuracy and speed deteriorated. However, a distinctive robustness to interference for the recognition of /a/ sounds was observed, while a lack of robustness was observed for /u/. An acoustical model highlighted the role of spectral envelope cues in sound recognition. In summary, these findings demonstrated that vocal recognition is not mandatorily more efficient compared to instrumental sound recognition. This calls for a revised concept of vocal processing, emphasizing the need for a comprehensive understanding of the various acoustic factors influencing both vocal and instrumental sound recognition.

4.2.9 References

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, *131*(5), 4124–4133. https://doi.org/10.1121/1.3701865

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: *Insights from noise. Neuron, 66(4), 610–618*. https://doi.org/10.1016/j.neuron.2010.04.014

Akça, M., Vuoskoski, J. K., Laeng, B., & Bishop, L. (2023). Recognition of brief sounds in rapid serial auditory presentation. *PloS One*, *18*(4), e0284396. https://doi.org/10.1371/journal.pone.0284396

Belin, P , Zatorre, R. J, Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Belin, P, Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Belin, P , Zatorre, R. J, & Ahad, P (2002). Human temporal-lobe response to vocal sounds. *Brain Research. Cognitive Brain Research*, *13*(1), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

Bigand, E., Delbé, C., Gérard, Y., & Tillmann, B. (2011). Categorization of extremely brief auditory stimuli: Domain-specific or domain-general processes? *PloS One*, *6*(10), e27024. https://doi.org/10.1371/journal.pone.0027024

Bigoni F, Dahl S. Timbre discrimination for brief instrument sounds. In: 19th International Society for Music Information Retrieval Conference. ISMIR 2018; 2018. p. 128–134

Bruyer, R., & Brysbaert, M. (2013). Combining Speed and Accuracy in Cognitive Psychology: Is the Inverse Efficiency Score (IES) a Better Dependent Variable than the Mean Reaction Time (RT) and the Percentage Of Errors (PE)? Psychologica Belgica, 51(1), 5.

Bürgel, M., Picinali, L., & Siedenburg, K. (2021). Listening in the Mix: Lead Vocals Robustly Attract Auditory Attention in Popular Music. *Frontiers in Psychology*, *12*, 769663. https://doi.org/10.3389/fpsyg.2021.769663

Bürgel, M., & Siedenburg, K. (2023). Salience of Frequency Micro-modulations in Popular Music. *Music Perception*, *41*(1), 1–14. https://doi.org/10.1525/mp.2023.41.1.1

Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters, 405(3), 164–167*. https://doi.org/10.1016/j.neulet.2006.06.053

d'Allessandro, C. & M. Castellengo (1991) Etudes, par la synthese, de la perception du vibrato vocal dans les transitions de notes. Bulletin d'audiophonologie 7: 551-564

Eipert, L., Selle, A., & Klump, G. M. (2019). Uncertainty in location, level and fundamental frequency results in informational masking in a vowel discrimination task for young and elderly subjects. *Hearing Research, 377, 142–152.* https://doi.org/10.1016/j.heares.2019.03.015

Fine PA, Moore BCJ. 1993 Frequency analysis and musical ability. *Music Percept*. 11, 39–53. . https://doi.org/10.2307/40285598

Gao, Z., & Oxenham, A. J. (2022). Voice disadvantage effects in absolute and relative pitch judgments. *The Journal of the Acoustical Society of America*, *151*(4), 2414. https://doi.org/10.1121/10.0010123

Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., & Morillon, B. (2020). Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. *PLoS Biology, 18(3), e3000207.* https://doi.org/10.1371/journal.pbio.3000207

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47(1-2), 103–138.* https://doi.org/10.1016/0378-5955(90)90170-T

Gordon, M., & Poeppel, D. (2002). Inequality in identification of direction of frequency change (up vs. down) for rapid frequency modulated sweeps. *Acoustics Research Letters Online*, *3*(1), 29–34. https://doi.org/10.1121/1.1429653

Gunji, A., Koyama, S., Ishii, R., Levy, D., Okamoto, H., Kakigi, R., & Pantev, C. (2003). Magnetoencephalographic study of the cortical activity elicited by human voice. *Neuroscience Letters*, *348*(1), 13–16. https://doi.org/10.1016/s0304-3940(03)00640-2

Hutchins, S., & Campbell, D. (2009). Estimating the time to reach a target frequency in singing. *Annals of the New York Academy of Sciences*, *1169*, 116–120. https://doi.org/10.1111/j.1749-6632.2009.04856.x

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception & Psychophysics*, *76*(8), 2522–2530. https://doi.org/10.3758/s13414-014-0732-1

Hutchins, S., Roquet, C., & Peretz, I. (2012). The Vocal Generosity Effect: How Bad Can Your Singing Be? *Music Perception*, *30*(2), 147–159. https://doi.org/10.1525/mp.2012.30.2.147

Isnard, V., Chastres, V., Viaud-Delmon, I., & Suied, C. (2019). The time course of auditory recognition measured with rapid sequences of short natural sounds. *Scientific Reports*, *9*(1), 8005. https://doi.org/10.1038/s41598-019-43126-5

Kannyo, I., & DeLong, C. M. (2011). The effect of musical training on auditory perception. In Proceedings of Meetings on Acoustics (p. 25002). Acoustical Society of America. https://doi.org/10.1121/1.4733850

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2008). Informational masking. In Yost, W., Popper, A. N., and Fay, R. R. (edt), Auditory perception of sound sources, p. 143–189. Springer. Heidelberg, Germany.

Larrouy-Maestri, P., & Pfordresher, P. Q. (2018). Pitch perception in music: Do scoops matter? *Journal of Experimental Psychology. Human Perception and Performance*, *44*(10), 1523–1541. https://doi.org/10.1037/xhp0000550

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. *Neuroreport*, *12*(12), 2653–2657. https://doi.org/10.1097/00001756-200108280-00013

Marin, C. M., & McAdams, S [S.] (1991). Segregation of concurrent sounds. Ii: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. *The Journal of the Acoustical Society of America*, *89*(1), 341–351. https://doi.org/10.1121/1.400469

Maurer, D. (2016). *Acoustics of the Vowel*. Peter Lang CH. https://doi.org/10.3726/978-3-0343-2391-8

McAdams, S [S.] (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *The Journal of the Acoustical Society of America*, *86*(6), 2148–2159. https://doi.org/10.1121/1.398475

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. Hearing Research, 219(1-2), 36–47. https://doi.org/10.1016/j.heares.2006.05.004

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America*, *128*(1), 435–443. https://doi.org/10.1121/1.3397384

Moskowitz, H. S., Lee, W. W., & Sussman, E. S. (2020). Response Advantage for the Identification of Speech Sounds. *Frontiers in Psychology*, *11*, 1155. https://doi.org/10.3389/fpsyg.2020.01155

Mueller, S. T., Alam, L., Funke, G. J., Linja, A., Ibne Mamun, T., & Smith, S. L. (2020). Examining Methods for Combining Speed and Accuracy in a Go/No-Go Vigilance Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1)*, 1202–1206. https://doi.org/10.1177/1071181320641286

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS One*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Murray, M. M., Camen, C., Gonzalez Andino, S. L., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *26*(4), 1293–1302. https://doi.org/10.1523/JNEUROSCI.4511-05.2006

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology : CB*, *32*(7), 1470-1484.e12. https://doi.org/10.1016/j.cub.2022.01.069

Ogg, M., Slevc, L. R., & Idsardi, W. J. (2017). The time course of sound category identification: Insights from acoustic features. *The Journal of the Acoustical Society of America*, *142*(6), 3459. https://doi.org/10.1121/1.5014057

Oxenham, A. J., Fligor, B. J., Mason, C. R., & Kidd, G. (2003). Informational masking and musical training. The Journal of the Acoustical Society of America, 114(3), 1543–1549. https://doi.org/10.1121/1.1598197

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’. *Speech Communication, 41(1), 245–255*. https://doi.org/10.1016/S0167-6393(02)00107-3

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America*, *57*(S1), S5-S5. https://doi.org/10.1121/1.1995329

Reuter, C., Czedik-Eysenberg, I., Siddiq, S., and Oehler, M. (2018). Formant distances and the similarity perception of wind instrument timbres. ICMPC15/ESCOM10, pages 367–371.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Communication*, *46*(3-4), 405–417. https://doi.org/10.1016/j.specom.2005.01.010

Santoro, R., Moerel, M., Martino, F. de, Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Computational Biology, 10(1), e1003412*. https://doi.org/10.1371/journal.pcbi.1003412

Siedenburg, K., & McAdams, S [Stephen] (2017). The role of long-term familiarity and attentional maintenance in short-term memory for timbre. *Memory (Hove, England)*, *25*(4), 550–564. https://doi.org/10.1080/09658211.2016.1197945

Siedenburg, K., Saitis, C., McAdams, S [Stephen], Popper, A. N., & Fay, R. R. (Eds.). (2019). *Springer Handbook of Auditory Research*. *Timbre: Acoustics, Perception, and Cognition*. Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4

Siedenburg, K., Schädler, M. R., & Hülsmeier, D. (2019). Modeling the onset advantage in musical instrument recognition. *The Journal of the Acoustical Society of America, 146(6), EL523*. https://doi.org/10.1121/1.5141369

Siedenburg, K., Röttges, S., Wagener, K. C., & Hohmann, V. (2020). Can You Hear Out the Melody? Testing Musical Scene Perception in Young Normal-Hearing and Older Hearing-Impaired Listeners. *Trends in Hearing, 24*, 2331216520945826. https://doi.org/10.1177/2331216520945826

Siedenburg, K., Jacobsen, S., and Reuter, C. (2021). Spectral envelope position and shape in orchestral instrument sounds. The Journal of the Acoustical Society of America, 149(6):3715– 3727, https://doi.org/10.1121/10.0005088.

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., & Pressnitzer, D. (2014). Auditory gist: Recognition of very short sounds from timbre cues. *The Journal of the Acoustical Society of America*, *135*(3), 1380–1391. https://doi.org/10.1121/1.4863659

Suied, C., Susini, P., McAdams, S [Stephen], & Patterson, R. D. (2010). Why are natural sounds detected faster than pips? *The Journal of the Acoustical Society of America*, *127*(3), EL105-10. https://doi.org/10.1121/1.3310196

Sundberg, J. (1999). The Perception of Singing. In *The Psychology of Music* (pp. 171–214). Elsevier. https://doi.org/10.1016/b978-012213564-4/50007-x

Sundberg, J. (2013). Perception of Singing. In *The Psychology of Music* (pp. 69–105). Elsevier. https://doi.org/10.1016/B978-0-12-381460-9.00003-1

Tanner, W. P. (1958). What is Masking? *The Journal of the Acoustical Society of America, 30(10), 919–921*. https://doi.org/10.1121/1.1909406

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: An event-related potential and behavioral study. Experimental Brain Research, 161(1), 1–10. https://doi.org/10.1007/s00221-004-2044-5

Townsend, James & Ashby, F. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), Cognitive theory. Vol. 3. (pp. 200-239). Hillsdale, N.J.: Erlbaum.

Vandierendonck, A. A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behav Res 49*, 653–673 (2017). https://doi.org/10.3758/s13428-016-0721-5

Weiss, M. W., Bissonnette, A.-M., & Peretz, I. (2021). The singing voice is special: Persistence of superior memory for vocal melodies despite vocal-motor distractions. *Cognition*, *213*, 104514. https://doi.org/10.1016/j.cognition.2020.104514

Weiss, M. W., & Peretz, I. (2019). Ability to process musical pitch is unrelated to the memory advantage for vocal music. *Brain and Cognition*, *129*, 35–39. https://doi.org/10.1016/j.bandc.2018.11.011

Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way she sings: Enhanced memory for vocal melodies. *Psychological Science*, *23*(10), 1074–1078. https://doi.org/10.1177/0956797612442552

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear Mixed Models*. Chapman and Hall/CRC. https://doi.org/10.1201/b17198

Zendel BR, Alain C. 2009 Concurrent sound segregation is enhanced in musicians. J. Cogn. Neurosci. 21, 1488–1498. https://10.1162/jocn.2009.21140

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330 A6. This research was also supported by a Freigeist Fellowship of the Volkswagen Foundation to KS.

Author Contributions

MB and KS designed the study. MB collected and analyzed the data. MB wrote the first draft of the manuscript. KS revised the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics Approval

The experiments were approved by the ethics committee of the University of Oldenburg and adhere to the ASA Ethical Principles (https://acousticalsociety.org/ethical-principles). The participants provided their written informed consent to participate in this study. Participation was compensated monetarily.

Data availability

The pseudonymized raw data and supplementary material are available at the GitHub account of MB under the following link: https://github.com/MichelBuergel/Data/vocalRecognition.

Supplementary Material

See supplementary material at SuppPub1.docx for an extended description of our methods and results.

4.3 Synopsis

The lack of a general effect for FMM observed in the results illustrates that auditory processing uses or weights different cues dependent on the task at hand, as illustrated here by differences for the recognition of short sounds and the detection of sounds within mixtures. Additionally, the pronounced differences emerging for /a/ and /u/ sounds, which were linked to spectral similarities, imply that being a sound produced by the human vocal system is not a guarantee for salience. However, the influence of similarity was reduced in excerpts of singing voices from pop music that did not strictly incorporate only one vowel. The results raise questions about what evokes the advantage seen in vocals and whether there is a single attribute that shapes this advantage across different realms of auditory perception, or if this advantage is triggered by a multitude of attributes that differently impact the various facets of auditory perception.

5. ENHANCED SALIENCE OF EDGE FREQUENCIES IN AUDITORY PATTERN RECOGNITION

5.1 Introduction

The final study presented in this dissertation is motivated by the characteristic of bass instruments to attract auditory attention to a particularly low degree and the notable salience of high voices in polyphonic music. The methodology of Studies 1 and 2 was adopted, but this time the musical mixtures consisted of pseudo-randomized pure-tone melodies presented in separated frequency bands to investigate whether melodies in certain bands exhibit increased salience.

5.2 Study 4

This chapter is currently in press at *Attention, perception & psychophysics* as: Bürgel, M., Mares, D., Siedenburg, K. (2024) Enhanced salience of edge frequencies in auditory pattern recognition. The content of this chapter is identical to the manuscript.

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, analyzed the data, and wrote the manuscript. Diana Mares formulated the research question, participated in the study design carried out the experiments and wrote the manuscript. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

______________________________ 20.07.2024

(name) Date

Supervisor

5.2.1 Abstract

Within musical scenes or textures, sounds from certain instruments capture attention more prominently than others, hinting at biases in the perception of multi-source mixtures. Besides musical factors, these effects might be related to frequency biases in auditory perception. Using an auditory pattern recognition task, we here studied the existence of such frequency biases. Mixtures of pure tone melodies were presented in six frequency bands. Listeners were instructed to assess whether the target melody was part of the mixture or not, with the target melody presented either before or after the mixture. In Experiment 1, the mixture always contained melodies in five out of the six bands. In Experiment 2, the mixture contained three bands that stemmed from the lower or the higher part of the range. As expected, Experiment 1 and 2 both highlighted strong effects of presentation order with higher accuracies for the target presented before the mixture. Notably, Experiment 1 showed that edge frequencies yielded superior accuracies compared to center frequencies. Experiment 2 corroborated this finding by yielding enhanced accuracies for edge frequencies irrespective of the absolute frequency region. Our results highlight the salience of sound elements located at spectral edges within complex musical scenes. Overall, this implies that neither the high voice superiority effect, nor the insensitivity to bass instruments observed by previous research can be explained by absolute frequency biases in auditory perception.

5.2.2 Introduction

Most acoustic scenes in the real world comprise sounds from multiple sources. Consider the realm of music, wherein the experience of listening to compositions featuring multiple instruments playing simultaneously is commonplace. In these complex scenes, certain instruments and their sounds and melodies often seem more prominent, standing out amidst the complex texture of overlapping sounds. Think of sung melodies, the lines of wind instruments, or guitar riffs – these elements frequently capture our attention. By contrast, the melodies of bass instruments less frequently stand out. This might suggest a potential perceptual bias that inhibits the recognition of low-frequency sound elements or melodies. Here we sought to study such spectral biases in auditory scene analysis (ASA).

Principles of ASA (Bregman & McAdams, 1994) are key to understanding how the human auditory system differentiates sounds within complex auditory scenes. ASA encapsulates a range of processes through which the auditory system organizes sound elements by segregation and grouping to craft coherent mental representations known as streams, following Gestalt principles. This is achieved by utilizing external sound features (bottom-up), and internal cognitive processes of the listener (top-down).

When dissecting a musical scene, bottom-up processing entails recognizing spectrally distinct sounds, differences in their continuity (onset and offset), or variations in spectrum. These factors contribute to the interpretation of distinct instruments and melodies within the scene. Conversely, top-down processing relies on learned musical patterns, expectations, and familiarity with specific instrumentations or musical arrangements, aiding in scene parsing.

The role of auditory attention is the topic of an ongoing discussion (for a review see Snyder et al., 2012; Sussman, 2017). Despite discrepancies regarding the extent to which attention affects ASA, studies indicate that attention facilitates the organization of sounds (Alain & Arnott, 2000; Sussman, 2006) and that it can emphasize otherwise hidden elements in auditory scenes (Sussman & Steinschneider, 2009). Investigating the audibility of instruments within musical scenes, Bürgel et al., (2021) conducted an instrument detection experiment using popular music excerpts, building upon a paradigm by Bey and McAdams (2002). Participants were tasked with detecting target instruments embedded in excerpts of popular music featuring mixtures of instruments. Target sounds included various instrument categories, such as bass instruments and singing voices. In half of the mixtures, the target was absent. To cue participants to the target, an isolated target track identical to the one embedded in the mixture was presented. The study varied whether the target cue was presented first or after the mixture: In cases where the target was played first, participants could use prior knowledge gained from the cue to direct attention towards the target in the mixture, making detection dependent on both its acoustical salience and selective attention. When the mixture was presented first, no cue was given, and performance relied much more on the salience of the target. The comparison of presentation orders allowed us to isolate the impact of selective attention from the target’s inherent acoustical salience, revealing the specific influence of attentional gain.

Results indicated that the presentation order considerably impacted detection, with superior detection observed when the target was presented first. Notably, the extent of this impact varied across instrument categories, with vocal sounds exhibiting almost no effect and bass instruments displaying the most significant decrease in detection accuracy among all studied sounds. This suggested that when attention was not directed towards the bass, it was less likely perceived in the musical scene, suggesting the notion of a spectral hierarchy. This diminished accuracy for bass sounds persisted even when the sounds were aligned in sound level. One interpretation of these findings is that the human auditory system exhibits spectral biases, which have the greatest influence when freely listening into an auditory scene.

Studies investigating polyphonic music with multiple independent voices have consistently reported a high-voice superiority effect (HVSE; Crawley et al., 2002), which could be related to potential spectral biases. The high-voice superiority effect refers to the phenomenon where, in the context of multiple simultaneous melodies, the melody with the highest pitch trajectory is more prominent in the cortical responses of listeners, rendering it more salient in polyphonic scenes (Fujioka et al., 2005). This effect has been observed even in infants (Marie & Trainor, 2013) and may stem from physiological factors within the human auditory system, influenced by the interplay of harmonic structures in tone complexes (Trainor et al., 2014). Research indicates that musical training in instruments within the soprano range enhances this effect, whereas training in the bass range doesn’t reverse the effect,but can lead to an equalization between lower and higher voices (Marie et al., 2012). Moreover, studies on instruments in the bass range have reported that while bass instruments may yield challenges in melody perception, they exhibit superior time perception (Hove et al., 2014). Taken together, these findings appear to imply that the human auditory system possesses mechanisms favoring melodies of higher voices in musical scenes.

