



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Explaining Spectrotemporal Differences
between Music and Language“

verfasst von / submitted by

Felix Haiduk, B.Sc. M.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

UA 794 685 437

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

Doctor of Philosophy-
Doktoratsstudium
Naturwissenschaften
(Lebenswiss.): UG2002 Biologie

Betreut von / Supervisor:

Univ.-Prof. Tecumseh Fitch, PhD

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Claus Lamm

to my family

ACKNOWLEDGEMENTS

This thesis wouldn't have been possible without the support of many people inside and outside of academia. I learned so much from all of them and can mention but a few. First, I want to thank Tecumseh Fitch for trusting me to do a PhD, for his encouragement and advice that put me on track again and again, for giving me the freedom and the resources to explore, and for his patience in reminding me that "allows to" is not proper English. I feel prepared to craft my niche and continue my academic journey. I am grateful to you. I want to thank Claus Lamm, who always offered to help even if we didn't end up working together. I want to deeply thank Clíodhna Quigley for her encouragement, academic and emotional support and the great conversations we had and will continue to have. I want to thank Philippe Albouy very much for hosting me in Québec and for allowing me to learn from and work closely with him despite the pandemic and the Atlantic Ocean. Many thanks to Marisa Hoeschele and Dan Mann of the Budgie Lab for their advice, support and the good vibes they spread. Of the Fitch lab, I want to especially thank Raffaella Lesch and Theresa Matzinger, who were helping me in many ways and greatly contributed to making this endeavour fun. Leo Fusani and his sexual selection group and Narly Golestani and her brain and language lab allowed me to be a guest in their lab meetings. I'm grateful for this chance to see beyond my nose and for the group support - you were fabulous hosts. Many thanks to Tudor Popescu for his trust in me, his help and for together spinning ideas and making them walk. I want to thank Jozsef Arato and Yogi Jäger for having me in their journal clubs and for the inspiring discussions we had and have. I want to thank Roger Mundry for his statistics teaching and help, a cornerstone in my education. I want to thank the DK organisers, especially Sarah Deventer and Petra Pesak, the whole team in the CogBio department and my fellow PhD fellows. Importantly, there are people outside of academia who were essential to me in this journey. I want to thank Markus Böckle for helping me to understand and trust myself, Sophia Aichinger for helping me rediscover my body, breath, and presence, my dear Scheinwerfer for the magic and love, and Semira Ali for the evening walks of conversations and laughter. I am deeply grateful to Julia Ohl for helping me grow in so many ways, for her constant support and for always being there - it means so much to me. And I want to thank my parents wholeheartedly for all their love, trust and support - I'm so glad to have you.

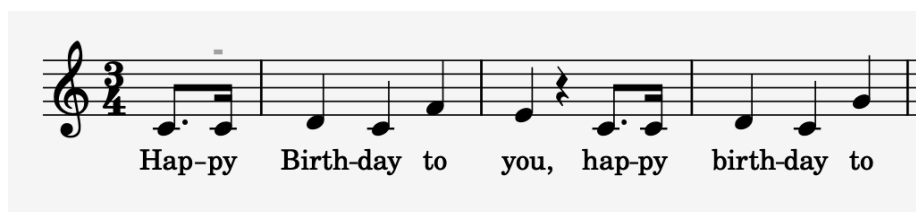
TABLE OF CONTENTS

INTRODUCTION	9
1 A first listen	9
2 Spectrotemporal differences as a useful focus	10
3 Speech and song - vocalisations as an ideal domain	12
4 Communication as grounded in social interaction	13
5 Attention, memory, and prediction as relevant cognitive processes	14
6 Marr & Tinbergen provide guiding heuristics	16
7 Aims of the Dissertation	17
References	18
CHAPTER 1	23
Understanding Design Features of Music and Language: The Choric/Dialogic Distinction	
Abstract	24
1 Introduction	25
2 Conceptual and Theoretical Framework	28
3 The ‘Goal’ Dimension: Propositional-Aesthetic Differences	32
4 The Novelty and Repetition Dimension	37
5 Interactivity: The Choric-Dialogic Dimension	41
6 Comparative Perspective	46
7 Conclusion	49
Additional Information	50
References	50
CHAPTER 2	61
Song Is More Memorable Than Speech Prosody: Discrete Pitches Aid Auditory Working Memory	
Abstract	62
1 Introduction	63
2 Study 1	66
2.1 Materials and Methods	66
2.2 Results	77
3 Study 2	80
3.1 Materials and Methods	80
3.2 Results	82
4 Study 3	84

4.1	Materials and Methods	84
4.2	Results	86
5	Post-hoc Rating Study	89
5.1	Materials and Methods	89
5.2	Results	90
6	Discussion	91
	Additional Information	96
	References	97
CHAPTER 3		101
Spectrotemporal Cues and Attention jointly modulate fMRI Network Topology for Sentence and Melody Perception		
	Abstract	102
1	Introduction	103
2	Results	105
3	Discussion	117
4	Materials and Methods	121
	Additional Information	127
	References	128
DISCUSSION		135
1	Three chapters in three perspectives	135
2	Additional helpful perspectives: active inference and semiotics	141
3	Interpreting the three chapters in the light of active inference and biosemiotics	146
4	Future research: Complementing analytical decomposition with recomposition	150
5	Conclusion	153
	References	154
ABSTRACT		161
ZUSAMMENFASSUNG		162

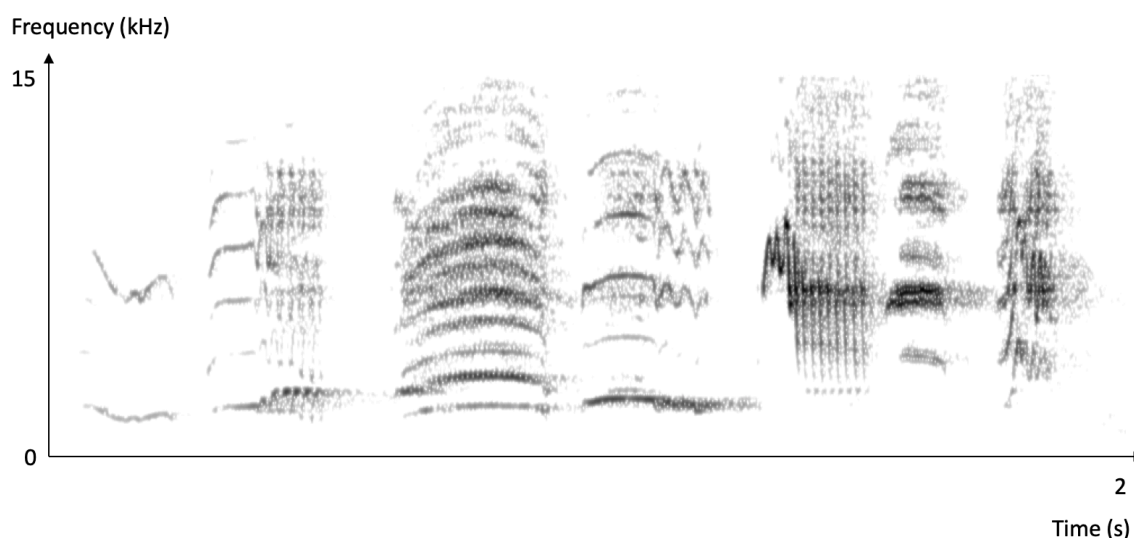
INTRODUCTION

1 A first listen



This song is probably very familiar to you, and no matter whether you can read the notes or not, you probably heard this song in your head when you saw the picture above. This imagery is fascinating by itself, but we can use it for a little experiment: listen again to the song in your head and pay attention to the pitch and the rhythm. Now imagine you are just reading the words: "Happy Birthday to you...". What changed in terms of pitch or terms of rhythm? Did you hear the same voice as before? Now imagine just the melody without lyrics. What is different from before? Did you still hear a voice or maybe an instrument playing it?

Compare the song above to another sound that is probably familiar:



This sonogram of a snippet from the song of a blackbird is an elaborate example of animal communication. Let us imagine that we tried to convert this song into notes and text-like syllables, as with Happy Birthday above. This seems not so straightforward (though it has been attempted), which is natural since notes and letters are human-made representations optimised for human communication, but also because it is still being determined how we could usefully separate pitch and sound quality in blackbird songs without taking things apart that form one entity (see Mann & Hoeschele, 2020, for a discussion). What appears so trivial is remarkable: as humans, we have two different communication systems that can be merged in some ways but remain separate, that are very

similar in their use of the vocal domain, pitch, sound quality and timing, yet different in the resulting elements and in the way they are used.

Human language and music are two of the most complex forms of communication we know of and, as such, as Tamariz (2017) puts it, "defining capacities of our species". Humans of all cultures engage in these two fundamental modes of communication, and most scholars would agree that we are also the only species with language and music, making these two systems very peculiar in the animal kingdom (Cross et al., 2013). This raises several questions: Why do we humans engage in language and music in the first place? How are these two phenomena related to each other? How do our biology and sociality shape them? How do they relate to communication in other species?

These questions share a common denominator: the origin of music and language. Finding answers entails inquiry into domains such as motivation and goals, precursors of communication, the role of general auditory cognition and physiology of hearing, cultural transmission, and whether and how all these aspects are shared among different species. The broadness of this scope suggests that the question about the origin of music and language needs to be broken down to be able to find answers, which later need to be put together again. In other words, it requires analytical decomposition (Fitch, 2015) and recomposition. In this respect, we can look at music and language as phenomena, as complex traits, as systems, as signals for communication, as sets of rules (a view from classical linguistics), or as signs (Lieck & Rohrmeier, 2021; Tomlinson, 2015), and each of these perspectives is valid for some questions. Music and language ultimately require explanations across several levels of investigation and from multiple perspectives (see also Giere, 2006; Massimi, 2022; Wimsatt, 2007). In the current thesis, I will therefore pursue analytic decomposition in several ways, each of which I will justify below: First, I will argue that spectrotemporal differences, as we encountered above, provide a useful starting point and suggest focusing on song and speech as vocal outcomes of language and music. I will then show why a perspective on music and language as rooted in social interaction is advantageous and that following from this perspective, attention, memory, and prediction are especially relevant cognitive traits. Finally, I will argue that Marr's (1982) and Tinbergen's (1963) heuristics provide useful guidance to address spectrotemporal differences in the three Chapters of this thesis.

2 Spectrotemporal differences as a useful focus

Spectrotemporal features are a useful starting point because they represent a large part of our interface with language and music, namely as spoken language and sung or instrumental music. Hence, spectrotemporal differences have been found to distinguish music and language cross-culturally (e.g., (Albouy et al., 2023; Ozaki et al., 2023)). After all, both music and spoken language consist of elements that partition the same acoustic dimensions (frequency, amplitude, formant space,

and duration). At the same time, the elements that are partitioning the spectrotemporal space are different: for example, both frequency components and temporality tend to be more regular in music (isochrony and discrete pitch). There is one circularity problem though: if spectrotemporal differences are the basis for how we define music and language, then explaining music and language by spectrotemporal differences leads nowhere. What is considered "music" or "language" can, for some instances, vary between cultures (Merker, 2000). There are, for example, certain linguistic forms that appear to make use of musical features, like talking drums (Ong, 1977) or whistled languages (Busnel & Classe, 1976). And under certain circumstances, spoken language can sound as if it was sung (Deutsch et al., 2011a; Tierney et al., 2018). This illustrates not only that music and language are culturally transmitted but also that they form a multilevel continuum where spectrotemporal cues are used for probabilistic inference depending on task goals. This latter insight will be discussed in Chapters 1 and 3.

Therefore, spectrotemporal differences have to be related to underlying cognitive processes and the behavioural contexts in which they are executed, especially social interaction. In this thesis, I will mainly focus on three interlinked cognitive processes in this respect: attention, memory, and prediction (see below). Music and spoken language tend to differ in temporal coordination (turn-taking in speech and concurrent production in music) and the performative context they are executed in (typically dyads in speech and larger groups in music, Haiduk & Fitch, 2022). This may suggest that these spectrotemporal and social-behavioural features are causally linked, a notion that we investigated in Chapter 1 of this thesis. Below I argue that regarding communication in terms of social interaction will be a valuable stance for this thesis.

Spectrotemporal differences between spoken language and music can also be considered on a more fine-grained level. First, the temporal modulation spectrum in speech appears to be focussed on faster modulations than in music: across several spoken languages, it peaks at about 5 Hz, compared to 2 Hz across several musical styles (Ding et al., 2017). Moreover, spoken language relies on fast temporal modulation as the acoustic basis for specific phonemes like plosives or clicks (see e.g. Arnal et al., 2016). On the other hand, the decoding of spectral differences with slower temporal modulation, especially note pitch, is crucial, at least for tonal music (an interesting boundary case might again be whistled languages or talking drums). In spoken language, such slower tonal aspects as intonation can often be omitted, for example, in whispering, without compromising at least the propositional semantic understanding of utterances significantly (which is not to deny that prosody can play a vital role in semantic understanding, Hellbernd & Sammler, 2016; Paulmann, 2016). These spectrotemporal differences have been related to the hemispheric lateralisation of language and music (Poeppel, 2003; Zatorre et al., 2002, but see Mcgettigan & Scott, 2012), following the idea that the left hemisphere focuses on fast temporal resolution while the right hemisphere preferentially processes slower

modulations with a focus on spectral information. Although lateralisation differences are strongly influenced by the semantic meaning of the perceived signal (McGettigan & Scott, 2012), homologous auditory regions have recently been described that encode temporal and spectral acoustic cues with the predicted lateralisation. Given that such a spectrotemporal lateralisation asymmetry has been observed in non-human animals as well (see Flinker et al., 2019), it seems possible that some aspects of spectrotemporal acoustic differences between spoken language and music exploit this lateralisation. Thus, Chapter 3 of this thesis will explore a decomposition into temporal and spectral aspects of a stimulus representing aspects of both music and spoken language: sung speech.

3 Speech and song - vocalisations as an ideal domain

I consider music and language as being evolutionarily grounded in general animal communication. This entails the possibility that at least some aspects of music and language differentiated from a common precursor (this is just one of several possibilities, see Cross et al., 2013), especially considering spectrotemporal aspects. I think this is a reasonable guiding assumption until falsified, because both music and language are dominantly executed in the audio-vocal domain, exploit the same acoustic dimensions, and make use of the same production system, the vocal apparatus, as speech and song.

There are several reasons to focus on song and speech as output modes of music and language rather than written or signed language or instrumental music. First, for most people, crossculturally, speech is the primary form of language use, and a substantial portion of the world's music is sung (Savage et al., 2015). Second, both song and speech are, therefore, easily accessible for people (even if they are not trained musicians) in both perception and production and across their lifespan. Third, the voice is an output domain that music and language share, enabling rigorously controlled comparisons, for example concerning timbre. Fourth, some brain regions primarily process vocal music and spoken language (Belin et al., 2011; Merrill et al., 2012), reflecting the salience of these vocalisations at the neural level. Fifth, vocalisations relate music and language to animal vocal communication, an essential medium across vertebrates. This embeds music and language in a larger, evolutionarily informed context, and indeed vocalisations have been proposed as one likely origin of music and language (Cross et al., 2013; Darwin, 1871; Fitch, 2013, but see Arbib et al., 2008; Tomlinson, 2015). Sixth, vocalisations enable the investigation of the interaction between bottom-up auditory perception mechanisms and top-down domain-specific processing for music and language simultaneously (see Chapter 3 of this thesis). Seventh, many technical tools are available that focus on the analysis and processing of vocal output (e.g., Boersma & Weenink, 2017), which aids stimulus creation and manipulation. Eighth, song and speech can be considered poles of a continuum, enabling us to consider and investigate intermediate forms and aspects of music and language (e.g., poetry, chant,

intonation, lexical tone, etc., see Chapter 1). This is also reflected by the aforementioned speech-to-song illusion, whereby repeated speech begins to sound like song, illustrating the deep connections between the two domains (Deutsch et al., 2011b). And finally, most of the key differences between language and music are manifested in the vocal domain (Fitch, 2006).

One might object that the audio-vocal modality is a mere surface feature used as a main communicative channel because of its readiness for innate vocalisations (laughter, cries etc.) and that a focus on the auditory-vocal modality would miss important aspects of language and music (Waciewicz & Żywiczyński, 2015). It is undoubtedly true that language and music are multimodal phenomena; both would likely not exist in their form without the constraints provided by bodily motion (Schiff & Schiavo, 2017; Vuust et al., 2022). Instrumental music does not use the vocal system, and sign language, writing, or many instances of contemporary dance do not make use of the auditory channel either. However, equally evident is that music and language can be understood without the visual presence of a physical body and that many features we observe in music and language seem to be shaped by the audiovocal channel (Trainor, 2018). Most features of music and language, bodily movement being a notable exception, can be studied using the auditory channel (but see Jacoby et al., 2020, for limitations of this approach). Thus, for the sake of simplicity, with all the disadvantages of reductionism, it seems acceptable as a first step to focus on speech and song. This thesis, therefore, focuses on speech and song as representative output modes for music and language.

4 Communication as grounded in social interaction

Human language often involves communication about external referents (Cross et al., 2013), but such communication is also clearly not confined to humans, as we famously see in bee waggle dance (Dyer, 2002; see also Scott-Phillips & Heintz, 2023). Even this notion of communication is too narrow, though, since it only comprises communicating something external to the organism - communicating emotional states appears to be even more abundant across animals (Filippi, 2016). Thus, if we want a biologically grounded way of describing communication, we need to go beyond human language; we need a more abstract way to describe communication conceptually. This also suggests that communication should be conceptualised as happening between at least two individuals. Scott-Phillips & Kirby (2013), therefore, proposed a model that regards communicative signals as acts of social influence on other individuals (see also Owren et al., 2010). In their definition, a signal is defined as a behaviour that (1) causes a reaction in another organism while (2) its function is to cause this reaction. In turn, (3) the function of the reaction itself is that it is caused by the behaviour. "Function" here means that aspect of a behaviour that can (at least in part) explain its reproduction in subsequent generations. This neatly includes an evolutionary perspective without specifying the evolutionary process itself (e.g., natural selection, cultural evolution etc). The model provided by

Scott-Philipps & Kirby (2013) is thus sufficiently wide to incorporate multimodal and multicomponent animal communication, music and language, dance and gestures, and even bacterial communication, and therefore well suited for a biologically grounded comparison of language and music.

Communication is used to guide social interactions (e.g. Seyfarth & Cheney, 2014), but in turn, social interaction can itself shape the communicative phenomena it produces. Theories that seek the basis for language learning in shared attention and shared intentionality, mediated by gestures between infants and caregivers, draw on this insight (Ibbotson & Tomasello, 2016; Pleyer & Hartmann, 2019). Children, for example, can develop novel gestures among each other through social interactions (Behne et al., 2014) and in great apes, social factors influence gestural usage learning (Fröhlich & Hobaiter, 2018; Halina et al., 2013). But also in the audio-acoustic domain, this is apparent in vocal convergence in humans (Nguyen & Delvaux, 2015), non-human primates (Candiotti et al., 2012; Fischer et al., 2020) or birds (Scarl & Bradbury, 2009), or in how human infant babbling is shaped by social interaction (Goldstein et al., 2003). In the last decade, it became generally apparent to what extent social factors influence cognition and behaviour in general, across different species but also including influence on linguistic structure (see Whiten et al., 2022, for an overview). This suggests that social interaction could provide the basis for spectrotemporal differentiation in human communication. We followed this line of thought in Chapter 1 of this thesis.

What is also suggested by a social interaction perspective on communication is the crucial role of basic cognitive processes underlying interaction itself: attention, memory, and prediction.

5 Attention, memory, and prediction as relevant cognitive processes

Attention, memory and prediction are cognitive processes fundamental for most animals but especially relevant regarding social interaction and communication.

Attentional processes are utilised in several ways. For signals to be effective, they need to be distinguished from noise. Especially in the auditory domain, this requires disentangling multiple sound sources. Auditory scene analysis (Bregman, 1994), a set of auditory processes shared across many species (Hulse, 2002), uses acoustic features to segregate and integrate sound sources. Some of these processes are preattentive (Schröger, 2007), but a large part of what is perceived depends on task goals and context and is mediated by attentional processes that provide gain to relevant signals (Parr & Friston, 2019). In turn, signals might cue attentional gain if they are learned or evolved to be perceived as relevant (REF). In the auditory domain, this salience is mediated by spectrotemporal structure, for example, periodicities in the signal (dynamic attending theory, Jones, 1976, 2018). The interaction of

spectrotemporal structure and task-related attention will be investigated in Chapter 3. In social interaction, especially joint action, interlocutors attending to a common object, each other, or each other's behaviours is crucial. It can additionally include that interlocutors know about each other's attentional focus (shared attention), whereby communicating this shared focus is essential (see Milward & Carpenter, 2018, for an overview). A social interaction perspective entails that signals such as musical or linguistic sequences can influence social interactions. A key aspect of this influence has been suggested to be attention manipulation (see Scott-Phillips & Heintz, 2023). As described in Chapter 1, the manipulation of attention is an essential basis for novelty, aesthetic appreciation, and signal coordination.

For signals to be recognised as such, some form of memory is relevant. In a more general sense, any process that can store information about an entity relatively reliably over a period can be considered to provide memory to an organism. Such storage processes can also comprise genetic and epigenetic information, connectivity of cells (see, e.g. Fitch, 2021), or evolved reactions to signals (see e.g. Scott-Phillips & Kirby, 2013). This generalised notion of memory illustrates the coupling of an organism and its niche. Consequently, some researchers have conceptualised both this general notion and cognition with the same models (e.g. Deacon, 2011; Ramstead et al., 2021). In the narrower cognitive sense, memory is relevant regarding signal usage, recognition and inference on learned signals. For music and language, the latter is most relevant. To what extent certain spectrotemporal features influence how a signal can be remembered is investigated in Chapter 2 of this thesis.

Finally, in social interaction, interlocutors must coordinate their behaviour, especially in joint action. Especially when interlocutors need to achieve high temporal or spatial coordination, as usually in music or spoken language, they rely on the mutual prediction of their actions for coordination (Sebanz & Knoblich, 2021). Social influence also entails that the signal receiver relies on prediction to make inferences on signal reference and relevance (Grice, 1975), especially when signals are learned. This requires both that signallers produce rich enough cues to their meaning and that receivers use signal context and background knowledge to infer signal meaning (both influencing attentional gain). In turn, a signal can be considered informative when it was not predicted, a notion from information theory (Shannon, 1948)). The mutual influence of predictability and spectrotemporal structure, especially considering the role of coordination, is investigated in Chapter 1.

Therefore, attention, memory and prediction are all closely related and fundamental for social interaction and communication. Consequently, they are pivotal elements in two frameworks that attempt to explain how organisms relate to each other and things in their environment: active inference (Parr et al., 2022) and biosemiotics sensu Deacon (Deacon, 2022b). The active inference framework

will become relevant in Chapter 1 of this thesis, and I will use both frameworks in the general discussion of this thesis to guide the integration of the three Chapters.

6 Marr & Tinbergen provide guiding heuristics

This thesis addresses why music and language differ in certain spectrotemporal features. This invokes Tinbergen's (1963) proposal that there are four kinds of causes in biology: individual development, evolution, function and mechanism. "Function" does not necessarily mean adaptation, as many biological traits are likely not adaptations in the narrow sense (see Fitch, 2012). Instead, following Bateson & Laland (2013), "function" can be considered more broadly as "current utility" to distinguish it from the processes that lead to its appearance. Such processes are not confined to adaptation by natural selection but can comprise flexible phenotypic responses to environmental conditions, cultural evolution, niche construction through behavioural change, and other processes (see also (Tomlinson, 2015)). Evolution and developmental processes are tightly connected (see Fitch, 2012; Jaeger & Crombach, 2012). Individual learning, for example, can constrain and alter the selective terrain and, thus, phenotypic variability (see Deacon, 2022a; but see Santos et al., 2015). Finally, "mechanism" can be explored at multiple levels, ranging from genetic expression to social cognition in certain environmental contexts. This brings in another heuristic that has been found useful in cognition research: Marr's proposal (Marr, 1982) of three levels of explanation to understand complex systems. These comprise the computational level, describing the task a system or organism has to solve; the algorithmic level, formulating how the task can be solved; and the implementational level, how it is neurally implemented. Tinbergen's and Marr's heuristics can thus be used to classify the three Chapters in this thesis, which all address Tinbergen's mechanistic level in different ways.

Chapter 1 focuses on Marr's goal and algorithmic levels: We present a framework to explain spectrotemporal differences between music and language as responses to social interaction goals, based on information theory (Shannon, 1948) as the algorithmic basis. In the paper, we take a comparative stance, examining evidence from both human and non-human communication. We conclude that critical features of music and language follow from their differential deployment regarding performative context, communicative goals, and auditory predictive structure. In this way, Chapter 1 is concerned with current utility in the contemporary reading of Tinbergen and proposes social interaction as a process of origin for spectrotemporal differences.

One of the spectrotemporal features investigated in Chapter 1 will be the focus in Chapter 2: pitch, which is typically discrete in music, gliding in spoken language and highly relevant in speech and song. The study analyses the influence of discrete and gliding pitch on auditory working memory. We thereby used synthesised vocal stimuli based on a continuum between song and speech, which

allowed us to minimise cultural biases. Memory is essential for learning in development, relating to Tinbergen's developmental level, and in this study represents the behavioural/cognitive goal in Marr's sense.

Study 3 (Chapter 3) focused on the interaction of acoustic spectrotemporal information and task-related attention in sung speech stimuli regarding hemispheric lateralisation for temporal or spectral acoustical cues (Poeppel, 2003; Zatorre et al., 2002). Using metrics from graph theory on fMRI-derived functional connectivity, we investigated multiple levels of implementation, from whole-brain connectivity to specific regions. Chapter 3 thus represents Marr's implementational level.

7 Aims of the Dissertation

My thesis is a cumulative dissertation with three main Chapters, in which I address spectrotemporal differences between language and music in two empirical projects and one review paper. I aim to contribute to the discussion about why music and language differ in critical spectrotemporal features, although both exploit the auditory-spectrotemporal space and, ultimately, why humans engage in these two modes of communication. Chapter 1, a published review article (Haiduk & Fitch, 2022), presents a novel theoretical framework based on information theory and a reverse-engineering perspective to explain several design feature differences between music and language. We propose that design feature differences follow naturally from a deployment along three dimensions - propositional or aesthetic goals, repetition or novelty, and dialogic or choric performance context. We examined evidence from humans and non-human animals and proposed several predictions and hypotheses concerning human music, language, and animal communication. The second study (Chapter 2, published as Haiduk et al., 2020), investigated the role of one key design feature of both song and speech - pitch - on a cognitive function that is fundamental to both language and music - auditory memory. We tested whether the discrete scale tones, characteristic of pitch in music, or the gliding intonation used in speech influence memory for vocal phrases, exploring whether the differential use of pitch in music and language could be rooted in different requirements for auditory memory. In the third study (Chapter 3), submitted for publication as Haiduk & Albouy, we tested to what extent hemispheric lateralisation was based on interactions of spectrotemporal acoustic cues and task-related attention.

The three studies address spectrotemporal differences from different angles and with different conceptual stances. Based on an evolutionarily informed perspective that relevant traits and aspects of musical and linguistic engagement are likely derived from non-human animals, my thesis focuses on cognitive and behavioural factors presumably relevant for spectrotemporal differentiation into musical and linguistic forms: communicative goals, performative context, spectrotemporal hemispheric

specialisation, and their interaction with attention, memory and prediction. I thereby considered behavioural-cognitive, phenomenal (audio-acoustical), and neural levels of explanation.

References

- Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., & Zatorre, R. J. (2023). Spectro-temporal acoustical markers differentiate speech from song across cultures. *BioRxiv*.
- Arbib, M. A., Liebal, K., & Pika, S. (2008). Primate Vocalization, Gesture, and the Evolution of Human Language. *Current Anthropology*, 49(6), 1053–1076. <https://doi.org/10.1086/593015>
- Arnal, L. H., Poeppel, D., & Giraud, A.-L. (2016). A neurophysiological perspective on speech processing in “The Neurobiology of Language.” In *Neurobiology of language* (pp. 463–478). Elsevier.
- Bateson, P., & Laland, K. N. (2013). Tinbergen’s four questions: An appreciation and an update. *Trends in Ecology and Evolution*, 28(12), 712–718. <https://doi.org/10.1016/j.tree.2013.09.013>
- Behne, T., Carpenter, M., & Tomasello, M. (2014). Young children create iconic gestures to inform others. *Developmental Psychology*, 50(8), 2049–2060. <https://doi.org/10.1037/a0037224>
- Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception. *British Journal of Psychology*, 102(4), 711–725. <https://doi.org/10.1111/j.2044-8295.2011.02041.x>
- Boersma, P., & Weenink, D. (2017). *Praat: doing phonetics by computer*. (Version 6.0. 36).
- Bregman, A. S. (1994). *Auditory Scene Analysis : The Perceptual Organization of Sound*. MIT Press.
- Busnel, R.-G., & Classe, A. (1976). *Whistled Languages*. Springer.
- Candiotti, A., Zuberbühler, K., & Lemasson, A. (2012). Convergence and divergence in Diana monkey vocalizations. *Biology Letters*, 8, 382–385.
- Cross, I., Fitch, W. T., Aboitiz, F., Iriki, A., Jarvis, E. D., Lewis, J., Liebal, K., Merker, B., Stout, D., & Trehub, S. E. (2013). Culture and Evolution. In M. A. Arbib (Ed.), *Language, Music, and the Brain* (pp. 541–562). The MIT Press. <https://doi.org/10.7551/mitpress/9780262018104.003.0021>
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. J. Murray.
- Deacon, T. W. (2011). *Incomplete nature: How mind emerged from matter*. WW Norton & Company.
- Deacon, T. W. (2022a). A degenerative process underlying hierarchic transitions in evolution. *BioSystems*, 222(August), 104770. <https://doi.org/10.1016/j.biosystems.2022.104770>
- Deacon, T. W. (2022b). Steps to a Semiotic Cognitive Neuroscience. In *The Routledge Handbook of Semiosis and the Brain*.
- Deutsch, D., Henthorn, T., & Lapidis, R. (2011a). Illusory transformation from speech to song. *Journal of the Acoustical Society of America*, 129(4), 2245–2252. <https://doi.org/10.1121/1.3562174>
- Deutsch, D., Henthorn, T., & Lapidis, R. (2011b). Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, 129(4), 2245–2252.