Given this background, we explored if ASA is subject to spectral biases, in which sounds within specific frequency regions possess a distinct salience in auditory pattern matching tasks. Specifically, we aimed to discern absolute and relative biases. Absolute biases would manifest as increased salience of bands in distinct frequency regions, whereas relative biases would manifest systematically at specific positions within the musical scene regardless of the absolute position in the auditory spectrum. To study this, the same task as in Bürgel et al. (2021) was used with acoustically more controlled stimuli where natural musical instruments were replaced by random pure tone melodies in different frequency bands.

We hypothesized that the presentation order between the target cue and mixture, as well as the frequency regions in which the target melodies appear, would impact detection accuracy. Specifically, due to the observed diminished performance of bass instruments, we expected the melodies in the lower frequencies to exhibit the poorest performance, with a significantly larger decline between the target-first and mixture-first presentation orders. Conversely, due to the reported high-voice superiority effect, we expected the melody in the highest frequency band to be more salient and outperform melodies in the low- and mid-frequency region. Additionally, we assumed that the high-voice superiority effect would lead to better performance for melodies in high frequency bands, highlighting biases towards specific absolute frequency regions.

5.2.3 Experiment 1

Methods

Participants

A total of 26 normal-hearing participants took part in the experiment. While formal power analyses were not conducted, the sample size was based on previous studies employing a similar detection paradigm (Bürgel et al., 2021, *Order effect - Experiment 1:* $\beta = -0.710$*, t = 15.542, p < 0.001; Bürgel & Siedenburg, 2023: Order effect - Experiment 1:* $\beta = -0.545$*, t = 9.356, p < 0.001)*. One participant was excluded due to obtaining a below-chance performance in the Target-Mixture condition. Of the remaining 25 subjects (mean age=26.5, SD=5.23; age range=19-38, diverse = 4, female = 7, male=14), 20 were categorized as musicians. In our study, musicianship was assigned based on whether an individual had received three or more years of formal training on a musical instrument, including voice. Additionally, participants' musical abilities were evaluated using questions from subscales of the Goldsmith Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014). The average Gold-MSI score for the Perceptual Abilities subscale was 53.0 (SD=8.52) for musicians and 39.8 (SD=11.26) for non-musicians and the one for the Musical Training subscale was 35.85 (SD=5.45) for musicians and 15.8 (SD=4.38) for non-musicians.

Stimuli

A schematic of the stimuli is presented in Figure 5.1. To mitigate energetic masking, each melody occupied a designated frequency space, maintaining sufficient frequency distances between melodies. Six frequency bands were used in the experiments, which were spaced on the equivalent rectangular bandwidth (ERB) scale (Moore & Glasberg, 1983) with center frequencies between 65 and 2080 Hz. The target was present in half of the trials. Because the absence of the target resulted in a less dense mixture, one random band was muted each time the target was present in the mixture.

The stimuli were created on the fly (open set design). Each frequency band contained eight pseudo-randomized pure-tone melodies, each lasting two seconds. Tone frequencies were drawn randomly from a uniform distribution on a logarithmic frequency scale with half-octave range, anchored at the center frequencies of the six ERB bands (65, 215, 441, 783, 1300, 2080 Hz). Frequencies were drawn with no constraints applied to musical intervals or semitone steps. For instance, melodies in the 1300 Hz band could encompass notes ranging from 1300 to 1838 Hz (1300 Hz + ½ octave). This approach ensured that frequencies spanned a sufficient frequency range to be discernible as melodies while maintaining sufficient spacing between bands. The resulting frequency distance between neighboring tones was at least 1 ERB to mitigate the potential effects of energetic masking. Tones had random durations, with onset and offset timepoints of the eight-tone sequence generated by drawing seven timepoints from a uniform distribution between 0 and 2. These randomly drawn timepoints were sorted in ascending order, and drawings were discarded that contained durations shorter or equal than 50 ms. Timepoint n would here serve as offset of tone n and onset of tone n+1; the onset of the first tone of the sequence was defined as t=0 ms and the offset of the last as t=2000 ms. Tones were then synthesized with the defined onset/offset times and separated by 10 ms offset and 10 ms onset cosine ramps.

To eliminate sound level cues all sounds were aligned in level using A-weighting and every band was presented at a level of 40 dB SPL (A). In the initial section of the experiment, the target was followed by the mixture (Target-Mixture condition), while in the final section, the order was reversed (Mixture-Target condition). The frequency conditions were randomized.

Procedure

First, participants were informed about the experiment and provided informed consent. A short training session followed, during which participants received feedback. Afterwards, the main experiment started, and no feedback was provided anymore. After the completion of the main experiment, each participant filled out a brief questionnaire comprising demographic data and components of the Gold-MSI. Participants were financially compensated.

Experiment 1 was divided into two blocks: Target-Mixture and Mixture-Target. Each block contained all six frequency conditions. The structure of a trial was as follows: presentation of a target sequence (2 s), short pause (1 s), presentation of a mixture (2 s). The second block maintained this structure but interchanged the presentation of the target with the one of the mixture. In both instances, the task was to detect whether the target was present in the mixture (yes/no task).

The Target-Mixture block comprised a total of 120 trials, with 20 trials allocated for each frequency condition. Following a short break, the experiment proceeded with the Mixture-Target block, throughout which 240 trials, with 40 trials per frequency condition, were presented. This second section included a break after the first 120 trials. In both blocks, the number of instances in which the target was present or absent from the mixture was counterbalanced between frequency conditions. The different number of stimuli between two order conditions was due to experience from pilot experiments: It was anticipated that detection performance would approach ceiling level for many participants in the Target-Mixture condition, while performance in the Mixture-Target condition was expected to be lower. Therefore, for the sake of precise performance estimates, we increased the number of stimuli in the Mixture-Target condition while the Target-Mixture condition acted as a control condition.

In the initial version of the experiment the mixture comprised six simultaneous melodies (one in each frequency band) when the target was present and five when the target was absent. There was a concern that participants might adopt a strategy of counting the number of melodies instead of focusing on detecting the target melody. To prevent this potential strategy, a second version of the experiment was conducted. In this version, when the target was present, one random non-target melody frequency band was muted. This adjustment ensured a consistent density of five melodies in the mixture, regardless of the target's presence. As there were no statistically significant differences found between the two versions, with a correlation of $R^2 = 0.94$ between the mean results of both variants of the experiment, the results of all participants were analyzed conjointly.

Apparatus

The experiment took place in a double-walled sound booth at the University of Oldenburg. Participants sat in a comfortable chair and interacted with the experiment using a touch screen attached to a movable arm in front of them. Stimuli were processed through an RME Fireface UCX soundcard at a 44.1 kHz sampling rate and presented on Sennheiser HD 650 headphones. Stimuli were synthesized in Matlab. Sound levels were measured with a Brüel & Kjær Type 2250 light sound-level meter and a Brüel & Kjær Type 4153 artificial ear to which the headphones were coupled.

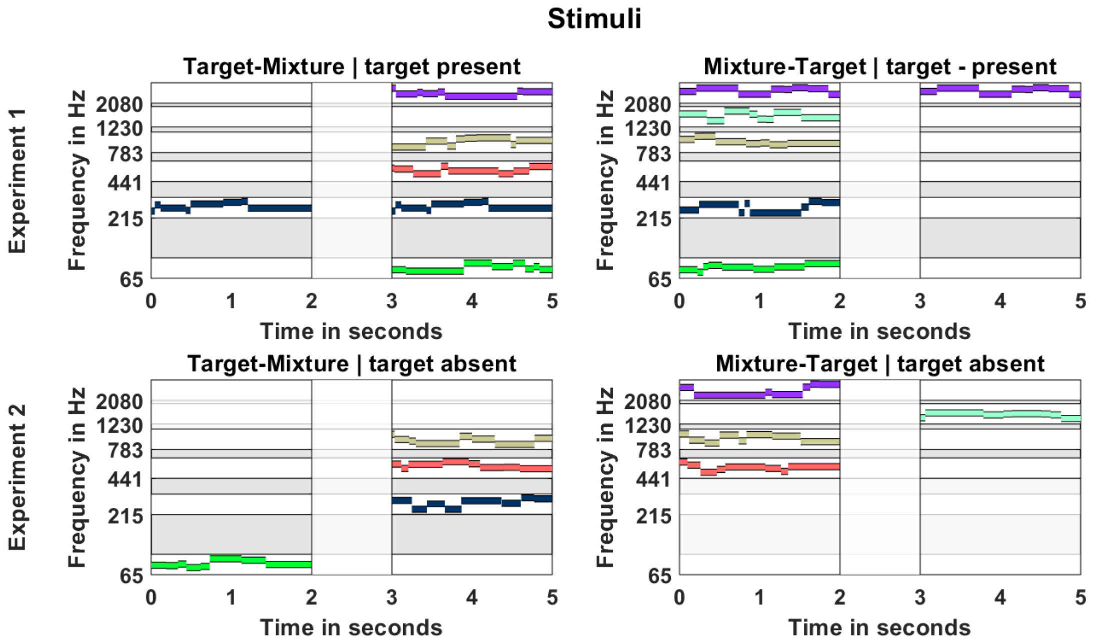

Figure 5.1: Illustration of the stimuli

Isolated 2-second target melodies were presented either before a 2-second mixture (Target-Mixture) or afterwards (Mixture-Target), with a 1-second pause in between. The target could be present or absent from the mixture. In Experiment 1, five melodies played in the mixture occurring in separated frequency bands. In Experiment 2, the frequency bands were split into a low and high range, each containing four frequency bands, and only three melodies were playing simultaneously. The colored lines represent example melodies in each band. The gray area represents the empty space where no melody could occur.

Results and discussion

The average performance in Experiment 1 is displayed as d-prime scores in Figure 5.2A (numerical values are available in supplementary Table 1). When comparing the presentation orders by averaging over frequency conditions, the Target-Mixture condition yielded clearly better detection scores, with an average score of d’ = 2.68, compared to the Mixture-Target condition with an average score of d’= 1.58 (-1.10).

Differences between melodies at the edge of the frequency range (lowest and highest) and those in the middle revealed an edge effect, with edge frequency bands yielding an average score of d'=2.54 compared to mid frequencies with an average score of d’=1.93 (-0.61). Edge effects were pronounced in both presentation orders, with scores for edges in the Target-Mixture condition of d’=2.95, compared to scores for mid frequencies of d’=2.55 (-0.40), and for edges in the Mixture-Target condition with a score of d’=2.12, compared to scores for mid frequencies of d’= 1.31 (-0.81). To further evaluate the observed effects, the difference between the performance of edge and middle frequencies was computed for each participant separately. The magnitude of individual edge effects is displayed in Figure 5.2B. Four participants showed differences close to zero (d’ < 0.1), wherein one participant exhibited better performance for the center frequencies with a reversed edge effect of d’ = -0.26. Overall, however, the majority of participants demonstrated clear edge effects.

A Linear Mixed Effects model (LME) was employed to analyse the data, incorporating random intercepts for each participant. Presentation order and categorization of whether targets appeared as edge frequencies vs. in middle frequencies were used as binary predictors. Musical sophistication scores and the frequency band in which the melody appeared were used as numerical predictors. The factors presentation order and edge frequency showed pronounced effects as well as an interaction (Order: β = 0.514, t = 13.479, p < 0.001; Edge: β = 0.303, t = 7.947, p < 0.001; Interaction: β = 0.106, t = 2.765, p = 0.005). The effects of frequency band and musical sophistication were negligible (frequency band: β = 0.042, t = 0.304, p = 0.761 ; Musical perception: β = 0.004, t = 0.268, p = 0.788; Musical training: β = 0.074, t = 0.401, p = 0.522). Further underlining the lack of impact of musical sophistication, correlations between participants' averaged d-prime scores and sophistication scores revealed R^2 values below 0.05 for both musical perception and musical training.

To further investigate the effect of local spectral edges caused by muting frequency bands adjacent to the target, we compared the results of detecting targets where a band adjacent to the target band was omitted and melodies where a band not adjacent to the target was omitted (see supplementary Figure 1). The results showed a close alignment in detection performance between the two conditions, with no considerable differences observed in the LME (main effect: β = -0.064, t = -0.169, p = 0.866; interaction with presentation order: β = -0.167, t = -0.699, p = 0.485). These findings suggest that there was no discernable impact of local edges resulting from the removal of aligned frequency bands. Additionally, the absence of differences between both conditions can be interpreted as an indication that energetic masking did not substantially contribute to the detection process. If energetic masking had a substantial effect, conditions with a missing neighbor would have performed better than those with neighbors present on both sides.

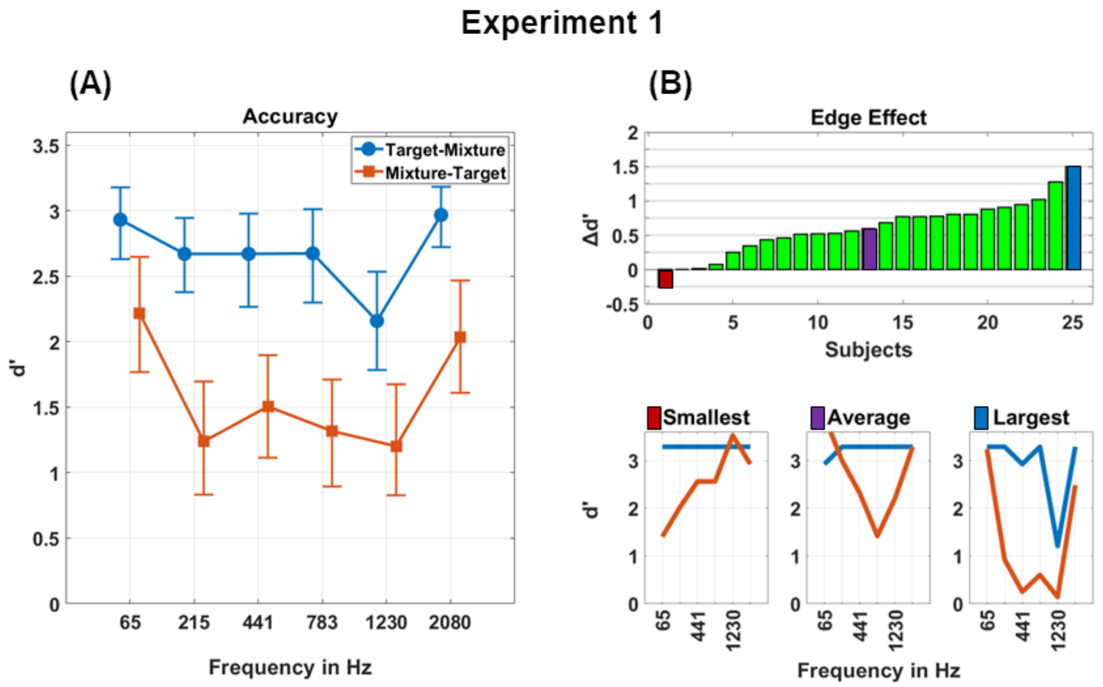

Figure 5.2: Detection accuracy in Experiment 1

(A) *Detection accuracy in Experiment 1 is represented as d' scores. The square and circle marks denote the mean scores for melodies in the specified frequency bands. The square marks indicate the presentation order "Target-Mixture," where the target cue was presented first followed by a mixture. The circle marks indicate the presentation order "Mixture-Target," where a mixture was presented first followed by the target cue. Error bars represent 95% CIs computed using a bootstrapping method.* ***(B)*** *For each subject the magnitude of the edge effect is displayed as the difference between mean d' scores for melodies in the first and last frequency band and mean d' scores for melodies between the first and last frequency band.* The bottom figures display individual d' scores for the subject with the smallest, the average, and the largest edge effect respectively.

Taken together, the observed order effect corroborates previous reports on the facilitating role of attention in auditory scene analysis (Alain & Arnott, 2000; Bey & McAdams, 2002; Woods & McDermott, 2015; Bürgel et al., 2021;). Prior knowledge could be used to direct auditory attention towards the target sound and thus follow and highlight auditory representations in the auditory scene. Contrary to our hypothesis , neither a general bias towards higher frequencies, nor an unawareness of activity in lower frequencies was observed. Instead, the results revealed a superior detection for both edges of the auditory scene. These results raise the question of whether an effect of relative frequency bias was at play, wherein the spectral edges of the acoustic scene were better recognized compared to sounds located closer to the frequency-center of the scene, or whether the resulting pattern was due to an absolute frequency bias in auditory pattern matching.

5.2.4 Experiment 2

In Experiment 2, we aimed to explore whether the previously observed edge effect persisted despite global changes in frequency. For this purpose, the frequency range was divided into two equal ranges. The first range comprised frequency bands one to four, and the second range comprised frequency bands three to six.

Participants

30 normal-hearing participants with a mean age of 24.7 years (SD=2.54, range=20-29; diverse = 2, female = 17, male=11) performed the second experiment. Following the predefined criterion, 14 were designated as musicians. No participant was excluded. The average Gold-MSI score for the Perceptual Abilities subscale was 53.07 (SD=7.26) for musicians and 44.31 (SD=7.11) for non-musicians and the one for the Musical Training subscale was 32.42 (SD=5.97) for musicians and 13.37 (SD=6.67) for non-musicians.

Stimuli and procedure

Experiment 2 consisted of four blocks: two Target-Mixture blocks, and two Mixture-Target blocks. One block of each order condition contained conditions in the lower frequency range (bands 1-4) while the other block contained the ones in the higher range (bands 3-6). The order of the frequency ranges was counterbalanced, such that each participant began either with the low or the high range in the Target-Mixture block. Each condition was presented 20 times in each Target-Mixture block, which summed up to a total of 160 trials for the first two blocks. The Mixture-Target blocks comprised twice the number of trials compared to the Target-Mixture block, with a total of 320 trials. After each block, a short break followed. The mixture density was constrained to three simultaneous frequency bands across all trial types. The training section involved eight Target-Mixture trials -- four in the lower frequency range and four in the higher frequency range.

Results and discussion

Results are displayed in Figure 5.3A (numerical values are available in supplementary Table 2). Similar to Experiment 1, when examining the difference between presentation orders by averaging over frequency and frequency range conditions, the Target-Mixture condition showed clearly higher scores (d’ = 2.75), compared to the Mixture-Target condition (d’= 2.07).

When examining the averages within each frequency range, the low frequency range exhibited a slightly better score of d’=2.41 compared to the high range with a score of d’=2.36 (-0.05). Differences between frequency conditions at the edge of the frequency ranges and those in the center of the ranges revealed an edge effect, with edge frequencies achieving better detection with an average score of d’=2.66 compared to mid frequencies with an average score of d’=2.10 (-0.56). Negligible differences within the edge effects between presentation orders or intervals were observed, with scores for the Target-Mixture condition in low and high ranges both being d’=2.93, and for the

Mixture-Target condition in low and high ranges being d’=2.42 and d’=2.36, respectively. To further evaluate the observed effects, the difference between the performance of edge and middle frequencies was computed for each participant separately. The magnitude of individual edge effects is displayed in Figure 5.3B. Three participants showed differences close to zero (d’ < 0.1), whereas two participant exhibited better performance for the center frequencies with reversed edge effects of Δd’ = -0.26 and Δd’ = -0.09. Overall, the majority of participants demonstrated an edge effect, whereby the most pronounced effect showed an enhancement of Δd’ = 1.22.