<https://doi.org/10.1121/1.3562174>

- Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, 81, 181–187. <https://doi.org/10.1016/j.neubiorev.2017.02.011>
- Dyer, F. C. (2002). THE BIOLOGY OF THE DANCE LANGUAGE. *Annual Review of Entomology*, 47(1), 917–949.
- Filippi, P. (2016). Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. *Frontiers in Psychology*, 7(SEP), 1–19. <https://doi.org/10.3389/fpsyg.2016.01393>
- Fischer, J., Wegdell, F., Trede, F., Pesco, F. D., Hammerschmidt, K., & Fischer, J. (2020). *Vocal convergence in a multi-level primate society : insights into the evolution of vocal learning*.
- Fitch, W. T. (2006). The biology and evolution of music : A comparative perspective. *Cognition*, 100, 173–215. <https://doi.org/10.1016/j.cognition.2005.11.009>
- Fitch, W. T. (2012). Evolutionary Developmental Biology and Human Language Evolution: Constraints on Adaptation. *Evolutionary Biology*, 39(4), 613–637. <https://doi.org/10.1007/s11692-012-9162-y>
- Fitch, W. T. (2013). Musical protolanguage : Darwin’s theory of language evolution revisited University of Vienna. In J. J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech, and language: exploring the evolution of mind and brain* (pp. 489–503). The MIT Press.
- Fitch, W. T. (2015). *Four principles of bio-musicology*. 10(3), 197–202.
- Fitch, W. T. (2021). Information and the single cell. *Current Opinion in Neurobiology*, 71, 150–157. <https://doi.org/10.1016/j.conb.2021.10.004>
- Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature Human Behaviour*, 3(4), 393–405. <https://doi.org/10.1038/s41562-019-0548-z>
- Fröhlich, M., & Hobaiter, C. (2018). The development of gestural communication in great apes. *Behavioral Ecology and Sociobiology*, 72(12). <https://doi.org/10.1007/s00265-018-2619-y>
- Giere, R. N. (2006). SCIENTIFIC PERSPECTIVISM. In *Movement Disorders*. The University of Chicago Press. <https://doi.org/10.1002/mds.26015>
- Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13), 8030–8035. <https://doi.org/10.1073/pnas.1332441100>
- Grice, H. P. (1975). Logic and Conversation. In D. Davidson & G. Harman (Eds.), *The Logic of Grammar* (pp. 64–153). Dickenson.
- Haiduk, F., & Fitch, W. T. (2022). Understanding Design Features of Music and Language : The Choric / Dialogic Distinction. *Frontiers in Psychology*, 13(April). <https://doi.org/10.3389/fpsyg.2022.786899>
- Haiduk, F., Quigley, C., & Fitch, W. T. (2020). Song Is More Memorable Than Speech Prosody: Discrete Pitches Aid Auditory Working Memory. *Frontiers in Psychology*, 11(December), 1–22. <https://doi.org/10.3389/fpsyg.2020.586723>

- Halina, M., Rossano, F., & Tomasello, M. (2013). The ontogenetic ritualization of bonobo gestures. *Animal Cognition*, 16(4), 653–666. <https://doi.org/10.1007/s10071-013-0601-7>
- Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88, 70–86. <https://doi.org/10.1016/j.jml.2016.01.001>
- Hulse, S. H. (2002). Auditory scene analysis in animal communication. *Advances in the Study of Behavior*, 31, 163–200. [https://doi.org/10.1016/s0065-3454\(02\)80008-0](https://doi.org/10.1016/s0065-3454(02)80008-0)
- Ibbotson, P., & Tomasello, M. (2016). What's universal grammar? Evidence rebuts Chomsky's theory of language learning. *Scientific American*, 1–19. <https://doi.org/10.1038/scientificamerican1116-70>
- Jacoby, N., Margulis, E. H., Clayton, M., Hannon, E., Honing, H., Iversen, J., Klein, T. R., Mehr, S. A., Pearson, L., Peretz, I., Perlman, M., Polak, R., Ravignani, A., Savage, P. E., Steingo, G., Stevens, C. J., Trainor, L., Trehub, S., Veal, M., & Wald-Fuhrmann, M. (2020). CROSS-CULTURAL WORK IN MUSIC COGNITION: CHALLENGES, INSIGHTS, AND RECOMMENDATIONS. *Music Perception*, 37(3), 185–195.
- Jaeger, J., & Crombach, A. (2012). Life's attractors understanding developmental systems through reverse engineering and in silico evolution. In *Advances in Experimental Medicine and Biology* (Vol. 751, pp. 93–119). https://doi.org/10.1007/978-1-4614-3567-9_5
- Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, 83(5), 323–355. <https://doi.org/10.1037/0033-295X.83.5.323>
- Jones, M. R. (2018). *Time will tell: A theory of dynamic attending*. Oxford University Press.
- Lieck, R., & Rohrmeier, M. (2021). Discretisation and continuity: The emergence of symbols in communication. *Cognition*, 215(December 2019), 104787. <https://doi.org/10.1016/j.cognition.2021.104787>
- Mann, D. C., & Hoeschele, M. (2020). Segmental units in nonhuman animal vocalization as a window into meaning, structure, and the evolution of language. *Animal Behavior and Cognition*, 7(2), 151–158. <https://doi.org/10.26451/abc.07.02.09.2020>
- Marr, D. (1982). *Vision*. Freeman.
- Massimi, M. (2022). Perspectival Realism. In *Encyclopedia of Educational Philosophy and Theory*. Oxford University Press. <https://doi.org/10.1093/oso/9780197555620.001.0001>
- Mcgettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: what's wrong, what's right, and what's left? *Trends in Cognitive Sciences*, 16(5), 269–276. <https://doi.org/10.1016/j.tics.2012.04.006.Cortical>
- Merker, B. (2000). Synchronous Chorusing and Human Origins. *The Origins of Music*, 1999, 315–327. <https://doi.org/10.1177/10298649000030S105>
- Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. D. (2012). Perception of words and pitch patterns in song and speech. *Frontiers in Psychology*, 3(MAR), 1–13. <https://doi.org/10.3389/fpsyg.2012.00076>
- Milward, S. J., & Carpenter, M. (2018). Joint action and joint attention: Drawing parallels between the literatures. *Social and Personality Psychology Compass*, 12(4), 1–27. <https://doi.org/10.1111/spc3.12377>

- Nguyen, N., & Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. *Journal of Phonetics*, 53, 46–54. <https://doi.org/10.1016/j.wocn.2015.08.004>
- Ong, W. J. (1977). African Talking Drums and Oral Noetics. *New Literary History*, 8(3), 411–429. <https://www.jstor.org/stable/468293> REFERENCES
- Owren, M. J., Rendall, D., & Ryan, M. J. (2010). Redefining animal signaling: Influence versus information in communication. *Biology and Philosophy*, 25(5), 755–780. <https://doi.org/10.1007/s10539-010-9224-4>
- Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J., Benetos, E., Proutskova, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., & Savage, P. E. (2023). Globally songs are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]. *Peer Community In Registered Reports, Preprint*:
- Parr, T., & Friston, K. J. (2019). Attention or salience? *Current Opinion in Psychology*, 29, 1–5. <https://doi.org/10.1016/j.copsyc.2018.10.006>
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference. The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press.
- Paulmann, S. (2016). The neurocognition of prosody. In *Neurobiology of language* (pp. 1109–1120). Elsevier.
- Pleyer, M., & Hartmann, S. (2019). Constructing a Consensus on Language Evolution? Convergences and Differences Between Biolinguistic and Usage-Based Approaches. *Frontiers in Psychology*, 10(November). <https://doi.org/10.3389/fpsyg.2019.02537>
- 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](https://doi.org/10.1016/S0167-6393(02)00107-3)
- Ramstead, M. J. D., Hesp, C., Tschantz, A., Smith, R., Constant, A., & Friston, K. (2021). Neural and phenotypic representation under the free-energy principle. *Neuroscience and Biobehavioral Reviews*, 120, 109–122. <https://doi.org/10.1016/j.neubiorev.2020.11.024>
- Santos, M., Szathmáry, E., Física, I. De, Carlos, D. S., Paulo, U. D. S., & Postal, C. (2015). Phenotypic Plasticity , the Baldwin Effect , and the Speeding up of Evolution : the Computational Roots of an Illusion. *Journal of Theoretical Biology*, 371, 127–136.
- Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, 112(29), 8987–8992. <https://doi.org/10.1073/pnas.1414495112>
- Scarl, J. C., & Bradbury, J. W. (2009). Rapid vocal convergence in an Australian cockatoo , the galah *Eolophus roseicapillus*. *Animal Behaviour*, 77(5), 1019–1026. <https://doi.org/10.1016/j.anbehav.2008.11.024>
- Schröger, E. (2007). Mismatch negativity: A microphone into auditory memory. *Journal of Psychophysiology*, 21(3–4), 138–146. <https://doi.org/10.1027/0269-8803.21.34.138>
- Schyff, D. Van Der, & Schiavio, A. (2017). *Evolutionary Musicology Meets Embodied Cognition : Biocultural Coevolution and the Enactive Origins of Human Musicality*. 11(September), 1–18. <https://doi.org/10.3389/fnins.2017.00519>

- Scott-Phillips, T. C., & Kirby, S. (2013). Information, influence and inference in language evolution. In U. Stegmann (Ed.), *Animal communication theory: Information and influence* (pp. 421–442). Cambridge University Press.
- Scott-Phillips, T., & Heintz, C. (2023). Animal Communication in Linguistic and Cognitive Perspective. *Annual Review of Linguistics*, 9(1), 1–27. <https://doi.org/10.1146/annurev-linguistics-030421-061233>
- Sebanz, N., & Knoblich, G. (2021). Progress in Joint Action Research. *Current Directions in Psychological Science*, 30(2).
- Seyfarth, R. M., & Cheney, D. L. (2014). The evolution of language from social cognition. *Current Opinion in Neurobiology*, 28, 5–9. <https://doi.org/10.1016/j.conb.2014.04.003>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Tamariz, M. (2017). Experimental Studies on the Cultural Evolution of Language. *Annual Review of Linguistics*, 3(1), 389–407. <https://doi.org/10.1146/annurev-linguistics-011516-033807>
- Tierney, A., Patel, A. D., & Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, 147(6). <https://doi.org/10.1037/xge0000455>
- Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschr. Tierpsychol.*, 20, 410–433.
- Tomlinson, G. (2015). *A million years of music: The emergence of human modernity*. MIT Press.
- Trainor, L. J. (2018). The origins of music: Auditory scene analysis, evolution, and culture in musical creation. In H. Honing (Ed.), *The origins of musicality* (pp. 81–112). MIT Press.
- Vuust, P., Heggli, O. A., Friston, K. J., & Kringelbach, M. L. (2022). Music in the brain. *Nature Reviews Neuroscience*, 23(5), 287–305. <https://doi.org/10.1038/s41583-022-00578-5>
- Waciewicz, S., & Żywiczyński, P. (2015). Language Evolution: Why Hockett’s Design Features are a Non-Starter. *Biosemiotics*, 8(1), 29–46. <https://doi.org/10.1007/s12304-014-9203-2>
- Whiten, A., Biro, D., Bredeche, N., Garland, E. C., & Kirby, S. (2022). The emergence of collective knowledge and cumulative culture in animals, humans and machines. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1843). <https://doi.org/10.1098/rstb.2020.0306>
- Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings*. Harvard University Press.
- Zatorre, R. 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](https://doi.org/10.1016/S1364-6613(00)01816-7)

CHAPTER 1

Understanding Design Features of Music and Language: The Choric/ Dialogic Distinction

Felix Haiduk¹ and W. Tecumseh Fitch^{1,2}

¹Department of Behavioral and Cognitive Biology, University of Vienna, Vienna, Austria

²Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria

Published (22 April 2022)
in



frontiers

Research Topic

Relationship of Language and Music, Ten Years After: Neural Organization, Cross-domain Transfer and Evolutionary Origins

Citation:

Haiduk, F., & Fitch, W. (2022). Understanding Design Features of Music and Language: The Choric/Dialogic Distinction. *Frontiers in Psychology*, 13, 786899.

Abstract

Music and spoken language share certain characteristics: both consist of sequences of acoustic elements that are combinatorically combined, and these elements partition the same continuous acoustic dimensions (frequency, formant space and duration). However, the resulting categories differ sharply: scale tones and note durations of small integer ratios appear in music, while speech uses phonemes, lexical tone, and non-isochronous durations. Why did music and language diverge into the two systems we have today, differing in these specific features? We propose a framework based on information theory and a reverse-engineering perspective, suggesting that design features of music and language are a response to their differential deployment along three different continuous dimensions. These include the familiar propositional-aesthetic ('goal') and repetitive-novel ('novelty') dimensions, and a dialogic-choric ('interactivity') dimension that is our focus here. Specifically, we hypothesize that music exhibits specializations enhancing coherent production by several individuals concurrently—the 'choric' context. In contrast, language is specialized for exchange in tightly coordinated turn-taking—'dialogic' contexts. We examine the evidence for our framework, both from humans and non-human animals, and conclude that many proposed design features of music and language follow naturally from their use in distinct dialogic and choric communicative contexts. Furthermore, the hybrid nature of intermediate systems like poetry, chant, or solo lament follows from their deployment in the less typical interactive context.

1 Introduction

Music and language are two human cognitive and communicative systems that are similar in a variety of ways: the vocal-auditory domain is typically the primary modality, but it is not the only one (writing, sign, or dance are others). Both utilise the same vocal apparatus, and similar motor systems and perceptual physiology. Their respective neural underpinnings have major shared portions. Both consist of elements combined in a hierarchical manner by certain, culture-specific rules. Both systems are learned, but have biological components shared with other species. Despite these many similarities, this paper is concerned with the differences between the two systems. Why should two universal human systems, that share so much, nonetheless exhibit consistent differences?

It is clear that there is a great variety of music and language within and across cultures, and what is termed ‘music’ varies within a culture (see Trehub et al., 2015; Thompson et al., 2019), fulfilling a broad range of psychological purposes that influence their acoustic features. For example, while dance music will engage motor systems, lullabies are used for soothing infants, and this translates into consistent acoustic differences cross-culturally (Mehr et al., 2018). Similarly, language changes when playing with young infants, reciting a mantra in a ritual, or engaging in political discussions. However, despite this variety, certain features seem to differentiate many instances of music and language (which we will term ‘typical’ in this paper). Hockett (1960) and Fitch (2006) termed these prototypical properties ‘design features’ of language and music, respectively (see Table 1). Nonetheless, borders between language and music are not clear cut (as in the case of poetry or religious chanting), and particular instantiations of music and language can be ‘more musical’ or ‘more linguistic’ than prototypical instances.

Explaining Spectrotemporal Differences between Music and Language

Design Feature	Language		Music		Definition
	V	S	V	I	
Vocal auditory channel	+	–	+	–	Signal sequences are patterns of sounds produced by the vocal tract and articulators
Broadcast transmission	+	+?	+	+	Signal sequences are detectable by anyone within given distance/line of sight
Rapid fading	+	+	+	+	Signal sequences dissipate when signalling stops
Interchangeability	+	+	+	–	Individuals can be both sender and receiver
Total feedback	+	+ ¹	+	+? ¹	Senders themselves perceive what they signal
Specialisation	+	+	+	+	A signal sequence does not directly trigger a specific behaviour in the receiver
Productivity	+	+	+	+	Ability to produce novel signal sequences
Discreteness	+	+	+	+	Signalling units are functionally distinct
Cultural transmission	+	+	+	+	The signalling system is transmitted between individuals <i>via</i> learning and teaching
Movement ²	+	+	+	+	Movements of body (–parts) accompany movements that create the signal itself
Transposability	+	+	+	+	The relationships between signal units rather than absolute features identify a signal sequence (a sentence is considered the same regardless of who spoke/signed it, a melody regardless of instrument, voice or absolute pitch)
Duality of Patterning	+	+	–	–	Signal sequences can be analysed both as units of signalling (cenemes) and meaning-bearing units (pleremes)
Generativity	+	+	+	+	Signal units are recombined according to rules
Semanticity	+	+	–	–	Fixed associations exist between meaning-bearing units and states or properties of the world/environment
Arbitrariness	+	+	–	–	The content of most meaning-bearing units is unrelated to features of signalling units
Displacement	+	+	–	–	Meaning-bearing units refer to entities outside their spatial and temporal context
Discrete pitches	–	–	+	+	Allowed pitches are based on a scale of tones related by intervals
Isochronic	–	–	+	+	Regular periodic pulse providing a reference framework for other temporal features of the signal sequence
Performative context	–	–	+	+	Classes of signal sequences (e.g. songs or styles) recur in specific social contexts
Repeatable (repertoire)	–	–	+	+	Signal sequences are distinguishable (pieces), exactly repeatable and repeated in certain contexts
A-referentially expressive	–	+?	+	+	Higher order relations of a signal sequence are cognitively mapped to movement and affective responses

These design features concern speech (including sign) or musical acts that we label as ‘typical’, e.g., spoken conversations or musical ensemble playing. V=vocal, S=signed, I=instrumental.

¹Sensorimotor.

²Added by the authors.

Table 1. Design features differing between language and music, updated from Fitch (2006).

In this paper, we propose a framework that aims to explain design features differentiating music and language as responses along three continuous dimensions. (1) the **goal** of the linguistic or musical act, with a more propositional or more aesthetic focus; (2) the repetitiveness or **novelty** of the events within a linguistic or musical sequence and (3) crucially, the **interaction** and temporal coordination between individuals participating in linguistic or musical acts, the poles of which we term ‘choric’ and ‘dialogic’. While the first two dimensions are widely recognised and discussed in comparisons of music and language, the last one has more often been neglected. We think that the interaction dimension is a crucial addition for understanding design feature differences between language and music, because major acoustic differences between spoken language and music are rooted in social cognition and interaction.

We derive these dimensions by applying a ‘reverse-engineering’ approach, based on information theory, starting from the observed design features. This framework supports predictions about the changes in design features expected for ‘nontypical’ instances of music and language, thus

laying the foundations for a more fine-grained and continuous analysis of music and language when used for different psychological and social purposes.

Our comparison of music and language focusses on social interactions and starts from the auditory domain, based on the premise that written communication is a derived form. However, both systems go beyond the purely acoustic domain (Cross et al., 2013; Levinson and Holler, 2014; Honing, 2018). For example, both music and language incorporate body movement in the form of dance and co-speech gestures, or mime and sign languages (which are typically silent). Although our framework takes the auditory domain as a starting point, we expect that it can also be applied more generally to movement-based communication, predicting changes in movement-based communication and incorporating movement into speech or song. We thus think our framework might also be useful for analysing animal communication, both acoustic and multimodal.

Our framework is in principle compatible with various hypotheses about the evolutionary relation of language and music (see Cross et al., 2013). We assume only that variation in acoustics occurs based on social and perceptual goals, pointing at fundamental relevant traits, but remain agnostic with regards to the evolutionary processes involved (biological and/or cultural) and/or the origin states of language and music (e.g., from a common audio-vocal precursor system, as Darwin, 1871 proposed). However, our framework does assume a pivotal role for audio-vocal communication at some point in evolution, thus incorporating the phylogenetically unusual trait of vocal learning (Fitch, 2006; Jarvis, 2019), which is shared by both systems. Crucially, our framework avoids dichotomous conceptions of music and language as either fully distinct or fully indissociable faculties. This notion of the differences along a continuum follows naturally both from neural evidence and from the existence of styles intermediate between music and language (poetry, rap, lament and others).

The paper is structured as follows: the first section presents the conceptual and theoretical foundations for our framework: a reverse engineering approach allows us to derive three dimensions from design features differing between language and music. The three dimensions described—goal, novelty, and interactivity—create a space within which both prototypical and non-canonical forms of both music and language can be situated. Information theory makes these design features predictable. The three sections that follow discuss the three dimensions in more detail, arguing that the characteristic design features of music and language can be understood as a function of their deployment within this three-dimensional space. The last section opens the door to comparative cognition, arguing that some vocal communication in non-human animals can also be fruitfully understood using our framework, and ends with predictions and suggestions for questions to be addressed in future empirical research.

2 Conceptual and Theoretical Framework

2.1 Design Stance and Reverse Engineering

In addition to investigations of neural and cognitive processes, individual development and cultural specifics, a deeper understanding of both language and music and how they differ requires inquiry into their evolutionary origin(s). Various hypotheses have been proposed regarding the origin of music, often concerned with finding an adaptive value (see Mehr et al., 2021; Savage et al., 2021 for recent reviews of the debate). We will not focus on possible adaptive values in this paper, nor will we investigate the causal roles of the many possible evolutionary, cultural or developmental processes involved. Rather, we will take a design stance and a ‘reverse engineering’ approach, using the design features proposed by Hockett (1960) and Fitch (2006; see Table 1) as a starting point for our framework.

The ‘design’ stance has a long tradition in biology and relies on the idea that under certain constraints evolutionary processes act to refine and optimize traits as would an engineer (Hockett, 1960; Krebs and Davies, 1997; Maynard Smith, 2000; Csete and Doyle, 2002; Richardson, 2003; Tooby and Cosmides, 2005); the use of ‘design’ in this context implies natural selection and has no association with unscientific notions of ‘intelligent design’. This allows us to ask what constraints on concrete linguistic or musical acts could plausibly yield the observed design features differentiating music and language. We will conceptualise these constraints as poles of continuous dimensions, creating a multidimensional conceptual space. Crucially, this continuous space allows us to predict how design features of non-typical instances of language and music, such as poetry or rap, should vary as a response of their deployment along the dimensions proposed.

First, note that the kind of elements that make up language and music differ. Language consists of phonemes that are the building blocks for meaning-bearing units like morphemes and words, which in turn are combined to yield sentences. This organisation rests on the need to convey propositional meaning, which is a key characteristic of prototypical language use, but not of prototypical music. Accordingly, the design feature of semanticity and those derived from it (arbitrariness, displacement and duality of patterning) discriminate prototypical language from prototypical music. Although sung music that uses lyrics is common, there is no requirement to perceive lyrics in order to recognise a sound sequence as music, and much music is purely instrumental. Music instead has stronger links to movement, and to emotional and aesthetic appraisal (Huron, 2006; see also Thompson et al., 2019 for cross-cultural perspectives). Fitch (2006) subsumes the expressive mappings of musical form to movement and emotions under the design feature ‘a-referentially expressive’.

These contrasting design feature differences between prototypical language and music suggest a trade-off between a primary goal of conveying semantic meaning for language (which we term ‘propositional’) and a goal of aesthetic appraisal (in a broad sense, see Huron, 2016) for music. We suggest that many observed design feature differences can be explained by interlocutors following either aesthetic or propositional goals. Notably, both aesthetic and propositional goals require predictive cognitive processes, but in different ways, as reviewed below. But simply categorizing music as aesthetic and language as propositional is also incomplete—some ways of speaking also pursue aesthetic goals, as in poetry, while some music has propositionality, like humming ‘Happy Birthday’, to indicate gift-giving, or songs mimicking birdsong. It is thus useful to conceive of music and language as lying on a propositional-aesthetic continuum, where language typically tends towards the propositional side while music tends towards the aesthetic side, but with some instances between these poles. We will term this continuous axis the ‘**goal**’ dimension.

A second dimension further partially differentiates language and music. Conversational language typically conveys a large amount of novel semantic information (Grice, 1975) and exact repetition is unusual. Music in contrast is typically characterised by repetition at multiple levels, from single tones or chords, motifs, and melodies, up to repeated performances of entire musical pieces. This is supported by two further contrasting design features: while language has gliding intonation, flexible lexical tone and continuously variable syllable durations, music typically consists of tones of fixed pitches organised in scales, and is prototypically characterised by rigorous timing based on isochronous meter (for exceptions see Savage et al., 2015). Thus, in music, both the temporal and the spectral acoustic dimensions relate their elements by small integer ratios. Repeatability is further related to the design feature of performative context, where certain kinds of music are repeated in specific cultural situations (e.g., lullabies to soothing babies). Repeating the same phrases does occur in language, but mostly in specific cultural situations like religious or artistic acts (e.g., prayers or poems). Typically, however, repetition is uncharacteristic of everyday conversations but abundant in music making (Savage et al., 2015). This repetitive-novel continuum is thus another dimension where music and language have a different focus, although again certain instances occupy the middle ground along the continuum. We will call this the ‘**novelty**’ dimension.

Both the goal and the novelty dimensions are widely known and discussed, and both involve predictive cognitive processes. However, although language and music can be deployed at several points along these dimensions, the predictive cues they provide differ (e.g., music has a much smaller set of possible temporal and frequency constituents than speech). We will argue that these differences make sense only when a third dimension is added, involving the timing of individual performances in a dyad or group. As Brown (2007) has argued, an important difference between music and language is their temporal coordination. Language prototypically exhibits sequential turn-taking, where speakers typically have little overlap in their utterances. In music, simultaneity is both possible and typical:

music is often performed by several people simultaneously. We will adapt the term ‘concurrent’ to refer to individuals simultaneously performing (vocalising, playing), specifically when these signals are coupled (causally related) and coordinated (thus excluding two unconnected conversations at the same party). Concurrence does not necessarily imply the same events happening at precisely the same time (which we term ‘synchronous’, following Ravignani et al., 2014). We dub the end of this dimension that involves turn-taking and alternation ‘**dialogic**’, and the pole featuring concurrent performance ‘**choric**’ (from the Greek *choros* meaning ‘chorus’). We choose these novel terms to specifically imply joint action: deliberate coordination within a common representational framework (see Sebanz and Knoblich, 2009, 2021; for music, e.g., Keller et al., 2014; for language, e.g., Tomasello, 2010). While turn-taking requires cues to predict the end of the current speaker’s phrase, concurrence requires much more fine-grained ongoing predictions about subsequent events in a vocal sequence. Again, this **choric/dialogic axis** defines a continuum, and there are intermediate cases of dialogic form in music, for example exchanging solos in jazz or call-and-response songs, and concurrence in language, such as group chanting or recitation. We call this axis the ‘**interactivity**’ dimension.

The purpose of these three dimensions (see Table 2) is to conceptualise a continuous space that can account for both prototypical instances of language and music and instances that are not considered typical, and to explain their design features as a consequence of the deployment along these dimensions.

Dimension	Pole 1		Pole 2	
	Name	Example	Name	Example
Goal	Propositional	Discussing the week’s events with a friend Singing ‘Happy Birthday’	Aesthetic	Shakespeare sonnets Listening to your favourite Beatles album
Novelty	Novelty	Listening to a conference talk Variation and recombination of melodic modules in BaAka music (Lewis, 2021)	Repetition	Word repetition for emphasis (‘I did not break the dish. I did not break the dish. I repeat, I did not break the dish’) Choruses in songs
Interactivity	Choric	Religious ensemble chanting Ensemble music Plain-tailed wren mating display (within sex)	Dialogic	Conversational speech Call-and-response song Animal antiphonal calling

Table 2. Overview of the three proposed dimensions of our framework, with examples from music, language, and animal communication.

Hockett’s design stance as applied to language has been criticised for neglecting cognition, being biased concerning the modality of transmission (auditory-vocal) and focussing on surface aspects of the linguistic code rather than its content (Waciewicz and Żywiczyński, 2015). However, these criticisms are less telling regarding music, and our approach attempts to overcome any such limitations. For example, we start our comparison of language and music assuming an auditory-vocal modality, but emphasize that it can also be applied to signed languages or mime, and incorporate facial expressions, gestures and body language, as long as information trajectories can be measured in the

target domain (see Table 3). Crucially, cognition plays a central role in our framework, *via* the mutually predictive role of the participants in temporally unfolding musical and linguistic acts, which require complex multi-time scale cognitive processes.

Assumption	Measure/method	References
Information is an adequate model of prediction, plausible to happen in the brain	Predictive coding and similar accounts	Friston, 2010; McDonnell et al., 2011; Pearce and Wiggins, 2012; Crupi et al., 2018; Koelsch et al., 2019
Entropy and information can be measured at multiple levels of the signal sequence concurrently, and their interaction can be modelled	Models based on statistical learning and using a multiple viewpoint approach	Pearce and Wiggins, 2012; Forth et al., 2016; see also Rohrmeier and Koelsch, 2012
The information/entropy trajectories of the different levels can be compared	Mutual information measures for multivariate time series (transfer entropy, partial information decomposition, etc.)	Hlaváčková-Schindler et al., 2007; Williams and Beer, 2010 (preprint); Williams and Beer, 2011 (preprint)
Context (e.g., discourse context, conceptual knowledge, etc.) can be modelled using information theory	Conditional entropy (e.g., with n -gram models)	Plantadosi et al., 2012; Mahowald et al., 2013; see also Kuperberg and Jaeger, 2016; see also Venhuizen et al., 2019
Information theory can be applied to both discrete or continuous (or discretisable) sequences, e.g., for body movement and gesturing	Discretisation of continuous signals Sample entropy, multiscale entropy	Glowinski et al., 2013; Zbili and Rama, 2021 Glowinski et al., 2010; Glowinski and Mancini, 2011

Table 3. Assumptions and measures of information theory.

2.2 Information Theory

All three of our dimensions centrally involve predictive processes. In language (goal dimension), inferring propositional meanings involves prediction at the level of semantics, while aesthetic experiences exploit the interplay between fulfilment of expectations and deviation from predictions. Repetition entails high predictability, while novelty implies low predictability. Finally, for the interactivity dimension, coordinating events in time between several individuals requires either prediction or reaction, with prediction being faster and more flexible.

Prediction involves estimates of probability: at any given point during the musical or linguistic act, possible subsequent events are assigned a probability given the current context, influencing the perceiver’s expectations about what happens next. Predictions always have a degree of uncertainty, allowing some possibility of other events to occur instead (even a highly familiar event may be corrupted by noise or mistakes). Thus, in order to support successful prediction, a signaller should decrease the uncertainty about subsequent events. Signals can in this way be analysed in terms of the change in their information over time.

The scientific field dealing with reduction in uncertainty is information theory, originally formulated by Shannon (1948), and we will use information theory as our theoretical foundation when analysing the deployment of language and music along the goal, novelty, and interactivity dimensions. The common currency is information, which is simply reduction in uncertainty, quantified in bits. If an event in a sequence is highly predictable, that event’s information content—should it occur—is low. Unexpected events are surprising and have a high information content, hence information content

is also termed ‘surprisal’. Information theory has developed considerably since Shannon’s fundamental insights, and now provides a rich toolbox for analysing a variety of phenomena (see Table 3; Crupi et al., 2018). In a crucial addition, the uncertainty of predictions themselves, i.e., the confidence in or precision of one’s own predictions (Koelsch et al., 2019), can also be quantified as the expected value of the information, or entropy (see, e.g., Hansen and Pearce, 2014). For concision, Table 3 lists some of the central assumptions we will adopt, and provides references to the methods and measures used to implement information theory in our framework.

Computational models have been successfully used to manipulate and analyse the information dynamics of sequences (e.g., Hansen et al., 2021). Most such models are probabilistic: they can capture multiple streams of musical features (see Table 3), and relying on the Markov assumption (see Rohrmeier and Koelsch, 2012), they predict local dependencies. However, predictions for musical and linguistic sequences can span more than just the next event, especially when syntax or harmonic schemas are considered (Rohrmeier and Koelsch, 2012), indicating the need for hierarchical processing across multiple related time scales (see Zuidema et al., 2018). As long as predictions for events with given probabilities are generated these can in principle be used for measuring information and entropy. Our framework will be discussed based on the prediction of the next, discrete event in a sequence, acknowledging that specific models and measures will need to take long-distance dependencies into account.

With these preliminaries in hand, we now turn to a more detailed consideration of the three axes of our framework, applying them to both prototypical song or speech, but also considering atypical or intermediate cases like poetry.

3 The ‘Goal’ Dimension: Propositional-Aesthetic Differences

The goal dimension concerns the broader purpose of linguistic or musical sequence productions, whether to convey semantic messages, or to elicit and modulate aesthetic responses in a broad sense (including emotional appraisal, pleasure, movement expressiveness, etc., see Huron, 2016). Both poles of this continuum involve predictions at multiple levels, but the poles differ in how the levels interact.

3.1 Propositionality in Language

The main goal of linguistic acts is arguably to convey propositional meaning: they enable a comprehender to infer the message the speaker intends to convey (Seifert et al., 2013; Kuperberg and Jaeger, 2016). Although speech acts can often convey social relationship, status, sex, origin, etc., paralinguistically (Ladd, 2014), propositionality is nonetheless at the core of language. From an

information theoretical perspective this entails reduction of uncertainty about the propositional content transmitted using the current context.

Applying a framework of reverse-engineering and information theory to language, Mahowald et al. (2020) argue that word length, word frequencies, and sequences of phonemes are all designed to optimise the lexicon in order to efficiently communicate, by optimally balancing complexity and informativity. This holds true over a wide variety of languages, and involves tight interactions between multiple linguistic levels. Using a comprehension model that implements both linguistic experience and world knowledge, Venhuizen et al. (2019) showed that entropy reduction is high in propositional words (reducing uncertainty in meaning), and surprisal (information) decreases towards the end of sentences, when the intended message becomes incrementally clearer. However, linguistic sequences involve multiple levels of representation (semantic, syntactic, phonological, etc.), and prediction takes place at all levels (Levinson and Torreira, 2015; Kuperberg and Jaeger, 2016 for a multimodal perspective see Holler and Levinson, 2019; for a critical review see Huettig and Mani, 2016). These levels have also been shown to interact. The hypothesis of uniform information density of a communicative act suggests a constant information rate per unit time (see, e.g., Aylett and Turk, 2004; Piantadosi et al., 2011), and studies show that speakers can actively manipulate information rate at different levels by altering for example phonetic cues, syntactic cues or word length (Mahowald et al., 2013). Specifically, enhanced prosodic prominence or longer durations are used when syllables cannot be predicted well (that is when entropy is high) based on syntactic, semantic or pragmatic contexts (Aylett and Turk, 2004). Comprehenders also use the current context for disambiguation to infer the conveyed message, and higher predictability given current contextual information yields shorter word lengths (Piantadosi et al., 2012; Gibson et al., 2019). This body of language research shows the direct interaction of information and acoustic features given a propositional goal, but also illustrates how conversational situations can be naturally implemented in an information-theoretic framework.

Thus, it appears that propositionality, specifically prediction and inference of encoded messages, profoundly affects the design of languages. The meaning-bearing level is of primary importance, and variations in predictability at the propositional level are balanced by changes in elements within non-propositional supporting levels, like phonology and word choice. These elements vary to enhance predictability (e.g., from context) or to alter the information rate (e.g., by changing in duration), supporting successful decoding of the propositional message. Thus, part of the attested prosodic variability of speech, e.g., in syllable duration or voice pitch, is an effective response that allows variable rates of information and predictability at lower levels, in support of the propositional goal.

3.2 Aesthetics and Reward in Music

Key components of the human reward system relate to prediction (expectancy) and surprise (expectancy violation; Schultz et al., 1997). When an outcome is better than expected, dopamine release is increased, resulting in positive emotions and supporting positive reinforcement learning. Worse than expected outcomes lead to decreased dopaminergic firing, negative emotions and learned avoidance. Dopaminergic firing also predicts the timing of rewarding events (Hollerman and Schultz, 1998). The difference between expected and actual outcomes is termed reward prediction error (Schultz, 2017), and involves predictions about how rewarding a future event will be, as distinguished from sensory predictions about which event will occur (de Fleurian et al., 2019; Koelsch et al., 2019). The extent to which these two predictive contexts—reward prediction error and sensory prediction error—provide appropriate explanatory frameworks for musical pleasure is debated (Colombo and Wright, 2017; Hansen et al., 2017; de Fleurian et al., 2019), but fundamental to either account is the ability to make predictions regarding sequences of sonic events. This ongoing or ‘on-line’ predictive processing is reflected in many theories of musical meaning based on tension-relaxation dynamics (e.g., Meyer, 1956; Narmour, 1990; Huron, 2006; Lerdahl and Krumhansl, 2007; see also Rohrmeier and Koelsch, 2012). However, we note that not all kinds of music rely on expectancy dynamics in order to fulfil their purposes (e.g., *Musique concrète*).

What design features allow a sequence of sounds to generate expectations, hence to be predictable, but also allow pleasant surprises and (reward) prediction errors? To generate expectations, there must be stable probabilistic relations between elements of a sound sequence, so the probability of particular events occurring concurrently or adjacent to another sound should be higher than random chance levels. Thus, regularity extraction is the foundation of statistical learning in music (Temperley, 2007), and if these relations span multiple time scales, a hierarchical structure of relations can occur (Rohrmeier and Koelsch, 2012; Rohrmeier et al., 2015).

Learned regularities regarding the temporal and spectral relations between events enable probabilistic expectations about which events are likely to occur when (Temperley, 2007). Musical pleasure has been shown to be highest when either prospective uncertainty is low and retrospective surprise is high, or vice versa (Cheung et al., 2019). Since, in music, no one level of the signal is primary *per se* (no single meaning-bearing level of the signal must be unambiguously inferred by a receiver), elements at different levels (e.g., tone frequencies, durational patterns, motifs, etc.) are not constrained to support any one primary level. Thus, both uncertainty and surprise can vary independently of each other at multiple levels, and fulfilment of predictions and surprise can occur concurrently (Rohrmeier and Koelsch, 2012)—think of a certain melodic motif where the expected last tone occurs at the expected time, but within a different harmonic context. This less constrained

design allows music to exploit the human reward system very effectively, supporting predictability at some levels and pleasant surprises at others (Zatorre, 2018).

How then can pleasure be gained from repetitive encounters with the same musical piece? Salimpoor et al. (2011) found that for familiar, liked musical pieces, dopamine is released in the striatum both in response to expectations of peak-pleasure events, and to the peak-pleasure events themselves, but in different striatal subregions. This partly explains why, even under low surprise conditions, pleasure can be gained from musical expectations being fulfilled. Representations of musical features might be sparse and decline over time, such that upon repeated listenings new predictions and prediction errors can be generated (Salimpoor et al., 2015). Furthermore, familiar music may remain rewarding upon repeated hearings if its structure is surprising in relation to other pieces of the same genre, that is when it deviates from schema-like representations (Zatorre, personal communication; Salimpoor et al., 2015). Similarly, liking familiar music can even go as far as disliking variant versions of the same song. Repeated listening to a musical piece can also allow listeners to redirect attention to levels not previously attended to and thus to discover new relations between events, again supporting novelty and surprise even in a highly familiar context (Margulis, 2014). Such attentional shifts allow music to occupy a highly rewarding sweet spot between fulfilling the prediction entirely and a total mismatch (i.e., too much information/surprise, see Zatorre, 2018).

In summary, music prototypically enables fulfilment of aesthetic goals while maintaining predictability by preserving the independence of multiple levels of the sound sequence, allowing concurrent surprise and fulfilment of predictions, as well as independent variation of prospective uncertainty and retrospective surprise. Thus, musical design solutions effectively exploit the basic mammalian dopaminergic reward system (Blood and Zatorre, 2001; Ferreri et al., 2019). Hierarchical relations between sounds in a sequence generate expectations in both music and language, but the aesthetic goal alone does not fully explain why particular design features of music arise. This becomes clearer when looking at atypical examples of language and music.

3.3 Aesthetics in Language and Propositionality in Music

Unless lyrics are present (implying a meaning-bearing linguistic layer), music rarely conveys propositional meaning. Exceptions include melodies that themselves stand for messages (e.g., whistling ‘Happy Birthday’ could convey the message of pleasant birthday wishes), ‘songlines’ that encode pathways across landscapes, connected to mythological stories (e.g., by Australian native peoples, Chatwin, 1987), or music that imitates natural sounds (e.g., birdsong). Whistled speech or ‘drum languages’ (cf. Busnel and Classe, 1976) encode propositional meaning in a superficial form, for example using pitch as a replacement of formants or phonemic tone from spoken language.

In such cases, propositional content is woven into the musical structure, and we would expect that exact repeatability plays a crucial role, because surprises would increase the uncertainty of the conveyed meaning. Altering the rhythm of ‘Happy Birthday’ substantially will make it unrecognisable, and keeping the melodic contour but changing the intervals will make it disconcerting or irritating. Imagine someone playing ‘Happy Birthday’ to you in a minor key—would you perceive this as sarcastic or ironic? It seems that in cases of propositionality in music, the acceptable variability of the musical structures is reduced, even more than in speech acts, because here the propositional message is encoded in several levels of the whole musical structure (e.g., pitch and rhythm), not primarily at a single semantic level. Such propositional musical pieces are thus more similar to words than sentences. On the other hand, adding a surprising context could make the piece aesthetically more interesting, thus shifting the goal toward the aesthetic pole.

What is predicted when language is deployed in a mainly aesthetic context? Language can also exploit the human reward system *via* generation of expectations, *via* its hierarchical structure of elements. When the goal is propositional, variations in semantic predictability are balanced by changes of elements within non-propositional levels to maintain a roughly uniform information density (see above). Thus, prospective prediction and retrospective information are tightly linked. But with an aesthetic goal, this constraint can be released, with levels of the signal becoming more independent. Enhancing the predictability of content words is no longer necessary, more variability in predictive uncertainty and surprise become possible, and attention can be focussed on other levels of the sequence. For example, in poetry intonation, phonology (rhyme), durations, stress patterns, etc., appear to vary more independently of propositional content. Propositional content is often not straightforward in poetry, and ambiguity and multiple possible interpretations are frequent. Indeed, some poetry in art movements like Dada, such as Kurt Schwitters’ ‘Ur-Sonata’ (see Schwitters, 1973), focusses on sound quality rather than propositional content (despite, in historical context, ‘conveying the message’ of ignoring artistic bourgeois conventions). Re-reading or re-hearing a poem can also yield new ways of interpretation similar to re-listening to a musical piece (but see Margulis, 2014). Increased independence of hierarchical levels might allow greater embodiment and/or a more musical perception, for example in a Shakespearean sonnet versus rap.

Infant-directed speech is another example of speech moving toward the aesthetic pole (e.g., Thiessen et al., 2005), although distress in young children is reduced more in response to infant-directed song than infant-directed speech (Corbeil et al., 2016), even for unfamiliar songs (Cirelli and Trehub, 2020). This might be related to the discreteness (high predictability) of pitch and especially duration in music. Our conception of flexibility along the propositional-aesthetic dimension could readily be applied to theatre and opera, both of which have to fulfil both propositional and aesthetic goals concurrently. We predict that predictability is traded off such that passages perceived as highly aesthetic are lower in information content, and vice versa.

To sum up, both language and music can be deployed in atypical propositional and aesthetic contexts, and similar responses follow: with more propositional goals, the multiple levels of the speech or musical sequence are more interdependent, and vary their information density to support successful inference of propositional content. For aesthetic goals, independent variation across levels enables more unconstrained variation in uncertainty and surprise, effectively exploiting the human reward system. However, given that music and speech can both be deployed in the nontypical context, aesthetic versus propositional goals alone cannot explain why certain design features characterize most music (e.g., discrete pitches or isochronous meter) but not speech (e.g., gliding intonation and variable syllable durations). This implies that further dimensions are necessary to explain these design differences.

4 The Novelty and Repetition Dimension

The novelty-repetition dimension is closely linked to the propositional-aesthetic dimension. This dimension involves the repeatability of elements and their relations at different scales (from single elements to entire pieces) and at multiple levels of musical or linguistic sequences, and their balance in use with novel elements and relations. Generally, repetition enhances predictability, whereas novelty is unpredictable and thus high in information.

4.1 Repetition in Music

One of the design features distinguishing prototypical music from language cross-culturally is that music is characterised by repetition at multiple levels (Fitch, 2006; Savage et al., 2015). Repetition can involve single notes, melodic motifs, chord progressions, rhythmic patterns, and the entire musical piece. Repetitiveness in music seems to be also a foundational perceptual principle: the speech-to-song illusion is a striking phenomenon in psychological research on music and language, whereby repetition of speech phrases leads to them being perceived as sung speech (Deutsch et al., 2011). Certain speech phrases, especially when characterised by relatively flat within-syllable pitch contours and less variability in tempo, are more prone to be judged as musical by Western listeners (Tierney et al., 2018). The repetition effect has recently been generalised to repetitions of random tone sequences (Margulis and Simchy-Gross, 2016) and of environmental sounds. These were judged as more musical by Western listeners (Rowland et al., 2019), suggesting that repetition leads to the inference of structural relationships between repeated sounds (cf. Winkler et al., 2009), which are then cognitively interpreted as ‘musical’.

What specific features of music allow or select for repeatability? Prior to recording technology, repetition entailed that a sound sequence be remembered and reproduced. To be remembered a sequence must be distinguishable from other, similar sequences (e.g., related melodies or rhythmic patterns), and learnable by establishing relationships between the constituent events. The

existence of sound categories and hierarchical rules to combine them (Herff et al., 2021; see also Rohrmeier and Pearce, 2018a,b) enables this. The musical design solutions in this respect are discrete tones in scales (in a hierarchical relation), and durations related in a simple fashion. From an information theoretical perspective, this means that the possible uncertainty about forthcoming musical events is reduced from the outset by adopting a smaller ‘alphabet’. This allows a lower number of plausible continuations of a sound sequence than if frequency and temporal dimensions were unconstrained. Because hierarchical relations exist between tones this factor also constrains plausible continuations among distant elements. Reduced alphabet size also supports statistical learning and the application of Gestalt principles, both relevant for prediction in music (Snyder, 2000; Morgan et al., 2019).

Repeatability in music seems to be particularly related to the fact that the temporal dimension in music is also hierarchically structured—durational patterns are related to an underlying meter. First, meter supports embodiment *via* beat extraction and entrainment (Kotz et al., 2018), adding a strong motoric component that may increase the memorability of musical sequences (Brown and Palmer, 2012). Second, meter can also function as a kind of glue between multiple levels of a musical sequence by enforcing relations among them, including higher-order levels like chord progressions, motifs etc. The auditory system is able to make predictions and track deviations at multiple levels at the same time (Vuust et al., 2011). High uncertainty in memory at one level of the musical signal (e.g., in melodic arrangement of pitches) can be countered by low uncertainty in another (e.g., rhythm), reducing the joint uncertainty of both levels and enhancing the confidence in the prediction of the ongoing musical sequence (‘I remember that this particular pitch followed with this rhythm’).

Is repeatability sufficient to explain the occurrence of discrete pitches on scales and meter in music? Rapid learning of auditory events is even possible for arbitrary sounds that are repeated within a stream of random sounds (Agus et al., 2010), suggesting that the auditory system is capable of finding repetition in the auditory stream irrespective of discreteness. This observation is consistent with our claim that specific design solutions for repeatability in music are not strictly necessary for perception, but relate to (re-)production. However, humans are easily capable of reproducing sound sequences that are not characterised by a reduced alphabet in the frequency and/or temporal domain. This suggests that repeatability is not a sufficient explanation for these design solutions of music. What seems to be crucial, we will argue below, is the interactivity between individuals in a group, when making music together in a choric context.

To summarise, repeatability in musical performances involves a reduction in the alphabet in multiple dimensions. This enables higher predictability and structural relations in a hierarchical manner between elements. In music, meter allows strong temporal predictions, enforcing predictive relations in higher-order levels and enabling a strong link to motoric processing. Scales in melody

allow equally strong frequency predictions, since the pitch of possible following notes is strictly circumscribed.

4.2 Novelty in Language

As emphasised above, language is mainly concerned with the primary goal of transmitting propositional meaning. These messages conveyed should be relevant and informative, and thus (typically) novel (Grice, 1975; Sperber and Wilson, 1986). The novelty typifying language acts is therefore closely linked to propositionality. What design features enable novelty in language? Crucially, language is characterised by duality of patterning (Hockett, 1960), and can be analysed both as an arrangement of meaning-bearing units (morphemes and words) supported by a lower-level arrangement of meaningless phonemes. Meaning-bearing units can be rearranged to convey new messages, which is termed productivity (Ladd, 2014). This productive layer is the main one that realises novelty (although neologisms can also enable novelty at the phonological level). Even repetition of propositional content is typically realised by a different arrangement of words or morphemes.

In language, repetition as a structural relationship of (relatively) categorical sound elements does occur at the phonemic level, where learned structural relationships between phonemes hold within a particular language. This is comparable to reduction of sound categories in music: a finite set of phonemes and specific restrictions on their combinations reduces the uncertainty of which phoneme could follow in a sequence. Words are also repeated (although the size of the lexicon is vast). Indeed, long-term memory for melodies has been proposed to be comparable to the word lexicon (Peretz et al., 2009). Language therefore can be interpreted as balancing novelty and repetition, prototypically by differentially deploying them at different levels of the linguistic stream—phonological repeatability enables morphosyntactic and semantic novelty. Thus, in prototypical conversational language, novelty is realized at the morphosyntactic and semantic levels, with phonology and the rote-memory lexicon playing a supporting role.

4.3 Repetition in Language and Novelty in Music

What happens when repetition in language occurs at the productive level, that is with morphemes, words, and sentences? Some instances of repetition are relevant in a propositional sense: repetition of the same word or morpheme (reduplication) can be used for emphasis, or serve grammatical functions like plural marking (Hurch and Mattes, 2005). Repetition might also encourage the receiver to seek different interpretations of the phrase that are not apparent at the first glance, to resolve ambiguity (Knox, 1994).

Some situations however require the repetition of entire speech phrases, for example in ritualised contexts. When memorability needs to be enhanced, this is achieved by emphasising structural relationships in other levels of the speech phrase like intonation, stress, using rhyme or specific repeated syllabic patterns (e.g., poetic forms). This can also be observed in infant-directed speech which is very repetitive (Margulis, 2014). A link to memory might be that attention allocation seems to be related to surprising events (Forth et al., 2016; Koelsch et al., 2019). In the event-related potential, a mismatch negativity, indexing unpredicted and thus surprising events, is usually followed by a P3a component, associated with attention allocation (Schröger et al., 2015). More independence of levels of the speech signal would enable more surprising events due to possible unexpected interactions between levels, emphasising the structural relationships between them. On the other hand, predictive cues can also guide attention to a specific stimulus or stimulus feature (Gazzaley and Nobre, 2012), enhancing memory encoding. Our framework predicts that actions with an aesthetic goal, where we expect a greater independence of representational levels of the sequence and more variety in predictability, should be remembered better. In line with this, Margulis (2014) proposes that memory for music, poetry or utterances with schematic form, like jokes, is based more on acoustic surface structure than in conversational speech: speech involves attention allocation towards propositional content. Note that this enables paraphrasing the same propositional content with different words, which is more difficult for musical structure with notes or chords.

Turning to novelty, because attention is drawn to surprising events (Forth et al., 2016; Koelsch et al., 2019), listening to music that is highly predictable and unsurprising could lead to attentional shift and boredom. Thus, an additional pressure for music is to include a degree of novelty. One design solution to balance both novelty and repetition is meter (hierarchical relation of durational patterns relative to a beat). Meter provides a predictive framework within which novelty—unexpected and surprising events—is well defined (e.g., syncopation). Because multiple levels of the signal allow for predictability within and across levels by means of probabilistic relationships between their elements (tones, intervals, chord progressions, etc.), each level also allows for surprise. In repeated performances novelty can be provided by slight shifts in performance style, tempo, expression, etc., making the interpretation of familiar pieces a common focus of Western classical music concerts or opera. Concerning recordings, the possibility of attentional allocation to different levels of the piece with each repeated listening could be interpreted as listener-generated ‘novelty’, since new, potentially surprising, relations might be perceived. Thus, music also balances novelty and repetition in multiple ways, but they are quite distinct from prototypical conversational language.

In summary, both music and language balance repetition and novelty, but in different ways. While language usually allocates repeatability to the phonological level and novelty occurs at the morphosyntactic and semantic levels (related to propositionality), music typically allows both novelty and repeatability across all levels of the musical sequence, and meter seems to be especially crucial as

a predictive layer throughout, enabling both prediction and surprise. However, language can also be repeatable at the word and sentence level. Thus, despite the clear validity and value of the two traditional dimensions on which music and language are differentiated — goal and novelty — certain design features are still not fully explained. We therefore suggest that understanding the design differences of language and music require a further explanatory dimension, to which we now turn.

5 Interactivity: The Choric-Dialogic Dimension

Our proposed interactivity dimension concerns the temporal coordination of linguistic or musical productions of multiple participants. Both concurrent production, in choric mode, and dialogic turn-taking involve joint actions that are causally coupled, but they pose different constraints on predictability in sequences.

5.1 Dialogic Contexts in Speech

Two speakers talking at the same time constitute noise for each others' speech signals: overlapping signals make propositional content harder to decode (Fargier and Laganaro, 2019). Therefore (among possible roots in cooperative social interaction, see Levinson, 2016; Pika et al., 2018; Pougnauld et al., 2022; but see Ravignani et al., 2019), dialogic contexts favour the avoidance of overlap and the coordination of turn-taking behaviour (Levinson and Torreira, 2015; Levinson, 2016), and signals should be designed such that receivers can predict the ending of an utterance (see, e.g., Castellucci et al., in press). Information should therefore be low (and predictability high) at the end of signal sequences. Given the requirement to reduce uncertainty in conveyed propositional messages, this should lead to high information density during most of the signal sequence (to optimally exploit the speech channel capacity, see above), with a decrease in information towards its end.

In dialogue, turn completions seem to be predicted based on both prosodic cues (Bögels and Torreira, 2015) and lexicosyntactic content (de Ruiter et al., 2006; Torreira et al., 2015), whereby semantic content seems to be more important in predicting the end of a speaker's phrase than syntax (Riest et al., 2015; see Jongman, 2021, for an overview). On the other hand, content prediction might be used to enable early response planning in parallel with comprehension of the current turn (Levinson and Torreira, 2015; Corps et al., 2018), which helps to avoid large gaps between turns that could themselves be interpreted as meaningful (e.g., Pomerantz and Heritage, 2013). Accordingly, Castellucci et al. (in press) proposed separate pathways for turn-timing and response planning. Since response planning requires neural resources that might compete with those utilised for comprehension (Bögels et al., 2015, 2018, see also Knudsen et al., 2020 for the role of backchannels, fillers and particles in this regard), it should start once the semantic uncertainty is low enough, and preferentially happen in places along the sequence that are low in information. Once the remainder of the sequence

is highly predictable, interlocutors can exchange the roles of sender and receiver: taking turns. The next utterance will again start high in uncertainty, requiring informative events, until it nears its end.

The information density trajectory must be perceivable if the current receiver is to be able to predict when uncertainty is low enough to take the floor. This requires an ongoing monitoring of the information density in the received signal and a continuous prediction of the amount of uncertainty reduction that will follow from later events. That is, listeners need to predict when new events do not reduce uncertainty much further, probably based on the semantic uncertainty of the conveyed message and taking several past events into account to capture the general information trajectory. If the time point of turn-taking is marked by low information density, then this should be unambiguously distinguishable from local information density minima that may occur in the signal before. In the case of language, some words are more informative than others even when the speech utterance is not finished yet (Venhuizen et al., 2019). Because a speech act consists of multiple interacting and integrated levels—phoneme level, morpheme and word level, prosodic intonation, stress, lexical tone, etc., as well as paralinguistic information like facial or body expressions (see e.g., Holler et al., 2018; Wohltjen and Wheatley, 2021) and contextual cues in the environment (Ladd, 2014; Kuperberg and Jaeger, 2016; Holler and Levinson, 2019), each level can play its own part in reducing uncertainty about the propositional semantic content that should be conveyed. In line with this, phrase-final lengthening (a prosodic cue that can signal turn-ending, see Wightman et al., 1992) would decrease information per unit time, while in turn speakers accelerate their speech rate and thus information per unit time when they want to continue their utterance (Walker, 2010). We would predict that in order to mark the ending of a speech phrase and to take turns, the end of a speech phrase should be highly predictable across all levels of the sequence, even if one feature (like falling intonation) might preferentially mark the ending of the speech phrase at a prosodic level (see de Ruiter et al., 2006). Evidence seems to confirm that prosodic, lexical and syntactic levels interact to mark turn-taking (reviewed in Forth et al., 2016). This makes sense: if only one single level, such as falling intonation or lengthening, predicted the end of the current speaker's utterance, it would be highly surprising if an unpredicted, highly informative event occurred at another level (for example a highly surprising word). Information density would locally peak and the receiver would likely stop their preparation to take turns and re-allocate attention. Precisely this cross-level effect is used in investigations of speech processing by event-related potentials such as the N400, an evoked potential component which deflects negatively when target words in sentences are semantically unexpected (e.g., Grisoni et al., 2017; but see Maess et al., 2016 for a differential effect for verbs and nouns), even if there is no surprise at all other levels of the speech stream (e.g., intonation, syntax, phonology, etc.). Syntactic violations in contrast are indexed by earlier evoked components like the ELAN (Hahne and Friederici, 1999), illustrating that the brain uses prediction at multiple levels of the speech stream in parallel.

The example mentioned above also illustrates an interaction with the propositional-aesthetic dimension. If there is propositional content, the representational levels of the sequence that carry this content are of primary importance, while other levels support the semantic predictability and are thus less free to vary in their information trajectories than when propositional content is absent (as in nonsense speech or music). An event high enough in information at a supporting level might both alter the semantic understanding and disturb the turn-taking process. Thus, robust semantic understanding under the constraint of efficiency (Gibson et al., 2019) might facilitate successful turn-taking as well. In contrast, the less focus is on the propositional content, the less a need for hierarchy among levels exists, which means information density among levels should be freer to vary, possibly converging only towards the end of the signal sequence to enable turn-taking.

Interestingly, if there are more than two participants in the conversation, information density alone cannot be used to coordinate who will start the next speech act. In order to achieve this, paralinguistic information like pointing gestures, naming or rules about who speaks next need to apply (e.g., Mondada, 2013). The empirical prediction would be that the larger the group, the higher the danger of overlap between former receivers' initiation of speech acts, and the more the requirement for paralinguistic coordination (or an individual designated to choose the next speaker, e.g., the chair of a meeting). However, such overlap should occur only after one interlocutor ends their speech phrase, since all receivers can predict the end of the current speaker's turn.

To sum up, dialogic contexts require that endings of sequences are perceivable by a decrease in information density (which means an increase in predictability) across levels of the sequence, such that later events have on average lower information than former events. Both language and music are designed to fulfil this requirement in dialogic contexts. The propositional focus in most spoken dialogues adds an additional constraint that non-propositional levels should be subordinate to the levels conveying propositional content.