Participants with higher musical sophistication scores showed better detection accuracy. Correlations based on musical perception scores yielded an R2 of 0.16, while those based on musical training scores yielded an R2 of 0.36. Differences between order conditions were apparent, with an R2 of 0.06 for perception and an R2 of 0.18 for training in the target-mixture condition.

Larger correlations were observed in the mixture-target condition, reaching an R2 of 0.21 for perception and an R2 of 0.43 for training. The LME used in Experiment 2 utilized the same fixed effects as in Experiment 1, with the addition of a binary variable determining whether the stimulus appeared in the low or high-frequency range. Effects were pronounced for presentation order and whether frequencies appeared on the edges of a frequency range, as well as for musical training scores (Order: β = 0.353, t = 13.064, p < 0.001; Edge: β = 0.267, t = 9.790, p < 0.001; Musical training: β = 0.028, t = 3.084, p = 0.003; Frequency band: β = 0.031, t = 1.164, p = 0.245; Frequency range: β = 0.023, t = 0.767, p = 0.443; Musical perception: β = 0.015, t = 1.226, p = 0.297).

As a consequence of the experimental design, there were trials where the melody in the first or last frequency band was muted, resulting in the target melody in the second or third band taking the role of an edge frequency. For example, if frequency band 1 was muted, band number 2 became the lowest frequency in the mixture. In the reported results above, only instances where target melodies were truly embedded in the center of the musical scene, with melodies in frequency bands above and below the target frequency band, were analyzed. To investigate how the relative position of frequency bands impacted the detection of melodies, a separate analysis was conducted. This involved analyzing separate hit rates for both variants using a generalized linear effects model to account for the constrained nature of hit rates. The model employed the same effects as the Linear Mixed-Effects Model (LME), except hit rates were used as a response variable. As observed for frequency bands on the edge of the frequency range, an edge effect was evident even within the same frequency band (F = 13.059, p < 0.001). Precisely, melodies on the edge outperformed melodies in the center, achieving hit rates of 89 [0.82 – 0.94] percentage points compared to the 75 [0.66 – 0.83] percentage points achieved by melodies in the center (see supplementary Figure 2).

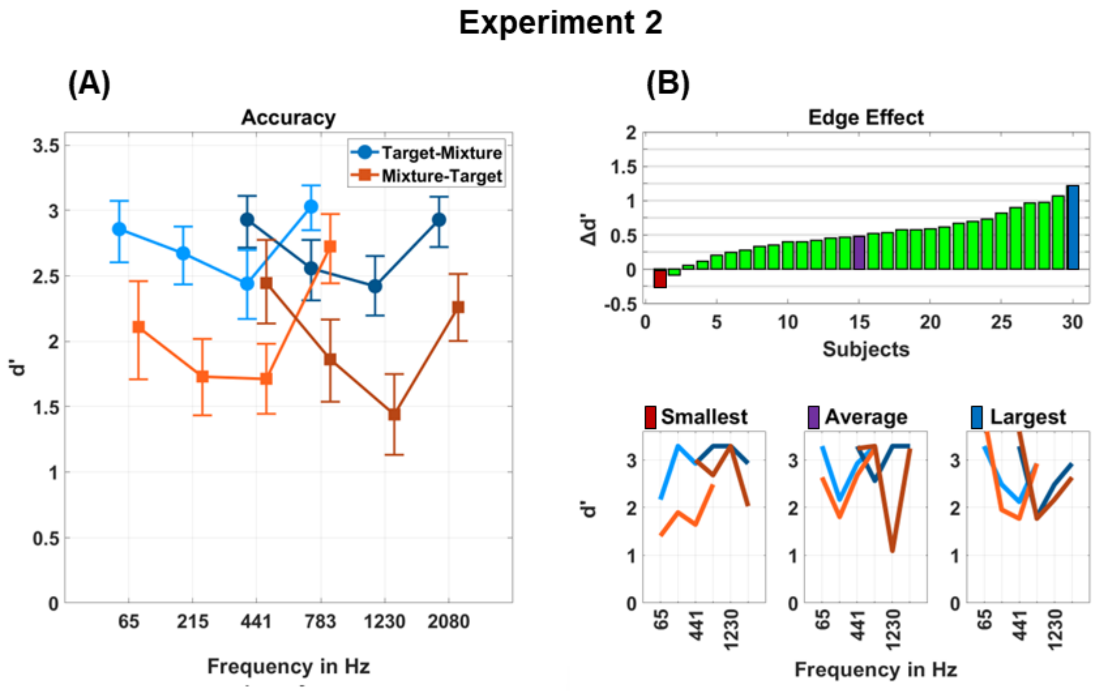

Figure 5.3 Detection accuracy in Experiment 2

(A) *Detection accuracy in Experiment 2 is represented as d' scores. Three melodies were presented simultaneously within either a low or high frequency range. The brighter color indicates the low-frequency range (65 - 783 Hz), and the darker color represents the high-frequency range (441 - 2080 Hz). The square and circle marks denote the mean scores for target melodies in the specified frequency bands. The square marks indicate the presentation order 'Target-Mixture,' where the target cue was presented first followed by a mixture. The circle marks indicate the presentation order 'Mixture-Target,' where a mixture was presented first followed by the target cue. Error bars indicate 95% CIs computed using a bootstrapping method. (B) For each subject the magnitude of the edge effect is displayed as the difference between mean d' scores for melodies in the first and last frequency bands and mean d' scores for melodies in the second and third frequency bands. The bottom figures display individual d' scores for the subject with the smallest, the average, and the largest edge effect respectively.*

Overall, Experiment 2 revealed pronounced edge effects across different frequency regions. Furthermore, a weak positive effect of musical training was evident, suggesting that individuals with higher musical sophistication scores also have improved abilities for the detection of melodies in complex acoustical scenes. The absence of this impact observed in Experiment 1 may be attributed to the participant pool, primarily consisting of musicians with a small range of musical sophistication scores.

5.2.5 General Discussion

In line with our hypotheses and consistent with previous research (Bey & McAdams, 2002; Bürgel et al., 2021), the presentation order of the target played a crucial role in target detection, highlighting the influence of top-down processing on ASA. Presented with the target before the mixture, listeners were able to leverage this information to selectively direct attention towards the melody in the target frequency band, which resulted in a higher detection accuracy compared to the order where the target was presented after the mixture. Our hypothesis regarding perceptual biases towards specific absolute frequency regions, on the other hand, was clearly refuted. Instead of an inferior detection of low frequency bands or superior detection of high frequency bands, detection was facilitated for frequencies that appeared on the edges of the acoustical scene, implying biases towards relative frequency regions. This edge effect was consistent for melodies on the relative outer frequency bands of the acoustical scene, regardless of absolute frequency. The effect was even more pronounced in the more difficult Mixture-Target condition. We interpret this effect as a salience phenomenon, where the melodies at the spectral edges attract auditory attention, thereby detracting from melodies in between. This effect particularly shapes perception when listening to the scene holistically without prior target (Mixture-Target). In line with these findings, participants informally stated that they had used the outer melodies as landmarks to subsequently hear out melodies between these. It would be interesting to explore whether the occurrence of edge effects is influenced by acoustic cues within the melodies, such as tone onsets and frequency trajectories, or if the edge effects can be fully attributed to the frequency bands. This could be investigated by examining whether such effects persist even when the melodies are replaced with static tones within the individual frequency bands.

MSI scores had a positive effect on melody detection (Müllensiefen et al., 2014). While this effect was not observed in the studies by Bey (2002) and Bürgel (Bürgel et al., 2021), it aligns with numerous reports in the literature where musical training scores are positively associated with performance in music related tasks. It is reported that individuals with higher scores exhibited better perceptual abilities, such as pitch and rhythm discrimination (Micheyl et al., 2006; Marozeau et al., 2010 ;Kannyo & DeLong, 2011), as well as the ability to recognize melodies and instruments in musical scenes (Crawley et al., 2002; Slater & Marozeau, 2016; Siedenburg et al., 2020). Musicians' auditory skills may have enabled them to better isolate individual melodies, allowing them to search the scene for the target melody more efficiently. Nevertheless, better recognition by musically trained individuals did not compensate for the order effect or the edge effects, indicating that these phenomena are fundamental even for high-performing individuals.

The enhanced detection of edges, where no difference was observed between the lower and upper edges, may initially appear at odds with the high-voice superiority effect (HVSE; Fujioka et al., 2005; Marie et al., 2012; Marie & Trainor, 2013; Marie & Trainor, 2014; Hove et al., 2014). However, studies on the HVSE suggest that it is grounded in the nonlinear processing of harmonics in tone complexes, where the harmonics of the higher voice are more likely to mask those of the lower voice (Trainor et al., 2014). In our experiment, the use of pure tones without harmonic overtone structures thus may account for the absence of this perceptual hierarchy. Comparing this study with HVSE studies suggests additional discrepancies. HVSE studies typically involve only two simultaneous melodies, which could be interpreted as a comparison between the lower and upper edges, which exhibited no significant differences in our experiment. However, such an interpretation would disregard the potential influence of other melodies in the center of the presented ranges. Furthermore, the edge effect stands in striking contrast to our findings using acoustic excerpts of pop music that comprised sounds from real musical instruments, wherein bass instruments had the most pronounced detection differences between both presentation orders and also the lowest detection score among tested instruments (Bürgel et al., 2021), even though the sound levels of the instruments were equalized. The present edge effect, however, challenges the notion that the diminished performance of bass instruments is solely due to spectral biases (whether absolute or relative). Several factors may contribute to these discrepancies. Unlike the pure tones used in our experiment, sounds of bass instruments comprise tone complexes that spectrally overlap with other instruments in the musical scene. This overlap might result in the lower voice being more easily masked by the higher ones, as explained by HVSE. Importantly, in contrast to the randomly generated melodies in our experiment, which had no systematic relationship to each other, instruments and their lines or melodies in musical compositions are often created in a deliberate relationship to each other, both in pitch and in time. Bass instruments, in particular, often provide the harmonic basis of tonal music, thus support other melodies besides providing rhythmic accents, which associates them more with rhythm and groove perception (Hove et al., 2014). However, these musical properties also make bass instruments more likely to be part of the musical background and avoid standing out.

The results obtained in our two experiments are concordant with research investigating the importance of individual frequency components to the perceived loudness of multitone complexes (Leibold & Jesteadt, 2007; Oberfeld et al., 2012; Jesteadt et al., 2017). By using a method called perceptual weight analysis, these studies have typically presented sound ranges with constant overall loudness but random trial-by-trial frequency level variation, and subsequently obtained weights by computing the correlation between these variations and the responses (Joshi et al., 2016). Consistently, these studies have shown that higher weights are given to the lower and the higher frequencies than to the middle frequencies, indicating a more substantial contribution of the edges to the overall loudness of a complex. Moreover, similar to the conclusions of our second experiment, the increased saliency of the edges seems to depend not on the absolute value of these frequencies, but rather on their relative position in a complex.

Beyond the auditory domain, contrasting and analogous effects emerge in the visual domain (for reviews see Marisa Carrasco, 2011; Marisa Carrasco, 2018). Unlike the observed facilitated recognition of melodies at the edges of the acoustic scene, spatial resolution in the visual domain adheres to a contrasting hierarchy, wherein central objects are identified more accurately and quickly, and in more detail than objects in the periphery or edges (Rijsdijk et al., 1980; Cannon, 1985). A decisive influence of attention has been well reported that allows for perceptual higher resolutions and thus the recognition of finer details at attended locations (Lee et al., 1997; Dosher & Lu, 2000; L. Huang & Dobkins, 2005). Moreover, directing attention through prior cues has been shown to compensate for the preferential focus on central objects in the visual scene, enabling objects at the periphery to be brought into focus (M. Carrasco & Yeshurun, 1998). This parallels observations in the acoustical realm (Alain & Arnott, 2000, Bey & McAdams, 2002) and mirrors our results, where auditory cues could be leveraged to spotlight specific frequency regions.

In conclusion, the results of our study suggest the absence of distinct biases towards absolute frequency regions. Instead, we observed a relative frequency bias with a specific salience of edge frequencies in auditory pattern matching. Future work should probe the generality of these findings and the ways in which they extrapolate to naturalistic acoustic scenes.

5.2.6 References

Bregman, A. S. & McAdams, S. Auditory scene analysis: The perceptual organization of sound. *The Journal Acoustical Society America* 95, 1177–1178 (1994). DOI 10.1121/1.408434.

Snyder, J. S., Gregg, M. K., Weintraub, D. M. & Alain, C. Attention, awareness, and the perception of auditory scenes. *Frontiers psychology* 3, 15 (2012). DOI 10.3389/fpsyg.2012.00015.

Sussman, E. S. Auditory scene analysis: An attention perspective. *Journal speech, language, hearing research : JSLHR* 60, 2989–3000 (2017). DOI 10.1044/2017_JSLHR-H-17-0041.

Alain, C. & Arnott, S. R. Selectively attending to auditory objects. *Frontiers bioscience : a journal virtual library* 5, D202–12 (2000). DOI 10.2741/alain.

Sussman, E. S. Multiple mechanisms of auditory attention. *The Journal Acoustical Society America* 119, 3416 (2006). DOI 10.1121/1.4786818.

Sussman, E. & Steinschneider, M. Attention effects on auditory scene analysis in children. *Neuropsychologia* 47, 771–785 (2009). DOI 10.1016/j.neuropsychologia.2008.12.007.

Bürgel, M., Picinali, L. & Siedenburg, K. Listening in the mix: Lead vocals robustly attract auditory attention in popular music. *Frontiers psychology* 12, 769663 (2021). DOI 10.3389/fpsyg.2021.769663.

Bey, C. & McAdams, S. Schema-based processing in auditory scene analysis. *Perception & psychophysics* 64, 844–854 (2002). DOI 10.3758/bf03194750.

Crawley, E. J., Acker-Mills, B. E., Pastore, R. E. & Weil, S. Change detection in multi-voice music: The role of musical structure, musical training, and task demands. *Journal Experimental Psychology: Human Perception Performance* 28, 367–378 (2002). DOI 10.1037/0096-1523.28.2.367.

Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R. & Pantev, C. Automatic encoding of polyphonic melodies in musicians and nonmusicians. *Journal cognitive neuroscience* 17, 1578–1592 (2005). DOI 10.1162/089892905774597263.

Marie, C. & Trainor, L. J. Development of simultaneous pitch encoding: infants show a high voice superiority effect. *Cerebral cortex (New York, N.Y. : 1991)* 23, 660–669 (2013). DOI 10.1093/cercor/bhs050.

Trainor, L. J., Marie, C., Bruce, I. C. & Bidelman, G. M. Explaining the high voice superiority effect in polyphonic music: evidence from cortical evoked potentials and peripheral auditory models. *Hearing research* 308, 60–70 (2014). DOI 10.1016/j.heares.2013.07.014.

Marie, C., Fujioka, T., Herrington, L. & Trainor, L. J. The high-voice superiority effect in polyphonic music is influenced by experience: A comparison of musicians who play soprano-range compared with bass-range instruments. *Psychomusicology: Music, Mind, Brain* 22, 97–104 (2012). DOI 10.1037/a0030858.

Hove, M. J., Marie, C., Bruce, I. C. & Trainor, L. J. Superior time perception for lower musical pitch explains why bass-ranged instruments lay down musical rhythms. *Proceedings National Academy Sciences United States America* 111, 10383–10388 (2014). DOI 10.1073/pnas.1402039111.

Glasberg, B. R. & Moore, B. C. Derivation of auditory filter shapes from notched-noise data. *Hearing research* 47, 103–138 (1990). DOI 10.1016/0378-5955(90)90170-T.

Müllensiefen, D., Gingras, B., Musil, J. & Stewart, L. The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one* 9, e89642 (2014). DOI 10.1371/journal.pone.0089642.

Woods, K. J., & McDermott, J. H. (2015). Attentive Tracking of Sound Sources*. Current Biology, 25(17)*, 2238–2246. https://doi.org/10.1016/j.cub.2015.07.04

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A. & Schröger, E. Pitch discrimination accuracy in musicians vs nonmusicians: an event-related potential and behavioral study. *Experimental brain research* 161, 1–10 (2005). DOI 10.1007/s00221-004-2044-5.

Micheyl, C., Delhommeau, K., Perrot, X. & Oxenham, A. J. Influence of musical and psychoacoustical training on pitch discrimination. *Hearing research* 219, 36–47 (2006). DOI 10.1016/j.heares.2006.05.004.

Kannyo, I. & DeLong, C. M. The effect of musical training on auditory perception. Proceedings of Meetings on Acoustics, 025002 (Acoustical Society of America, 2011). DOI 10.1121/1.4733850.

Marozeau, J., Innes-Brown, H., Grayden, D. B., Burkitt, A. N. & Blamey, P. J. The effect of visual cues on auditory stream segregation in musicians and non-musicians. *PloS one* 5, e11297 (2010). DOI 10.1371/journal.pone.0011297.

Slater, K. D. & Marozeau, J. The effect of tactile cues on auditory stream segregation ability of musicians and nonmusicians. *Psychomusicology: Music, Mind, Brain* 26, 162–166 (2016). DOI 10.1037/pmu0000143.

Siedenburg, K., Röttges, S., Wagener, K. C. & Hohmann, V. Can you hear out the melody? testing musical scene perception in young normal-hearing and older hearing-impaired listeners. *Trends hearing* 24, 2331216520945826 (2020). DOI 10.1177/2331216520945826.

Leibold, L. J. & Jesteadt, W. Use of perceptual weights to test a model of loudness summation. *The Journal Acoustical Society America* 122, EL69 (2007). DOI 10.1121/1.2761918.

Oberfeld, D., Heeren, W., Rennies, J. & Verhey, J. Spectro-temporal weighting of loudness. *PloS one* 7, e50184 (2012). DOI 10.1371/journal.pone.0050184.

Jesteadt, W. *et al.* Relative contributions of specific frequency bands to the loudness of broadband sounds. *The Journal Acoustical Society America* 142, 1597 (2017). DOI 10.1121/1.5003778.

Joshi, S. N., Wróblewski, M., Schmid, K. K. & Jesteadt, W. Effects of relative and absolute frequency in the spectral weighting of loudness. *The Journal Acoustical Society America* 139, 373–383 (2016). DOI 10.1121/1.4939893.

Carrasco, M. Visual attention: the past 25 years. *Vision research* 51, 1484–1525 (2011). DOI 10.1016/j.visres.2011.04.012.

Carrasco, M. How visual spatial attention alters perception. *Cognitive processing* 19, 77–88 (2018). DOI 10.1007/s10339- 018-0883-4.

Rijsdijk, J. P., Kroon, J. N. & van der Wildt, G. J. Contrast sensitivity as a function of position on the retina. *Vision research* 20, 235–241 (1980). DOI 10.1016/0042-6989(80)90108-x.

Cannon, M. W. Perceived contrast in the fovea and periphery. *Journal Optical Society America. A, Optics image science* 2, 1760–1768 (1985). DOI 10.1364/JOSAA.2.001760.

Lee, D. K., Koch, C. & Braun, J. Spatial vision thresholds in the near absence of attention. *Vision research* 37, 2409–2418 (1997). DOI 10.1016/s0042-6989(97)00055-2.

Dosher, B. A. & Lu, Z. L. Mechanisms of perceptual attention in precuing of location. *Vision research* 40, 1269–1292 (2000). DOI 10.1016/s0042-6989(00)00019-5.

Huang, L. & Dobkins, K. R. Attentional effects on contrast discrimination in humans: evidence for both contrast gain and response gain. *Vision research* 45, 1201–1212 (2005). DOI 10.1016/j.visres.2004.10.024.

Carrasco, M. & Yeshurun, Y. The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal Experimental Psychology: Human Perception Performance* 24, 673–692 (1998). DOI 10.1037//0096- 1523.24.2.673.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330 A6. This study was also supported by a Freigeist Fellowship of the Volkswagen Foundation to Kai Siedenburg.

Author contributions statement

All authors conceived the experiment. DM and MB set up the experiment, DM collected the data. DM, MB and KS analyzed the data. DM and MB wrote the first manuscript draft. All authors reviewed and edited the manuscript.

Additional information

Data availability

All experimental data and analysis scripts can be downloaded under the following link: https://github.com/Music-Perception-and-Processing/SpectralPreferences

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics Approval

The experiments were approved by the ethics committee of the University of Oldenburg and adhere to the ASA Ethical Principles (https://acousticalsociety.org/ethical-principles). The participants provided their written informed consent to participate in this study. Participation was compensated monetarily.

5.3 Synopsis

Contrary to expectations, both the lowest and highest melodies exhibited pronounced salience. When the spectral boundaries of the mixture were moved, the salient frequency bands also shifted towards the relatively lowest and highest melodies occurring in the mixture. This finding demonstrates that the auditory system does not inherently exhibit a reduced perception of bass voices, but rather that perception is shaped by the acoustic edges. Investigating how this behavior affects more naturalistic sounds than pure tones and examining the connection to the lack of salience in bass instruments provides an intriguing opportunity for future research.

6. CONCLUSION

6.1 Summary

Exploring salience in musical scenes

This dissertation analyzes the salience of sound sources within auditory scenes, focusing on musical mixtures wherein multiple sounds appear simultaneously. Salience refers to the attribute of a sound that draws auditory attention towards the sound, making it more prominent than other sounds within the mixture. Listening experiments were conducted, and the behavioral responses of participants were investigated and related to acoustical analyses of the sound stimuli. Understanding the acoustic underpinnings that contribute to salience within musical scenes is crucial for understanding auditory scene perception and has practical implications, such as the development of hearing aids and sound engineering in general.

Study 1 - Listening in the mix: Lead vocals robustly attract auditory attention in popular music

To explore the trajectory of auditory attention in musical scenes, three online detection experiments with 40 to 47 young participants each were conducted. Participants were asked to detect single cued instruments or vocals in musical mixtures with multiple instruments. The mixture consisted of monophonic 2-second multi-track excerpts of popular music and a cue signal that incorporated an isolated target instrument or vocal track from the mixture. Half of the time, the target was missing in the mixture. To test the effects of selective attention, the presentation order varied among participants. For one half of the participants, the target cue preceded the mixture, allowing listeners to use this prior knowledge of the target to direct their attention to search for it in the mixture. For the second half of the participants, the target was presented after the mixture, so that the detection depended strongly on the salience of the target in the mixture. Detection accuracies were measured and data on the participants musical sophistication was acquired using subscales of the Gold-MSI.

In the first experiment, excerpts were directly extracted from the pop music pieces without any spectral or level modification of the individual tracks. Results showed no significant correlations between recognition accuracy and musical sophistication scores, which continued throughout all three experiments. Target category-dependent detection accuracy was observed, with vocals showing the best performance, outperforming all instrument categories. The presentation order also highly influenced detection, with better results when target information was given prior to the mixture. This effect was especially pronounced for bass instruments, with detection accuracies dropping close to chance level without prior cue, suggesting that instruments in the low-frequency range especially lack salience. Fascinatingly, while this was true for almost all target categories, the vocals stood out, showing only minor differences between the orders, implying that vocals always attracted auditory attention even when no prior cue was given.

The acoustic origin of this unique vocal salience was then investigated in the following two experiments. In the second experiment, it was investigated if this salience was caused by spectral masking. Filtering was used to create a spectral composition in which all targets would appear in the same spectral region within the mixture. All non-target sounds were suppressed in this region, creating an isolated spectral region for the target. While overall detection accuracy improved, the pronounced differences for all targets except the lead vocals remained, suggesting that vocal salience is not caused by masking effects. In the third experiment, it was investigated if vocal salience can be explained solely by sound level advantage, as an acoustical analysis of the music database revealed that vocals had the highest sound level among all tested categories. However, even when the sound levels for all targets were aligned, a clear order effect for all instruments except for the lead vocals persisted, suggesting that the higher sound levels of vocals are not the origin of lead vocal salience. This novel finding of vocal salience in musical mixtures provided the impetus for further research.

Study 2 - Salience of frequency micro-modulations in popular music

Continuing the search for the origins of vocal salience, two online detection experiments were conducted with 67 young participants each. While the detection paradigm remained the same as in Study 1, the extent to which phonological cues, musical features of the main melody, or frequency micro-modulation (FMM) contribute to salience and specifically to vocal salience was investigated. FMM in this study refers to frequency modulation inherent in singing voices, caused by the imperfect intonation of the human voice, which is usually smaller than one note and is particularly strong in note onsets, offsets, and transitions. The sound levels were aligned for all targets, and sounds were presented monophonically. As in Study 1, sound excerpts were extracted from a multi-track database consisting of pop music, though a different database was used. Detection accuracies were measured and data on the participants musical sophistication was acquired using subscales of the Gold-MSI.

In the first experiment, the effect of FMM was investigated by producing an additional set of vocal excerpts and eliminating their FMM, creating an auto-tune-like robotic voice effect. The modified excerpts were then either used as stimuli for a singing voice without FMM category or processed to test the effect of their melody. For this, the pitch trajectory of the excerpts was estimated and recreated in various virtual instruments that were tested as target categories.

Results revealed that detection accuracy was influenced by presentation order for all targets except the unmodified vocals, replicating the results observed in Study 1 and implying robustness of vocal salience, as it emerged across different music database. Playing the main melody did not eliminate the order effect. The removal of FMM for the vocal excerpts introduced an order effect, stripping them of their unique vocal salience, even though phonological cues persisted. This suggests that neither the phonological cues that could facilitate the processing of speech-like sounds nor the musical features of the main melody are sufficient to drive vocal salience. No significant correlations were found between recognition accuracy and musical sophistication scores for both experiments.

In the second experiment, the excerpts with eliminated FMM were modified by resynthesizing the original FMM of the singing voices to both the pitch-quantized vocal excerpts and the recreated instrument excerpts that played the vocal melody. Transferring FMM from vocals to the targets reduced the magnitude of the order effect considerably. A negative correlation between FMM intensity and order effect revealed that sounds with more intense FMM also had heightened salience. In summary, the results suggest that FMM is an important acoustical feature contributing to vocal salience in musical mixtures.

Study 3 - Fast recognition of voice and instrument sounds in musical scenes

Motivated by the influence of frequency micro-modulations (FMM) on the detection of instruments within musical mixtures, Study 3 aimed to investigate the influence of FMM on the recognition of vocal and instrument sounds. The study comprised of three experiments with thirty young participants each, half with musical training. In the experiments, participants were asked to classify the sounds into vocal or instrument sounds in a go/no-go task as fast as possible. Stimuli consisted of 250 ms sound excerpts, originating from recorded samples and single-source excerpts of popular music featuring various short sung vowels and vocal sounds alongside instrumental sounds. Each sound was presented in separate versions with and without FMM to analyze the influence of frequency modulation. Additionally, sounds were presented in isolation or accompanied by a spatially separated piano interferer to test the sounds' salience in the form of robustness to interference. Accuracy and reaction times were measured as well as data on the participants musical sophistication using subscales of the Gold-MSI. An acoustical analysis was performed by investigating spectral similarity between the sounds and quantifying the FMM intensity.

In the first experiment, various orchestral sound samples from different target categories were utilized, including sung vowels /a/ and /u/ as well as string and wind sounds. Recognition performance varied across sound categories; recognition of vowel sounds was not superior to the recognition of instrumental sounds, and no effects of FMM were observed. When presented with interfering sounds, all sounds exhibited degradation in recognition. Notably, /a/ sounds showed distinct robustness to interference, displaying less degradation than other sounds, whereas /u/ sounds lacked this robustness and were surpassed by instrumental sounds. Acoustical analysis revealed a lack of correlation between FMM intensity and performance but highlighted correlations for spectral similarities between the target sounds and interferers, which were particularly strong for /u/ and wind sounds.

Therefore, a second experiment was conducted with a reduced stimulus set, removing the wind sound altogether to reduce the effects of spectral similarity. While the recognition in the isolated presentation improved for /u/ sounds, no recognition advantage for vowel sounds was observed. In the presentation with an interferer, the robustness for /a/ and the lack of robustness for /u/ persisted, which, as the acoustic analysis implied, was likely caused by relatively small and high similarity, respectively.

For the third experiment, sound excerpts of popular music were used, incorporating female and male vocals as well as string and wind sounds extracted from the same multi-track database as in Study 2, where effects of FMM were apparent. Results still lacked a recognition advantage for vocal sounds in the isolated condition, while a distinct robustness in the interferer condition persisted, underscoring the robustness of the effect. However, no effect of FMM was found, and no notable correlation between recognition and intensity of FMM was observed. No notable correlations between musical sophistications were found in the experiments. Taken together, the findings demonstrate that vocal recognition is not inherently more efficient compared to instrumental sound recognition and does not possess properties that make it impervious to confusion with spectrally similar instrumental sounds. The lack of impact of FMM on the recognition of sounds suggests that cues related to FMM are not exploited or do not affect the recognition of short musical sounds. These results emphasize the necessity of a comprehensive understanding of various acoustic factors influencing both vocal and instrumental sound recognition.

Study 4 - Enhanced salience of edge frequencies in auditory pattern recognition

Motivated by the bass instruments' lack of attention attraction and reports of a high voice superiority effect, a study was conducted to investigate whether the auditory system operates under a frequency bias that dampens the detection of sounds in low-frequency regions and fosters the detection of sounds in higher frequency regions. Using the same detection paradigm with varying presentation orders of cue and mixture as in Studies 1 and 2, two on-site experiments were conducted with 26 and 30 young participants.

The mixtures consisted of pseudo-randomized pure tone melodies with eight notes. These melodies appeared in five out of six possible spectrally separated frequency regions, with a spectral distance between the regions of one ERB-filter bandwidth to mitigate spectral masking. The sound level of melodies was aligned to 40 dB(A) to compensate for frequency-specific hearing thresholds. The target was absent in half of the stimuli. All participants were tested on stimuli in both presentation orders, and each experiment followed a sequence starting with a presentation order where the target cue was presented before the mixture, followed by an order where the mixture was played first. Detection accuracies and data on musical sophistication using subscales of the Gold-MSI were recorded.

Results showed pronounced effects for presentation order, with higher detection accuracies when the target cue was presented first. This implies that even in scenes with spectrally distinct melodies, detection is highly influenced by prior information about the target, replicating the results from the first two studies. However, both the melodies in the lowest and highest frequency regions showed the least impact of the order effect, with similar performance in both regions, outperforming those in between. These findings question the notion of bass inferiority, implying a salience for sounds at the edges of the mixture.

To further confirm this finding, a second experiment was conducted to investigate if the enhanced salience of edge frequencies is caused by the absolute frequencies utilized or shifts to the relative frequency edges of a mixture. For this, the mixture was reduced to three melodies that either appeared in the first four or last four frequency bands of the first experiment. The results corroborated the findings by yielding enhanced accuracies for edge frequencies irrespective of the absolute frequency region.

While no notable correlations between musical sophistication and performance were observed in the first experiment, likely due to a participant pool primarily consisting of musicians, moderate positive correlations were seen in the second experiment, where participants with and without musical training were recruited in a balanced manner. The correlation between musical sophistication and performance implies that individuals with higher musical sophistication have better abilities to search the mixture for the target.

The overall presence of the salience of edge frequencies refutes the assumption that the auditory system is driven by a frequency bias that favors higher frequencies and inhibits lower frequencies. Thus, the explanation for the effects of bass inferiority and high voice superiority observed in other studies likely stems from the interplay of complex harmonic structures of sounds outside of pure tones, such as those produced by realistic instruments.

6.2 Implications

The findings from the studies presented in this dissertation have several important implications for our understanding of attention and salience in musical mixtures. These implications extend across various aspects of auditory perception and contribute to the broader field of psychoacoustics.

6.2.1 Attention and salience in musical scenes

A constant finding throughout the studies is the remarkable human ability to extract acoustic cues, even in rich monophonic musical mixtures containing various instruments, and successfully detect individual instruments within these mixtures. This highlights the exceptional capabilities of the human auditory system in auditory scene analysis abilities (Bregman, 1990). The provision of prior information about a specific sound significantly enhanced the detection of melodies, instruments, or vocals. This finding aligns with numerous reports that top-down processing can facilitate stream formation (Luce & Green, 1978; Mondor & Bregman, 1994;Sussman, Winkler, et al., 1999; Bey & McAdams, 2002; Sussman, 2006; Fritz et al., 2007). Notably, this effect was particularly pronounced for bass instruments, which were otherwise obscured by other elements, as indicated by a detection accuracy close to chance level when no prior information was presented. This aligns with the notion, that top-down processing is capable of highlighting otherwise obscure sounds within acoustic scenes (Botte et al., 1997; Eramudugolla et al., 2005; Sussman & Steinschneider, 2009).

In the absence of prior information, detection accuracy diminished. Nevertheless, the auditory system was still able to disentangle the scene to form separate auditory streams and achieve detection above chance levels. This ability persisted even under acoustically disadvantageous conditions where important cues were reduced, such as the lack of distinctive timbre in mixtures of pure tones, lower sound levels for the target, or a general lack of spatial separation (Huron, 1989; Mondor & Zatorre, 1995, Darwin & Hukin, 1999). This supports the notion of an attentive system governed by global (background) and local (foreground) organization (for a review see Sussman, 2017). Rather than merely gating out unattended sounds and enhancing attended ones, this system allows multiple sound organizations to be held in memory. In musical mixtures, listeners can access the interplay of instruments within the mixture (global) and also focus selectively on individual instruments (local). However, attention remains a limited resource. Therefore, allocating resources to local organizations detracts from global processing capabilities, particularly in complex acoustic scenes where the demand on finite cognitive resources is higher. This is reflected in studies that decoded neural activity in a speaker scenario, where one speaker had to be attended to while ignoring the other, revealing stronger neural representations of the attended speaker compared to the ignored speaker (Mesgarani & Chang, 2012; Zion Golumbic et al., 2013; Bednar & Lalor, 2020). Similar observations were studied in musical mixtures, where one instrument had to be attended among various instruments (Treder et al., 2014; Cantisani et al., 2019). This dual capability aligns with theories of auditory attention that propose simultaneous the perception of a large-scale auditory scene information alongside the selection of specific sounds (Bigand et al., 2000; Shinn-Cunningham & Best, 2008). Recent findings further support this view, demonstrating that unattended sounds are still processed through stream formation. Once segregated, these sounds become gradually more attenuated in deeper layers of auditory processing (Puschmann et al., 2024).

This organization is shared across modalities as studies in visual attention show commonalities (Shinn-Cunningham, 2008): In a visual scene, objects are formed by an interplay of bottom-up processes operating on properties such as size, brightness or spatial position, and top-down processes such as prior knowledge, expectations, and specific task demands (Desimone & Duncan, 1995). Multiple objects can be formed within a scene, but they compete for visual attention, as only one object can be focused, mirroring the notion of global and local organizations in auditory attention. As attentional resources are limited, the greater the attentional resources devoted to focus one target object, the less the processing can be used for other non-target objects (Kastner &

Ungerleider, 2000). Salience is created by both inhomogeneities or sudden changes such as a red dot popping out of a field of green dots, or by new objects appearing in a scene. Similar to the auditory domain, prior knowledge of an object's properties such as the shape, color or location can be used to search the visual scene (Bundesen, 1990; Pestilli et al., 2007). In an analogy between spectral locations in auditory scenes and spatial locations in visual scenes, a complementary effect regarding the salience of edges is observed. Central objects in the visual field are identified more accurately, quickly, and in greater detail than objects located in the periphery or at the edges (Rijsdijk et al., 1980; Cannon, 1985).

6.2.2 Salience of edge frequencies

Interestingly striking differences appeared for the presentation of mixtures containing pure tones melodies and real instrument. In one study pure tone melodies showed an enhanced salience for the outer spectral melodies (lowest and highest). This salience of edge frequencies suggests that the auditory system utilizes frequency-based cues to parse complex auditory environments, potentially prioritizing sounds that mark the boundaries of perceptual streams. However, this seems to contradict the lack of salience seen in the bass instruments, which showed to be particularly bad at attracting attention when no attention was directed towards it. These differences can be partially attributed to the many spectro-temporal differences between pure-tones and natural instruments that provide acoustical cues exploitable for ASA, but also differences within the set of sounds used within the experiments. As the pure tones used provide no level, or timbral cues, the auditory system relied on temporal cues (synchrony in onset and offset) and differences in frequency. Both are known to provide enough information for enabling stream segregation (Vliegen et al., 1999). Spectral distance was mapped in a way that melodies always had a minimum distance of at least one equivalent rectangular bandwidth (ERB, Glasberg & Moore, 1990) so that the degree of overlap of excitation patterns evoked in the cochlea was minimized, shown to be a particularly strong cue for stream segregation (Rose & Moore, 2000).

This was further supported by a melody discrimination study conducted by Brochard and colleagues (Brochard et al., 1999), which demonstrated that a distance of only one ERB was sufficient to focus on a single cued pure tone melody embedded in mixtures of up to four simultaneously presented pure tone melodies. Reflecting the edge effects found in Study 4, differences were found in the ability to judge whether a focused melody differed from a cued target melody. Specifically, inner melodies showed worse performance than both the lowest and highest melodies. Additionally, a second experiment was conducted to investigate the spectral spacing needed so that performance is independent of the target melody. This experiment again revealed better performance for outer melodies and inner melodies, with considerably increased ERB distance required for inner melodies to achieve comparable results to outer melodies. Interestingly, musically trained participants also demonstrated better results than those without training, another similarity between both studies. Regarding the enhanced salience, Brochard and colleagues (Brochard et al., 1999) argued that his phenomenon is caused by the fact that the attentional system either needs to segregate two or three streams: Focusing on an outer sequence, creates two perceptual streams, one composed of the attended stream and another composed of all other sequences either above or below the attended sequence, conversely focusing on an inner sequence, divides the acoustic scene into three streams: an attended stream composed of the sequence in the center, a second stream composed of the sequences below the attended stream and a third stream composed of the sequence above the attended stream which demands higher processing resources. These findings were partially seen in another study which focused on musical mixtures consisting of three spectrally distinct piano melodies or pure tone melodies (Palmer & Holleran, 1994). The results contradicted the salience of edges partially as a hierarchy arises, in which the highest melody outperformed all outer melodies, and the lowest melody outperformed the center melody. They argue that the particularly sensitive perception of the high voice is caused by its frequency range which lies in the optimal range for vocal melodies. However, this is in strong contrast to the findings presented in this dissertation, as both in Study 1, a lack of salience was observed for instruments playing in the average spectral region of vocals, as well as in Study 2 a lack of salience was observed for instruments playing vocal melodies. Interestingly, the results of the study by Palmer and Holleran (Palmer & Holleran, 1994) demonstrated differences between melodies played by piano and pure tones that were associated with more distinct voices for the pure tones due to the absence of overlapping harmonics of the piano tones. The harmonics of the piano tones interfere with another, and thus hinder the segregation of both sounds. The differences seen reflect observations of differences in stream formation between pure-tones and naturalistic syllables, in which enhanced distance in fundamental frequency between two syllables was needed to enable a stable stream segregation compared to pure-tones (Gustafson et al., 2020).