5.2 Choric Contexts in Music

Turning to the concurrent, choric pole, successful concurrent performance requires that signals do not disrupt processing when they overlap. One design feature that avoids masking by concurrent sounds (which is more effective when frequencies are more similar, see Moore, 2014) is to make them discrete and related by small integer ratios, as are the tones on musical scales. One example is the octave, whose existence across cultures is a statistical universal, and which enables all members of a group to sing in unison even when males' vocal range lowers after pubertal vocal change (Harries et al., 1998). In line with this, octave equivalence (perceiving two pitches as categorically the same when they are an octave apart) seems not to be perceived in a culture where individuals rarely sing together (Jacoby et al., 2019). This 'simple ratios' constraint interacts with another melodic design feature:

reduction of the set of possible tones by limiting them to a small set (a ‘scale’). Again, in information theoretic terms, establishing pitches on scales with strict tonal relations involves a reduction of the alphabet of allowed symbols along the fundamental frequency dimension (for a proposal for the roots of tonality in the physiology of hearing see Trainor, 2018). This limited set of possible tones allows individuals to join a music making chorus, match the produced sound sequences and/or complement them (for example in BaAka polyphonic singing, see Lewis, 2021), thus contributing to a unified sound entity in a coherent performance (a joint action with the deliberate coordination of actions, see Sebanz and Knoblich, 2009, 2021; Tomasello, 2010; Keller et al., 2014) rather than generating a set of sounds that are not causally coupled (cf. Ravignani et al., 2014). If scale tones are hierarchically related, the continuation of a melody can be predicted with a limited uncertainty by the participating individuals, allowing them to contribute in an ongoing manner, as well as allowing for variation and thus individuality in their contribution (cf. Savage et al., 2021).

The choric context also requires that events in separate sound streams should be tightly coordinated in time. Uncertainty in timing would lead to disintegration of concurrency and coordinated joint action. Therefore, the signal should be designed to enable high-precision temporal predictability throughout. One key design feature that enables such predictability is isochrony. Tight coordination however requires participants to attend to the other participants’ actions, since ongoing coordination means prediction and monitoring on a moment-to-moment basis (see, e.g., Keller et al., 2014). If the next element can be precisely predicted in time, then the temporal information gain from each event is low, which lowers attentional demands (Koelsch et al., 2019). On the other hand, unpredicted events capture attention. How can these two requirements—high predictability and ongoing attention allocation—be aligned? If isochrony happens not at the level of each individual event, but on a meta-level, providing a scaffolding which still enables novelty (see the section about the novelty dimension), then both requirements can be fulfilled. The design solution satisfying these constraints is the concept of hierarchical meter, which allows certain placements in time and forbids others, but which also gives room for variability to create novelty since not each possible slot needs to be filled by an event. Again, meter represents a reduction in the alphabet, in this case a small set of possible onset and duration patterns relative to the beat.

Unlike in spoken language, there is much less noise and thus much less uncertainty when different participants in a choric performance contribute with different events to each of the levels of the musical sequence. We would therefore expect more degrees of freedom in terms of what event—which tone or chord—occurs than in speech deployed in a choric context (see below). However, the timing, that is when events occur in choric performances, is crucial and needs to be coordinated in a precise manner. Note that with a meter, events do not need to be played all at the same time (synchrony) nor be evenly spaced in time (isochrony). Rather, the burden is to keep events within the metrical scaffolding, so that the contributions of the participants relate to each other in the moment, or

there would no longer be one coherent performance. Even in cases of notated music where it is clear which note must be played when, there is still the need for coordination among musicians, and in the absence of strict isochrony, a coherent performance needs to be otherwise synchronised, for example by using participants' body motion.

In summary, we argue that meter is a crucial design feature that develops in music deployed in a choric context, with the goal of balancing coordination and attention, while still permitting variation or improvisation. Meter provides a predictive fabric throughout an ongoing performance. This is complemented by a reduction of the alphabet in the tonal domain: a limited set of hierarchically related pitches allows accurate predictions of possible continuations of an ongoing acoustic performance, and of multiple different complementary event streams, without compromising the coherence of the overall performance as a joint action.

5.3 Dialogic Contexts in Music and Choric Contexts in Speech

Dialogic contexts also can occur in music, for example when several musicians take turns soloing in jazz, or in call-and-response singing. Here the same constraints must apply as for spoken dialogue, and we expect musical phrases to show lower information density towards the end. Thus, there must be a means to increase predictability at phrase endings. This aligns well with the notion of musical closure in harmony and melody, or the tendency of melodies to be shaped like an arch (Huron, 2006). Again, in music we expect that phrase endings tend to have low information density across all levels of the sequence on average. There are conventions which time or mark musical phrase endings, like the number of beats a soloist has available to perform their solo, or certain rhythmical or musical motifs, but we predict that even in these cases there should be decreasing information density towards musical phrase endings. An example illustrating how the prediction of a phrase ending can be disturbed if one level is high in information, at the level of harmony, are deceptive cadences in Western classical music, where a surprise chord replaces the highly predicted tonic as final chord, disrupting an expected sense of closure.

Again, we would not expect that one signal level is primary over another, at least if there is no propositional content (as typical in music). Rather, we expect that the levels of the sequence have more degrees of freedom to vary in their information content, as long as the phrase ending remains predictable. For example, a soloist in jazz might introduce a harmonic modulation (with high information content at the harmonic signal level) at the end of a phrase, while keeping melodic and rhythmic levels highly predictable and thus inviting a turn-taking event after which the next solo now occurs in a new key. That musical phrases in a more general sense exist, for example phrases in a solo Lied (song), could thus be an abstraction of deployment in a dialogic context.

What design features are predicted for language in a choric context? The particular tendency of overlapping speech stimuli to act as each other's noise (thus increasing uncertainty) means that simultaneity can only occur if precisely the same words or syllables are uttered at the same time, as happens for example in simultaneous speaking in religious or theatre performances (chanting). The prediction here is that attention is much more focussed on coordination than in dialogic speech acts, and that an isochronic and/or metrical scaffolding should develop (cf. Bowling et al., 2013). Since word order in such a sequence must be pre-given, and thus the information density of the word and phoneme level would be constant and very low, we would expect suprasegmental or paralinguistic levels to vary more in information density and to be used to reduce uncertainty regarding timing. That means body motion, facial expressions or prosodic intonation should be more pronounced in a spoken choric context.

6 Comparative Perspective

We suggest that the perspective of deployment of sound sequences in a three-dimensional quality space (goal: propositional-aesthetic, novelty: repetition-novelty, interactivity: dialogic-choric) along with the information theoretic concept of reducing uncertainty can also be used in bioacoustics research. We are aware that transferring a concept directly from human language and music might not work for animal communication, but we think that, especially for complex vocal displays in birds or whales, our framework may provide some insight, since these 'animal songs' bear some structural similarities to language and music (Rohrmeier et al., 2015). Information theory has been employed in animal communication research, although the term 'information' has often been used in a colloquial sense or inconsistently (Stegmann, 2013; Fitch, 2014). We hope that applying our framework provides some useful insights regarding complex vocal displays, along with call combinations and sequences (Engesser and Townsend, 2019). The information theoretic framework also encourages us to consider non-vocal levels like body motion that might be especially relevant for mating displays (Mitoyen et al., 2019).

When talking about 'goals' in animal communication, it is necessary to consider that goals other than propositional or aesthetic ones (e.g., social bonding by vocal convergence, providing information about sex or status, etc.) might use the vocal domain independently of the dimensions we derived for human language and music. Often it is unclear what animals communicate in their vocalisations, and some researchers question the existence of communicative content at all, proposing instead that animals manipulate others by means of their signals (Owren et al., 2010; Stegmann, 2013). On the other hand, there is evidence in some cases that animal calls can be functionally referential, reliably co-occurring with external entities (Seyfarth et al., 1980; Price et al., 2015), but little evidence that complex vocalisations like bird song or whale song have functional referential meaning (Engesser and Townsend, 2019). Analysing information trajectories across multiple levels of the sequence might

give additional insight into this important question, but this requires that several such levels can be disentangled in the first place, which might not be easily the case in animal vocalisations.

Duality of patterning appears to exist only in human language (Bowling and Fitch, 2015). When vocalisations have a propositional goal (in the sense of referring to external entities and eliciting reliable behavioural responses), we would therefore expect this content to be encoded in a whole structure across levels of the communicative signal, similar to music when deployed in a propositional context. We would also expect high predictability within the signal (since surprising elements could potentially add uncertainty to the inferred content) and probably a relatively uniform information density to enhance transmission. In turn, given that reward systems in other animals (Connell and Hofmann, 2011) would also be related to prediction and surprise (Schultz, 2016), we would expect that sound sequences with independence in information variation between different levels are more likely to fulfil an aesthetic rather than a propositional goal. We speculate that, in mating displays, constrained surprise rather than complexity is what makes displaying individuals attractive for mating, accounting for the common occurrence of individually distinctive songs and song repertoires.

Animal vocalisations like bird or whale song consist of subunits that can be structured in a hierarchical manner, and thus bear some structural similarities to human language and music (Payne and McVay, 1971; Rohrmeier et al., 2015). Repetition of subunits and their recombination often characterise such complex vocal displays. We would expect that subunits used in repetition are categorical to reduce the alphabet and enhance predictability. We would also expect a tendency towards hierarchical organisation in the temporal domain for longer vocalisations relative to shorter ones within comparable species.

What about novelty? Novelty might be realised by using new sounds with high surprisal. For example, the best documented example of ‘vocal’ learning in chimpanzees involved the addition of a lip buzz or ‘raspberry’ at the end of pant-hoot sequences (Marshall et al., 1999). However, since such new sounds would not reduce uncertainty in decoding a message for the receiver, unless an association to some external entity is learned, they would probably not have ‘content’ in the sense of functional referentiality. Novelty can also be realised by rearranging subunits that are structurally highly predictable, as is often characteristic of bird song (Kroodsma, 1978) and somewhat structurally similar to human song (Lomax, 1968).

Repetition can also have referential relevance in call combinations or sequences (see Engesser and Townsend, 2019), for example in chickadee call repeats (e.g., Hailman and Ficken, 1987). Some of these calls appear to be categorical (in the sense of being discriminable) while others are higher in variability. Moreover, some of these combinations are rather short. It seems that such call combinations may not be straightforwardly accounted for by our three-dimensional framework.

However, we expect that for stand-alone calls, uncertainty should be higher than when another call (same or different) is appended, as evaluated by changes in attention of receivers. Calls that occur only in fixed combinations might however not induce surprise since there is little uncertainty in referential content when encountering them.

Using the predictions derived from deployment of human language and music in choric or dialogic contexts could reveal whether animal vocalisations show design features based on differential temporal coordination of signals (Ravignani et al., 2019). Coordinated vocal displays, both concurrent and turn-taking, are widespread in animal communication. Duetting for example is widely observed in bird species. Investigating the information trajectories could reveal whether and how individuals relate to each other in their vocalisations. Interlocking vocalisations between two pair-bonded or courting birds for example could be investigated to see whether it is based on decreasing information density after one phrase. This would indicate a dialogic deployment. Competitive vocal displays, for example between two males in territorial contexts, might show dialogic design features with decreasing information density at the end of an individual's vocal phrase. If the display involves masking the competitor, we would predict that overlap is done in moments of high information. On the other hand, duets might be based on a coordination of each event with predictability enabled by some isochronous scaffolding, similar to a musical piece where performers do not play each note concurrently but nonetheless contribute to a unified musical piece. Such 'hocketting' would be indicative of a choric deployment. Bird species differ in their preference for overlap or overlap avoidance and in their flexibility depending on social and environmental context (Pika et al., 2018). Starlings seem to be a particularly interesting model species, showing both turn-taking and overlapping vocalisations depending on social context, and varying in their proportion of either by sociality of subspecies (Henry et al., 2015). We would predict that in all these cases information trajectories are perceived by receivers and used to coordinate their own vocalisation in relation to that of the other individual(s).

Other vocal displays that might be interesting to investigate in this context are antiphonal calling in elephants (Soltis et al., 2005) or bats (Carter et al., 2008) or duets in gibbons (Geissmann, 2002) or indris (Gamba et al., 2014), as well as group calling in meerkats (Demartsev et al., 2018). Castellucci et al. (in press) suggest singing mice as a model species for coordinated vocal timing. Bottlenose dolphins can switch between alternating vocalisations and simultaneous duetting (Lilly and Miller, 1961). We would predict that their alternating vocalisations are coordinated by decrease in information density at the end of phrases, while their duetting might be coordinated on an event-by-event basis as in choric deployment. A particularly interesting case is the group territorial display of plain-tailed wrens (*Thryothorus euophrys*) where group choruses including multiple males and multiple females are used to jointly defend a territory. In this species simultaneous performance occurs within sexes, with turn-taking between sexes (Mann et al., 2006) within a chorus. We expect within-

sex performance to be characterised by a tendency toward an isochronous scaffolding, comparable to meter, and turn-taking between sexes to be coordinated by terminal decreases in information density.

Animal group communicative (choric) displays often appear uncoordinated, for example in howler monkeys (Sekulic, 1982) or, at the other extreme, highly synchronised, as in some fireflies (see Ravignani et al., 2014). Group coordination can either be based on predictive or reactive behaviour, and we suggest that information trajectories could be examined to address this issue. If individuals in a group (for example howling wolves) vocalise in a coordinated manner, we expect each individual's contribution to reduce uncertainty about event timing for the other individuals. This would be indicative that the performance aims at creating a coherent entity, implying choric design features.

7 Conclusion

In this paper we have proposed a framework based on information theory, adopting a design stance to investigate differences in language and music. We suggested that some key design features of music and language can be explained as responses to their deployment between dialogic and choric poles of a continuum rooted in interactive performative constraints. This interactivity (choric-dialogic) dimension complements the widely recognised goal (propositional-aesthetic) and novelty (repetition-novelty) dimensions, forming a three-dimensional framework within which different forms of music and language can be placed, and their design differences understood.

We argued that the goal and novelty dimensions alone are not sufficient to explain differences in design features between music and language: the interactivity between individuals is crucial. For dialogic contexts, the only coordinative constraint concerns the timing of the turn-taking between individuals, which should be indexed by a lower information density towards the end of phrases, across all levels of the sonic stream. When there is also a propositional goal, non-propositional levels should be constrained in their variability to support the decoding of propositional content. Information rate should be high—realised mainly by novelty at the propositional level—until turn-taking is indicated. Conversational speech acts and turn-taking are the prototypical features that fulfil these requirements.

In contrast, choric performance requires tight temporal coordination of all contributing individuals, as well as avoidance of masking by simultaneous sound events, enabled by high predictability in timing and frequency of sonic events. When there is also a pressure for novelty, isochronous meter and discrete pitches in scales are design solutions that enable a group of participants to join in making a coherent sound sequence, allowing both novelty and repeatability. This contributes to the independence of multiple levels in the sonic stream with regard to surprise and uncertainty, making these independent levels well suited to exploit the human reward system. The prototypical form of choric performance is joint music making, but our framework also encompasses

non-canonical forms of music and language like chant, poetry, or exchange of musical solos, thereby avoiding an overly simplistic dichotomy between language and music. Furthermore, our framework supports comparisons of different forms of communication across distinct modalities and species and can help to generate new hypotheses about optimal design of signals satisfying multiple different requirements. We hope that this framework will also be fruitfully employed in animal communication research, broadening the scope of comparisons with music and/or language.

Additional Information

Author Contributions: All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding: This article was supported by the Austrian Science Fund (FWF) number W1262-B29 (DK Grant Cognition & Communication). FH was also supported by a VDS CoBeNe final fellowship. The funders had no role in decision on the article content, decision to publish, or preparation of the manuscript.

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.

Acknowledgments: The authors thank Tudor Popescu, Sabrina Engesser, and Gregg Castellucci for helpful discussions and comments, and two reviewers for helpful comments on the manuscript.

References

- 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>
- Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis : A Functional Explanation for Relationships between Hypothesis : A Functional Explanation for Relationships between Redundancy , Prosodic Prominence , and Duration in Spontaneous Speech*. *Language and Speech*, 47(1), 31–56. <https://doi.org/10.1177/00238309040470010201>
- Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20), 11818–11823. <https://doi.org/10.1073/pnas.191355898>
- Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52, 46–57. <https://doi.org/10.1016/j.wocn.2015.04.004>
- Bowling, D. L., & Fitch, W. T. (2015). Do Animal Communication Systems Have Phonemes? *Trends in Cognitive Sciences*, 19(10), 555–557. <https://doi.org/10.1016/j.tics.2015.08.011>
- Bowling, D. L., Herbst, C. T., & Fitch, W. T. (2013). Social origins of rhythm? Synchrony and temporal regularity in human vocalization. *PLoS ONE*, 8(11), e80402. <https://doi.org/10.1371/journal.pone.0080402>

- Brown, R. M., & Palmer, C. (2012). Auditory – motor learning influences auditory memory for music. *Memory & Cognition*, 40, 567–578. <https://doi.org/10.3758/s13421-011-0177-x>
- Brown, S. (2007). Contagious heterophony: A new theory about the origins of music. *Musicae Scientiae*, 11(1), 3–26. <https://doi.org/10.1177/102986490701100101>
- Busnel, R.-G., & Classe, A. (1976). *Whistled Languages*. Berlin, Heidelberg, Germany: Springer.
- Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *The Journal of the Acoustical Society of America*, 45(3), 694–703.
- Carter, G. G., Skowronski, M. D., Faure, P. A., & Fenton, B. (2008). Antiphonal calling allows individual discrimination in white-winged vampire bats. *Animal Behaviour*, 76, 1343–1355. <https://doi.org/10.1016/j.anbehav.2008.04.023>
- Castellucci, G. A., Guenther, F. H., & Long, M. A. (in press). A Theoretical Framework for Human and Nonhuman Vocal Interaction. *Annual Review of Neuroscience*, 45. <https://doi.org/https://doi.org/10.1146/annurev-neuro-111020-094807>
- Chatwin, B. (1987). *The songlines*. New York, NY, USA: Viking.
- Cheung, V. K. M., Harrison, P. M. C., Meyer, L., Pearce, M. T., Haynes, J., Cheung, V. K. M., ... Haynes, J. (2019). Uncertainty and Surprise Jointly Predict Musical Pleasure and Amygdala , Hippocampus , and Auditory Cortex Activity. *Current Biology*, 29(23), 4084–4092.e4. <https://doi.org/10.1016/j.cub.2019.09.067>
- Cirelli, L. K., & Trehub, S. E. (2020). Familiar songs reduce infant distress. *Developmental Psychobiology*, 56(5), 861–868. <https://doi.org/https://doi.org/10.1037/dev0000917>
- Colombo, M., & Wright, C. (2017). Brain and Cognition Explanatory pluralism : An unrewarding prediction error for free energy theorists. *Brain and Cognition*, 112, 3–12. <https://doi.org/10.1016/j.bandc.2016.02.003>
- Connell, L. A. O., & Hofmann, H. A. (2011). The Vertebrate Mesolimbic Reward System and Social Behavior Network : A Comparative Synthesis. *The Journal of Comparative Neurology*, 519, 3599–3639. <https://doi.org/10.1002/cne.22735>
- Corbeil, M., Trehub, S. E., & Peretz, I. (2016). Singing Delays the Onset of Infant Distress. *Infancy*, 21(3), 373–391. <https://doi.org/10.1111/inf.12114>
- Cross, I., Fitch, W. T., Aboitiz, F., Iriki, A., Jarvis, E. D., Lewis, J., ... Trehub, S. E. (2013). Culture and Evolution. In M. A. Arbib (Ed.), *Language, Music, and the Brain* (pp. 541–562). Cambridge, Massachusetts; London, England: The MIT Press. <https://doi.org/10.7551/mitpress/9780262018104.003.0021>
- Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized Information Theory Meets Human Cognition: Introducing a Unified Framework to Model Uncertainty and Information Search. *Cognitive Science*, 42(5), 1410–1456. <https://doi.org/10.1111/cogs.12613>
- Csete, M. E., & Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, 295(5560), 1664–1669. <https://doi.org/10.1126/science.1069981>
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London, UK: J. Murray.
- de Fleurian, R., Harrison, P. M. C., Pearce, M. T., & Quiroga-Martinez, D. R. (2019). Reward prediction tells us less than expected about musical pleasure. *Proceedings of the National Academy of Sciences*, 116(42), 20813–20814. <https://doi.org/10.1073/pnas.1913244116>

- de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language*, 82(3), 515–535. <https://doi.org/10.1353/lan.2006.0130>
- Demartsev, V., Strandburg-Peshkin, A., Ruffner, M., & Manser, M. (2018). Vocal Turn-Taking in Meerkat Group Calling Sessions. *Current Biology*, 28(22), 3661–3666.e3. <https://doi.org/10.1016/j.cub.2018.09.065>
- Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, 129(4), 2245–2252. <https://doi.org/10.1121/1.3562174>
- Engesser, S., & Townsend, S. W. (2019). Combinatoriality in the vocal systems of nonhuman animals. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(4), e1493. <https://doi.org/https://doi.org/10.1002/wcs.1493>
- Ferreri, L., Mas-Herrero, E., Zatorre, R. J., Ripollés, P., Gomez-Andres, A., Alicart, H., ... Rodriguez-Fornells, A. (2019). Dopamine modulates the reward experiences elicited by music. *Proceedings of the National Academy of Sciences of the United States of America*, 116(9), 3793–3798. <https://doi.org/10.1073/pnas.1811878116>
- Fitch, W. T. (2006). The biology and evolution of music : A comparative perspective. *Cognition*, 100, 173–215. <https://doi.org/10.1016/j.cognition.2005.11.009>
- Fitch, W. T. (2014). Information considered harmful in animal communication. *Current Biology*, 24(1), R8–R10. <https://doi.org/10.1016/j.cub.2013.11.020>
- Forth, J., Agres, K., Purver, M., & Wiggins, G. A. (2016). Entraining IDyOT : Timing in the Information Dynamics of Thinking. *Frontiers in Psychology*, 7(1575), 1–19. <https://doi.org/10.3389/fpsyg.2016.01575>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Gamba, M., Torti, V., Bonadonna, G., Guzzo, G., & Giacomini, C. (2014). Overlapping And Synchronization In The Song Of The Indris (Indri indri). In E. A. Cartmill, S. Roberts, H. Lyn, & H. Cornish (Eds.), *The Evolution of Language: Proceedings of the 10th international conference (EvoLangX)* (pp. 90–97). Vienna, Austria. Singapore: World Scientific Publishing Co. https://doi.org/https://doi.org/10.1142/9789814603638_0010
- Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, 16(2), 129–135. <https://doi.org/10.1016/j.tics.2011.11.014>
- Geissmann, T. (2002). Duet-splitting and the evolution of gibbon songs. *Biological Reviews*, 77, 57–76. <https://doi.org/10.1017/S1464793101005826>
- Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 23(5), 389–407. <https://doi.org/https://doi.org/10.1016/j.tics.2019.02.003>
- Glowinski, D., Coletta, P., Volpe, G., Camurri, A., Chiorri, C., & Schenone, A. (2010). Multi-scale entropy analysis of dominance in social creative activities. In A. del Bimbo, S.-F. Chang, & A. Smeulders (Eds.), *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference* (pp. 1035–1038). New York, NY,USA: Association for Computing Machinery. <https://doi.org/10.1145/1873951.1874143>

- Glowinski, D., & Mancini, M. (2011). Towards real-time affect detection based on sample entropy analysis of expressive gesture. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective Computing and Intelligent Interaction* (Vol. 6974 LNCS, pp. 527–537). Berlin, Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-642-24600-5_56
- Glowinski, D., Mancini, M., Cowie, R., Camurri, A., Chiorri, C., & Doherty, C. (2013). The movements made by performers in a skilled quartet: A distinctive pattern, and the function that it serves. *Frontiers in Psychology*, 4(841), 1–9. <https://doi.org/10.3389/fpsyg.2013.00841>
- Gordon-Salant, S. (1985). Some perceptual properties of consonants in multitalker babble. *Perception & Psychophysics*, 38(1), 81–90. <https://doi.org/10.3758/BF03202928>
- Grice, H. P. (1975). Logic and Conversation. In D. Davidson & G. Harman (Eds.), *The Logic of Grammar* (pp. 64–153). Encino, CA,USA: Dickenson.
- Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, 11(2), 194–205. <https://doi.org/10.1162/089892999563328>
- Hansen, N. C., Dietz, M. J., & Vuust, P. (2017). Commentary : Predictions and the brain : how musical sounds become rewarding. *Frontiers in Human Neuroscience*, 11(168), 10–12. <https://doi.org/10.3389/fnhum.2017.00168>
- Hansen, N. C., Kragness, H. E., Vuust, P., Trainor, L., & Pearce, M. T. (2021). Predictive Uncertainty Underlies Auditory Boundary Perception. *Psychological Science*, 32(9), 1416–1425. <https://doi.org/10.1177/0956797621997349>
- Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, 5(1052), 1–17. <https://doi.org/10.3389/fpsyg.2014.01052>
- Harries, M., Hawkins, S., Hacking, J., & Hughes, I. (1998). Changes in the male voice at puberty: Vocal fold length and its relationship to the fundamental frequency of the voice. *Journal of Laryngology and Otology*, 112(5), 451–454. <https://doi.org/10.1017/s0022215100140757>
- Henry, L., Craig, A., Lemasson, A., & Hausberger, M. (2015). Social coordination in animal vocal interactions. Is there any evidence of turn-taking? The starling as an animal model. *Frontiers in Psychology*, 6(1416), 39–59. <https://doi.org/doi:10.3389/fpsyg.2015.01416>
- Herff, S. A., Harasim, D., Cecchetti, G., Finkensiep, C., & Rohrmeier, M. A. (2021). Hierarchical syntactic structure predicts listeners' sequence completion in music. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, pp. 276–281).
- Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., & Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1), 1–46. <https://doi.org/10.1016/j.physrep.2006.12.004>
- Hockett, C. (1960). The origin of speech. *Scientific American*, 203(3), 88–97.
- Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1(4), 304–309. <https://doi.org/10.1038/1124>
- Honing, H. (2018). On the biological basis of musicality. *Annals of the New York Academy of Sciences*, 1423(2018), 51–56. <https://doi.org/10.1111/nyas.13638>
- Hurch, B., & Mattes, V. (2005). *Studies on reduplication*. Berlin, Germany: de Gruyter Mouton. <https://doi.org/https://doi.org/10.1515/9783110911466>

- Huron, D. (2006). *Sweet Anticipation*. Cambridge, Massachusetts: MIT Press.
- Huron, D. (2016). Aesthetics. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford Handbook of Music Psychology* (2 ed., pp. 233–245). Oxford, UK: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198722946.013.19>
- Jacoby, N., Undurraga, E. A., McPherson, M. J., & McDermott, J. H. (2019). Universal and Non-universal Features of Musical Pitch Perception Revealed by Singing. *Current Biology*, 29, 3229–3243. <https://doi.org/10.1016/j.cub.2019.08.020>
- Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science*, 366(6461), 50–54. <https://doi.org/10.1126/science.aax0287>
- Keller, P. E., Novembre, G., & Hove, M. J. (2014). Rhythm in joint action: Psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 20130394. <https://doi.org/10.1098/rstb.2013.0394>
- Knox, L. (1994). Repetition and relevance: Self-repetition as a strategy for initiating cooperation in nonnative/native speaker conversations. *Advances in Discourse Processes*, 47, 195–206.
- Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive Processes and the Peculiar Case of Music. *Trends in Cognitive Sciences*, 23(1), 63–77. <https://doi.org/10.1016/j.tics.2018.10.006>
- Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The Evolution of Rhythm Processing. *Trends in Cognitive Sciences*, 22(10), 896–910. <https://doi.org/10.1016/j.tics.2018.08.002>
- Krebs, J. R., & Davies, N. B. (Eds.). (1997). *BEHAVIOURAL ECOLOGY: An Evolutionary Approach* (4th ed.). Hoboken, New Jersey, USA: Wiley-Blackwell.
- Kroodsma, D. E. (1978). Continuity and versatility in bird song: support for the monotony–threshold hypothesis. *Nature*, 274(5672), 681–683.
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59. <https://doi.org/10.1080/23273798.2015.1102299>
- Ladd, D. R. (2014). *Simultaneous structure in phonology*. Oxford, UK/New York: Oxford University Press.
- Lerdahl, F., & Krumhansl, C. L. (2007). Modeling Tonal Tension. *Music Perception: An Interdisciplinary Journal*, 24(4), 329–366. <https://doi.org/10.1525/mp.2007.24.4.329>
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731), 10–26.
- Lewis, J. (2021). Why music matters: social aesthetics and cultural transmission. In S. Cottrell (Ed.), *Music, Dance, Anthropology*. Oxford, UK: Sean Kingston Publishing.
- Lilly, J. C., & Miller, A. M. (1961). Sounds Emitted by the Bottlenose Dolphin. *Science*, 133(3465), 1689–1693.
- Lomax, A. (1968). *Folk song style and culture*. Washington D.C: American Association for the Advancement of Science.
- Maess, B., Mamashli, F., Obleser, J., Helle, L., & Friederici, A. D. (2016). Prediction signatures in the brain: Semantic pre-activation during language comprehension. *Frontiers in Human*