This interplay of complex spectral structures caused by naturalistic instruments is likely to be the key in understanding the differences between the pronounced edge effect and the lack of ability to attract attention for the bass resulting in a failure of stream segregation. In naturalistic musical mixtures, the mix of instruments create an acoustic scene in which each instrument dependent specto-temporal patterns overlap in both the temporal and spectral dimension. While timbre cues are known to foster stream segregation (Bregman, 1990; Huron, 2001), the scene can be so dense that timbre and pitch of the individual instruments tempo-spectrally superimpose another. This phenomenon also manifests in the high voice superiority effect (HVSE; Fujioka et al., 2005), which is caused by the interplay of complex spectral tone structure caused by naturalistic timbre, in which the energy of tones in the higher melody superimposes harmonic structures of the tones in the lower melody (Trainor et al., 2014). Interestingly Trainor and colleagues even assumed a lack of the effect when pure tones are used, which reflects the results of Study 4, as otherwise a correlation between increasing frequency band and increasing detection accuracy would have emerged. However, this is somewhat in conflict with Huron (Huron, 1989) who also reported that in realistic polyphonic music excerpts, which contain multiple independent melodies, detection of simultaneous melodies smears for those in between the outer edges. The term "independent" here adds a perspective that represents a stark contrast between the tested musical scenes: the pure-tone melodies represented independent melodies with equal structural weight. In contrast, in pop music, the primary role of bass instruments is not to play a lead melody but to strengthen groove perception by providing a rhythmic foundation. While the perception of groove and timing is enhanced for bass instruments (Hove et al., 2014), it appears to be more abstract than the perception of individual melodic elements within a musical mixture. This abstract perception contributes to bass instruments being perceived rather unconsciously and not as separate auditory streams within the mixture. Furthermore, the main melody in a musical mixture is usually arranged to have distinct acoustic cues, leading to its perception as more salient than the accompanying parts of the mixture (C. K. Madsen, 1997; Ragert et al., 2014).

6.2.3 Vocal salience

Another consistent finding emerged for lead vocals in the musical mixtures: vocals robustly attracted auditory attention, irrespective of prior cueing, implying an inherent salience of vocal sounds in musical mixtures. Additionally, singing voice excerpts showed robust recognition when presented simultaneously with interfering sounds, unmatched by instrumental sounds. These findings align with reports of enhanced cortical responses observed when isolated singing voices and isolated instrumental sounds were presented sequentially (Levy et al., 2001; Gunji et al., 2003). Furthermore, speech sounds have been observed to be processed faster (Parviainen et al., 2005) and robustly trigger distinct neural mechanisms associated with specialized voice-specific cortical areas (Belin et al., 2000; Belin et al., 2002; Murray et al., 2006; Agus et al., 2017; Moskowitz et al., 2020). An interpretation of this salience could be indicative of a specialized processing mechanism for vocal features during the segregation of auditory objects in a musical scene. Therefore, when the auditory system segregates sound into mental representations of distinct streams, vocal features are able to trigger a prioritized voice-specific processing. This specialized processing might enable better isolation of vocal sounds within the auditory scene, contributing to better detection. This vocal salience aligns with the notion of timbral salience (Chon & McAdams, 2012), in which different instruments exhibit a unique degree of salience. The vocal salience could imply such a hierarchical structure where the timbre of singing voice dominates at the top. However, as indicated by the results of Study 3, namely the lack of salience for the vowel /u/, not all human vowels have a guaranteed superior level of salience. Rather, it seems as if the salience is an effect of unique features that can be produced by vocal sounds, helping them stand out from other instruments within musical mixtures.

To explore the origins of vocal salience, various acoustic manipulations were conducted. An analysis of musical mixtures provided insights that the vocals are mixed at a higher sound level than any other sound within the mixture and are spectrally favorably positioned in a large spectral region where they can pass through the mixture unmasked. These advantages could provide pivotal cues that foster stream segregation, making it easier for vocals to stand out amidst other sounds within the mixture.

Therefore, the contribution of differences between vocals and instruments in sound level and spectral filtering was investigated. While equalization in level and spectral filtering enhanced the prominence of instruments aiding their overall detection. The ability to attract attention, regardless of prior direction, still lacked for all sounds except vocals. Additionally, even when the vocals were decreased to an unfavorable level or spectral position within the mixture, vocal salience still persisted, highlighting that those cues cannot be the driving factors enabling vocal salience.

Another origin of vocal salience was assumed to be the role of the vocals in a musical mixture, as in pop music the vocals usually sing the main melody. Experiments have demonstrated that the main melody within a musical mixture is perceived to be more salient (C. K. Madsen, 1997; Ragert et al., 2014). Furthermore, vocals singing the main melody are shown to be better represented in memory (Weiss et al., 2012). To explore the influence of the main melody, the vocals singing the main melody in pop music excerpts were replaced by instruments playing the same melody. This substitution enhanced the prominence of the instruments in the musical mixture, aiding their overall detection. One probable explanation for this is that the main melody is arranged in a way that makes it stand out from other sounds within the musical mixture due to more distinct acoustic cues like differences in rhythm, playing a higher pitch range, and pitch variability (C. K. Madsen & Geringer, 1990; Uhlig et al., 2013; Ragert et al., 2014). However, like the manipulation of sound level and spectral filtering, these cues were not sufficient to recreate vocal salience on their own. This underscores the superior memory for sung melodies, excluding those played by instruments (Weiss et al., 2012; Weiss et al., 2021). Regarding the notion of vocal salience (Chon & McAdams, 2012), one might expect that the absence of vocals in mixtures, where vocals were replaced with instruments, would create a perceptual gap filled by other instruments emerging as particularly salient. However, the absence of such an effect indicates some degree of perceptual balance between the instruments within the mixture, such that no other instrument systematically attracted the listeners’ attention.

Experiments investigating the detection of phonological sounds containing lexical words, pseudo-words, and non-phonological sounds revealed better performance for phonological sounds, implying that an advantage is caused by phonological cues alone (Signoret et al., 2011). Additionally, detection was even more enhanced for words compared to pseudo-words, revealing an additional influence of lexical knowledge. This suggests hierarchical differences between lyrical singing with meaningful words and vocalizations of nonsense syllables like “la la la”. Underlining the impact of phonological cues is the assumption that the perception of pitched vocalizations creates heightened interconnectivity between brain hemispheres in otherwise lateralized human brain processes (Riecker et al., 2000; Zatorre et al., 2002; S. Norman-Haignere et al., 2015), These processes are associated with either timing and speech processing (left) or spectral and music processing (right). This interconnectivity, which could foster stream segregation, is observed in the perception and production of singing (Schön et al., 2005; Callan et al., 2006), and in the processing of tonal languages (Sammler et al., 2015; Chien et al., 2020). Also, remotely in assumption of the importance of phonetical cues, stronger cortical responses were observed for singing than for humming (Ozdemir et al., 2006). However, it should be noted that this view of lateralization is subject to ongoing debate, with contradictory results emphasizing overlaps in the processing of language and music (see Koelsch, 2011). Nonetheless, phonological cues are a distinct feature of the singing voice and a likely candidate for enabling their unique salience. Despite this, the results of Study 2 present a different picture. While these cues are likely candidates for the overall enhanced detection of vocal sounds within musical mixtures, the findings indicate that phonological cues alone are not sufficient to enable vocal salience. The elimination of frequency micro-modulation (FMM) resulted in the loss of this salience, even though phonological cues were kept intact, refuting phonological cues as the sole contributor to vocal salience.

6.2.4 The role of FMM

At the same time, this result demonstrated the importance of FMM, which was further emphasized in a follow-up experiment, as the reintroduction of FMM significantly increased salience. The ability of FMM to reduce the order effect and maintain vocal salience suggests that micro-modulation may be a key element in the auditory system's mechanism for distinguishing vocal sounds from other auditory stimuli. Further underscoring the effect of FMM, a correlation was observed between a sound's ability to attract attention and FMM intensity.

To understand what makes FMM of singing voices such an important candidate for creating salience and how FMM enhances the perceptual distinctiveness of vocals, an inspection of its characteristics is needed: FMM of singing voices are characterized by subtle frequency variations caused by the imperfect pitch control of human singing (Hutchins et al., 2014), most extensively in the transition toward or away from a target pitch (Larrouy-Maestri et al., 2014). While these frequency modulations are likely too subtle to be perceived as pitch errors in singing voices (Hutchins et al., 2012; Sundberg et al., 2013; Gao & Oxenham, 2022), they are still detected by the auditory system (Gockel et al., 2001; Lyzenga et al., 2004; Larrouy-Maestri & Pfordresher, 2018), and are associated with natural singing (Merrill & Larrouy-Maestri, 2017). These modulations can be intentionally enhanced by singers to add expressiveness (Sundberg et al., 2013) and emotional prosody (Larrouy-Maestri et al., 2024). This enhancement is able to trigger automatic processes that prioritize sounds with emotional prosody within vocalizations, similar to the salience of emotional faces in the visual modality (Liebenthal et al., 2016). This processing adds top-down cues that guide attention towards the emotional vocalization (Vuilleumier, 2005), thus highlighting the singing voice within the musical mixture. Besides impacting emotional prosody, the existence of pitch transitions between notes enhances the perceived continuity of singing voices (Larrouy-Maestri & Pfordresher, 2018; Weiss & Peretz, 2019), while its constant frequency modulations also add irregularities to the signal, both providing additional acoustical cues that foster the segregation of singing voices.

In speech, FMM has demonstrated to be an important factor in enhancing speech detection and recognition. FMM enhances the prominence of voices (McAdams, 1989), plays a crucial role in speech intelligibility (Wingfield et al., 1984; Miller et al., 2010), and aids speech recognition by helping to identify and isolate a voice embedded in auditory scenes with various sounds (Strelcyk & Dau, 2009; Dupuis & Pichora-Fuller, 2014) or multiple competing speakers (Culling & Summerfield, 1995; Zeng et al., 2005; Parthasarathy et al., 2019). Additionally, memory advantages for vocalizations with emotional prosody have been demonstrated (Armony et al., 2007). Of course, this does not mean that vocal sounds or vocalizations are the only sounds that can create emotional prosody, as it is undeniable that non-vocal sounds such as instrumental music are capable of evoking emotional prosody (Koelsch, 2014). Rather, this supports the assumption that the auditory system is specialized towards naturalistic FMM in vocals, as evidenced by advantages for speech with natural FMM compared to no FMM (Dupuis & Pichora-Fuller, 2014), or speech with overly decreased or exaggerated modulations (Miller et al., 2010). This aligns with findings that the processing advantage of vocals is highly specific to features of naturalistic vocal sounds and is absent for non-vocal sounds, even when they are matched in various acoustic cues (Bélizaire et al., 2007; Agus et al., 2012; Agus et al., 2017).

Interestingly, the FMM showed no effect on the recognition of short sounds presented outside of musical mixtures as observed in Study 3. Several reasons may explain this. Firstly, the use of stationary tones: The sounds used were mostly extracts of stationary tones with no transitions between notes and were truncated at the end. Studies on FMM have demonstrated that FMM are particularly intense in note transitions and perceptually more important at the end of the note compared to the beginning, where they are commonly associated with poor singing abilities (Larrouy-Maestri & Pfordresher, 2018). Additionally, the perception of FMM is particularly prominent in the context of preceding or following notes and thus likely not exploited in single stationary tones (Pearce & Wiggins, 2006; Larrouy-Maestri & Pfordresher, 2018). Moreover, the duration of the short sounds may have been too brief to extract meaningful FMM information. Studies on emotional prosody in vocalizations have demonstrated that a minimum exposure of roughly 200 milliseconds is needed for prosody extraction (Liebenthal et al., 2016). This aligns with the ‘Asymmetric Sampling in Time theory’ (Poeppel, 2003), that picks up on the notion that the brain has specialized and lateralized processes (Zatorre et al., 2002) that distinctly evaluate short time windows in one hemisphere that tracks what happens ‘now’ and evaluate long time windows to compare events that change over time in the other hemisphere (Poeppel, 2003; Santoro et al., 2014; Giroud et al., 2020). The estimated time window for the change over time which could process FMM is approximately 200 milliseconds, similar to those found in emotional prosody. This implies that the perceptual advantage of FMM is relatively slow, too slow to explain the reported recognition advantages of the voice even for extremely short sound events (Suied et al., 2014; Isnard et al., 2019), suggesting that the vocals are endowed with additional cues that account for its unique salience.

In summary, FMM is shown to be an important component contributing to vocal salience as it provides acoustical cues about a sound’s continuity and adds emotional prosody, facilitating the detection and processing of vocal sounds. Investigations of other characteristics of the voice, such as favorable sound levels, spectral arrangement and dissimilarity, phonological information, and playing in the main melody showed further positive influences on the prominence of the voice in musical mixtures. Therefore, it is likely that the multitude of unique features possessed by vocals collectively contribute to vocal salience in auditory scenes. The observation that some vocal sounds do not exhibit enhanced salience highlights that vocal salience is not inherently produced by being a vocal sound, but rather is based on the interplay of acoustic cues.

6.2.5 Musical sophistication

The impact of musical sophistication was mostly lacking in all experiments, with one exception. Higher levels of musical sophistication are associated with improved results in various ASA tasks (e.g., Micheyl et al., 2006; Marozeau et al., 2010; Başkent et al., 2018; Siedenburg et al., 2020; Hake et al., 2023). However, a lack of correlation between better performance and musical sophistication, as measured with subsets of the Gold-MSI (Müllensiefen et al., 2014), indicates that the effects revealed within the experiments are of a fundamental nature and irrespective of musical sophistication. Specifically, the vocal salience effect emerged in both groups, making it similar to other salience effects observed in musical mixtures, like the high voice superiority effect, which has been demonstrated to exist in both musicians and non-musicians, and even infants (Marie & Trainor, 2013; Marie & Trainor, 2014). Furthermore, the lack of effect for musical sophistication highlights that the segregation of sound sources within musical mixtures is trained to such a degree that it is independent of musical sophistication, supporting the notion that everyday exposure is sufficient to develop some musically relevant abilities without the need for explicit musical training (Bigand & Poulin-Charronnat, 2006). Additionally, the lack of correlation aligns with the observations in the melody detection study by Bey & McAdams (2002), which served as a basis for the detection paradigm in the experiments.

Interestingly, no correlation with musical sophistication was found in the recognition and differentiation of instrumental and vocal sounds. One might argue that this task should favor musically trained listeners, as familiarity with the presented timbres is known to modulate recognition (Siedenburg & McAdams, 2017b), and musically sophisticated listeners have demonstrated advantages in timbre discrimination (Kannyo & DeLong, 2011; Martins et al., 2022). Yet, the lack of correlation is likely due to the demands of the task — the categorization between vocal and non-vocal sounds — which is a more fundamental ability already acquired through implicit training, and thus does not benefit from musical sophistication. This finding aligns with other studies that found no difference between both groups in timbre discrimination tasks (Allen & Oxenham, 2014; Bigoni & Dahl, 2018). Bigoni & Dahl (2018) reasoned that the lack of an effect can be attributed to the observation that inter-individual differences in timbre perception play a larger role than musical training, adding another possible explanation.

Conversely, a positive correlation with musical sophistication was observed in the detection of pure-tone melodies. The structural differences between the musical scenes, such as pop musical scenes structured with a defined leading melody and instrumental accompaniment, contrast with the structure of the pure-tone melodies, where each melody was independent and equal, lacking a clear hierarchy. These structural differences demand different capabilities from the auditory system, explaining the observed differences. This correlation aligns with studies that found better detection of melodies and instruments in complex auditory environments and musical mixtures among musically sophisticated listeners (Marozeau et al., 2010; Slater & Marozeau, 2016; Siedenburg et al., 2020). Part of the reason for these ambiguous results could relate to the number of test subjects and the insufficient spread of different levels in their musical abilities, as positive correlations were also found in a large-scale study employing the same target-in-mixture detection paradigm but with a total of 525 normal-hearing listeners (Hake et al., 2023).

Taken together, these contradictory results emphasize that the question of differences between auditory skills associated with higher levels of musical sophistication is not as clear-cut as it might seem. Although the results presented in this dissertation demonstrate that everyday exposure to music is sufficient to elicit individual sounds in musical mixtures, the contradictions also highlight the need for more extensive studies involving a larger number of participants covering a wide range of musical sophistication levels.

6.3 Final conclusion and outlook

6.3.1 Final conclusion

The four studies incorporating a total of ten experiments presented in this dissertation provide a comprehensive examination of factors contributing to auditory attention and salience in musical mixtures. On the basis of these experiments, several key findings have emerged, enhancing our understanding of how certain sounds stand out in complex acoustic mixtures.

The experiments demonstrated the significant interplay between bottom-up and top-down processes in complex auditory scenes such as musical mixtures. Acoustic properties of sounds (bottom-up processes) play a crucial role in their detectability, while prior knowledge (top-down processes) also shapes the perception of the musical mixture. An inferior detection of bass instruments was observed, revealing that bass instruments lacked the ability to attract auditory attention. Investigations into whether the auditory system possesses perceptual biases that could explain this bass inferiority, utilizing mixtures of spectrally distinct pure-tone melodies, rejected this assumption. Instead, an enhanced salience of sounds at the spectral edges of auditory scenes appeared, implying that the bass inferiority results from musical structures with distinct lead and accompaniment roles and from spectral patterns evoked by naturalistic instruments.

Conversely, the lead vocals attracted listeners’ auditory attention to a degree unmatched by any other sound within the mixture. This finding was consistent across multiple experiments, highlighting a unique vocal salience that manifested in musical mixtures and when presented alongside an interfering sound. This ability to attract auditory attention more effectively than other instruments suggests that the human singing voice possesses inherent acoustic properties that make it particularly prominent in auditory scenes. However, not all vocalizations showed such enhanced salience, revealing that being a vocal sound does not automatically produce this salience. This emphasizes that multiple acoustic factors contribute to enabling vocal salience. Still, vocal salience persisted across different sets of sounds and experimental conditions, including variations in spectral masking and sound levels. Additionally, this unique salience could not be reinforced by other instruments playing vocal melodies.

The frequency micro-modulations (FMM) characterized by pitch variations inherent in natural singing were isolated as contributing to this salience, equipping the vocals with additional acoustical cues and features that are extracted for emotional prosody processing, thus facilitating their segregation by triggering additional cortical resources. As the FMM had no effect on the recognition of short single-note vocal excerpts, which nevertheless exhibited salience, it appears that other features of the voice also contribute to vocal salience. This is further supported by the observation that all attributes transferred from voice to instruments increased the prominence of the instruments, indicating that typical vocal sounds are equipped with a multitude of advantageous features that help attract auditory attention.

6.3.2 Outlook

The above findings contribute to the broader field of psychoacoustics and open new avenues for further research into the intricate processes underlying auditory scene analysis. However, with every answered question, new ones arise in the pursuit of understanding human perception.

The studies presented in this dissertation could be further expanded: For example, to better understand the interplay of stream segregation and attention, and the extent to which global organizations are monitored, the detection paradigm used in Study 1 could be complemented with an additional listening condition. In this condition, instead of cueing the target before the presentation of the mixture, a random instrument within the mixture is cued, and participants are asked if they have perceived a different non-cued target instrument within the scene. Utilizing this condition would distract auditory attention away from the target sound. An analysis of the detection accuracy of correctly identifying whether the non-cued target was playing may provide additional insights into the ability of the wrongfully cued instrument to distract from other sounds within the mixture. Additionally, by directing the listeners' attention towards an irrelevant sound, this experimental condition would further study how global organizations within musical scenes are tracked by the auditory system.

With respect to the salience of edge frequencies and the inability of the bass to attract auditory attention, several possibilities arise. To bridge the gap between artificial and naturalistic musical stimuli, pure tones could be replaced by tone complexes, ensuring no variation between the melodies within a scene to de-emphasize timbre cues. This could support the hypothesis that the differences between edge salience and the high voice superiority effect (HVSE; Fujioka et al., 2005) are based on harmonic structures. Alternatively, a melody recognition task could be used instead of a detection task. This would require participants to focus on subtleties in the melody, engaging more complex cognitive processes and potentially enhancing the differences between the melodies. Additionally, this approach would prevent participants from simply identifying the target by comparing the frequency band of the target with the frequency band missing from the mixture, which would increase the informative potential of the experiment.

Investigating individual differences in auditory perception, such as the effects of musical training, age, or hearing impairment, could offer valuable insights. Experiments could be designed to explore how these factors influence the detection and processing of salient sounds. For example, as musicians show pronounced cortical responses when hearing their trained instruments (Pantev et al., 2001; Shahin, Roberts, & Trainor, 2008; Strait et al., 2012), it would be interesting to study if the experience in playing the instrument also shapes the perception of the musical scene in making the trained instrument more salient, potentially even counteracting the vocal salience. Another aspect could include investigating individual preferences for singing voices or even the dislike of certain voices (Bruder et al., 2024) and how this would shape vocal salience. As both age and hearing impairment are associated with decreasing musical scene analysis abilities (Hake et al., 2023), the investigation of these factors is a highly important topic. The knowledge gained about the salience of different sound features can be further tested in hearing-impaired individuals and can inform the design of more effective hearing aids and auditory prosthetics. For example, incorporating algorithms that boost FMM could improve the user's ability to focus on these sounds in noisy environments and, if desired, in musical mixtures, enhancing the overall auditory experience for users. Additionally, experiments with hearing-impaired participants could add a new perspective by exploring whether the same salience effects are observed between normal hearing and hearing-impaired listeners. The effect of FMM would be particularly interesting, as hearing impairment is also associated with difficulties in the detection of frequency modulations (Moore & Skrodzka, 2002).

Another direction for future research is multi-sensory integration. Auditory perception does not occur in isolation; it is often influenced by other sensory modalities. Future experiments could explore how visual or tactile cues interact with auditory information to influence salience and attention. For example, investigating how visual cues of a singer or instrumentalist impact the auditory perception of their sound could reveal important insights into multi-sensory integration. Exploring how vibro-tactile stimulation influences auditory perception and how particularly salient features can be exploited to enhance auditory perception for hearing aid users would be valuable. While it has been observed that vibro-tactile stimulation can enhance the music listening experience (Siedenburg et al., 2023), the extent to which such modalities improve musical scene analysis abilities remains open to be explored.

To complement behavioral experiments, neurophysiological studies using techniques such as EEG or MEG could be conducted to investigate the neural correlates of auditory salience and investigate if a neural correlate for the vocal salience in musical mixtures can be found. Understanding the brain mechanisms underlying the detection and processing of salient sounds could provide deeper insights into the cognitive and neural processes involved in auditory scene analysis.

Future research could also aim to increase the ecological validity of experiments by using more naturalistic and contextually rich auditory scenes. Studies could involve live music performances, real-world environments, or virtual reality setups to better understand how auditory salience operates in everyday listening situations. One approach to this was tested by Bürgel et al.(2024), in which the audience of a concert was tasked with recognizing intentionally inserted errors, exemplifying one of the many possibilities to bridge the gap between laboratory findings and real-world applications.

BIBLIOGRAPHY

Agus, T. R., Paquette, S., Suied, C., Pressnitzer, D., & Belin, P. (2017). Voice selectivity in the temporal voice area despite matched low-level acoustic cues. *Scientific Reports*, *7*(1), 11526. https://doi.org/10.1038/s41598-017-11684-1

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, *131*(5), 4124–4133. https://doi.org/10.1121/1.3701865

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, *66*(4), 610–618. https://doi.org/10.1016/j.neuron.2010.04.014

Akça, M., Vuoskoski, J. K., Laeng, B., & Bishop, L. (2023). Recognition of brief sounds in rapid serial auditory presentation. *PloS One*, *18*(4), e0284396. https://doi.org/10.1371/journal.pone.0284396

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience : A Journal and Virtual Library*, *5*, D202-12. https://doi.org/10.2741/alain

Alain, C., & Bernstein, L. J. (2015). Auditory Scene Analysis. *Music Perception*, *33*(1), 70–82. https://doi.org/10.1525/mp.2015.33.1.70

Alain, C., & Woods, D. L [D. L.] (1997). Attention modulates auditory pattern memory as indexed by event-related brain potentials. *Psychophysiology*, *34*(5), 534–546. https://doi.org/10.1111/j.1469-8986.1997.tb01740.x

Allen, E. J., & Oxenham, A. J [Andrew J.] (2014). Symmetric interactions and interference between pitch and timbre. *The Journal of the Acoustical Society of America*, *135*(3), 1371–1379. https://doi.org/10.1121/1.4863269

Armony, J. L., Chochol, C., Fecteau, S., & Belin, P. (2007). Laugh (or cry) and you will be remembered: Influence of emotional expression on memory for vocalizations. *Psychological Science*, *18*(12), 1027–1029. https://doi.org/10.1111/j.1467-9280.2007.02019.x

Barrett, K. C., Ashley, R., Strait, D. L., Skoe, E., Limb, C. J., & Kraus, N. (2021). Multi-Voiced Music Bypasses Attentional Limitations in the Brain. *Frontiers in Neuroscience*, *15*, 588914. https://doi.org/10.3389/fnins.2021.588914

Başkent, D., Fuller, C. D., Galvin, J. J., Schepel, L., Gaudrain, E., & Free, R. H. (2018). Musician effect on perception of spectro-temporally degraded speech, vocal emotion, and music in young adolescents. *The Journal of the Acoustical Society of America*, *143*(5), EL311. https://doi.org/10.1121/1.5034489

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bayanova, L., Chichinina, E., & Aslanova, M. (2024). The association between music training and executive function in 6–7-year-old children. *Frontiers in Education*, *9*, Article 1333580. https://doi.org/10.3389/feduc.2024.1333580

Bednar, A., & Lalor, E. C. (2020). Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. *NeuroImage*, *205*, 116283. https://doi.org/10.1016/j.neuroimage.2019.116283

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Belin, P., Zatorre, R. J [Robert J.], & Ahad, P [Pierre] (2002). Human temporal-lobe response to vocal sounds. *Brain Research. Cognitive Brain Research*, *13*(1), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

Belin, P., Zatorre, R. J [R. J.], Lafaille, P., Ahad, P [P.], & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Bélizaire, G., Fillion-Bilodeau, S., Chartrand, J.-P., Bertrand-Gauvin, C., & Belin, P. (2007). Cerebral response to 'voiceness': A functional magnetic resonance imaging study. *Neuroreport*, *18*(1), 29–33. https://doi.org/10.1097/WNR.0b013e3280122718

Benz, S., Sellaro, R., Hommel, B., & Colzato, L. S. (2015). Music Makes the World Go Round: The Impact of Musical Training on Non-musical Cognitive Functions-A Review. *Frontiers in Psychology*, *6*, 2023. https://doi.org/10.3389/fpsyg.2015.02023

Best, V., Ozmeral, E. J., Kopco, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(35), 13174–13178. https://doi.org/10.1073/pnas.0803718105

Bey, C., & McAdams, S [Stephen] (2002). Schema-based processing in auditory scene analysis. *Perception & Psychophysics*, *64*(5), 844–854. https://doi.org/10.3758/bf03194750

Bidelman, G. M., Krishnan, A., & Gandour, J. T. (2011). Enhanced brainstem encoding predicts musicians' perceptual advantages with pitch. *The European Journal of Neuroscience*, *33*(3), 530–538. https://doi.org/10.1111/j.1460-9568.2010.07527.x

Bigand, E., Delbé, C., Gérard, Y., & Tillmann, B. (2011). Categorization of extremely brief auditory stimuli: Domain-specific or domain-general processes? *PloS One*, *6*(10), e27024. https://doi.org/10.1371/journal.pone.0027024

Bigand, E., McAdams, S [S.], & Forêt, S. (2000). Divided attention in music. *International Journal of Psychology*, *35*(6), 270–278. https://doi.org/10.1080/002075900750047987

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100–130. https://doi.org/10.1016/j.cognition.2005.11.007

Bigoni F., & Dahl S. (2018). *Timbre Discrimination for Brief Instrument Sounds.* https://doi.org/10.5281/zenodo.1492361

Botte, M. C., Drake, C., Brochard, R., & McAdams, S [S.] (1997). Perceptual attenuation of nonfocused auditory streams. *0031-5117*, *59*(3), 419–425. https://doi.org/10.3758/BF03211908

Bregman, A. S [Albert S.] (1978). Auditory streaming is cumulative. *0096-1523*, *4*(3), 380–387. https://doi.org/10.1037/0096-1523.4.3.380

Bregman, A. S [Albert S.]. (1990). *Auditory Scene Analysis*. The MIT Press. https://doi.org/10.7551/mitpress/1486.001.0001

Bregman, A. S [Albert S.], Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *0031-5117*, *37*(5), 483–493. https://doi.org/10.3758/bf03202881

Bregman, A. S [Albert S.], & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*(2), 244–249. https://doi.org/10.1037/h0031163

Bregman, A. S [Albert S.], Liao, C., & Levitan, R. (1990). Auditory grouping based on fundamental frequency and formant peak frequency. *Canadian Journal of Psychology*, *44*(3), 400–413. https://doi.org/10.1037/h0084255

Bregman, A. S [Albert S.], & McAdams, S [Stephen] (1994). Auditory Scene Analysis: The Perceptual Organization of Sound. *The Journal of the Acoustical Society of America*, *95*(2), 1177–1178. https://doi.org/10.1121/1.408434

Bregman, A. S [Albert S.], & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, *32*(1), 19–31. https://doi.org/10.1037/h0081664

Brochard, R., Drake, C., Botte, M.-C., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *0096-1523*, *25*(6), 1742–1759. https://doi.org/10.1037/0096-1523.25.6.1742

Bruder, C., Poeppel, D., & Larrouy-Maestri, P. (2024). Perceptual (but not acoustic) features predict singing voice preferences. *Scientific Reports*, *14*(1), 8977. https://doi.org/10.1038/s41598-024-58924-9

Bruyer, R., & Brysbaert, M. (2013). Combining Speed and Accuracy in Cognitive Psychology: Is the Inverse Efficiency Score (IES) a Better Dependent Variable than the Mean Reaction Time (RT) and the Percentage Of Errors (PE)? *Psychologica Belgica*, *51*(1), 5. https://doi.org/10.5334/pb-51-1-5

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547. https://doi.org/10.1037/0033-295x.97.4.523

Bürgel, M., Mencke, I., Benjamin, A., Dechert, M., Derks, D., Gerdes, K., Hake, R., Jacobsen, S., & Siedenburg, K. (2024). Unifying concert research and science outreach. *Musicae Scientiae*, *28*(1), 187–191. https://doi.org/10.1177/10298649231182078

Bürgel, M., Picinali, L., & Siedenburg, K. (2021). Listening in the Mix: Lead Vocals Robustly Attract Auditory Attention in Popular Music. *Frontiers in Psychology*, *12*, 769663. https://doi.org/10.3389/fpsyg.2021.769663

Bürgel, M., & Siedenburg, K. (2023). Salience of Frequency Micro-modulations in Popular Music. *Music Perception*, *41*(1), 1–14. https://doi.org/10.1525/mp.2023.41.1.1

Callan, D. E., Tsytsarev, V., Hanakawa, T., Callan, A. M., Katsuhara, M., Fukuyama, H., & Turner, R. (2006). Song and speech: Brain regions involved with perception and covert production. *NeuroImage*, *31*(3), 1327–1342. https://doi.org/10.1016/j.neuroimage.2006.01.036

Cannon, M. W. (1985). Perceived contrast in the fovea and periphery. *Journal of the Optical Society of America. A, Optics and Image Science*, *2*(10), 1760–1768. https://doi.org/10.1364/JOSAA.2.001760

Cantisani, G., Essid, S., & Richard, G. (2019). EEG-Based Decoding of Auditory Attention to a Target Instrument in Polyphonic Music. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 80–84). IEEE. https://doi.org/10.1109/WASPAA.2019.8937219

Carrasco, M [M.], & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 673–692. https://doi.org/10.1037//0096-1523.24.2.673

Carrasco, M [Marisa] (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

Carrasco, M [Marisa] (2018). How visual spatial attention alters perception. *Cognitive Processing*, *19*(Suppl 1), 77–88. https://doi.org/10.1007/s10339-018-0883-4

Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters*, *405*(3), 164–167. https://doi.org/10.1016/j.neulet.2006.06.053

Cherry, E. C., & Taylor, W. K. (1954). Some Further Experiments upon the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *26*(4), 554–559. https://doi.org/10.1121/1.1907373

Chien, P.-J., Friederici, A. D., Hartwigsen, G., & Sammler, D. (2020). Neural correlates of intonation and lexical tone in tonal and non-tonal language speakers. *Human Brain Mapping*, *41*(7), 1842–1858. https://doi.org/10.1002/hbm.24916

Chon, S. H., & McAdams, S [Stephen] (2012). Investigation of timbre saliency, the attention-capturing quality of timbre. *The Journal of the Acoustical Society of America*, *131*(4_Supplement), 3433. https://doi.org/10.1121/1.4708879

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience : A Journal and Virtual Library*, *13*, 148–169. https://doi.org/10.2741/2666

Cowan, N., Blume, C. L., & Saults, J. S. (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(3), 731–747. https://doi.org/10.1037/a0029687

Crawley, E. J., Acker-Mills, B. E., Pastore, R. E., & Weil, S. (2002). Change detection in multi-voice music: The role of musical structure, musical training, and task demands. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 367–378. https://doi.org/10.1037/0096-1523.28.2.367

Culling, J. F., & Summerfield, Q. (1995). The role of frequency modulation in the perceptual segregation of concurrent vowels. *The Journal of the Acoustical Society of America*, *98*(2 Pt 1), 837–846. https://doi.org/10.1121/1.413510

Cusack, R., & Roberts, B [B.] (2000). Effects of differences in timbre on sequential grouping. *0031-5117*, *62*(5), 1112–1120. https://doi.org/10.3758/BF03212092

d'Alessandro, C., & Castellengo, M. (1994). The pitch of short-duration vibrato tones. *The Journal of the Acoustical Society of America*, *95*(3), 1617–1630. https://doi.org/10.1121/1.408548

Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: The role of interaural time differences. *0096-1523*, *25*(3), 617–629. https://doi.org/10.1037//0096-1523.25.3.617

Davis, S. (2006). Implied Polyphony in the Solo String Works of J. S. Bach: A Case for the Perceptual Relevance of Structural Expression. *Music Perception*, *23*(5), 423–446. https://doi.org/10.1525/mp.2006.23.5.423

Demany, L. (1982). Auditory stream segregation in infancy. *Infant Behavior and Development*, *5*(2-4), 261–276. https://doi.org/10.1016/S0163-6383(82)80036-2

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Devergie, A., Grimault, N., Tillmann, B., & Berthommier, F. (2010). Effect of rhythmic attention on the segregation of interleaved melodies. *The Journal of the Acoustical Society of America*, *128*(1), EL1-7. https://doi.org/10.1121/1.3436498

Disbergen, N. R., Valente, G., Formisano, E., & Zatorre, R. J [Robert J.] (2018). Assessing Top-Down and Bottom-Up Contributions to Auditory Stream Segregation and Integration With Polyphonic Music. *Frontiers in Neuroscience*, *12*, 121. https://doi.org/10.3389/fnins.2018.00121

Dosher, B. A., & Lu, Z. L. (2000). Mechanisms of perceptual attention in precuing of location. *Vision Research*, *40*(10-12), 1269–1292. https://doi.org/10.1016/s0042-6989(00)00019-5

Drayna, D., Manichaikul, A., Lange, M. de, Snieder, H., & Spector, T. (2001). Genetic correlates of musical pitch recognition in humans. *Science (New York, N.Y.)*, *291*(5510), 1969–1972. https://doi.org/10.1126/science.291.5510.1969

Dubinsky, E., Wood, E. A., Nespoli, G., & Russo, F. A. (2019). Short-Term Choir Singing Supports Speech-in-Noise Perception and Neural Pitch Strength in Older Adults With Age-Related Hearing Loss. *1662-4548*, *13*, 1153. https://doi.org/10.3389/fnins.2019.01153

Dupuis, K., & Pichora-Fuller, M. K. (2014). Intelligibility of Emotional Speech in Younger and Older Adults. *Ear & Hearing*, *35*(6), 695–707. https://doi.org/10.1097/AUD.0000000000000082

Eipert, L., Selle, A., & Klump, G. M. (2019). Uncertainty in location, level and fundamental frequency results in informational masking in a vowel discrimination task for young and elderly subjects. *Hearing Research*, *377*, 142–152. https://doi.org/10.1016/j.heares.2019.03.015

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology*, *7*(6), e1000129. https://doi.org/10.1371/journal.pbio.1000129

Eramudugolla, R., Irvine, D. R. F., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates 'change deafness' in complex auditory scenes. *Current Biology : CB*, *15*(12), 1108–1113. https://doi.org/10.1016/j.cub.2005.05.051

Fine, P. A., & Moore, B. C. J [Brian C. J.] (1993). Frequency Analysis and Musical Ability. *Music Perception*, *11*(1), 39–53. https://doi.org/10.2307/40285598

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention--focusing the searchlight on sound. *Current Opinion in Neurobiology*, *17*(4), 437–455. https://doi.org/10.1016/j.conb.2007.07.011