- Neuroscience*, 10(591), 1–11. <https://doi.org/10.3389/fnhum.2016.00591>
- Mahowald, K., Dautriche, I., Braginsky, M., & Gibson, E. (2020). Efficient communication and the organization of the lexicon. *PsyArXiv*, 1–46. <https://doi.org/10.31234/osf.io/4an6v>
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info / information theory : Speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313–318.
- Mann, N. I., Dingess, K. A., & Slater, P. J. B. (2006). Antiphonal four-part synchronized chorusing in a Neotropical wren. *Biology Letters*, 2, 1–4. <https://doi.org/10.1098/rsbl.2005.0373>
- Margulis, E. H. (2014). *On Repeat: How Music Plays the Mind*. Oxford, UK/New York: Oxford University Press.
- Margulis, E. H., & Simchy-Gross, R. (2016). Repetition Enhances the Musicality of Randomly Generated Tone Sequences. *Music Perception*, 33(4), 509–514.
- Marshall, A. J., Wrangham, R. W., & Arcadi, A. C. (1999). Does learning affect the structure of vocalizations in chimpanzees? *Animal Behaviour*, 58(4), 825–830. <https://doi.org/10.1006/anbe.1999.1219>
- Maynard Smith, J. (2000). The concept of information in biology. *Philosophy of Science*, 67(2), 177–194. <https://doi.org/10.1017/CBO9780511778759.007>
- McDonnell, M. D., Ikeda, S., & Manton, J. H. (2011). An introductory review of information theory in the context of computational neuroscience. *Biological Cybernetics*, 105(1), 55–70. <https://doi.org/10.1007/s00422-011-0451-9>
- Mehr, S. A., Krasnow, M. M., Bryant, G. A., & Hagen, E. H. (2021). Origins of music in credible signaling. *Behavioural and Brain Sciences*, 44, E60. <https://doi.org/doi:10.1017/S0140525X20000345>
- Mehr, S. A., Singh, M., York, H., Glowacki, L., & Krasnow, M. M. (2018). Form and Function in Human Song. *Current Biology*, 28(3), 356–368. <https://doi.org/10.1016/j.cub.2017.12.042>
- Mitoyen, C., Quigley, C., & Fusani, L. (2019). Evolution and function of multimodal courtship displays. *Ethology*, 125(8), 503–515. <https://doi.org/10.1111/eth.12882>
- Moore, B. C. J. (2014). Psychoacoustics. In *Springer handbook of acoustics* (pp. 475–517). Berlin, Heidelberg, Germany: Springer.
- Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and Gestalt-like principles predict melodic expectations. *Cognition*, 189, 23–34. <https://doi.org/10.1016/j.cognition.2018.12.015>
- Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. University of Chicago Press.
- Owren, M. J., Rendall, D., & Ryan, M. J. (2010). Redefining animal signaling: Influence versus information in communication. *Biology and Philosophy*, 25(5), 755–780. <https://doi.org/10.1007/s10539-010-9224-4>
- Payne, R. S., & McVay, S. (1971). Songs of humpback whales. *Science*, 173(3997), 585–597.
- Pearce, M. T., & Wiggins, G. A. (2012). Auditory Expectation: The Information Dynamics of Music Perception and Cognition. *Topics in Cognitive Science*, 4(4), 625–652. <https://doi.org/10.1111/j.1756-8765.2012.01214.x>

- Peretz, I., Gosselin, N., Belin, P., Zatorre, R. J., Plailly, J., & Tillmann, B. (2009). Music lexical networks: The cortical organization of music recognition. *Annals of the New York Academy of Sciences*, 1169, 256–265. <https://doi.org/10.1111/j.1749-6632.2009.04557.x>
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526–3529. <https://doi.org/https://doi.org/10.1073/pnas.1012551108>
- Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3), 280–291. <https://doi.org/10.1016/j.cognition.2011.10.004>
- Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns : bridging the gap between human and animal communication. *Proceedings of the Royal Society B: Biological Sciences*, 285, 20180598. <https://doi.org/http://dx.doi.org/10.1098/rspb.2018.0598>
- Price, T., Wadewitz, P., Cheney, D., Seyfarth, R., Hammerschmidt, K., & Fischer, J. (2015). Vervets revisited : A quantitative analysis of alarm call structure and context specificity. *Scientific Reports*, 5(13220), 1–11. <https://doi.org/10.1038/srep13220>
- Ravignani, A., Bowling, D., & Fitch, W. T. (2014). Chorusing, synchrony and the evolutionary functions of rhythm. *Frontiers in Psychology*, 5(1118), 1–15. <https://doi.org/10.3389/fpsyg.2014.01118>
- Richardson, R. C. (2003). Engineering Design and Adaptation. *Philosophy of Science*, 70(5), 1277–1288. <https://doi.org/10.1086/377407>
- Riest, C., Jorschick, A. B., & de Ruiter, J. P. (2015). Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology*, 6(89), 62–75. <https://doi.org/10.3389/fpsyg.2015.00089>
- Rohrmeier, M. A., & Koelsch, S. (2012). Predictive information processing in music cognition. A critical review. *International Journal of Psychophysiology*, 83(2), 164–175. <https://doi.org/10.1016/j.ijpsycho.2011.12.010>
- Rohrmeier, M., & Pearce, M. (2018a). Musical Syntax I: Theoretical Perspectives. In R. Bader (Ed.), *Springer Handbook of Systematic Musicology* (pp. 473–486). Berlin, Heidelberg, Germany: Springer Verlag. https://doi.org/10.1007/978-3-662-55004-5_25
- Rohrmeier, M., & Pearce, M. (2018b). Musical Syntax II: Empirical Perspectives. In R. Bader (Ed.), *Springer Handbook of Systematic Musicology* (pp. 473–486). Berlin, Heidelberg, Germany: Springer Verlag.
- Rohrmeier, M., Zuidema, W., Wiggins, G. A., & Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 370(1664), 20140097-. <https://doi.org/10.1098/rstb.2014.0097>
- Rowland, J., Kasdan, A., & Poeppel, D. (2019). There is music in repetition : Looped segments of speech and nonspeech induce the perception of music in a time-dependent manner. *Psychonomic Bulletin & Review*, 26, 583–590.
- Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Publishing Group*, 14(2), 257–262. <https://doi.org/10.1038/nn.2726>
- Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., & McIntosh, A. R. (2015). Predictions and the brain: How musical sounds become rewarding. *Trends in Cognitive Sciences*, 19(2), 86–91. <https://doi.org/10.1016/j.tics.2014.12.001>

- Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, 112(29), 8987–8992. <https://doi.org/10.1073/pnas.1414495112>
- Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioural and Brain Sciences*, 44, E59.
- Schröger, E., Marzecová, A., & Sanmiguel, I. (2015). Attention and prediction in human audition: A lesson from cognitive psychophysiology. *European Journal of Neuroscience*, 41(5), 641–664. <https://doi.org/10.1111/ejn.12816>
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>
- Schultz, Wolfram. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, 18(1), 23–32. <https://doi.org/10.31887/DCNS.2016.18.1/wschultz>
- Schultz, Wolfram. (2017). Reward prediction error. *Current Biology*, 27, 369–371. <https://doi.org/10.1016/j.cub.2017.02.064>
- Schwitters, K. (1973). *Das literarische Werk, 5: Manifeste und kritische Prosa*. (K. Schwitters & F. Lach, Eds.). Köln, Germany: DuMont Schauberg.
- Sebanz, N., & Knoblich, G. (2009). Prediction in Joint Action: What, When, and Where. *Topics in Cognitive Science*, 1(2), 353–367. <https://doi.org/10.1111/j.1756-8765.2009.01024.x>
- Sebanz, N., & Knoblich, G. (2021). Progress in Joint Action Research. *Current Directions in Psychological Science*, 30(2).
- Seifert, U., Verschure, P. F. M. J., Arbib, M. A., Cohen, A. J., Fogassi, L., Fritz, T. H., ... Rickard, N. (2013). Semantics of internal and external worlds. In M. A. Arbib (Ed.), *Language, music, and the brain: A mysterious relationship* (pp. 203–232). Cambridge, Massachusetts; London, England: The MIT Press.
- Sekulic, R. (1982). The function of howling in red howler monkeys (*Alouatta seniculus*). *Behaviour*, 81(1), 38–54.
- Seyfarth, B. Y. R. M., Cheney, D. L., & Marler, P. (1980). Vervet Monkey Alarm Calls: Semantic Communication in a Free-ranging Primate. *Animal Behaviour*, 28, 1070–1094.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Snyder, B. (2000). *Music and Memory. An Introduction*. Cambridge, Massachusetts: MIT Press.
- Soltis, J., Leon, K., & Savage, A. (2005). African elephant vocal communication I : antiphonal calling behaviour among affiliated females. *Animal Behaviour*, 70, 579–587. <https://doi.org/10.1016/j.anbehav.2004.11.015>
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford, UK & Cambridge, USA: Blackwell.
- Stegmann (Ed.), U. (2013). *Animal communication theory: information and influence*. Cambridge, UK: Cambridge University Press.
- Temperley, D. (2007). *Music and Probability*. Cambridge, Massachusetts: MIT Press.

- Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, 7(1), 53–71. https://doi.org/10.1207/s15327078in0701_5
- Thompson, W. F., Sun, Y., & Fritz, T. (2019). Music across Cultures. In P. J. Rentfrow & D. J. Levitin (Eds.), *Foundations in Music Psychology*. Cambridge, Massachusetts: MIT Press.
- Tierney, A., Patel, A. D., & Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, 147(6). <https://doi.org/10.1037/xge0000455>
- Tomasello, M. (2010). *Origins of Human Communication*. Cambridge, Massachusetts: MIT Press. https://doi.org/10.9793/elsj.27.2_567
- Tooby, J., & Cosmides, L. (2005). Conceptual Foundations of Evolutionary Psychology. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (1st ed.). Hoboken, New Jersey: John Wiley & Sons, Inc. <https://doi.org/10.1007/s12041-017-0789-4>
- Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology*, 6(284), 135–145. <https://doi.org/10.3389/fpsyg.2015.00284>
- Trainor, L. J. (2018). The origins of music: Auditory scene analysis, evolution, and culture in musical creation. In H. Honing (Ed.), *The origins of musicality* (pp. 81–112). Cambridge, Massachusetts: MIT Press.
- Trehub, S. E., Becker, J., & Morley, I. (2015). Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*. <https://doi.org/10.1098/rstb.2014.0096>
- Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Semantic entropy in language comprehension. *Entropy*, 21(1159), 1–21. <https://doi.org/10.3390/e21121159>
- Vuust, P., Brattico, E., Glerean, E., Seppänen, M., Pakarinen, S., Tervaniemi, M., & Näätänen, R. (2011). New fast mismatch negativity paradigm for determining the neural prerequisites for musical ability. *Cortex*, 47(9), 1091–1098. <https://doi.org/10.1016/j.cortex.2011.04.026>
- Waciewicz, S., & Żywiczyński, P. (2015). Language Evolution: Why Hockett’s Design Features are a Non-Starter. *Biosemiotics*, 8(1), 29–46. <https://doi.org/10.1007/s12304-014-9203-2>
- Walker, G. (2010). The phonetic constitution of a turn-holding practice: rush-throughs in English talk-in-interaction. In D. Barth-Weingarten, E. Reber, & M. Selting (Eds.), *Prosody in Interaction. Studies in discourse and grammar* (23). (pp. 51–72). Amsterdam, Philadelphia: John Benjamins Publishing Company. <https://doi.org/https://doi.org/10.1075/sidag.23>
- Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3), 1707–1717.
- Williams, P. L., & Beer, R. D. (2010). Nonnegative Decomposition of Multivariate Information, 1–14. Retrieved from <http://arxiv.org/abs/1004.2515>
- Williams, P. L., & Beer, R. D. (2011). Generalized Measures of Information Transfer, 1–6. Retrieved from <http://arxiv.org/abs/1102.1507>
- Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, 13(12), 532–540. <https://doi.org/10.1016/j.tics.2009.09.003>

- Zatorre, R. J. (2018). Why Do We Love Music ? *Cerebrum: The Dana Forum on Brain Science*, 1–12. Retrieved from http://www.dana.org/Cerebrum/2018/Why_Do_We_Love_Music/ Zatorre
- Zbili, M., & Rama, S. (2021). A Quick and Easy Way to Estimate Entropy and Mutual Information for Neuroscience. *Frontiers in Neuroinformatics*, 15(596443), 1–17. <https://doi.org/10.3389/fninf.2021.596443>
- Zuidema, W., Hupkes, D., Scharff, C., & Rohrmeier, M. (2018). Formal Models of Structure Building in Music, Language, and Animal Song. In H. Honing (Ed.), *The Origins of Musicality* (pp. 253–286). Cambridge, Massachusetts; London, England: The MIT Press. <https://doi.org/10.7551/mitpress/10636.003.0017>

CHAPTER 2

Song Is More Memorable than Speech Prosody: Discrete Pitches Aid Auditory Working Memory

Felix Haiduk¹, Clíodhna Quigley^{1,2,3} and W. Tecumseh Fitch^{1,2}

¹Department of Behavioral and Cognitive Biology, University of Vienna, Vienna, Austria

²Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria

³Konrad Lorenz Institute of Ethology, University of Veterinary Medicine Vienna, Vienna, Austria

Published (10 December 2020)
in

 **frontiers**
in Psychology | Auditory Cognitive Neuroscience

Citation: Haiduk, F., Quigley, C., & Fitch, W. T. (2020). Song Is More Memorable Than Speech Prosody: Discrete Pitches Aid Auditory Working Memory. *Frontiers in psychology*, 11, 586723.

Abstract

Vocal music and spoken language both have important roles in human communication, but it is unclear why these two different modes of vocal communication exist. Although similar, speech and song differ in certain design features. One interesting difference is in the pitch intonation contour, which consists of discrete tones in song, vs. gliding intonation contours in speech. Here, we investigated whether vocal phrases consisting of discrete pitches (song-like) or gliding pitches (speech-like) are remembered better, conducting three studies implementing auditory same-different tasks at three levels of difficulty. We tested two hypotheses: that discrete pitch contours aid auditory memory, independent of musical experience (“song memory advantage hypothesis”), or that the higher everyday experience perceiving and producing speech make speech intonation easier to remember (“experience advantage hypothesis”). We used closely matched stimuli, controlling for rhythm and timbre, and we included a stimulus intermediate between song-like and speech-like pitch contours (with partially gliding and partially discrete pitches). We also assessed participants' musicality to evaluate experience-dependent effects. We found that song-like vocal phrases are remembered better than speech-like vocal phrases, and that intermediate vocal phrases evoked a similar advantage to song-like vocal phrases. Participants with more musical experience were better in remembering all three types of vocal phrases. The precise roles of absolute and relative pitch perception and the influence of top-down vs. bottom-up processing should be clarified in future studies. However, our results suggest that one potential reason for the emergence of discrete pitch—a feature that characterises music across cultures—might be that it enhances auditory memory.

1 Introduction

Human communication is a fundamental behaviour defining how we see ourselves, particularly regarding our apparent uniqueness compared to other species. Two communicative systems are found in all human cultures, namely language and music, and both have key components executed in the vocal domain. Vocal music (song) and spoken language (speech) constitute two distinct but related modes of learned volitional vocalisations. The question of why humans engage in these two modes of vocalisations goes back to Darwin who proposed that they have a common origin (Darwin, 1871). They share the same vocal output system, which might be fundamental to the origin of both systems (see Levinson, 2013).

To clarify why humans engage in two modes of vocalisations—speech and song—a useful starting point is to break down these complex systems into components accessible for empirical study, in order to investigate the corresponding perceptual differences and their cognitive and behavioural consequences. Song and speech have multiple components in common. Both make use of the vocal apparatus (Lindblom and Sundberg, 2014) and share widely overlapping vocal production networks (Zarate, 2013; Pisanski et al., 2016). Both are complex auditory signals that rely on learned rules and are volitionally produced [for which the capacity of vocal learning is necessary; see Fitch and Jarvis (2013)]. Both consist of elements (for example notes or phonemes) that are generatively combined into sequences and group at different levels hierarchically, which exerts high demands on auditory memory. Finally, both modes of vocalisation are culturally transmitted and culturally variable (see e.g., Trehub, 2013; Mehr et al., 2019). However, song and speech are also perceptually clearly distinguishable, and not simply discriminated in arbitrary ways by different cultures (see Fritz et al., 2013). Aspects of the structure of the vocal spectrotemporal signal presumably influence this distinction. Furthermore, since these structural aspects are probably indicative of the underlying brain systems that shape the production and perception of song and speech (see e.g., Poeppel, 2011), signal structure offers a useful starting point to investigate their underlying cognitive and behavioural mechanisms.

Several structural aspects distinguishing music from language, termed “design features of music,” have been proposed by Fitch (2006). One of these design features concerns pitch, the perceptual component mostly based on the fundamental frequency (f_0) of a sound. Pitch is an acoustic feature fundamental to both music (in song melody) and language (in speech prosody). Crucially, the variation of pitch over time, known as pitch trajectory or pitch contour, differs between the two modes, offering a well-defined perceptual component that contrasts between song and speech. In song, pitch contours typically consist of tones that are discrete in both time and frequency. The frequencies of these tones map onto culturally transmitted scales that themselves add structure to a musical phrase by relating pitches in particular ratios (intervals, see e.g., Krohn et al., 2007). Pitches and intervals

provide the building blocks of melodies that can convey both affective and/or symbolic meaning to a culturally informed listener (see Seifert et al., 2013). In contrast, in speech, pitch contours are smoother and follow gliding up and down patterns. Speech intonation does not intentionally match specific pitches of a musical scale. This remains true in tonal languages, where either the direction of pitch shift (i.e., the contour) or the general relative level [high, low, medium pitch etc., see Bradley (2013)] is relevant, but not the specific absolute pitch. Nonetheless, speech intonation conveys considerable information in all languages, expressing emotional states, pragmatic and communicative intention, lexical meanings, etc (see Paulmann, 2016). Humans perceive and produce speech intonation constantly, starting in early childhood, and even prenatal exposure shapes neonates' vocal intonation production (Mampe et al., 2009). The interval relations between single tones in song have been suggested to utilise a fine-grained intonation perception system exclusive to music, while both song and speech would share a more coarse-grained perceptive system for the overall directional course of the intonation contour (Zatorre and Baum, 2012; see also Merrill, 2013). In summary, pitch supports a rich and overlapping body of communicative purposes in both song and speech, making it puzzling why these two distinct modes of utilising pitch would emerge in human vocalisations in the first place. A general auditory mechanism might become domain-specific by being utilised differently based on the nature of the input signal (phenotypic plasticity; Fritz et al., 2013). In turn a perceptual system that evolved for some other purpose might be exploited by a given signal structure, yielding a certain behavioural outcome (exaptation, Lloyd and Gould, 2017). It is also possible that a signal structure like discreteness of pitch in song is simply an epiphenomenon, without any behavioural consequences, but emerges due to sensory biases (see e.g., ten Cate and Rowe, 2007). Relating signal structure to cognitive consequences can therefore be informative about the underlying cognitive system, especially if it concerns a fundamental design feature that separates one signal type from others. Thus, it is worthwhile to investigate whether use of discrete vs. gliding pitches has functional consequences in auditory cognition.

Music and language both require auditory memory since both types of signals have variable content, and extend over time. Learned vocal phrases are important for song and spoken conversation. However, since music is not linguistically propositional (Fritz et al., 2013) it seems likely that auditory memory for song centres on spectrotemporal features like pitch. As for many studies comparing music and language, results concerning auditory working memory are mixed (see Schulze et al., 2018, for an overview). It has been suggested that remembering song and speech utilises two distinct systems of subvocal rehearsal: a phonological loop for verbal stimuli, and a tonal loop for musical sounds. Evidence for such a distinction comes from patient studies: Tillmann et al. (2015) reviewed pitch memory-deficits in congenital amusia (CA), concluding that there is a distinction between verbal and pitch-related memory systems, with only the latter being impaired in CA patients. However, CA patients do have difficulties in prosodic pitch perception as well (Tillmann et al., 2015), and it is unclear whether pitch-related memory is clearly divided into music-specific and speech-specific

networks. Moreover, studies using intervening distractor sounds in a memory task found that pitch similarity of target (tones or words) and distractor sounds (words or tones) influences performance, suggesting an overlap in pitch-related working memory for song and speech prosody (Semal et al., 1996; Ueda, 2004). There is also evidence from fMRI studies that neural populations active in perception are recruited for subvocal rehearsal for working memory (Hickok et al., 2003), for both speech and melodies (see Buchsbaum, 2016). Hickok et al. (2003) also compared working memory maintenance of non-sense speech and piano music in these regions and found nearly identical time courses of activation. This might suggest substantial mechanistic overlap, and thus no substantial difference between memorability of sung and spoken stimuli.

However, a study by Verhoef et al. (2011) suggested that under demands for memory in a transmission task, random intonation contours morph into distinct elements that are then utilised in a combinatorial way. Using a slide whistle in an iterated learning task (where stimuli played by one participant were passed to another to imitate, and so on in a chain for many participants), pitch contours gradually appeared that were distinct and easy to memorise. However, no distinct scale tones emerged. Reasons might lie in the use of a non-vocal instrument, the structure of the initial contours in the iterative chains, or simply because there is no memory advantage for pitch contours consisting of discrete pitches over gliding intonation. Thus, it is not clear whether the pitch trajectory—discrete or gliding—by itself has an effect on memory for vocal phrases. This is the issue we aimed to address in the current studies.

We conducted three studies to test whether identical vocal phrases can be remembered better when consisting of discrete (song-like) or gliding (speech-like) pitches. We presented these stimuli in a same-different paradigm using a two-alternative forced-choice task, with an auditory distractor stimulus interposed between the target and test phases. We took care to minimise cultural biases, and we used various types of distractor sounds for each study, to vary task difficulty and to interfere with subvocal rehearsal. We also quantified musical experience in order to investigate its effects on auditory memory, since studies comparing musicians and non-musicians in terms of auditory memory often reveal advantages for musicians (see Schulze et al., 2018).

The existing literature reviewed above suggests several hypotheses and predictions. We first hypothesised that discrete pitches are a spectrotemporal property that enhances auditory memory (“song memory advantage”). This hypothesis predicts that vocal phrases consisting of discrete pitches should be remembered better than vocal phrases consisting of gliding pitches. Such a result would indicate that a fundamental and widespread acoustic feature of music has clear cognitive effects beyond surface-level perception. An alternative hypothesis is that gliding speech pitch might be remembered better, because humans perceive and produce speech prosody with high abundance from early childhood (“experience advantage”). The high familiarity of speech prosody might therefore

make the gliding pitches of speech the more salient, and thus better-remembered, type of signal. Finally, based on previous literature we hypothesised that musical training might enhance memory effects, but we had no strong prediction about whether this advantage would apply either across stimulus classes, or specifically only for songlike stimuli with discrete pitch. Either way, the results should provide insight into the memory systems underlying these two systems.

2 Study 1

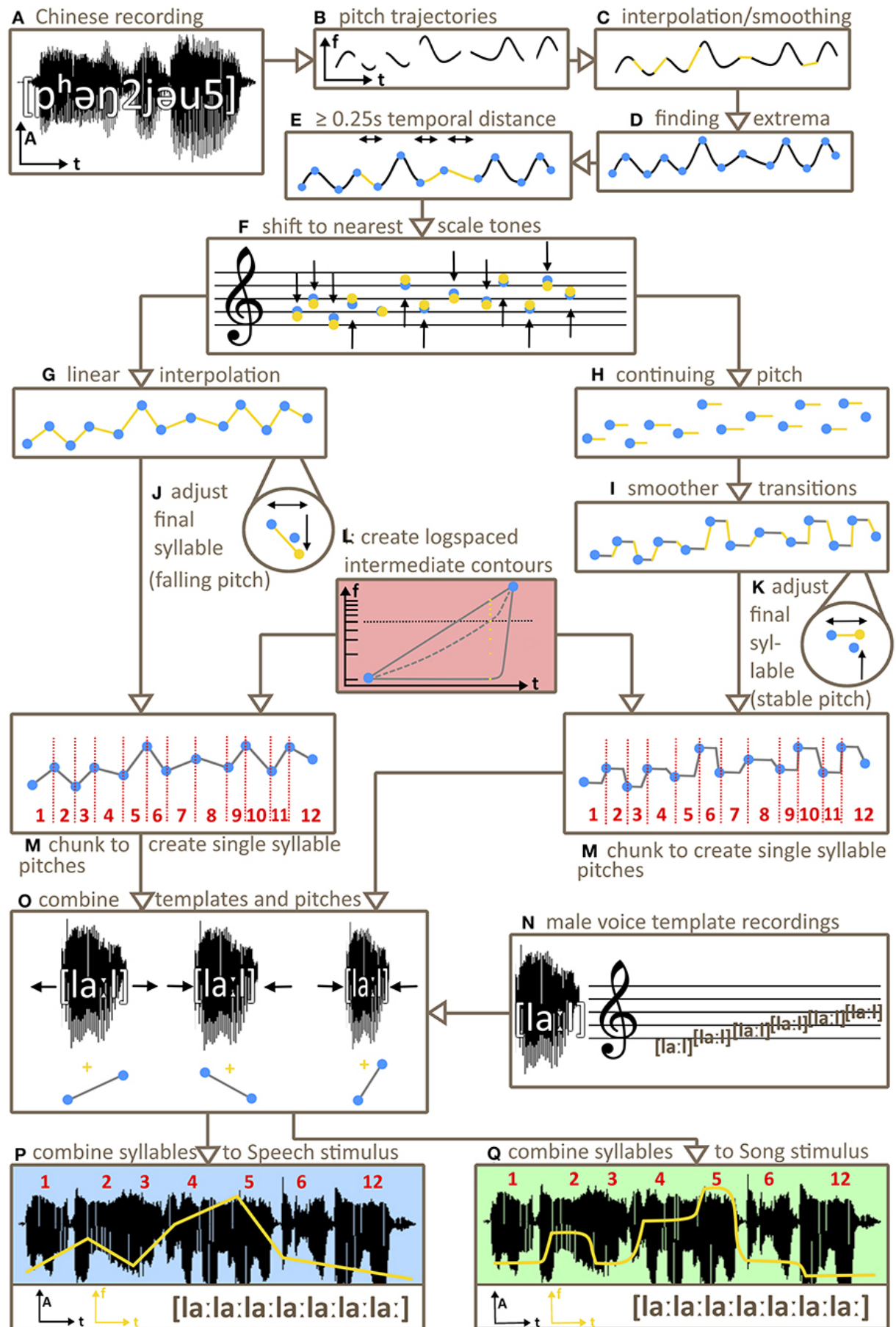
2.1 Materials and Methods

2.1.1 Stimuli

Our stimuli were based on natural speech samples, from which the pitch trajectory was extracted (see Figure 1). In order to minimize cultural influences due to familiarity these were Mandarin speech samples, spoken by an adult male¹. For the same reason pitch tones were adjusted to lie on an unfamiliar Bohlen-Pierce-Scale. From each of the pitch trajectories new pitch contours were derived to keep the sentence-level prosodic movement. Note that Mandarin Chinese shows both sentence-level intonation and syllable-level lexical tone. Our manipulation removed syllable-level but maintained sentence-level intonation. Sentence-level intonation supports both semantic and phonological processing in native speakers of Mandarin Chinese, but not in non-speakers (Tong et al., 2005). However, as non-speakers of Mandarin Chinese still show behavioural and neuronal effects of prosodic processing (Tong et al., 2005), using Mandarin sentence-level prosody seems a reasonable choice to minimize effects of cultural familiarity. The pitch contours derived from the Mandarin speech samples were the basis for stimuli of two variants: one variant consisting of discrete tones (as in a song melodic intonation) and another variant consisting of gliding tones (as in speech prosodic intonation). Additionally, seven intermediate variants were created of which one was used as a stimulus halfway between discrete and gliding variants. The reason for this was to examine to what extent potential effects are based on bottom-up processing of pitch as opposed to top-down categorical processing (hearing the stimuli “as song” or “as speech”). We hypothesised that an intermediate stimulus would neither be perceived as clear song nor as clear speech prosody. If the predicted effect that discrete pitch aids memory was mediated by a categorical top-down perception of a stimulus “as song” we would expect the intermediate stimulus to show very minor effects only. If gliding pitch was more memorable, we also would expect only minor effects for intermediate stimuli since they would not represent clear speech prosody. If however any effects were based on bottom-up pitch perception, we would expect intermediate stimuli to show medium effects with a magnitude between our Song

¹ Mandarin phrases were downloaded from <http://www.101languages.net/chinese/chinese-vocabulary-lessons/> (provided by 50languages.com under the creative commons license CC BY-NC-ND 3.0 US, <https://creativecommons.org/licenses/by-nc-nd/3.0/us/>).

and our Speech stimuli, since the pitch is partly gliding and partly discrete. Finally, all pitch contours were re-synthesised using a one-syllable constant-pitch recording of [la:l] of a male singer. Final stimuli consisted of seven syllables for study 1 and study 3 and ten syllables for study 2.



Previous page: **Figure 1.** Schema of stimulus creation. From recordings of Mandarin Chinese phrases spoken by a male (**A**) pitch trajectories were extracted (**B**), interpolated and smoothed (**C**). After finding the local frequency extrema (minima/maxima) of the trajectories (**D**), new pitch contours were derived by shifting the temporal distance between these extrema to create intervals of at least 0.25 s (**E**) and by subsequently shifting the extrema in frequency to the nearest Bohlen-Pierce or diatonic scale tones (**F**). The type of interpolation between extrema was determined by stimulus category. For speech stimuli, Praat interpolated linearly between extrema (**G**). For Song stimuli, pitch values at extrema were continued (**H**), including a smooth transition right before the next minimum/maximum to avoid unnaturally abrupt pitch changes (**I**). Pitch values for last syllables of each contour were adjusted to be 0.3s in duration, with falling pitch for Speech stimuli (**J**) and stable pitch for Song stimuli (**K**). Seven intermediate interpolation variants on a continuum between Song and Speech stimuli were created (**L**). Chunking pitch contours of all variants (Song, Speech, intermediates) using extrema as borders resulted in single pitch contour chunks (**M**). One-syllable template sounds, recorded by a male on the syllable [la:l] at all Bohlen-Pierce scale tones (**N**) were selected to match the initial pitch of a pitch contour chunk. The respective chunk's duration and pitch contour were then carried over to the respective template sound (**O**) to derive stimulus syllables. Finally, the first six (studies 1 and 3) or nine (study 2) stimulus syllables and the respective last syllable of each contour and variant were concatenated to generate the final stimuli (**P,Q**).