Fujioka, T., Trainor, L. J., Ross, B [Bernhard], Kakigi, R., & Pantev, C [Christo] (2005). Automatic encoding of polyphonic melodies in musicians and nonmusicians. *Journal of Cognitive Neuroscience*, *17*(10), 1578–1592. https://doi.org/10.1162/089892905774597263

Gao, Z., & Oxenham, A. J [Andrew J.] (2022). Voice disadvantage effects in absolute and relative pitch judgments. *The Journal of the Acoustical Society of America*, *151*(4), 2414. https://doi.org/10.1121/10.0010123

Gervain, J., & Geffen, M. N. (2019). Efficient Neural Coding in Auditory and Speech Perception. *Trends in Neurosciences*, *42*(1), 56–65. https://doi.org/10.1016/j.tins.2018.09.004

Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., & Morillon, B. (2020). Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. *PLoS Biology*, *18*(3), e3000207. https://doi.org/10.1371/journal.pbio.3000207

Glasberg, B. R., & Moore, B. C. J [Brian C. J.] (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1-2), 103–138. https://doi.org/10.1016/0378-5955(90)90170-T

Gockel, H., Moore, B. C., & Carlyon, R. P. (2001). Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. *The Journal of the Acoustical Society of America*, *109*(2), 701–712. https://doi.org/10.1121/1.1342073

Gordon, M., & Poeppel, D. (2002). Inequality in identification of direction of frequency change (up vs. down) for rapid frequency modulated sweeps. *Acoustics Research Letters Online*, *3*(1), 29–34. https://doi.org/10.1121/1.1429653

Gunji, A., Koyama, S., Ishii, R., Levy, D., Okamoto, H., Kakigi, R., & Pantev, C [Christo] (2003). Magnetoencephalographic study of the cortical activity elicited by human voice. *Neuroscience Letters*, *348*(1), 13–16. https://doi.org/10.1016/s0304-3940(03)00640-2

Gustafson, S. J., Grose, J., & Buss, E. (2020). Perceptual organization and stability of auditory streaming for pure tones and /ba/ stimuli. *The Journal of the Acoustical Society of America*, *148*(2), EL159. https://doi.org/10.1121/10.0001744

Hake, R., Bürgel, M., Nguyen, N. K., Greasley, A., Müllensiefen, D., & Siedenburg, K. (2023). Development of an adaptive test of musical scene analysis abilities for normal-hearing and hearing-impaired listeners. *Behavior Research Methods.* Advance online publication. https://doi.org/10.3758/s13428-023-02279-y

Haywood, N. R., & Roberts, B [Brian] (2010). Build-up of the tendency to segregate auditory streams: Resetting effects evoked by a single deviant tone. *The Journal of the Acoustical Society of America*, *128*(5), 3019–3031. https://doi.org/10.1121/1.3488675

Herholz, S. C., & Zatorre, R. J [Robert J.] (2012). Musical training as a framework for brain plasticity: Behavior, function, and structure. *Neuron*, *76*(3), 486–502. https://doi.org/10.1016/j.neuron.2012.10.011

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception & Psychophysics*, *81*(7), 2288–2303. https://doi.org/10.3758/s13414-019-01846-w

Hove, M. J., Marie, C., Bruce, I. C., & Trainor, L. J. (2014). Superior time perception for lower musical pitch explains why bass-ranged instruments lay down musical rhythms. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), 10383–10388. https://doi.org/10.1073/pnas.1402039111

Huang, L., & Dobkins, K. R. (2005). Attentional effects on contrast discrimination in humans: Evidence for both contrast gain and response gain. *Vision Research*, *45*(9), 1201–1212. https://doi.org/10.1016/j.visres.2004.10.024

Huang, N [Nicholas], & Elhilali, M. (2017). Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, *141*(3), 2163. https://doi.org/10.1121/1.4979055

Huron, D. (1989). Voice Denumerability in Polyphonic Music of Homogeneous Timbres. *Music Perception*, *6*(4), 361–382. https://doi.org/10.2307/40285438

Huron, D. (2001). Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception*, *19*(1), 1–64. https://doi.org/10.1525/mp.2001.19.1.1

Hutchins, S., & Campbell, D. (2009). Estimating the time to reach a target frequency in singing. *Annals of the New York Academy of Sciences*, *1169*, 116–120. https://doi.org/10.1111/j.1749-6632.2009.04856.x

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception & Psychophysics*, *76*(8), 2522–2530. https://doi.org/10.3758/s13414-014-0732-1

Hutchins, S., Roquet, C., & Peretz, I. (2012). The Vocal Generosity Effect: How Bad Can Your Singing Be? *Music Perception*, *30*(2), 147–159. https://doi.org/10.1525/mp.2012.30.2.147

Isnard, V., Chastres, V., Viaud-Delmon, I., & Suied, C. (2019). The time course of auditory recognition measured with rapid sequences of short natural sounds. *Scientific Reports*, *9*(1), 8005. https://doi.org/10.1038/s41598-019-43126-5

Iverson, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *0096-1523*, *21*(4), 751–763. https://doi.org/10.1037/0096-1523.21.4.751

Janata, P., Tillmann, B., & Bharucha, J. J. (2002). Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive, Affective & Behavioral Neuroscience*, *2*(2), 121–140. https://doi.org/10.3758/CABN.2.2.121

Jesteadt, W., Walker, S. M., Ogun, O. A., Ohlrich, B., Brunette, K. E., Wróblewski, M., & Schmid, K. K. (2017). Relative contributions of specific frequency bands to the loudness of broadband sounds. *The Journal of the Acoustical Society of America*, *142*(3), 1597. https://doi.org/10.1121/1.5003778

Jones, M. R., Jagacinski, R. J., Yee, W., Floyd, R. L., & Klapp, S. T. (1995). Tests of attentional flexibility in listening to polyrhythmic patterns. *0096-1523*, *21*(2), 293–307. https://doi.org/10.1037//0096-1523.21.2.293

Jones, M. R., Kidd, G [G.], & Wetzel, R. (1981). Evidence for rhythmic attention. *0096-1523*, *7*(5), 1059–1073. https://doi.org/10.1037//0096-1523.7.5.1059

Joshi, S. N., Wróblewski, M., Schmid, K. K., & Jesteadt, W. (2016). Effects of relative and absolute frequency in the spectral weighting of loudness. *The Journal of the Acoustical Society of America*, *139*(1), 373–383. https://doi.org/10.1121/1.4939893

Kalinli, O., & Narayanan, S. (2009). Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(5), 1009–1024. https://doi.org/10.1109/tasl.2009.2014795

Kannyo, I., & DeLong, C. M. (2011). The effect of musical training on auditory perception. In *Proceedings of Meetings on Acoustics* (p. 25002). Acoustical Society of America. https://doi.org/10.1121/1.4733850

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*, 315–341. https://doi.org/10.1146/annurev.neuro.23.1.315

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*, 327. https://doi.org/10.3389/fnhum.2014.00327

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1714). https://doi.org/10.1098/rstb.2016.0101

Kaya, E. M., Huang, N [Nicolas], & Elhilali, M. (2020). Pitch, Timbre and Intensity Interdependently Modulate Neural Responses to Salient Sounds. *Neuroscience*, *440*, 1–14. https://doi.org/10.1016/j.neuroscience.2020.05.018

Keller, P. E., & Burnham, D. K. (2005). Musical Meter in Attention to Multipart Rhythm. *Music Perception*, *22*(4), 629–661. https://doi.org/10.1525/mp.2005.22.4.629

Kidd, G [Gerald], Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational Masking. In R. R. Fay, A. N. Popper, & W. A. Yost (Eds.), *Springer Handbook of Auditory Research. Auditory Perception of Sound Sources* (Vol. 29, pp. 143–189). Springer US. https://doi.org/10.1007/978-0-387-71305-2_6

Koelsch, S. (2011). Toward a neural basis of music perception - a review and updated model. *Frontiers in Psychology*, *2*, 110. https://doi.org/10.3389/fpsyg.2011.00110

Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nature Reviews. Neuroscience*, *15*(3), 170–180. https://doi.org/10.1038/nrn3666

Kothinti, S. R., & Elhilali, M. (2023). Are acoustics enough? Semantic effects on auditory salience in natural scenes. *Frontiers in Psychology*, *14*, 1276237. https://doi.org/10.3389/fpsyg.2023.1276237

Krentz, U. C., & Corina, D. P. (2008). Preference for language in early infancy: The human language bias is not speech specific. *Developmental Science*, *11*(1), 1–9. https://doi.org/10.1111/j.1467-7687.2007.00652.x

Larrouy-Maestri, P., Magis, D., & Morsomme, D. (2014). Effects of melody and technique on acoustical and musical features of western operatic singing voices. *Journal of Voice : Official Journal of the Voice Foundation*, *28*(3), 332–340. https://doi.org/10.1016/j.jvoice.2013.10.019

Larrouy-Maestri, P., & Pfordresher, P. Q. (2018). Pitch perception in music: Do scoops matter? *Journal of Experimental Psychology. Human Perception and Performance*, *44*(10), 1523–1541. https://doi.org/10.1037/xhp0000550

Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2024). The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, 17456916231217722. https://doi.org/10.1177/17456916231217722

Lee, D. K., Koch, C., & Braun, J. (1997). Spatial vision thresholds in the near absence of attention. *Vision Research*, *37*(17), 2409–2418. https://doi.org/10.1016/s0042-6989(97)00055-2

Leibold, L. J., & Jesteadt, W. (2007). Use of perceptual weights to test a model of loudness summation. *The Journal of the Acoustical Society of America*, *122*(3), EL69. https://doi.org/10.1121/1.2761918

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. *Neuroreport*, *12*(12), 2653–2657. https://doi.org/10.1097/00001756-200108280-00013

Liebenthal, E., Silbersweig, D. A., & Stern, E. (2016). The Language, Tone and Prosody of Emotions: Neural Substrates and Dynamics of Spoken-Word Emotion Perception. *1662-4548*, *10*, 506. https://doi.org/10.3389/fnins.2016.00506

Luce, R. D., & Green, D. M. (1978). Two tests of a neural attention hypothesis for auditory psychophysics. *0031-5117*, *23*(5), 363–371. https://doi.org/10.3758/BF03204138

Lyzenga, J., Carlyon, R. P., & Moore, B. C. J. (2004). The effects of real and illusory glides on pure-tone frequency discrimination. *The Journal of the Acoustical Society of America*, *116*(1), 491–501. https://doi.org/10.1121/1.1756616

Madsen, C. K. (1997). Focus of Attention and Aesthetic Response. *Journal of Research in Music Education*, *45*(1), 80–89. https://doi.org/10.2307/3345467

Madsen, C. K., & Geringer, J. M. (1990). Differential Patterns of Music Listening: Focus of Attention of Musicians versus Nonmusicians. *Bulletin of the Council for Research in Music Education*(105), 45–57. http://www.jstor.org/stable/40318390

Madsen, S. M. K., Marschall, M., Dau, T., & Oxenham, A. J [Andrew J.] (2019). Speech perception is similar for musicians and non-musicians across a wide range of conditions. *Scientific Reports*, *9*(1), 10404. https://doi.org/10.1038/s41598-019-46728-1

Madsen, S. M. K., Whiteford, K. L., & Oxenham, A. J [Andrew J.] (2017). Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds. *Scientific Reports*, *7*(1), 12624. https://doi.org/10.1038/s41598-017-12937-9

Marie, C., Fujioka, T., Herrington, L., & Trainor, L. J. (2012). The high-voice superiority effect in polyphonic music is influenced by experience: A comparison of musicians who play soprano-range compared with bass-range instruments. *Psychomusicology: Music, Mind, and Brain*, *22*(2), 97–104. https://doi.org/10.1037/a0030858

Marie, C., & Trainor, L. J. (2013). Development of simultaneous pitch encoding: Infants show a high voice superiority effect. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(3), 660–669. https://doi.org/10.1093/cercor/bhs050

Marie, C., & Trainor, L. J. (2014). Early development of polyphonic sound encoding and the high voice superiority effect. *Neuropsychologia*, *57*, 50–58. https://doi.org/10.1016/j.neuropsychologia.2014.02.023

Marin, C. M., & McAdams, S [S.] (1991). Segregation of concurrent sounds. Ii: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. *The Journal of the Acoustical Society of America*, *89*(1), 341–351. https://doi.org/10.1121/1.400469

Marozeau, J., Innes-Brown, H., Grayden, D. B., Burkitt, A. N., & Blamey, P. J. (2010). The Effect of Visual Cues on Auditory Stream Segregation in Musicians and Non-Musicians. *PloS One*, *5*(6), e11297. https://doi.org/10.1371/journal.pone.0011297

Martins, I., Lima, C. F., & Pinheiro, A. P. (2022). Enhanced salience of musical sounds in singers and instrumentalists. *Cognitive, Affective & Behavioral Neuroscience*, *22*(5), 1044–1062. https://doi.org/10.3758/s13415-022-01007-x

Maurer, D. (2016). *Acoustics of the Vowel*. Peter Lang CH. https://doi.org/10.3726/978-3-0343-2391-8

McAdams, S [Stephen] (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *The Journal of the Acoustical Society of America*, *86*(6), 2148–2159. https://doi.org/10.1121/1.398475

McAdams, S [Stephen]. (2019). Timbre as a Structuring Force in Music. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Springer Handbook of Auditory Research. Timbre: Acoustics, Perception, and Cognition* (Vol. 69, pp. 211–243). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_8

McAdams, S [Stephen], & Bertoncini, J. (1997). Organization and discrimination of repeating sound sequences by newborn infants. *The Journal of the Acoustical Society of America*, *102*(5 Pt 1), 2945–2953. https://doi.org/10.1121/1.420349

Merrill, J., & Larrouy-Maestri, P. (2017). Vocal Features of Song and Speech: Insights from Schoenberg's Pierrot Lunaire. *Frontiers in Psychology*, *8*, 1108. https://doi.org/10.3389/fpsyg.2017.01108

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. https://doi.org/10.1038/nature11020

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J [Andrew J.] (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*(1-2), 36–47. https://doi.org/10.1016/j.heares.2006.05.004

Micheyl, C., & Oxenham, A. J [Andrew J.] (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing Research*, *266*(1-2), 36–51. https://doi.org/10.1016/j.heares.2009.09.012

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America*, *128*(1), 435–443. https://doi.org/10.1121/1.3397384

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J [Andrew J.], Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods.* Advance online publication. https://doi.org/10.3758/s13428-020-01514-0

Molloy, K., Griffiths, T. D., Chait, M., & Lavie, N. (2015). Inattentional Deafness: Visual Load Leads to Time-Specific Suppression of Auditory Evoked Responses. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(49), 16046–16054. https://doi.org/10.1523/JNEUROSCI.2931-15.2015

Mondor, T. A., & Bregman, A. S [A. S.] (1994). Allocating attention to frequency regions. *0031-5117*, *56*(3), 268–276. https://doi.org/10.3758/BF03209761

Mondor, T. A., & Zatorre, R. J [R. J.] (1995). Shifting and focusing auditory spatial attention. *0096-1523*, *21*(2), 387–409. https://doi.org/10.1037/0096-1523.21.2.387

Monir, R., Kostrzewa, D., & Mrozek, D. (2022). Singing Voice Detection: A Survey. *Entropy (Basel, Switzerland)*, *24*(1). https://doi.org/10.3390/e24010114

Moore, B. C. J [Brian C. J.], & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, *74*(3), 750–753. https://doi.org/10.1121/1.389861

Moore, B. C. J [Brian C. J.], & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1591), 919–931. https://doi.org/10.1098/rstb.2011.0355

Moore, B. C. J [Brian C. J.], & Skrodzka, E. (2002). Detection of frequency modulation by hearing-impaired listeners: Effects of carrier frequency, modulation rate, and added amplitude modulation. *The Journal of the Acoustical Society of America*, *111*(1 Pt 1), 327–335. https://doi.org/10.1121/1.1424871

Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60. https://doi.org/10.1080/17470215908416289

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., & Ullén, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science*, *25*(9), 1795–1803. https://doi.org/10.1177/0956797614541990

Moskowitz, H. S., Lee, W. W., & Sussman, E. S. (2020). Response Advantage for the Identification of Speech Sounds. *Frontiers in Psychology*, *11*, 1155. https://doi.org/10.3389/fpsyg.2020.01155

Mueller, S. T., Alam, L., Funke, G. J., Linja, A., Ibne Mamun, T., & Smith, S. L. (2020). Examining Methods for Combining Speed and Accuracy in a Go/No-Go Vigilance Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1202–1206. https://doi.org/10.1177/1071181320641286

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS One*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Murray, M. M., Camen, C., Gonzalez Andino, S. L., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *26*(4), 1293–1302. https://doi.org/10.1523/JNEUROSCI.4511-05.2006

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, *88*(6), 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology : CB*, *32*(7), 1470-1484.e12. https://doi.org/10.1016/j.cub.2022.01.069

Oberfeld, D., Heeren, W., Rennies, J., & Verhey, J. (2012). Spectro-temporal weighting of loudness. *PloS One*, *7*(11), e50184. https://doi.org/10.1371/journal.pone.0050184

Ogg, M., Slevc, L. R., & Idsardi, W. J. (2017). The time course of sound category identification: Insights from acoustic features. *The Journal of the Acoustical Society of America*, *142*(6), 3459. https://doi.org/10.1121/1.5014057

Oxenham, A. J [Andrew J.], Fligor, B. J., Mason, C. R., & Kidd, G [Gerald] (2003). Informational masking and musical training. *The Journal of the Acoustical Society of America*, *114*(3), 1543–1549. https://doi.org/10.1121/1.1598197

Ozdemir, E., Norton, A., & Schlaug, G. (2006). Shared and distinct neural correlates of singing and speaking. *NeuroImage*, *33*(2), 628–635. https://doi.org/10.1016/j.neuroimage.2006.07.013

Palmer, C., & Holleran, S. (1994). Harmonic, melodic, and frequency height influences in the perception of multivoiced music. *0031-5117*, *56*(3), 301–312. https://doi.org/10.3758/bf03209764

Pannese, A., Herrmann, C. S., & Sussman, E. S. (2015). Analyzing the auditory scene: Neurophysiologic evidence of a dissociation between detection of regularity and detection of change. *Brain Topography*, *28*(3), 411–422. https://doi.org/10.1007/s10548-014-0368-4

Pantev, C [C.], Roberts, L. E., Schulz, M., Engelien, A., & Ross, B [B.] (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *0959-4965*, *12*(1), 169–174. https://doi.org/10.1097/00001756-200101220-00041

Parthasarathy, A., Hancock, K. E., Bennett, K., DeGruttola, V., & Polley, D. B. (2019). *Neural signatures of disordered multi-talker speech perception in adults with normal hearing.* https://doi.org/10.1101/744813

Parviainen, T., Helenius, P., & Salmelin, R. (2005). Cortical differentiation of speech and nonspeech sounds at 100 ms: Implications for dyslexia. *Cerebral Cortex (New York, N.Y. : 1991)*, *15*(7), 1054–1063. https://doi.org/10.1093/cercor/bhh206

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in Melody: The Influence of Context and Learning. *Music Perception*, *23*(5), 377–405. https://doi.org/10.1525/mp.2006.23.5.377

Pestilli, F., Viera, G., & Carrasco, M [Marisa] (2007). How do attention and adaptation affect contrast sensitivity? *Journal of Vision*, *7*(7), 9.1-12. https://doi.org/10.1167/7.7.9

Petkov, C. I., Kang, X., Alho, K [Kimmo], Bertrand, O., Yund, E. W., & Woods, D. L [David L.] (2004). Attentional modulation of human auditory cortex. *Nature Neuroscience*, *7*(6), 658–663. https://doi.org/10.1038/nn1256

Pisoni, D. B. (1979). On the perception of speech sounds as biologically significant signals. *Brain, Behavior and Evolution*, *16*(5-6), 330–350. https://doi.org/10.1159/000121875

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’. *Speech Communication*, *41*(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America*, *57*(S1), S5-S5. https://doi.org/10.1121/1.1995329

Puschmann, S., Baillet, S., & Zatorre, R. J [Robert J.] (2019). Musicians at the Cocktail Party: Neural Substrates of Musical Training During Selective Listening in Multispeaker Situations. *Cerebral Cortex (New York, N.Y. : 1991)*, *29*(8), 3253–3265. https://doi.org/10.1093/cercor/bhy193

Puschmann, S., Regev, M., Fakhar, K., Zatorre, R. J [Robert J.], & Thiel, C. M. (2024). Attention-Driven Modulation of Auditory Cortex Activity during Selective Listening in a Multispeaker Setting. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *44*(15). https://doi.org/10.1523/JNEUROSCI.1157-23.2023

Ragert, M., Fairhurst, M. T., & Keller, P. E. (2014). Segregation and integration of auditory streams when listening to multi-part music. *PloS One*, *9*(1), e84085. https://doi.org/10.1371/journal.pone.0084085

Reuter, C., Czedik-Eysenberg, I., Siddiq, S., & Oehler, M. (2018). *Formant Distances and the Similarity Perception of Wind Instrument Timbres.*

Reuter, Christoph & Czedik-Eysenberg, Isabella & Siddiq, Saleh & Oehler, Michael (Ed.) (2018). *Formant Distances and the Similarity Perception of Wind Instrument Timbres*.