The software package Praat (Version 6.0.36, Boersma and Weenink, 2017) was used to create the stimuli (see Figure 1). In a first step 200 pitch trajectories were isolated from 200 mandarin phrases, spoken by a male [function “To Pitch (ac),” for details on the parameters: see Supplementary Material].

Pitch trajectories were then smoothed using the function “pitchsmoothing” from Praat Vocal Toolkit (80%; Corrette, 2012) and unvoiced gaps were linearly interpolated using the praat standard function “interpolate.” Frequency values at local extrema (minima/maxima) of the smoothed pitch trajectory, along with the onset and offset frequencies of the whole pitch trajectory, were used as markers for deriving discrete pitches. These extrema were shifted in time, if necessary, to be at least 0.25 s distant from each other to later avoid unnaturally short tones/syllables. The frequency values at these extrema were shifted to the nearest Bohlen-Pierce scale tone, again to avoid any culture-specific biases. For a second variant frequency values at extrema were shifted to the nearest tone of the familiar western diatonic scale including both major and minor thirds. This resulted in new pitch contours, each in both a Bohlen-Pierce and a diatonic variant, both variants being equally tempered. Both the Bohlen-Pierce scale and the diatonic scale were based on 80 Hz as lowest tonic frequency and comprised 14 possible tones (see Table 1).

#	Bohlen-pierce scale tones	Diatonic scale tones	Difference in cents
1	80	80	0
2	87.1	89.8	52.85
3	94.7	95.1	7.3
4	103.1	100.8	-39.06
5	112.2	106.8	-85.39
6	122.1	119.9	-31.48
7	132.8	134.5	22.02
8	144.5	151	76.18
9	157.3	160	29.46
10	171.2	179.6	82.93
11	186.3	190.3	36.78
12	202.7	201.6	-9.42
13	220.6	213.6	-55.83
14	240	239.7	-2.17

Equal tempered Bohlen-Pierce tones are about 146 cents apart from each other (ratio of $^{13}\sqrt{3}$), while equal tempered chromatic tones of a diatonic scale differ by 100 cents ($^{12}\sqrt{2}$). The diatonic scale contained both major and minor thirds.

Table 1. Possible scale tones used in stimulus creation (both equal tempered, in Hz).

To derive the discrete variant of each contour (henceforth Song stimuli), pitch samples between two adjacent extrema were set to the respective Bohlen-Pierce/diatonic tone of the first of the two extrema. This way stable pitches emerged between every two extrema. To ensure naturalistic sounding stimuli with no abrupt changes of pitch (that would have led to a yodelling sound), transitions between these stable pitches were smoothed by transitioning to the next pitch from 1/8 of the duration between two extrema before the next extremum. To derive the gliding contours, pitch samples between extrema were linearly interpolated. We henceforth refer to these stimuli as Speech stimuli for simplicity, noting that we investigated speech intonation, not full-fledged speech. Extrema were used as cut-offs to chunk whole contours into syllable-length PitchTiers. These chunks were later combined with the one-syllable male voice recordings [la:l]. The last chunk of a Speech contour was altered such that it always had a falling pitch from the final minimum/maximum and a duration of 0.3 s, while the last chunk of Song contours was altered such that it always had a sustained pitch and a

duration of 0.3 s. This mimicked phrase-final lengthening that occurs in both natural speech and song (Arnold and Jusczyk, 2002). If the last minimum/maximum was near 80 Hz this led to some pitch samples below 80 Hz for the last syllables of some Speech stimuli (see Supplementary Table 1 for details).

To create stimulus variants between Speech and Song, we synthesised a continuum of seven intermediate stimuli between Song and Speech variants of each contour. To this end, frequency values (in Hz) of a stimulus' Song and the Speech variant at each given timepoint were converted to semitone values ("semitones re 1 Hz" in Praat). This way the difference between both values was converted from logspace (Hz) to linear space (semitones). This difference was then logarithmically divided into seven steps, resulting in seven values for the stimulus variants between Song and Speech. The logspacing was such that steps closer to Song were smaller than steps closer to Speech. Stimuli located perceptually halfway between Speech and Song would be categorised by a listener as Song (and not as Speech) 50% of times (which is called the Point of Subjective Equality, see Kingdom and Prins, 2016). To obtain an empirical estimate of this category boundary, we presented stimuli of all variants and contours to one participant (author FH) in a two-alternative forced-choice Speech/Song categorisation task, using a staircase method. We chose the variant closest to the Point of Subjective Equality as intermediate stimulus for study 1 (henceforth Intermediate). This was step 6, with Speech being step 1 and Song being step 9.

To obtain a naturalistic timbre, a male singer recorded sustained tones on the syllables [la:l] at each scale tone of the Bohlen-Pierce scale as well as the diatonic scale in the octave from 80 to 160 Hz (henceforth template syllables). Recording was done using a Zoom H4n recording device (16 bit, 44.1 kHz). In order to avoid template syllables at higher frequencies sounding more distant than template syllables at lower frequencies in the final stimuli and therefore disrupting the perception of a single speaker/singer (Bregman, 1994; Zahorik et al., 2005; Bregman et al., 2016), template syllables were not equalised in intensity. For each discrete Bohlen-Pierce /diatonic PitchTier chunk of each contour, the template syllable nearest in frequency to the respective pitch was adjusted in duration to match the duration of the respective PitchTier chunk, whereby the first and last 0.02 s of each template syllable remained unstretched in time in order to keep the liquid consonant [l] unchanged. The pitch contour was then combined with the spectral information of the template syllable and resynthesised to a syllable [la:l] following the current pitch contour. This procedure resulted in the final syllable sounds (henceforth syllables) for each contour and variant (i.e., Song, Speech and the seven Intermediates). The first six and the last syllable of each contour and variant were then concatenated with an overlap of 0.03 s, and the resulting stimulus was ramped using a trapezoid with 0.01 s rise and fall. Thus, the same syllable sequence with the same durations for each syllable was used for all variants of each contour (i.e., Song, Speech, and the Intermediate stimuli). The mean duration of syllables across all stimuli used was 0.360 s (SD = 0.168). The mean pitch acceleration rate of syllables across the speech

prosody stimuli was 0.106 Hz/ms ($SD = 0.081$), thus much lower than for lexical tones in Mandarin Chinese (see Krishnan et al., 2010). Stimuli were scaled in intensity to 73 dB SPL (Praat: “Scale intensity”; i.e., the root-mean-square amplitude of the stimuli was changed to 73 dB above 0.00002 Pa). All stimuli were monophonic, 16 bit and with a sampling rate of 44.1 kHz.

For study 1 and 3, only stimuli with a duration between 2 and 3 s and with 7 syllables were used. For study 2, only stimuli between 3 and 4 s of duration and with 10 syllables were used. Note that syllable durations were ultimately based on the natural speech signals we used (Mandarin Chinese) and that therefore our stimuli were not isochronous or metrical in any sense.

Stimuli that were too similar to each other were excluded (evaluated aurally by author FH), resulting in 94 different stimulus contours per scale (diatonic/Bohlen-Pierce), each as Song, Speech and Intermediate variants.

To create the deviant stimuli for the same/different memory task, the pitch of one randomly chosen syllable (target syllable) of each contour, excluding the first and the last syllable, was shifted to another of the scale tones of the Bohlen-Pierce/diatonic scales specified above. The direction of the shift reversed the direction of pitch change between the target syllable and the syllable preceding it, such that a rising pitch interval between target syllable and preceding syllable became falling after the shift and vice versa. That way the overall pitch trajectory, that is the global up and down of the pitch contour, always changed. Exceptions were cases when the shift would have resulted in pitches of target syllables outside the Bohlen-Pierce/diatonic scales specified above. In such cases the pitch interval between two consecutive syllables changed in magnitude but not in direction (falling or rising). This was the case for 280 out of the 3,720 trials that made up the sample for study 1. The amount of pitch shift was random within the constraints just mentioned (see below in the statistical methods for details on the pitch deviation). The mean absolute pitch deviation was 3.8 semitones ($SD = 5.0$).

Distractor sounds matching the temporal structure of the tonal stimuli were constructed from pink noise bursts of random duration between 0.2 and 0.3 s in Praat (decreasing by 6 dB SPL per octave, $F0 = 100$ Hz) and concatenated with 0.01 s overlap, such that the total duration of the distractor was 2 s. Two hundred and eighty two different distractors were created to allow unique distractors to be used for each variant of each contour.

Distractor sounds and stimuli were presented such that in the same-different task participants would hear an original stimulus contour first, then a distractor sound and then either the original stimulus contour again (same) or the respective deviant stimulus contour (different).

2.1.2 Procedure

Thirty three participants (21 females, 12 males, age range 18 to 56 years, mean = 24.3 years) took part in the study. Throughout all three studies, participants were recruited via posters at the university venue and via Facebook and most of them were university students from the University of Vienna. All participants in the three studies were rewarded with 5 € per half hour of participation (approx. study duration was 60 min for all three studies). All participants in the three studies gave written informed consent to participate. All three studies were approved by the ethics committee of the University of Vienna (# 00361).

Participants were welcomed by the experimenter (author FH) and led to one of two adjacent testing rooms for human participants where they were seated in front of a computer. To avoid expectancy effects participants were given minimal, written instructions in German or English². They were told simply that their task was to make decisions about sounds by clicking on circles on a screen. Words such as “song,” “speech,” “language,” “music” were not mentioned.

Python 2.7 was used to present stimuli (scripts based on the script-building program “Experimenter_GUI, version 0.1” by Pinker, 1997). Sounds were played via Sennheiser HD 201 headphones at about 73 dB SPL. The study was run on a Macbook Pro laptop (Retina display, 15-inch screen, Mid 2015). Except during the training phase, the experimenter left the testing room so that each participant executed the tasks alone.

The experiment consisted of four phases: a training phase for the same-different task (about 10 min duration), the same-different task itself (about 30 min), a test of spectral or holistic hearing (Schneider et al., 2005) and the Gold-MSI musicality questionnaire (Müllensiefen et al., 2014). We did not include the results from the test of spectral or holistic hearing as results were heavily skewed towards holistic hearing (holistic:mix:spectral hearing 16:15:2 in study 1) and therefore uninformative. Instructions were provided in written form on the computer screen. Trials were self-paced so participants could take breaks between trials and after each part of the experiment.

In the initial training phase, participants were asked to judge whether a “sound example” (German: “Klangbeispiel”) changed or stayed the same and were instructed that their goal was to gain as many points as possible by answering correctly. They could start each trial by clicking on a small white filled circle in the middle of a grey screen and made decisions by clicking on one of two rectangular response boxes on the left and right side of the screen (side randomised between trials). One response box contained the word “anders” (or “different”), the other one “gleich” (or “same”). In

² The experiment was prepared in both German and English. All but one participant chose the German version.

order to minimise lapse errors due to random change of the sides where two boxes would appear, boxes were consistently outlined by rectangular lines of yellow (“same”) or blue (“different”). Participants were informed that their current points score would be shown on the screen after each trial. After reading these instructions, participants were allowed to ask clarification questions. No information concerning the task purpose, the nature of the stimuli or the type of change in difference stimuli was given at any point during the study.

In the training phase, both Song and Speech variants of ten different pitch contours randomly chosen from the pool of diatonic pitch contours (thus 20 stimuli) were presented in random order in the two-alternative forced choice design. Pitch contours used in the training phase were not used in the later test phase to avoid any memory transfer effects. Participants first heard one stimulus, followed by the distractor sound, and then either the stimulus again (same) or its deviant version (different). Thus, in cases when deviants occurred they were always played after the original (unaltered) stimulus. After clicking on one of the two response boxes, the participants received auditory and visual feedback. Correct choices were rewarded by displaying a gain of five points and a short bell-like sound. Incorrect choices were punished by the screen turning red for 500 ms along with playing an aversive sound and by displaying a loss of five points. Failed trials were repeated until successful. The current number of points (which could only be positive) was also displayed after each trial. During this initial training phase, the experimenter stayed in the room to allow the participants to ask questions about the task.

The second phase of the same-different task was the test phase and involved 120 forced-choice trials. Participants first received instructions written on the computer screen. The instructions given were the same as in the training phase, except that participants were told that the following task would consist of three blocks with an approximate duration of 10 min each and the possibility to take a break of minimum 30 s between blocks. Additionally, participants were informed that points they gained would now only be displayed after each block (no feedback after trials). After the instruction, participants were allowed to ask clarification questions. After the instructions the experimenter left the room until the participant was finished. The same setup was used as in the training phase. This time, 40 different pitch contour pairs, randomly drawn from the pool of Bohlen-Pierce stimuli individually for each participant, were presented in each the Speech variant, the Song variant, and the Intermediate variant (in sum 120 trials). Half of the stimuli were difference stimuli (i.e., included a pitch change on one syllable during the second presentation of the pitch contour after the distractor noise) and half of the stimuli were without change. Presentation order of the stimuli was random. In total 120 trials were presented. After each block of 40 trials, the run was interrupted for 30 s and written instructions appeared on the screen, asking the participant to take a short break and continue by clicking the white circle. To ensure a high level of motivation throughout the task an arbitrary current point score was displayed as well, showing 135 points after the first block and 260 points after the second block.

Participants were also instructed to self-assess their current concentration abilities after each block on a 7-point Likert scale. This was done to assess the reliability of the results for each block, whereby our a-priori criterion was to drop blocks for which concentration was below 3. During the test phase, no feedback was given after each trial and no trials were repeated. After the participants were finished, the experimenter re-entered the room.

Part three of the study consisted of the test on holistic vs. spectral hearing from Schneider et al. (2005), which we did not use in the analysis because of biased results. Part four of the study consisted of the Gold-MSI test to assess musicality of the participants (Müllensiefen et al., 2014), and a post-experimental questionnaire asking about strategies the participants used, any speculations about the purpose of the study and the languages they spoke (including level of proficiency).

Two speakers of Mandarin Chinese were excluded to avoid any familiarity confounds due to the original contours being Chinese.

Afterwards, participants gave post-experimental consent for using their data, were thanked, paid, and accompanied to the exit. The total duration of part 1 of the experiment ranged from 50 to 90 min (two participants needed more than 60 min due to the self-paced procedure).

2.1.3 Statistical Methods

Our goal was to quantify whether vocal contours are remembered better when consisting of discrete rather than gliding intonation contour. Our memory task required a binary response (original and target contour judged to be different or same), given two possible states of the target signal (original and target contour deviate or do not deviate). Our task therefore can be analysed in the framework of signal detection theory (see Kingdom and Prins, 2016, for an overview).

To this end, we used a logistic GLMM with logit link function. The response variable was the response given by the participants (“Response Given”). Fixed effects entered were the response that would have been correct (“Stimulus State”) and the variant of the contour (“Variant” Song, Speech, Intermediate) along with their interaction. We z-transformed the musicality score of the Gold-MSI in order to obtain model estimates for participants of average musicality (“z.Musicality”) and included this as a fixed effect as well. As random slopes we entered “Stimulus State,” “Variant” and “z.Musicality” within “Contour” (the 94 different stimulus pitch trajectories created) and “Stimulus State” and “Variant” within “Participant.” Interactions of random intercepts and random slopes were not included. We also did not include semitone deviation as a predictor in the model since it did not differ significantly between Variants ($\text{Median}_{\text{Speech}}: 6.67; \text{Median}_{\text{Intermediate}}: 8.6; \text{Median}_{\text{Song}}: 7.88$; Kruskal Wallis Test: $H = 0.51, df = 2, p = 0.78$; note that the semitone differences were derived in Praat as real numbers, not integers). We did not include the results of the test of spectral vs. holistic

hearing as its results were heavily skewed towards holistic hearing (only 1 out of 31 participants in the analysis sample scored as spectral hearer). Individual participants (“Participant”) and stimulus contours (“Contour”) were included as random intercepts. Dummy coding was done by a custom-built function. Optimizer “bobyqa” with maximum 100,000 function iterations was used to assess the maximum likelihood.

All statistical analyses were done in R (R Core Team, 2019). We fitted the model (and all following models) using the function `glmer` (R package “lme4”; version 1.1-21; Bates et al., 2015). Collinearity was assessed by fitting a linear model without random effects and interactions between fixed effects and applying the function “`vif`” (R package “car”; version 3.0-2; Fox et al., 2011). The function revealed no collinearity issues. Overdispersion was tested using a custom-built function. The model was not overdispersed, and in fact appeared to be slightly underdispersed ($\chi^2 = 2831.574$, $df = 3699$, $p = 1$, dispersion parameter = 0.77), so no further action was required. To assess how robust the model estimates would be given changes in the predictors we assessed model stability using a custom-built function. This function excludes participants one at a time from the data. Model estimates of the reduced and the full datasets are then compared (please see Nieuwenhuis et al., 2012 for a comparable approach). The model turned out to be stable (for ranges of parameter estimates of fixed effects please see Supplementary Table 2 and Supplementary Figure 1). In order to test the effect of all predictors as a whole we compared the full model with a null model lacking the predictors of interest, i.e., including only the random factors. Models were compared using a likelihood ratio test, applying the function “`anova.glm`” (argument `test` set to “Chisq”; Dobson, 2002; Forstmeier and Schielzeth, 2011). To test the effect of single predictors we used likelihood ratio tests, comparing the full model with a model reduced by one of the effects of interest (R function “`drop1`” with argument “`test`” set to “Chisq”). We first assessed the effect of the three-way interaction “Stimulus State:Variant:z.Musicality.” To test the effect of all two-way interactions, including the interaction of interest “Stimulus State:Variant,” we fitted the model without the three-way interactions and executed likelihood ratio tests as above. The sample size for this model was 31 individuals with 120 observations each, thus a total of 3,720 observations. We derived 0.95 confidence intervals using the function “`bootMer`” (package “lme4”) with 1,000 parametric bootstraps.

After applying the analyses mentioned above, we additionally did a repeated-measures Anova (R function “`aov_ez`,” from package “afex,” version 0.26-0; Singmann et al., 2020) on classically calculated d-primes for informative purposes (correction for hit rates and false alarm rates of 0 and 1: $\pm 1/(2N)$). A hit was defined as correctly detecting a difference (deviation), a false alarm accordingly was a response as “different” while there was no change in the stimulus. Corrections had to be done for 17 of 93 cells for this analysis). Note that this approach goes along with a loss of information in the data and is therefore not the ideal way of dealing with a data set with binary responses. We included “Variant” as repeated measures factor.

2.2 Results

2.2.1 Post-experiment Questionnaire

Twelve participants reported that they had attempted to repeat the stimuli silently or by humming to remember them, some of these participants additionally tapped their fingers. One participant reported to have visualised the contours, one used only finger tapping as strategy. Four participants guessed that the study was about memory. This suggested that the distractor sound we used was not distracting enough which was one reason to replicate the study under more difficult conditions.

2.2.2 Statistical Results

After fitting the full model (model coefficients see Table 2) we first assessed whether it differed significantly from the null model (comprising only the random factors). We found that all predictors of interest as a whole explained the data significantly better than the null model ($\chi^2 = 117.436$, $df = 11$, $p < 0.001$). Figure 2 and Table 3 show the probability estimates of the full model.

Term	log odds	SE	CI _{lower}	CI _{upper}	z	p
Intercept	-2.678	0.275	-3.229	-2.166	-9.735	
Stimulus state diff	2.860	0.254	2.391	3.378	11.281	0.000
Intermediate	0.197	0.209	-0.240	0.643	0.941	0.347
Song	0.264	0.210	-0.175	0.692	1.257	0.209
z.Musicality	0.295	0.247	-0.217	0.768	1.191	0.234
Stimulus state diff: Intermediate	0.593	0.239	0.127	1.079	2.477	0.013
Stimulus State diff: Song	0.397	0.236	-0.084	0.882	1.680	0.093
Stimulus state diff:z.Musicality	0.016	0.234	-0.442	0.488	0.070	0.945
Intermediate:z.Musicality	-0.309	0.173	-0.683	0.084	-1.787	0.074
Song:z.Musicality	-0.086	0.174	-0.481	0.324	-0.490	0.624
Stimulus state diff: Intermediate:z.Musicality	0.616	0.225	0.142	1.119	2.742	0.006
Stimulus state diff:Song:z.Musicality	-0.010	0.223	-0.468	0.476	-0.044	0.965

The model estimated detection of a difference between original sound and target sound (reference level: Stimulus State same) as a function of Stimulus Variant (Speech, Song, Intermediate. Reference level: Speech) and Musicality as assessed by the GoldMSI questionnaire and z-transformed (that way the estimates refer to an individual with average musicality).

Table 2. Estimates of model coefficients (log odds), standard errors (SE), lower (CI_{lower}) and upper (CI_{upper}) 0.95 confidence intervals, z statistics (z-value) and associated p-values of the model fitted for study 1.

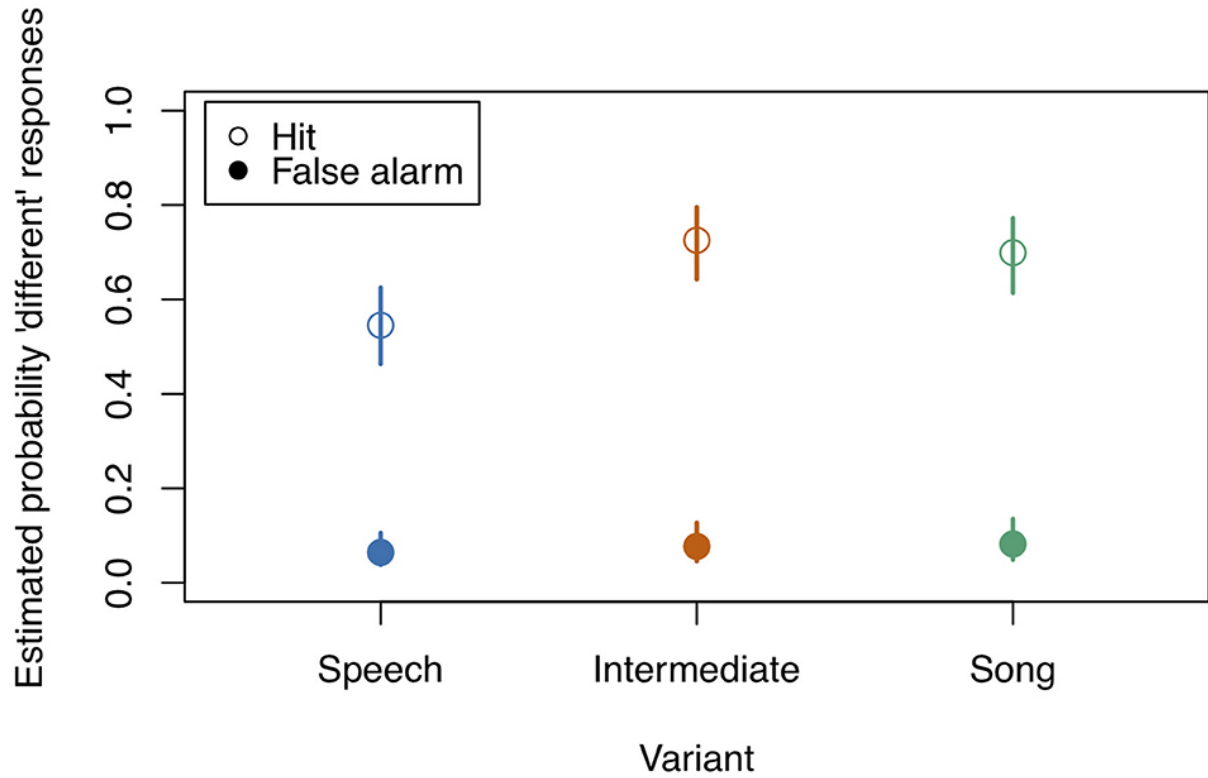


Figure 2. Model estimates for the probability of detecting a deviant in study 1 (see also Table 3). Empty circles: Hits, filled circles: False Alarms. Bars indicate 0.95 confidence intervals. $N = 31$.

Stimulus state	Speech			Intermediate			Song		
	Estimate	CI _{lower}	CI _{upper}	Estimate	CI _{lower}	CI _{upper}	Estimate	CI _{lower}	CI _{upper}
Diff	0.545	0.463	0.625	0.725	0.643	0.795	0.699	0.614	0.772
Same	0.064	0.039	0.105	0.077	0.046	0.127	0.082	0.049	0.135

Derived from the R-package "effects", version 4.1-0 (Fox et al., 2019).

Table 3. Probability estimates and 0.95 confidence intervals for Hits and False Alarms as depicted in Figure 2.

We next tested the three-way interaction “Stimulus State:Variant:z.Musicality,” which was significant ($LRT = 9.686$, $AIC = 3,556.861$, $df = 2$, $p = 0.008$, Nagelkerke's $R^2 = 0.046$). Musicality enhanced memorising much more for Intermediate stimuli than for Speech or Song stimuli. Note that therefore the interpretation of the two-way interactions was not straightforward anymore (see Figure 3 and Table 4 for detailed results on the two-way interactions). We found that the interaction “Stimulus State:Variant” was significant, that is, participants could memorise both Song and Intermediate stimuli better than Speech stimuli (see also Figure 2), with the magnitude depending on their Musicality score.

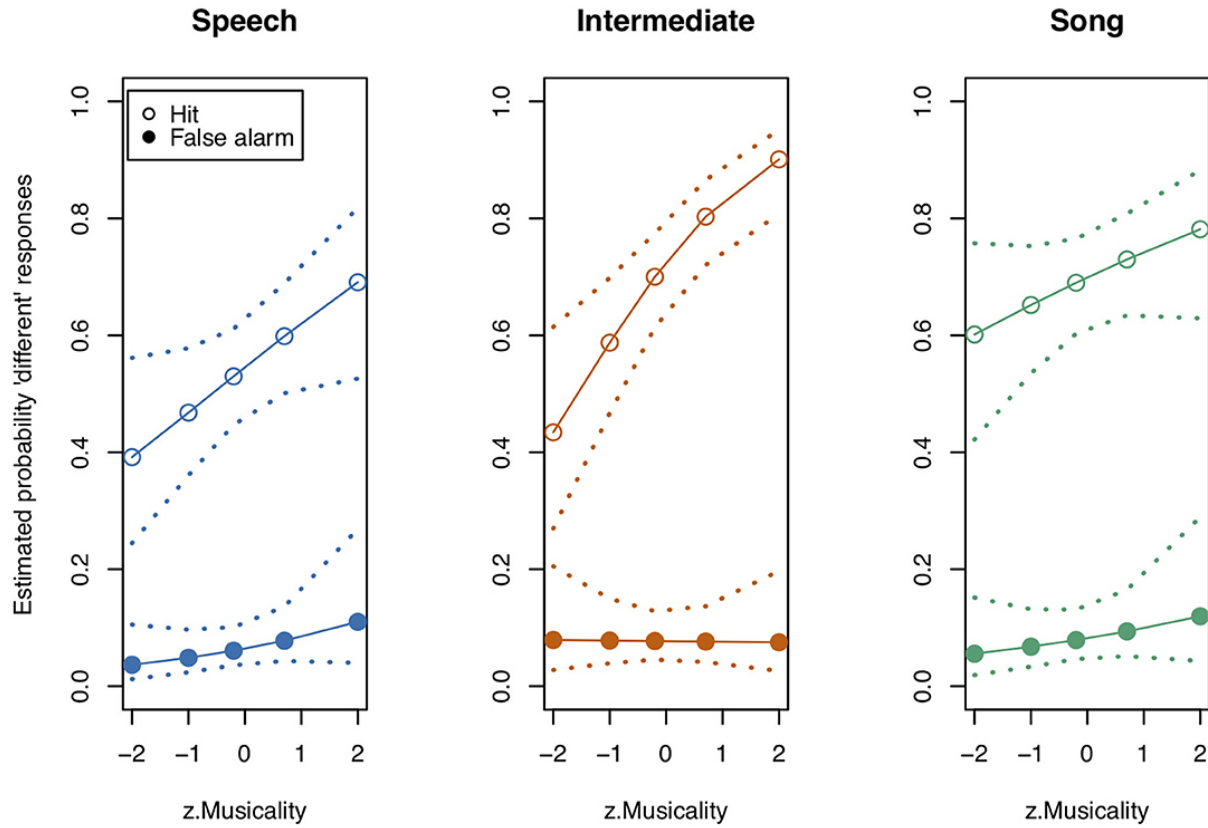


Figure 3. Model estimates for Musicality on the probability of detecting a deviant in study 1. Musicality had been assessed with the Gold-MSI and z-transformed for analysis. Empty circles: Hits, filled circles: False Alarms. Dotted lines represent 0.95 confidence intervals. Musicality showed the strongest impact for Intermediate stimuli. $N = 31$.

Term	Df	AIC	LRT	Pr(Chi)	Nagelkerke's R^2
Stimulus state:Variant	2	3559.364	6.503	0.039	0.043
Stimulus State:z.Musicality	1	3556.025	1.164	0.281	0.035
Variant:z.Musicality	2	3554.465	1.604	0.448	0.017

Table 4. Degrees of freedom (Df), Akaike Information Criteria (AIC), Likelihood Ratio test statistic (LRT), associated p -value [Pr(Chi)] and Nagelkerke's R^2 , based on likelihood ratio tests for two-way interactions of the model fitted in study 1.

For the Anova, we assessed the assumption of normality of residuals using a qqplot and histogram of the residuals, which showed no obvious deviation from normality. The assumption of sphericity had not been violated (Mauchly's Test of Sphericity: $W = 0.984$, $p = 0.793$). We obtained a trend for a difference between Variants [$F_{(2, 60)} = 2.5$, $p = 0.09$, generalised $\eta^2 = 0.03$]. For an overview of the distribution of d-primes as function of Stimulus Variant please see Supplementary Figure 2, for the raw hit and false alarm rates please see Supplementary Table 3.

3 Study 2

Study 1 showed the predicted effect of discrete pitch on memory. To test how robust and far-reaching the effect would be, and in particular whether it would be present in a much more difficult task, we replicated study 1 with more challenging settings. To increase difficulty, we decreased the magnitude of deviation in the target stimuli, increased the number of syllables in each stimulus and changed the distractor sound. Moreover, given that we found that Intermediate stimuli had similar effects to Song stimuli, we replicated study 1 with an Intermediate stimulus closer to Speech stimuli. We hypothesised that the effects of Intermediate stimuli would then be between those for Song and Speech stimuli.

3.1 Materials and Methods

3.1.1 Stimuli

The same procedure for stimulus creation was used as for study 1 except that now 10 syllables were used per stimulus contour and the duration limits changed to 2–4 s (as opposed to 7 syllables and a duration between 2 and 3 s in study 1). To this end the first nine syllables and the last syllable of each contour were concatenated. Also, we reduced the pitch shift of the target syllable from the template stimulus in comparison to study 1 (mean = 1.1 semitones [absolute], SD = 1.4). Two variants were created of each deviant: one where the pitch change went in the same direction between target syllable and its predecessor as before the shift, and one where the change was in the opposite direction. Moreover, the distractor sound was changed such that white noise was band-pass filtered (Hann) with a random value between 50 and 300 Hz for each burst as the lower cut-off and 10 times this value as the upper cut-off (smoothing: 10 Hz). A unique distractor sound was assigned to each variant of each contour. Since in study 1 both Song and Intermediate stimuli showed very similar effects, we decided to use a stimulus variant more similar to Speech. Of the seven intermediate variants that made a continuum between Speech and Song we used the variant one step closer to Speech (i.e., step 5 as compared to step 6 in study 1).

3.1.2 Procedure

Forty participants (35 females, 5 males, mean age = 21.5 years, range 19 to 54 years) took part. The same experimenter as in study 1 ran all participants. The procedure was exactly the same as for study 1, except that study 2 took place in a sound proof chamber. The total duration of part 2 of the experiment ranged from 50 to 105 min. Three participants needed more than 60 min due to the self-paced procedure, one of which took 105 min to complete the experiment. To avoid that speakers of tonal languages would interpret stimuli differently on the basis of intonation than other participants,

we excluded speakers of tonal languages. Therefore, one speaker of Mandarin Chinese and one speaker of Vietnamese were excluded from the sample for statistical analysis.

3.1.3 Statistical Methods

The data showed very different effects in comparison to the data of study 1 (see Figure 4). Calculating d-primes the traditional way (see Supplementary Figures 2–4) showed much lower sensitivity in study 2 (mean d-prime = 0.48, var = 0.26) compared to study 1 (mean = 1.65, var = 0.49) (study 3 showed intermediate values: mean = 1.08, var = 0.65). We therefore concluded that this task was quite difficult to solve for the participants.

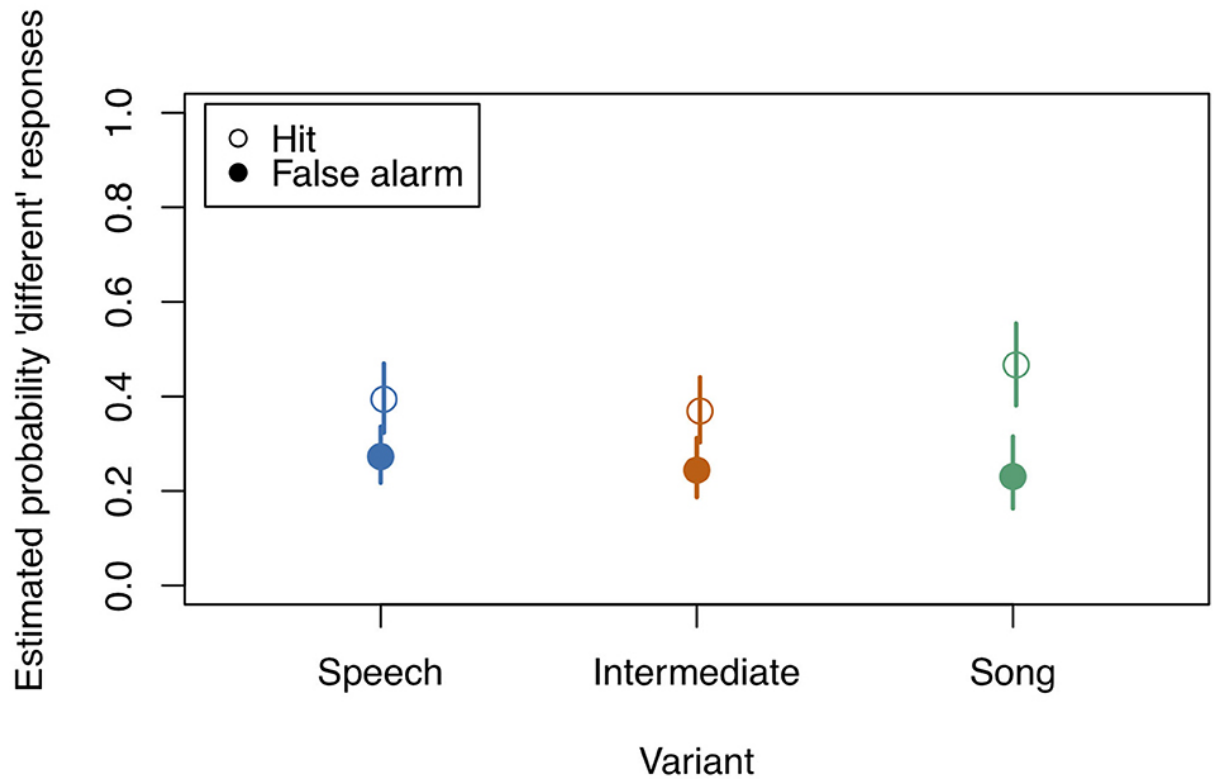


Figure 4. Model estimates for the probability of detecting a deviant in study 2. Empty circles: Hits, filled circles: False Alarms. Bars indicate 0.95 confidence intervals. Note that the probability of detection is overall very low. $N = 38$.

We fitted a logistic GLMM with logit link function with the same parameters as in study 1. Again, interactions of random intercepts and random slopes were not included. We attempted to limit semitone deviation in this study to 1 or 2 semitones and therefore did not include this factor in the model. Although the limitation worked on average (mean = 1.1 semitones [absolute], SD = 1.4), we found in a *post-hoc* analysis of stimuli that the deviation surpassed this limit for some contours, yielding a significant difference between stimulus variants in terms of deviation (Median_{Speech}: 2.27; Median_{Intermediate}: 2.29; Median_{Song}: 1.61; Kruskal Wallis Test: $H = 29.68$, $df = 2$, $p < 0.001$). However,

including absolute semitone deviation as an additional random factor in the model yielded slightly different results, but did not change the interpretation from the original model, so we only report the original model here. Trials of blocks for which participants indicated a concentration level below 3 (on a 7-point Likert scale with 1 being lowest concentration) were excluded from the data. This concerned three blocks in total, from two participants, excluding 120 out of 4,560 trials. Thus, 4,440 observations from 38 individuals remained as sample size for the model. Model stability was confirmed using a custom-built function (for ranges of parameter estimates of fixed effects please see Supplementary Table 4 and Supplementary Figure 5). Collinearity was assessed as for study 1 and no collinearity issues were found. Testing overdispersion revealed that the model was slightly underdispersed ($\chi^2 = 3685.337$, $df = 4019$, $p = 0.99$, dispersion parameter = 0.92), so no further action was required. Full-null model-comparison and tests of fixed predictors were done using likelihood ratio tests as in study 1. We derived 0.95-confidence intervals using the function “bootMer” (package “lme4”) with 1,000 parametric bootstraps. After fitting this model, we additionally did a repeated-measures Anova on classically calculated d-primes for informative purposes with the same parameters as in study 1 (Corrections for hit rates and false alarm rates of 0 and 1 had to be done for 3 of 114 cells for this analysis). *Post-hoc* tests were calculated using the package “lsmeans,” version 2.30-0 (Lenth, 2016).

3.2 Results

3.2.1 Post-experiment Questionnaire

Fourteen participants reported they attempted subvocal rehearsal, three of these participants also visualised the stimuli. Four additional participants reported visualising the stimuli. Two participants guessed that the study was about memory.

3.2.2 Statistical Results

Although the full model (model coefficients see Table 5) revealed significance in comparison to the null model ($\chi^2 = 40.959$, $df = 11$, $p < 0.001$), likelihood ratio tests revealed no significance of the three-way interaction “Stimulus State:Variant:z.Musicality” (LRT = 0.485, AIC = 4886.652, $df = 2$, $p = 0.785$). There was also no significant effect of any of the two-way interactions “Stimulus State:Variant” (LRT = 3.795, AIC = 4886.447, $df = 2$, $p = 0.450$), “Stimulus State:z.Musicality” (LRT = 2.865, AIC = 4887.517, $df = 1$, $p = 0.091$), and “Variant:z.Musicality” (LRT = 4.142, AIC = 4886.794, $df = 2$, $p = 0.126$). Figures 4, 5 show the probability estimates based on the model. We could therefore not replicate our findings from study 1. However, based on the overall low values of d' , we concluded that the task was too difficult for participants, which motivated study 3, in which we created a task of medium difficulty.

Explaining Spectrotemporal Differences between Music and Language

Term	log odds	SE	CI _{lower}	CI _{upper}	z	p
Intercept	-0.980	0.153	-1.292	-0.665	-6.411	
Stimulus state diff	0.550	0.169	0.231	0.887	3.247	0.001
Intermediate	-0.150	0.147	-0.437	0.142	-1.023	0.306
Song	-0.225	0.201	-0.614	0.160	-1.117	0.264
z.Musicality	-0.063	0.138	-0.333	0.207	-0.457	0.648
Stimulus state diff: Intermediate	0.044	0.228	-0.428	0.498	0.193	0.847
Stimulus state diff: Song	0.521	0.277	-0.022	1.072	1.885	0.059
Stimulus state diff: z.Musicality	0.117	0.145	-0.207	0.417	0.812	0.417
Intermediate: z.Musicality	0.012	0.125	-0.256	0.263	0.093	0.926
Song: z.Musicality	0.204	0.145	-0.066	0.478	1.410	0.159
Stimulus state diff: Intermediate: z.Musicality	0.125	0.186	-0.246	0.515	0.675	0.500
Stimulus state diff: Song: z.Musicality	0.032	0.191	-0.374	0.419	0.170	0.865

The model estimated detection of a difference between original sound and target sound (reference level: Stimulus State same) as a function of Stimulus Variant (Speech, Song, Intermediate, Reference level: Speech) and Musicality as assessed by the GoldMSI questionnaire and z-transformed (that way the estimates refer to an individual with average musicality).

Table 5. Estimates of model coefficients (log odds), standard errors (SE), lower (CI_{lower}) and upper (CI_{upper}) 0.95 confidence intervals, z statistics (z-value) and associated p-values of the model fitted for study 2.

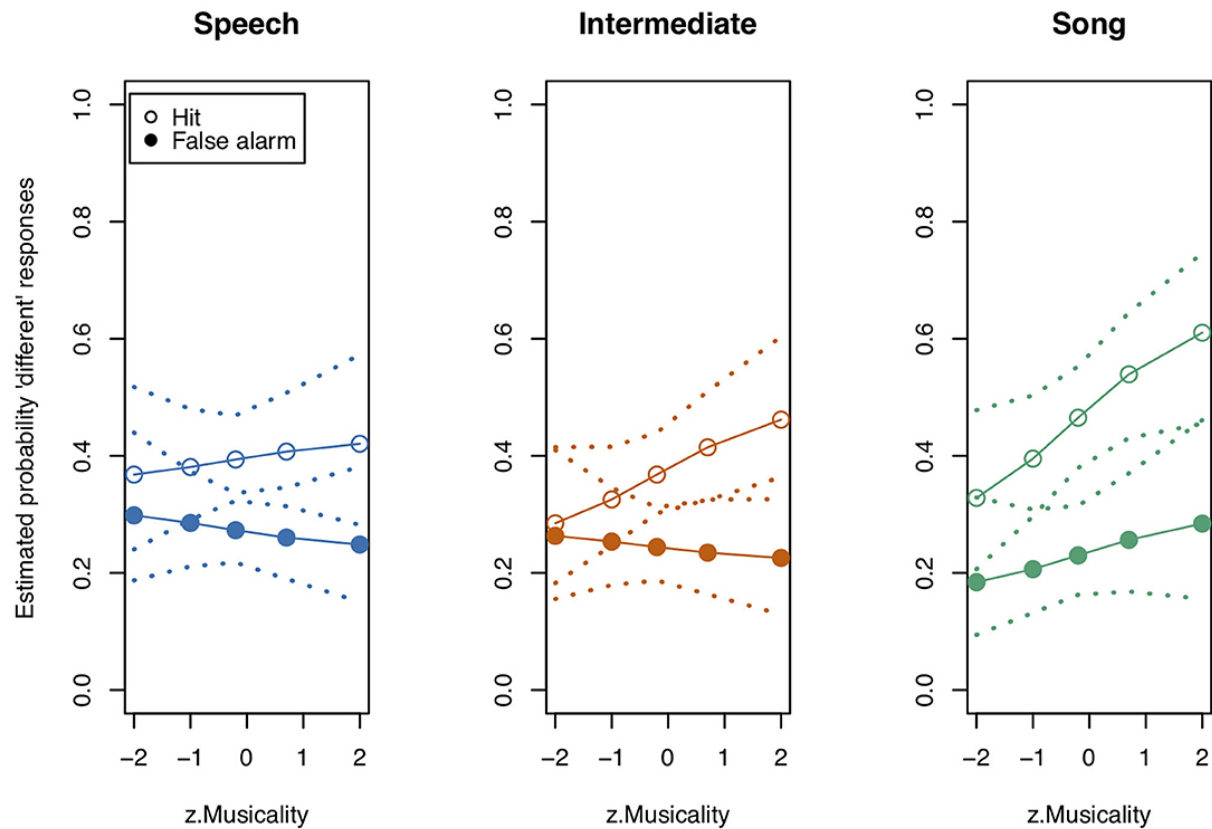


Figure 5. Model estimates for Musicality on the probability of detecting a deviant in study 2.

Musicality had been assessed with the Gold-MSI and z-transformed for analysis. Empty circles: Hits, filled circles: False Alarms. Dotted lines represent 0.95 confidence intervals. Musicality showed the strongest impact for Intermediate stimuli. $N = 38$.

For the Anova, a qqplot and histogram of the residuals showed no obvious deviation from normality. The assumption of sphericity had not been violated (Mauchly's Test of Sphericity: $W = 0.876, p = 0.092$). D-primes differed significantly between Variants [$F_{(2, 74)} = 3.89, p = 0.02$, generalised $\eta^2 = 0.05$]. Tukey-adjusted pairwise contrasts showed that the effect was driven by the contrast between Speech and Song ($t_{(74)} = -2.638, p = 0.027$). The contrast Intermediate–Song showed a trend [$t_{(74)} = -2.105, p = 0.096$], while the contrast Speech–Intermediate remained non-significant: $t_{(74)} = -0.532, p = 0.855$. For an overview of the distribution of d-primes as function of Stimulus Variant please see Supplementary Figure 3, for the raw hit and false alarm rates please see Supplementary Table 5.

4 Study 3

Since the task changes we did in study 2 to replicate study 1 appeared to have lead to difficulties for our participants to be successful at all, we attempted to create a task of medium difficulty. To this end we allowed a slightly greater magnitude of deviation within target stimuli and changed the number of syllables per stimulus back to seven. Since the distractor sound in study 2–narrowband white noise bursts with different centre frequencies–might have interfered more strongly with discrete pitches, we created new distractor sounds based on female voices.

4.1 Materials and Methods

4.1.1 Stimuli

The same stimulus contours as in study 1 were used, i.e., seven syllables per stimulus contour with a duration between 2 and 3 s. However, contours of which deviant syllables differed by more than six semitones from the original contour were removed (measured as the difference at 0.04 s from onset of a deviant syllable and the respective syllable of the original contour, using Praat). Thus, the mean absolute deviation was 1.6 semitones (SD = 1.9). We controlled the direction of deviation more precisely such that in study 3 a pitch interval from preceding syllable to target syllable always reversed the direction in comparison to the original stimulus, with no exception. This resulted in a pool of 45 contours in three variants (Song, Speech, Intermediate) each. As intermediate stimulus contour, step 5 of the song-speech continuum was used as in study 2. There were no significant differences between the three variants with regard to both directional (Median_{Speech}: 1.06; Median_{Intermediate}: 1.38; Median_{Song}: 1.73; Kruskal Wallis Test: $H = 0.35, df = 2, p = 0.84$) and absolute semitone deviations (Median_{Speech}: 4.10; Median_{Intermediate}: 3.95; Median_{Song}: 3.95; Kruskal Wallis Test: $H = 0.68, df = 2, p = 0.71$). Since the distractor sounds in study two consisted of narrowband noise chunks with stable frequency range for 0.2 to 0.3 s each they might have been perceived as being of stable pitch (however unprecise) by our participants. This way, study 2 would have been more like an n-back task with stronger interference for song variants. We therefore decided to create new distractor sounds consisting of a

mixture of female voices. To this end, voices of six females were recorded singing one syllable [la:l] on tones one octave above the tones used for the actual (male voiced) stimuli. A Zoom H4n recording device (16 bit, 44.1 kHz) was used for these recordings. The procedure was the same as for creating the original stimuli. PitchTiers from all variants of the continuum (from speech-like to song-like) that were originally created as basis of the stimulus sounds (male voiced) were shifted up by one octave. Female recordings closest to the pitch of these octave-shifted PitchTiers were then adjusted in duration to the respective PitchTiers. Afterwards, PitchTier and respective recording were combined to keep the spectral information of the female recording along with the desired pitch contour. These syllables were then ramped (trapezoid). That way, a pool of syllables of six female voices was created. From this pool syllables were drawn randomly for each female voice to result in a stream of syllables of maximum 2 s duration. Streams of all six female voices were then mixed together to mono. This way, the distractors contained voices with intermixed syllables of all variants on the continuum. One unique distractor sound was assigned to each variant of each stimulus contour.

4.1.2 Procedure

Twenty one participants (16 females, mean age = 24.8 years, range 19 to 46 years) took part in the study. The same experimenter as in study 1 and 2 ran all participants. The procedure was equal to that of study 2 except that the test of holistic vs. spectral hearing was left away. No participant needed more than 60 min to complete the study. None of the participants had to be excluded for statistical analysis.

4.1.3 Statistical Methods

We fitted a logistic GLMM with logit link function with the same parameters as in study 1. Interactions of random intercepts and random slopes were not included. Dummy coding was done using a custom-built function. Optimizer “bobyqa” with maximum 100.000 function iterations was used to assess the maximum likelihood. The model showed singular fit due to boundary estimates in the random effects structure. Since the removal of random effects might lead to inflated Type I error rates and singular fit does not influence the estimates of the fixed effects, we decided to keep the random effects in the model. Trials of blocks for which participants indicated a concentration level below 3 (on a 7-point Likert scale with 1 being lowest concentration) were excluded from the data. This was the case for 80 out of 2,520 trials, stemming from two blocks of one participant. Thus, 2,440 observations from 21 individuals remained as sample size for the model. Model stability was assessed using a custom-built function. On this basis we did not obtain any influential cases, thus the model was stable (for ranges of parameter estimates of fixed effects please see Supplementary Table 6 and Supplementary Figure 6). Collinearity was assessed as for study 1. There were no collinearity issues found. Testing overdispersion revealed that the model was slightly underdispersed ($\chi^2 = 2108.610$, df

= 2,419, $p = 1$, dispersion parameter = 0.87). Full-null model-comparison and tests of fixed predictors were done using likelihood ratio tests as in study 1. We derived 0.95-confidence intervals using the function “bootMer” (package “lme4”) with 1,000 parametric bootstraps. After these analyses, we additionally did a repeated-measures Anova on classically calculated d-primes for informative purposes with the same parameters as in study 1 (Corrections for hit rates and false alarm rates of 0 and 1 had to be done for 5 of 63 cells for this analysis). *Post-hoc* tests were calculated using the package “lsmeans,” version 2.30-0 (Lenth, 2016).

4.2 Results

4.2.1 Post-experiment Questionnaire

Three participants reported that they mentally visualised the contours in order to memorise them. Two participants reported using movements of their hands or feet for the same purpose. Six participants reported that they attempted to practice the stimulus silently during the distractor sound in order to not forget it. One participant guessed the study was about a comparison of prosody and melody.

4.2.2 Statistical Results

After fitting the full model (model coefficients see Table 6) we compared it to the null model (which was lacking the predictors of interest). All predictors of interest together explained the data significantly better than the null model ($\chi^2 = 67.342$, $df = 11$, $p < 0.001$). Figures 6 and Table 7 show the probability estimates derived from our full model.

Term	log odds	SE	CI _{lower}	CI _{upper}	z	p
Intercept	-1.478	0.199	-1.921	-1.119	-7.413	
Stimulus state diff	1.156	0.224	0.725	1.642	5.167	0.000
Intermediate	-0.137	0.204	-0.576	0.258	-0.670	0.503
Song	-0.169	0.201	-0.581	0.236	-0.842	0.400
z.Musicality	-0.218	0.194	-0.596	0.118	-1.126	0.260
Stimulus state diff:Intermediate	0.867	0.247	0.370	1.390	3.519	0.000
Stimulus state diff: Song	0.832	0.249	0.339	1.320	3.342	0.001
Stimulus state diff:z.Musicality	0.456	0.221	0.024	0.897	2.063	0.039
Intermediate:z.Musicality	-0.030	0.188	-0.365	0.346	-0.157	0.875
Song:z.Musicality	-0.175	0.191	-0.531	0.169	-0.915	0.360
Stimulus state diff: Intermediate:z.Musicality	0.054	0.247	-0.407	0.534	0.217	0.828
Stimulus state diff: Song:z.Musicality	0.297	0.248	-0.160	-0.160	1.195	0.232

The model estimated detection of a difference between original sound and target sound (reference level: Stimulus State same) as a function of Stimulus Variant (Speech, Song, Intermediate, Reference level: Speech) and Musicality as assessed by the GoldMSI questionnaire and z-transformed (that way the estimates refer to an individual with average musicality).

Table 6. Estimates of model coefficients (log odds), standard errors (SE), lower (CI_{lower}) and upper (CI_{upper}) 0.95 confidence intervals, z statistics (z-value) and associated p-values of the model fitted for study 3.

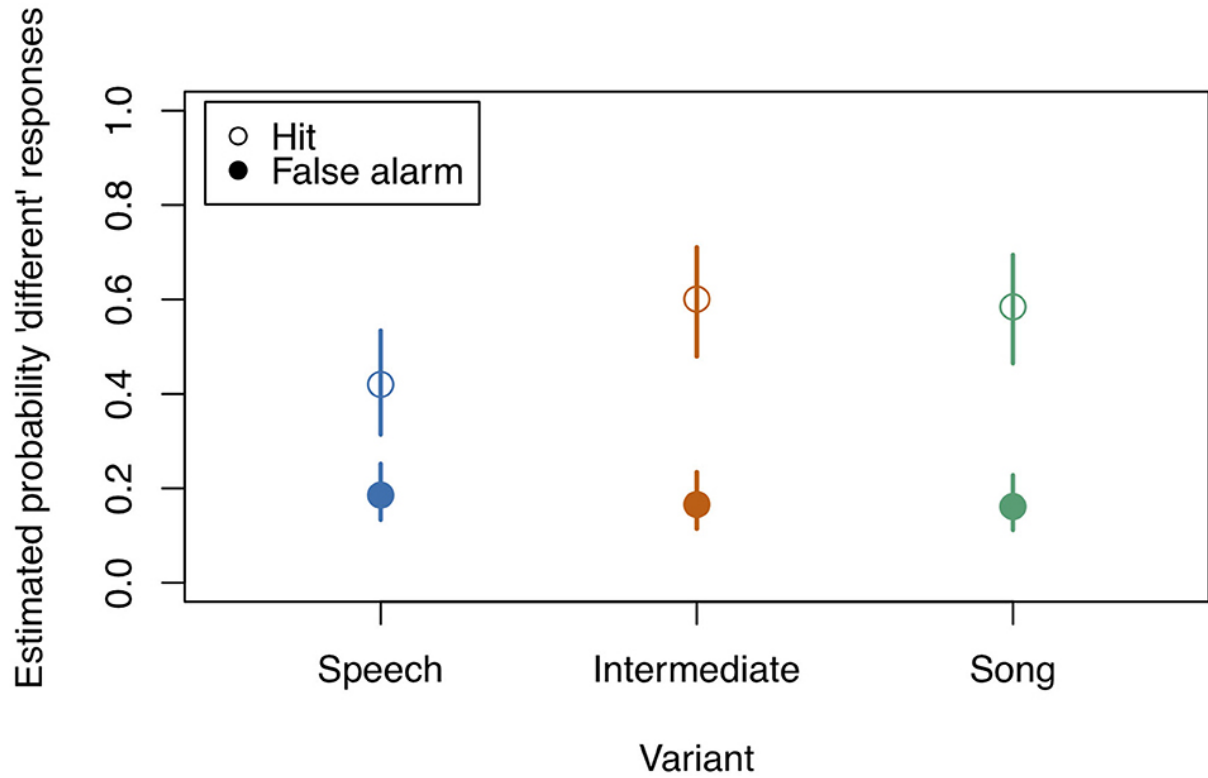


Figure 6. Model estimates for the probability of detecting a deviant in study 3 (see also Table 7). Empty circles: Hits, filled circles: False Alarms. Bars indicate 0.95 confidence intervals. $N = 21$.

Stimulus state	Speech			Intermediate			Song		
	Estimate	CI _{lower}	CI _{upper}	Estimate	CI _{lower}	CI _{upper}	Estimate	CI _{lower}	CI _{upper}
cliff	0.420	0.314	0.534	0.601	0.479	0.711	0.584	0.495	0.694
same	0.186	0.134	0.252	0.166	0.114	0.234	0.161	0.112	0.228

Derived from the R-package "effects," version 4.1-0 (Fox et al., 2019).

Table 7. Probability estimates and 0.95 confidence intervals for Hits and False Alarms as depicted in Figure 6.

Thereafter, we tested the effects of the interactions. In contrast to study 1, the three-way interaction “Stimulus State:Variant:z.Musicality” remained non-significant ($LRT = 1.764$, $AIC = 2723.899$, $df = 2$, $p = 0.414$). However, consistent with our hypothesis we found a significant interaction of “Stimulus State:Variant,” with Song and Intermediate enhancing memory with respect to Speech. Participants with a higher score in “Musicality” (assessed by the Gold-MSI) were better able to remember stimulus contours across all three Stimulus States (Song, Speech, Intermediate), suggested by a significant interaction of “z.Musicality” with “Stimulus State” (see Figure 7 and Table 8).

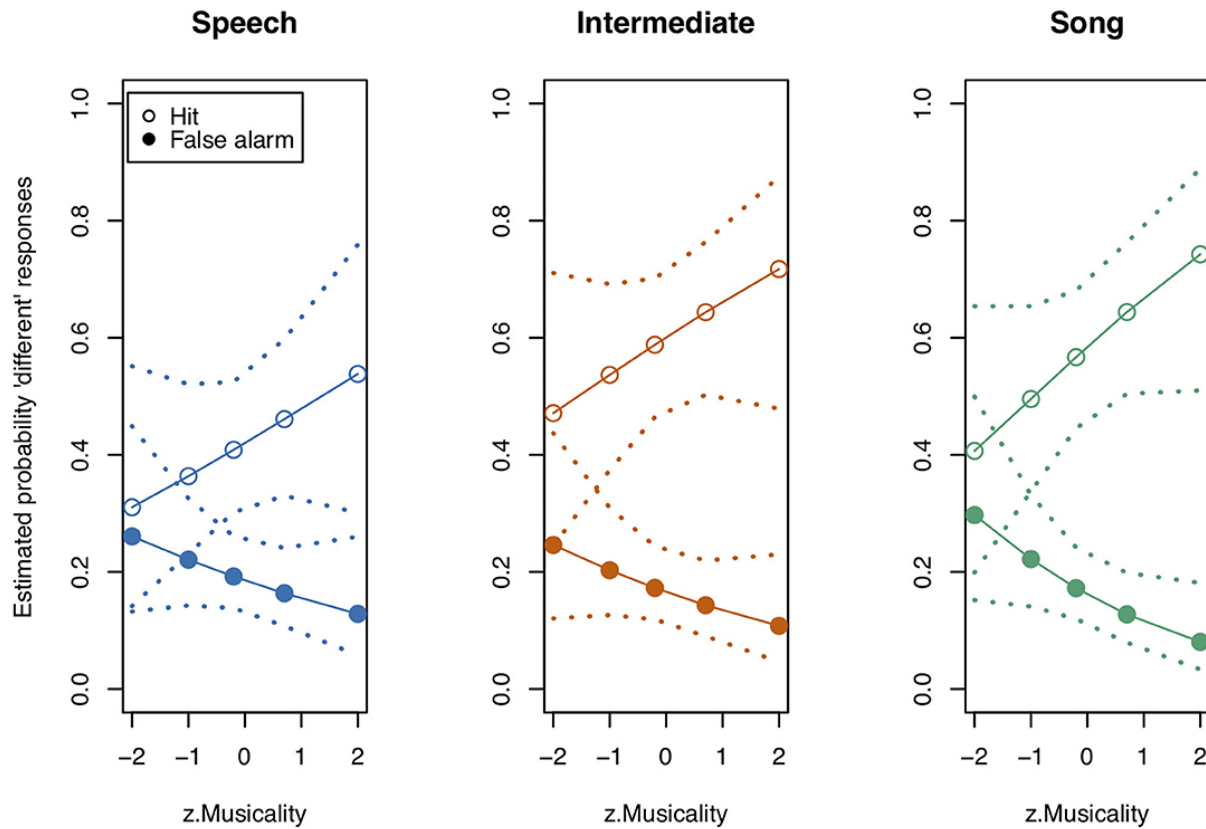


Figure 7. Model estimates for Musicality on the probability of detecting a deviant in study 3. Musicality had been assessed with the Gold-MSI and z-transformed for analysis. Empty circles: Hits, filled circles: False Alarms. Dotted lines represent 0.95 confidence intervals. $N = 21$.