Riecker, A., Ackermann, H., Wildgruber, D., Dogil, G., & Grodd, W. (2000). Opposite hemispheric lateralization effects during speaking and singing at motor cortex, insula and cerebellum. *0959-4965*, *11*(9), 1997–2000. https://doi.org/10.1097/00001756-200006260-00038

Rijsdijk, J. P., Kroon, J. N., & van der Wildt, G. J. (1980). Contrast sensitivity as a function of position on the retina. *Vision Research*, *20*(3), 235–241. https://doi.org/10.1016/0042-6989(80)90108-x

Rodriguez-Gomez, D. A., & Talero-Gutiérrez, C. (2022). Effects of music training in executive function performance in children: A systematic review. *Frontiers in Psychology*, *13*, 968144. https://doi.org/10.3389/fpsyg.2022.968144

Rose, M. M., & Moore, B. C. J [Brian C. J.] (2000). Effects of frequency and level on auditory stream segregation. *The Journal of the Acoustical Society of America*, *108*(3 Pt 1), 1209–1214. https://doi.org/10.1121/1.1287708

Ruggles, D. R., Freyman, R. L., & Oxenham, A. J [Andrew J.] (2014). Influence of musical training on understanding voiced and whispered speech in noise. *PloS One*, *9*(1), e86980. https://doi.org/10.1371/journal.pone.0086980

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Communication*, *46*(3-4), 405–417. https://doi.org/10.1016/j.specom.2005.01.010

Sala, G., & Gobet, F. (2020). Cognitive and academic benefits of music training with children: A multilevel meta-analysis. *Memory & Cognition*, *48*(8), 1429–1441. https://doi.org/10.3758/s13421-020-01060-2

Sammler, D., Grosbras, M.-H., Anwander, A., Bestelmeyer, P. E. G., & Belin, P. (2015). Dorsal and Ventral Pathways for Prosody. *Current Biology : CB*, *25*(23), 3079–3085. https://doi.org/10.1016/j.cub.2015.10.009

Santoro, R., Moerel, M., Martino, F. de, Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Computational Biology*, *10*(1), e1003412. https://doi.org/10.1371/journal.pcbi.1003412

Saults, J. S., & Cowan, N. (2007). A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *Journal of Experimental Psychology. General*, *136*(4), 663–684. https://doi.org/10.1037/0096-3445.136.4.663

Schellenberg, E. G., & Lima, C. F. (2024). Music Training and Nonmusical Abilities. *Annual Review of Psychology*, *75*, 87–128. https://doi.org/10.1146/annurev-psych-032323-051354

Schneider, P., Engelmann, D., Groß, C., Bernhofs, V., Hofmann, E., Christiner, M., Benner, J., Bücher, S., Ludwig, A., Serrallach, B. L., Zeidler, B. M., Turker, S., Parncutt, R., & Seither-Preisler, A. (2023). Neuroanatomical Disposition, Natural Development, and Training-Induced Plasticity of the Human Auditory System from Childhood to Adulthood: A 12-Year Study in Musicians and Nonmusicians. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *43*(37), 6430–6446. https://doi.org/10.1523/JNEUROSCI.0274-23.2023

Schön, D., Gordon, R. L., & Besson, M. (2005). Musical and linguistic processing in song perception. *Annals of the New York Academy of Sciences*, *1060*, 71–81. https://doi.org/10.1196/annals.1360.006

Shamma, S. A., & Fritz, J. (2014). Adaptive auditory computations. *Current Opinion in Neurobiology*, *25*, 164–168. https://doi.org/10.1016/j.conb.2014.01.011

Shamma, S. A., & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, *20*(3), 361–366. https://doi.org/10.1016/j.conb.2010.03.009

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification*, *12*(4), 283–299. https://doi.org/10.1177/1084713808325306

Siedenburg, K., & McAdams, S [Stephen] (2017a). Four Distinctions for the Auditory "Wastebasket" of Timbre. *Frontiers in Psychology*, *8*, 1747. https://doi.org/10.3389/fpsyg.2017.01747

Siedenburg, K., & McAdams, S [Stephen] (2017b). The role of long-term familiarity and attentional maintenance in short-term memory for timbre. *Memory (Hove, England)*, *25*(4), 550–564. https://doi.org/10.1080/09658211.2016.1197945

Siedenburg, K., & McAdams, S [Stephen] (2018). Short-term Recognition of Timbre Sequences. *Music Perception*, *36*(1), 24–39. https://doi.org/10.1525/mp.2018.36.1.24

Siedenburg, K., Michel, B., Özgür, E., Scheicht, C., & Töpken, S. (2023). *Vibrotactile enhancement of musical engagement.* https://doi.org/10.21203/rs.3.rs-3528141/v1

Siedenburg, K., & Müllensiefen, D. (2019). Memory for Timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Springer Handbook of Auditory Research. Timbre: Acoustics, Perception, and Cognition* (Vol. 69, pp. 87–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_4

Siedenburg, K., Röttges, S., Wagener, K. C., & Hohmann, V. (2020). Can You Hear Out the Melody? Testing Musical Scene Perception in Young Normal-Hearing and Older Hearing-Impaired Listeners. *Trends in Hearing*, *24*, 2331216520945826. https://doi.org/10.1177/2331216520945826

Siedenburg, K., Saitis, C., & McAdams, S [Stephen]. (2019). The Present, Past, and Future of Timbre Research. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Springer Handbook of Auditory Research. Timbre: Acoustics, Perception, and Cognition* (Vol. 69, pp. 1–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_1

Siedenburg, K., Schädler, M. R., & Hülsmeier, D. (2019). Modeling the onset advantage in musical instrument recognition. *The Journal of the Acoustical Society of America*, *146*(6), EL523. https://doi.org/10.1121/1.5141369

Signoret, C., Gaudrain, E., Tillmann, B., Grimault, N., & Perrin, F. (2011). Facilitated auditory detection for speech sounds. *Frontiers in Psychology*, *2*, 176. https://doi.org/10.3389/fpsyg.2011.00176

Slater, K. D., & Marozeau, J. (2016). The effect of tactile cues on auditory stream segregation ability of musicians and nonmusicians. *Psychomusicology: Music, Mind, and Brain*, *26*(2), 162–166. https://doi.org/10.1037/pmu0000143

Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, awareness, and the perception of auditory scenes. *Frontiers in Psychology*, *3*, 15. https://doi.org/10.3389/fpsyg.2012.00015

Strait, D. L., Chan, K., Ashley, R., & Kraus, N. (2012). Specialization among the specialized: Auditory brainstem function is tuned in to timbre. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *48*(3), 360–362. https://doi.org/10.1016/j.cortex.2011.03.015

Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *The Journal of the Acoustical Society of America*, *125*(5), 3328–3345. https://doi.org/10.1121/1.3097469

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., & Pressnitzer, D. (2014). Auditory gist: Recognition of very short sounds from timbre cues. *The Journal of the Acoustical Society of America*, *135*(3), 1380–1391. https://doi.org/10.1121/1.4863659

Suied, C., Susini, P., McAdams, S [Stephen], & Patterson, R. D. (2010). Why are natural sounds detected faster than pips? *The Journal of the Acoustical Society of America*, *127*(3), EL105-10. https://doi.org/10.1121/1.3310196

Sundberg, J. (2013). Perception of Singing. In *The Psychology of Music* (pp. 69–105). Elsevier. https://doi.org/10.1016/B978-0-12-381460-9.00003-1

Sundberg, J., Lã, F. M. B., & Himonides, E. (2013). Intonation and expressivity: A single case study of classical western singing. *Journal of Voice : Official Journal of the Voice Foundation*, *27*(3), 391.e1-8. https://doi.org/10.1016/j.jvoice.2012.11.009

Sundberg, J., Prame Eric, & Iwarsson J. (1996). *Vocal fold physiology: Controlling complexity and chaos*. *Vocal fold physiology series*. Singular Pub. Group.

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *The Journal of the Acoustical Society of America*, *117*(3 Pt 1), 1285–1298. https://doi.org/10.1121/1.1854312

Sussman, E. S. (2006). Multiple mechanisms of auditory attention. *The Journal of the Acoustical Society of America*, *119*(5_Supplement), 3416. https://doi.org/10.1121/1.4786818

Sussman, E. S. (2017). Auditory Scene Analysis: An Attention Perspective. *Journal of Speech, Language, and Hearing Research : JSLHR*, *60*(10), 2989–3000. https://doi.org/10.1044/2017_JSLHR-H-17-0041

Sussman, E. S., Bregman, A. S [A. S.], Wang, W. J., & Khan, F. J. (2005). Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *1530-7026*, *5*(1), 93–110. https://doi.org/10.3758/CABN.5.1.93

Sussman, E. S., Ritter, W., & Vaughan, H. G. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology*, *36*(1), 22–34. https://doi.org/10.1017/s0048577299971056

Sussman, E. S., & Steinschneider, M. (2009). Attention effects on auditory scene analysis in children. *Neuropsychologia*, *47*(3), 771–785. https://doi.org/10.1016/j.neuropsychologia.2008.12.007

Sussman, E. S., & Winkler, I [I.] (2001). Dynamic sensory updating in the auditory system. *Brain Research. Cognitive Brain Research*, *12*(3), 431–439. https://doi.org/10.1016/s0926-6410(01)00067-2

Sussman, E. S., Winkler, I [I.], Ritter, W., Alho, K [K.], & Näätänen, R [R.] (1999). Temporal integration of auditory stimulus deviance as reflected by the mismatch negativity. *Neuroscience Letters*, *264*(1-3), 161–164. https://doi.org/10.1016/S0304-3940(99)00214-1

Taher, C., Rusch, R., & McAdams, S [Stephen] (2016). Effects of Repetition on Attention in Two-Part Counterpoint. *Music Perception*, *33*(3), 306–318. https://doi.org/10.1525/mp.2016.33.3.306

Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PloS One*, *12*(10), e0186773. https://doi.org/10.1371/journal.pone.0186773

Tanner, W. P. (1958). What is Masking? *The Journal of the Acoustical Society of America*, *30*(10), 919–921. https://doi.org/10.1121/1.1909406

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: An event-related potential and behavioral study. *Experimental Brain Research*, *161*(1), 1–10. https://doi.org/10.1007/s00221-004-2044-5

Townsend, James, T., Ashby,Gregory, F., Castellan, N. J., JR, Restle, F., Birnbaum, M. H., Link, S. W., & Potts, G. R. (1978). *Cognitive theory*. Psychology Press.

Trainor, L. J., Marie, C., Bruce, I. C., & Bidelman, G. M. (2014). Explaining the high voice superiority effect in polyphonic music: Evidence from cortical evoked potentials and peripheral auditory models. *Hearing Research*, *308*, 60–70. https://doi.org/10.1016/j.heares.2013.07.014

Treder, M. S., Purwins, H., Miklody, D., Sturm, I., & Blankertz, B. (2014). Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *Journal of Neural Engineering*, *11*(2), 26009. https://doi.org/10.1088/1741-2560/11/2/026009

Uhlig, M., Fairhurst, M. T., & Keller, P. E. (2013). The importance of integration and top-down salience when listening to complex multi-part musical stimuli. *NeuroImage*, *77*, 52–61. https://doi.org/10.1016/j.neuroimage.2013.03.051

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Vliegen, J., Moore, B. C. J [Brian C. J.], & Oxenham, A. J [A. J.] (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *The Journal of the Acoustical Society of America*, *106*(2), 938–945. https://doi.org/10.1121/1.427140

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159–164. https://doi.org/10.1111/j.1467-7687.2007.00549.x

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585–594. https://doi.org/10.1016/j.tics.2005.10.011

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond " p < 0.05". *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Weiss, M. W., Bissonnette, A.-M., & Peretz, I. (2021). The singing voice is special: Persistence of superior memory for vocal melodies despite vocal-motor distractions. *Cognition*, *213*, 104514. https://doi.org/10.1016/j.cognition.2020.104514

Weiss, M. W., & Peretz, I. (2019). Ability to process musical pitch is unrelated to the memory advantage for vocal music. *Brain and Cognition*, *129*, 35–39. https://doi.org/10.1016/j.bandc.2018.11.011

Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way she sings: Enhanced memory for vocal melodies. *Psychological Science*, *23*(10), 1074–1078. https://doi.org/10.1177/0956797612442552

Weiss, M. W., Trehub, S. E., Schellenberg, E. G., & Habashi, P. (2016). Pupils dilate for vocal or familiar music. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1061–1065. https://doi.org/10.1037/xhp0000226

West, B. T. (2014). *Linear Mixed Models*. CRC Press.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear Mixed Models*. Chapman and Hall/CRC. https://doi.org/10.1201/b17198

Wingfield, A., Lombardi, L., & Sokol, S. (1984). Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation. *Journal of Speech and Hearing Research*, *27*(1), 128–134. https://doi.org/10.1044/jshr.2701.128

Winkler, I [István], Kushnerenko, E., Horváth, J., Ceponiene, R., Fellman, V., Huotilainen, M., Näätänen, R [Risto], & Sussman, E. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11812–11815. https://doi.org/10.1073/pnas.2031891100

Woods, K. J. P., & McDermott, J. H. (2015). Attentive Tracking of Sound Sources. *Current Biology : CB*, *25*(17), 2238–2246. https://doi.org/10.1016/j.cub.2015.07.043

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Zatorre, R. J [Robert J.], Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7

Zatorre, R. J [Robert J.], & McGill, J. (2005). Music, the food of neuroscience? *Nature*, *434*(7031), 312–315. https://doi.org/10.1038/434312a

Zendel, B. R., & Alain, C. (2009). Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience*, *21*(8), 1488–1498. https://doi.org/10.1162/jocn.2009.21140

Zendel, B. R., West, G. L., Belleville, S., & Peretz, I. (2019). Musical training improves the ability to understand speech-in-noise in older adults. *Neurobiology of Aging*, *81*, 102–115. https://doi.org/10.1016/j.neurobiolaging.2019.05.015

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., & Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2293–2298. https://doi.org/10.1073/pnas.0406460102

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zuk, J., Vanderauwera, J., Turesky, T., Yu, X., & Gaab, N. (2023). Neurobiological predispositions for musicality: White matter in infancy predicts school-age music aptitude. *Developmental Science*, *26*(5), e13365. https://doi.org/10.1111/desc.13365

LIST OF PUBLICATIONS BY THE AUTHOR

Bürgel, M., Picinali, L., & Siedenburg, K. (2021). Listening in the Mix: Lead Vocals Robustly Attract Auditory Attention in Popular Music. Frontiers in Psychology, 12, 769663. https://doi.org/10.3389/fpsyg.2021.769663

Bürgel, M., & Siedenburg, K. (2023). Salience of Frequency Micro-modulations in Popular Music. *Music Perception*, *41*(1), 1–14. https://doi.org/10.1525/mp.2023.41.1.1

Bürgel. M. & Siedenburg, K. (2024) Impact of Interference on Vocal and Instrument Recognition. In press at *The Journal of the Acoustical Society of America*.

Bürgel, M., Mares, D., Siedenburg, K. (2024) Enhanced salience of edge frequencies in auditory pattern recognition. In press in *Attention, perception & psychophysics*.

DECLARATION OF THE OWN CONTRIBUTION

I hereby confirm, that Michel Bürgel contributed to the aforementioned studies as stated below:

Article: Bürgel, M., Picinali, L., & Siedenburg, K. (2021). Listening in the Mix: Lead Vocals Robustly Attract Auditory Attention in Popular Music. *Frontiers in Psychology, 12, 769663. https://doi.org/10.3389/fpsyg.2021.769663*

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, carried out the experiments, analyzed the data and wrote the manuscript. Lorenzo Picinali provided the stimuli and revised the manuscript. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

Article: Bürgel, M., & Siedenburg, K. (2023). Salience of Frequency Micro-modulations in Popular Music. *Music Perception, 41(1), 1–14. https://doi.org/10.1525/mp.2023.41.1.1*

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, carried out the experiments, analyzed the data and wrote the manuscript. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

Article: Bürgel. M. & Siedenburg, K. (2024) Impact of Interference on Vocal and Instrument Recognition. In press at *The Journal of the Acoustical Society of America.*

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, carried out the experiments, analyzed the data and wrote the final paper. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

Article: Bürgel, M., Mares, D., Siedenburg, K. (2024) Enhanced salience of edge frequencies in auditory pattern recognition. In press at *Attention, perception & psychophysics*.

Author Contributions: Michel Bürgel formulated the research question, participated in the study design, analyzed the data, and wrote the manuscript. Diana Mares formulated the research question, participated in the study design carried out the experiments and wrote the manuscript. Kai Siedenburg formulated the research question, guided the study design and data analysis, and revised the manuscript.

______________________________ 20.07.2024

(name) Date

Supervisor

DECLARATION OF COMPLIANCE WITH GOOD SCIENTIFIC PRACTICE

I hereby declare, that I have completed the work independently and used the indicated facilities.

This dissertation is my own work. All the sources of information have been acknowledged by means of references.

This dissertation has neither as a whole nor in part been published or submitted for assessment in a doctoral procedure at another university.

This is to confirm that I am aware of the guidelines of good scientific practice of the Carl von Ossietzky University of Oldenburg and that I observed them.

This is to confirm that I have not availed myself of any commercial placement or consulting services in connection with my promotion procedure.

Michel Bürgel
(Name)

20.07.24
(Date)