Term	Df	AIC	LRT	Pr(Chi)	Nagelkerke's R^2
Stimulus state:Variant	2	2729.686	15.174	<0.001	0.039
Stimulus State:z.Musicality	1	2725.351	8.839	0.003	0.035
Variant:z.Musicality	2	2714.512	<0.001	1.000	0.022

Derived from the R-package "effects," version 4.1-0 (Fox et al., 2019).

Table 8. Degrees of freedom (Df), Akaike Information Criteria (AIC), Likelihood Ratio test statistic (LRT), associated p -value [Pr(Chi)] and Nagelkerke's R^2 , based on likelihood ratio tests for two-way interactions of the model fitted in study 3.

For the additional Anova, normality of residuals was assessed using a qqplot and histogram of the residuals. There was no obvious deviation from normality. The assumption of sphericity was not violated (Mauchly's Test of Sphericity: $W = 0.896$, $p = 0.353$). We found a significant difference between Variants [$F_{(2, 40)} = 5.648$, $p = 0.007$, $\eta^2 = 0.10$]. Contrasts using Tukey-adjusted p -values showed that the differences that drove the effect of "Variants" were between Speech and Song [$t_{(40)} = -2.476$, $p = 0.046$], and between Speech and Intermediate [$t_{(40)} = -3.206$, $p = 0.007$], but not between

Song and Intermediate [$t_{(40)} = 0.731, p = 0.747$]. For an overview of the distribution of d-primes as function of Stimulus Variant please see Supplementary Figure 4, for the raw hit and false alarm rates please see Supplementary Table 7.

5 Post-hoc Rating Study

To clarify whether our results are better explained by stimulus-driven bottom-up effects or by top-down decision making based on the categories “song” and “speech,” we carried out an online rating study with a new sample of participants.

5.1 Materials and Methods

5.1.1 Stimuli

We used all stimulus contours used in the three studies, with all three versions (Song, Speech, Intermediate), but without the distractor sounds. Additionally, we included the respective original Mandarin Chinese phrases that were the basis for the creation of our stimuli. These Mandarin phrases were lowpass-filtered with a cutoff of 600 Hz (100 Hz smoothing) to keep the pitch contours but reduce phonemic information. Stimuli had to be converted to mp3 format for technical reasons.

5.1.2 Procedure

The online study platform “Labvanced” (<https://www.labvanced.com>) was used for setting up the experiment and recruiting and testing participants. This platform operates under the General Data Protection Regulation of the European Union. Fifty seven participants took part in the study (38 females, mean age = 41.1 years, range 19 to 69 years). One participant had to be excluded because of not correctly filling out the consent form. All others gave informed consent. Participants were recruited world-wide. None of the participants stated that they spoke Mandarin Chinese.

First, a written instruction about the rating task was shown on the screen. Afterwards, participants had to do a headphone screening test implemented from the Labvanced study library to ensure that participants were wearing headphones. This test consisted of a three-alternative forced choice task where three sine tones of different loudness were presented per trial, with 12 trials in total. Only participants who successfully identified the quietest tones were allowed to continue. Subsequently, participants had to give informed consent that they understood the task, that their data would be used for analysis and that they were willing to participate. Afterwards, participants started the rating task, whereby each participant was presented with 13 randomly chosen stimuli from each of the four categories (Song, Speech, Intermediate, Lowpass-filtered Mandarin). The first 12 stimuli (three from each category) were considered training and not used in the analysis. During the display of

a white screen, the stimulus was played. After playing of the stimulus was finished, a text (“How does it sound like?”) and two rating bars appeared. The rating bars corresponded to a 9-point Likert scale, but no numbers were displayed. The upper rating bar displayed “Song” (corresponding to a Likert rating of 1) and “Speech” (9) as extremes (Song-Speech rating), the lower rating bar showed “natural” (1) and “artificial” (9) as the extremes (Natural-Artificial rating). For each trial the rating button was initially set in the middle. After a two second delay, a button appeared to enable the participants to continue with the next trial. After the rating task, the Gold-MSI as well as additional questions on age, gender and languages spoken were presented. This was followed by a written debriefing stating the purpose of the study.

5.1.3 Statistical Analysis

We analysed both ratings (Song-Speech and Natural-Artificial) with Friedman tests, using the mean rating per participant and stimulus class. As *post-hoc* tests, we applied Wilcoxon signed-rank tests. For the Song-Speech rating we compared the following stimulus variants: lowpass Mandarin vs. Speech, Song vs. Speech, Speech vs. Intermediate, and Song vs. Intermediate. We also calculated the differences in ratings for Speech vs. Intermediate and Song vs. Intermediate and compared those using a Wilcoxon test. For the Artificial-Natural rating, we compared the following stimulus variants: lowpass Mandarin vs. Speech, and Song vs. Speech.

5.2 Results

The results showed that when asked explicitly, participants rated all three stimulus classes used in our studies (Song, Speech, Intermediate) with a tendency towards artificially sounding, while the lowpass-filtered Mandarin was rated with a tendency towards natural sounding (see Supplementary Figure 7). Because of the significant Friedman test for the Natural-Artificial rating ($\chi^2 = 60.367$, $df = 3$, $p < 0.001$), we conducted Wilcoxon tests for pairwise comparison. Lowpass Mandarin was rated significantly more towards natural sounding than Speech (Median_{lowpass Mandarin} = 3.725, Median_{Speech} = 7.05, $z = 6.046$, $N = 53$, $p < 0.001$) while both Song and Speech did not differ significantly in their ratings towards artificially sounding (Median_{Song} = 7.3, Median_{Speech} = 7.05, $z = 0.188$, $N = 41$, $p = 0.855$; the Median for Intermediate was 7.0). These results suggest that although the stimuli sounded artificial, there was no great difference in this respect between stimulus categories.

For the Song-Speech rating, all three stimulus classes used in our studies (Song, Speech, Intermediate) tended to be rated as more song-like, while the lowpass-filtered Mandarin was rated as more speech-like (see Supplementary Figure 8). Since the Friedman test was significant ($\chi^2 = 125.330$, $df = 3$, $p < 0.001$), we could conduct *post-hoc* Wilcoxon tests. Among the three stimuli used in our studies, Song was rated significantly more song-like than Intermediate (Median_{Song} = 2.15, Median_{Intermediate} = 2.5, $z = 2.692$, $N = 48$, $p = 0.006$) and Intermediate was rated significantly more

song-like than Speech (Speech vs. Intermediate: $\text{Median}_{\text{Speech}} = 3.3$, $\text{Median}_{\text{Intermediate}} = 2.5$, $z = 5.476$, $N = 49$, $p < 0.001$). Song and Speech ratings differed significantly ($\text{Median}_{\text{Song}} = 2.15$, $\text{Median}_{\text{Speech}} = 3.3$, $z = 5.436$, $N = 48$, $p < 0.001$) as did the ratings for lowpass Mandarin and Speech ($\text{Median}_{\text{lowpass Mandarin}} = 8.6$, $\text{Median}_{\text{Speech}} = 3.3$, $z = 6.451$, $N = 55$, $p < 0.001$). The difference between ratings for Intermediate and Speech were significantly larger than those for ratings between Intermediate and Song ($\text{Median}_{\text{Speech-Intermediate}} = 0.75$, $\text{Median}_{\text{Intermediate-Song}} = 0.1$, $z = 4.952$, $N = 50$, $p < 0.001$). These results suggest that the memory advantage we observed in studies 1 and 3 was likely stimulus-driven rather than mediated by a top-down decision making guided by a “song vs. speech” decision.

6 Discussion

In the current studies we tested the hypothesis that the discrete pitches typifying song are a spectrotemporal property optimised for auditory memory. We thus predicted that vocal phrases are remembered better when consisting of discrete (song-like) pitches than when consisting of gliding (speech-like) pitches. To this end we conducted three studies in a same-different paradigm, using a forced-choice task where song-like, speech-like and intermediate stimuli were presented. To vary task difficulty among studies, we modified stimulus length, used a different type of distractor sound in each study, and slightly modified the difficulty of discrimination in each task. The results of study 1 are consistent with our hypothesis, showing that the use of discrete pitches in an auditory stimulus modelled after a vocal phrase indeed enhances memory of that stimulus. In study 2 we were not able to replicate these results, possibly due to overall task difficulty. In study 3 we replicated the effect of memory advantage for stimuli of discrete pitches that we found in study 1. Intermediate stimuli (partially discretised) could be remembered as well as song-like stimuli in studies 1 and 3. We also found that participants with more musical experience remembered the stimuli better, in both study 1 and study 3. In study 1 this effect occurred as a stronger advantage for more musically experienced participants for song-like stimuli, compared to the other stimuli. In study 3 we found that participants with more musical experience remembered all three types of stimuli (Song, Speech, Intermediate) better than participants with less musical experience. Our results are thus consistent with the hypothesis that discrete pitches are a spectrotemporal property aiding auditory memory (“song memory advantage” hypothesis), and do not support the alternative “experience advantage” hypothesis: that the abundance of speech perception and production throughout the human lifespan would make gliding pitch more salient for auditory memory than discrete pitch.

Our results indicate that one fundamental design feature of music, discrete pitches, has cognitive effects that go beyond surface-level perception. Our results cannot be based on timbre differences, since all stimuli were based on recorded syllables from the same male voice. Neither can the effects be explained by differences in rhythmic structure, isochrony or meter since all three versions of our stimuli (Song, Speech and Intermediate) used the same syllable timing. Syllable

durations were based on natural speech prosody and were additionally occasionally lengthened to have at least 250 ms duration, making them non-isochronous. Finally, we used the same set of pitch contours, that is the global upwards and downwards movement of the pitch, for all three stimulus types. Therefore, we conclude that one of the design features of song (and music more generally)—the use of discrete pitches—has a significant cognitive effect on auditory memory, and is not a non-functional side-effect of speech prosody.

We used Mandarin Chinese as the pitch contour template for our stimuli to minimise the influence of familiarity to our participants, who were not experienced with tonal languages. Although non-tonal language speakers are worse than native speakers of Mandarin in processing rapid pitch changes of lexical tone (Krishnan et al., 2010) and musical expertise may facilitate lexical tone processing (Alexander et al., 2005; Chandrasekaran et al., 2009), we emphasise that our stimulus creation process favoured the maintenance of sentence-level prosody in the stimuli over lexical tones. Moreover, the overall pitch contour was the same across all three stimulus types (see Supplementary Figures 9–14 for three sample stimuli and Supplementary Figures 15, 16 for a comparison to natural song and speech pitch contours). Sentence-level prosody of Mandarin Chinese has been shown to be unfamiliar but still behaviourally and neurally processable by non-tonal language speakers (Tong et al., 2005). It is therefore unlikely that our effects are related to difficulty in following rapid pitch changes in the speech prosody stimuli and are rather explained by a memory advantage for discrete pitch.

Our *post-hoc* rating study showed that when explicitly asked to rate the stimuli used in studies 1 to 3 (Song, Speech, Intermediate) as well as lowpass-filtered Mandarin phrases as more song-like or more speech-like, the former were rated as more song-like while the lowpass-filtered Mandarin phrases were rated as more speech-like. Lowpass-filtered Mandarin was also rated as more natural sounding, while Song, Speech and Intermediate stimuli were rated as more artificially sounding, with no significant differences among the latter. Even though the nature of the rating study was different from studies 1 to 3 (which did not include lowpass-filtered Mandarin phrases nor explicit mentioning of “song” and “speech” categories), this results suggests that our stimuli, although based on recordings of natural vocalisations, might not have represented natural speech and song very well. The memory effects we observed are therefore probably based on bottom-up processing of the pitch differences rather than categorical top-down processing. It is possible that the fact that including the lowpass-filtered Mandarin phrases induced that the Song, Speech and Intermediate stimuli were perceived as very similar and thus biased towards a song-like rating. However, among the three stimulus variants used in studies 1 to 3, Song stimuli were still rated as most song-like, and Speech stimuli as least song-like, with Intermediate stimuli in between, all differing significantly. It thus seems possible that actual natural speech and song might induce a stronger difference in auditory memory, or that the effect

could be replicated with non-vocal stimuli as well. These questions could be addressed by future studies.

We were unable to replicate our findings in study 2. D-primes were on average lowest for this study in comparison to studies 1 and 3. It seems therefore likely that the task in study 2 was too difficult for our participants, which masked any differences between stimulus variants. This could be explained by changes in the distractor stimuli we introduced to increase the difficulty of study 2. In comparison to the other two studies, we used narrowband white noise with different, variable centre frequencies as distractor sounds, while study 1 used simple pink noise bursts, and study 3 used a “cocktail party” mixture of female voices. We also decreased the pitch deviation of the target from the template sound among the studies (study 1: mean = 3.8 semitones, SD = 5.0; study 2: mean = 1.1 semitones, SD = 1.4; study 3: mean = 1.6 semitones, SD = 1.9). Moreover, we increased the number of syllables in each stimulus to 10 in study 2 (rather than 7 as in studies 1 and 3). This increased the duration of each stimulus contour to 3 to 4 s (studies 1 and 3: duration between 2 and 3 s). These three differences conspired to make study 2 a very challenging task for our participants.

Narrowband white noise stimuli with different centre frequencies can lead to a perception of discrete tones (see Hesse, 1982). This “pitchlike” quality of the distractor sounds in study 2 may have interfered most with discrete-pitched song stimuli and least with speech prosody stimuli. However, d-primes were also low for speech prosody stimuli in this study. Moreover, in study 3 we included discrete-pitched and gliding female voices as distractor sounds, which should have been most distracting throughout the three studies due to the close similarity to our target stimuli. In study 3 we nonetheless replicated the findings of study 1. The distractor sounds used in study 2 might therefore explain the low performance of our participants to only a small degree. Pitch deviations in target sounds however were smallest for study 3 in comparison to studies 1 and 2. Winkler et al. (2002) showed that auditory sensory memory representations are less stable for smaller pitch deviations after a 10-s silent interval. In line with these findings, visual inspection of the data of study 3, the study where our participants performed sufficiently well, and where pitch deviations were tightly controlled, shows that greater pitch deviation slightly enhanced d-primes (see Supplementary Figure 17). Nonetheless, it also shows that performance remained quite high for small pitch deviations. Small pitch deviations therefore seem insufficient to explain the low performance of our participants in study 2. Finally, the number of syllables per stimulus was 10 in study 2 (vs. 7 in the other studies). Williamson et al. (2010) showed in a tone sequence memory test that performance approached chance level when sequences consisted of more than seven tones. It seems plausible that this was the case for our study 2 participants as well. If this is indeed the main explanation for the low performance of our participants in study 2, it might suggest that our stimuli were remembered element-wise rather than as a global pitch contour. This seems natural concerning the task was to detect pitch deviations in a

single syllable. Future research should investigate limits of auditory memory capacity for syllables with discrete and gliding pitch more thoroughly and systematically.

We found that participants with higher musical experience remembered the stimuli better, in both study 1 and study 3. Among the three stimulus types this advantage was strongest for song stimuli in study 1. In study 3 there was no significant difference for the effect of musical experience between the three stimulus types. This suggests that the advantage musically experienced people show occurs for song as well as for speech prosody stimuli and the intermediate stimulus class, not specifically for song stimuli only. This is in line with cross-over effects of musical training into the linguistic domain (Patel and Morgan, 2017). Musicians have also been shown to process rapid pitch changes better than non-musicians (Chandrasekaran et al., 2009). Although the pitch changes in our stimuli were slower than in Mandarin Chinese lexical tone, this might have influenced the detection of deviants in musicians. A meta-analysis by Talamini et al. (2017) also suggested that musicians show a general advantage over non-musicians in auditory working memory, in both tonal stimuli and verbal stimuli, with tonal stimuli showing the largest effect size. However, our data show that participants with lower self-assessed musicality also remember vocal phrases better when they consist of discrete pitches. This indicates that discrete pitches are processed similarly in auditory memory by both musicians and non-musicians, and that no musical training is necessary to obtain an advantage for discretised pitch. Our stimuli were not based on western musical scales (we used a Bohlen-Pierce scale as our model). The memory advantage for discretised pitch might be stronger, and might be greater in musicians, if major and minor western scales are used. Based on these results, we conclude that musical training is not a necessary prerequisite for a memory advantage for discrete pitches, and that the average day-to-day exposure to music of any hearing person in modern society may be enough to elicit this memory advantage.

An unexpected finding was that vocal phrases intermediate between song- and speech-like intonation, that is with syllables with partially stable and partially gliding pitch, were remembered equally well as song-like (fully discrete) phrases. We derived the category boundary between speech prosody and song for study 1 (based on a 9-step continuum, using a staircase procedure with two-alternative forced choice categorisation as speech/song, see Kingdom and Prins, 2016) by estimating the point of subjective equality for one subject (author FH). For study 1 we then used the stimulus closest to that boundary (step 6), and later we used one step closer to Speech (step 5) as the intermediate stimulus in studies 2 and 3. We expected that stimuli falling between clear song and speech categories would elicit a moderate memory advantage at best, but in fact performance was comparable to the clear song category. However, the *post-hoc* rating study showed that the ratings differed significantly between all three stimulus variants used in studies 1 to 3, with Song being rated more song-like than Intermediate, and Intermediate being rated more song-like than Speech. This suggests that the intermediate stimulus was indeed perceived as being between Song and Speech

stimuli. Although the differences between ratings for Speech and Intermediate were significantly greater than those between Song and Intermediate, it remains debatable how well this explains the comparable memory effects for Song and Intermediate stimuli. To further probe the finding that our intermediate stimuli showed similar memory effects to our song stimuli, although being significantly discriminated from both speech prosody stimuli and song stimuli, it would be useful in future to use adaptive psychoacoustic methods to choose the stimulus material. In doing so, the stimulus space spanning between speech prosody and song could be explored in an efficient way. This could be informative about the extent that speech and song form two perceptual categories, based on pitch contour differences, potentially revealing ambiguous stimuli individually for each participant. Contextual cues that would change the categorical interpretation of these ambiguous stimuli as song- or speech-like would allow experimenters to isolate the influence of bottom-up perceptual vs. top-down categorical processes on auditory memory.

Overall, our findings indicate that song-like pitch trajectories, consisting of discrete pitches, yield a memory advantage over speech-like stimuli with gliding pitch. Since discrete pitch is a characteristic feature of song in comparison to speech across cultures, this suggests that the spectrotemporal structure of song as a vocal signal yields a behavioural outcome—remembering vocal sequences—that differs quantitatively from that elicited by speech intonation. The auditory memory system seems to make different use of these two signal types, or alternatively, there might be two distinct subsystems for memory based on pitch, echoing the distinction that Schulze et al. (2018) proposed for the subvocal rehearsal of verbal material (phonological loop) and tonal material (tonal loop), but on the level of pitch memory itself. In line with this, Zatorre and Baum (2012) suggested a distinction between coarse-grained and fine-grained mechanisms for auditory pitch contour analysis, in which fine-grained mechanisms are used exclusively to perceive musical sounds. Consistent with this hypothesis, cortical representations of pitch have been found to differ between speech and song (Merrill et al., 2012; Tierney et al., 2013). Across our stimulus types the memory advantage for song and intermediate stimuli, relative to speech prosody stimuli, is probably related to fine-grained pitch analysis, because the overall up-and-down movement of the pitch contour was kept constant. This is because each song pitch contour was matched with an equivalent intermediate- and speech-stimulus pitch contour, with only the pitch trajectory within each syllable differing (gliding, intermediate, or discrete, respectively).

If vocal phrases were remembered better when produced with discretised pitches, why do humans not sing all the time? Focusing just on pitch in speech and song, one could argue that speech prosody functions primarily in a context of temporal proximity, for example in a turn-taking context that is more typical of speech than music (Cross et al., 2013). This might make exact reproduction of pitch unnecessary. The use of phonetically distinct words to convey information about concepts, objects etc. and the ability to remember this information, independent of pitch, might further relax the

necessity to remember (gliding) speech prosodic contours exactly. Even in tonal languages, the phonetic information conveyed by pitch rests mainly on pitch contour (e.g., falling, rising) and very rough pitch range (e.g., high, low) differences rather than on exact pitch reproduction.

However, vocal phrases that need to be remembered for longer durations and more precisely might need to rely on another structural layer beyond such coarse-grained intonation contour, namely specific relations among pitches. In fact, there is a long tradition of using sung passages to learn, remember, and pass down verbal information, from the sung works of Homer (initially orally transmitted), to Viking sagas, folk ballads, or religious texts such as the Rig Veda. This use of sung text is both ancient historically, and widespread among human cultures. Indeed indigenous Australians use melodic contours to map landscapes and pass this information orally, and such “songlines” cross language barriers and are at the same time deeply connected to myth and group identity (see e.g., Norris and Harney, 2014). This suggests that discrete pitches have been used to remember vocal phrases across time and cultures. Thus, the emergence of discrete pitches may result from a necessity to remember vocal phrases, making use of absolute and relative pitch perception. If this was the case, it would suggest that discrete pitches emerged as a fundamental feature of music to solve a specific cognitive problem, rather than a by-product of speech prosody [as argued by Pinker (1997)].

In summary, we found that discrete pitches lead to a memory advantage for auditory sequences modelled after vocal phrases. This effect occurred in both more and less musically experienced people, although more musically experienced people showed a better general memory for vocal phrases (independent of gliding or discrete pitch). Our results show that at least one spectrotemporal design feature distinguishing music from spoken language has clear cognitive and behavioural consequences, making it unlikely to be an epiphenomenon of speech prosody. Rather, discrete pitch seems to be a feature enhancing auditory memory. Our results open the door for future studies to further explore the memory advantage of song over speech. Comparative studies involving non-human animals, including additional design features of music, might shed a light on their evolutionary origin.

Additional Information

Data Availability Statement: The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics Statement: The studies involving human participants were reviewed and approved by Ethics Committee of the University of Vienna. The patients/participants provided their written informed consent to participate in this study.

Author Contributions: F.H. and W.T.F. contributed to conception and design of the study. F.H. recruited and tested the participants. Under guidance of C.Q. and W.T.F., F.H. performed the statistical analysis and drafted the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding: This study was supported by the Austrian Science Fund (FWF) number W1262-B29 (DK Grant Cognition & Communication). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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.

Acknowledgments: We like to thank Roger Mundry for helpful statistical advice and for providing the following scripts that he has written: `overdisp.test` to test overdispersion, `glmm.model.stab` to test model stability, `xx.fe.re` to do dummy coding. We thank two reviewers for helpful comments on the manuscript.

Supplementary Material: The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.586723/full#supplementary-material>

References

- Alexander, J. A., Wong, P. C. M., & Bradlow, A. R. (2005). Lexical Tone Perception in Musicians and Non-musicians. In *Ninth European Conference on Speech Communication and Technology*. <https://doi.org/10.1080/0046760860150301>
- Arbib, M. A. (2013). *Language, Music, and the Brain: A Mysterious Relationship*. <https://doi.org/10.7551/mitpress/9780262018104.001.0001>
- Arnold, K. M., & Jusczyk, P. W. (2002). Text-to-tune Alignment in Speech and Song. *Speech and Prosody*, 135–138. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.7946&rep=rep1&type=pdf>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4.
- Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer.
- Bradley, E. (2013). Pitch perception in lexical tone and melody. *Reviews of Research in Human Learning and Music*, 1(May), 1–26. <https://doi.org/10.6022/journal.rhlm.2013002>
- Bregman, A. S. (1994). *Auditory Scene Analysis : The Perceptual Organization of Sound*. Cambridge, Massachusetts: MIT Press.
- Bregman, M. R., Patel, A. D., & Gentner, T. Q. (2016). Songbirds use spectral shape, not pitch, for sound pattern recognition. *Proceedings of the National Academy of Sciences*, 113(6), 1666–1671. <https://doi.org/10.1073/pnas.1515380113>
- Buchsbaum, B. R. (2016). Working Memory and Language. In *Neurobiology of language* (pp. 863–875). Elsevier.
- Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2009). Cortical Processing of Pitch Contours. *Brain and Language*, 108(1), 1–9. <https://doi.org/10.1016/j.bandl.2008.02.001> Relative

- Corrette, R. (2012). Praat vocal toolkit: A praat plugin with automated scripts for voice processing.
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London: J. Murray.
- Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Boca Raton: Chapman & Hall/CRC.
- Fitch, W. T. (2006). The biology and evolution of music : A comparative perspective, *100*, 173–215. <https://doi.org/10.1016/j.cognition.2005.11.009>
- Fitch, W. T., & Jarvis, E. D. (2013). Birdsong and Other Animal Models for Human Speech, Song, and Vocal Learning. In *Language, Music, and the Brain* (pp. 499–540). <https://doi.org/10.7551/mitpress/9780262018104.003.0020>
- Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner’s curse. *Behavioral Ecology and Sociobiology*, *65*(1), 47–55. <https://doi.org/10.1007/s00265-010-1038-5>
- Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., & Heilberger, R. (2011). Package “car”: Companion to applied regression. Thousand Oaks CA: Sage.
- Fox, J., Weisberg, S., Price, B., Friendly, M., Hong, J., Andersen, R., ... R Core Team. (2019). Package ‘effects’.
- Fritz, J., Poeppel, D., Trainor, L., Schlaug, G., Patel, A. D., Peretz, I., ... Parsons, L. M. (2013). The neurobiology of language, speech, and music. In *Music, language, and the brain*. MIT Press.
- Hesse, H.-P. (1982). The judgment of musical intervals. In *Music, Mind, and Brain* (pp. 217–225). Springer.
- Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience*, *15*(5), 673–682. <https://doi.org/10.1162/089892903322307393>
- Kingdom, F. A. A., & Prins, N. (2016a). Adaptive Methods*. *Psychophysics*, 119–148. <https://doi.org/10.1016/B978-0-12-407156-8.00005-0>
- Kingdom, F. A. A., & Prins, N. (2016b). *Psychometric Functions**. *Psychophysics*. <https://doi.org/10.1016/B978-0-12-407156-8.00004-9>
- Krishnan, A., Gandour, J. T., Smalt, C. J., & Gavin, M. (2010). Language-dependent pitch encoding advantage in the brainstem Is Not Limited To Acceleration Rates That Occur in Natural Speech. *Brain and Language*, *114*(3), 193–198. <https://doi.org/10.1016/j.bandl.2010.05.004>. Language-dependent
- Krohn, K. I., Brattico, E., Välimäki, V., & Tervaniemi, M. (2007). NEURAL REPRESENTATIONS OF THE HIERARCHICAL SCALE PITCH STRUCTURE. *Music Perception*, *24*(3), 281–296. Retrieved from <http://caliber.ucpress.net/doi/abs/10.1525/mp.2007.24.3.281>
- Lenth, R. (2016). Least-Squares Means. R Package ‘lsmeans’.
- Levinson, S. C. (2013). Cross-Cultural Universals and Communication Structures. In *Language, Music, and the Brain* (Vol. 10, pp. 67–81). <https://doi.org/10.7551/mitpress/9780262018104.003.0003>
- Lindblom, B., & Sundberg, J. (2014). The Human Voice in Speech and Singing. In T. D. Rossing (Ed.), *Springer Handbook of Acoustics* (pp. 703–746). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-0755-7_16
- Lloyd, E. A., & Gould, S. J. (2017). Exaptation Revisited: Changes Imposed by Evolutionary Psychologists and Behavioral Biologists. *Biological Theory*, *12*(1), 50–65. <https://doi.org/10.1007/s13752-016-0258-y>
- Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns’ Cry Melody Is Shaped by Their Native Language. *Current Biology*, *19*(23), 1994–1997. <https://doi.org/10.1016/j.cub.2009.09.064>

- Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., ... Glowacki, L. (2019). Universality and diversity in human song. *Science (New York, N.Y.)*, 366(6468). <https://doi.org/10.1126/science.aax0868>
- Merrill, J. (2013). *Song and Speech Perception – Evidence from fMRI, Lesion Studies and Musical Disorder*. PhD Thesis, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig.
- Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. D. (2012). Perception of words and pitch patterns in song and speech. *Frontiers in Psychology*, 3(MAR), 1–13. <https://doi.org/10.3389/fpsyg.2012.00076>
- 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). <https://doi.org/10.1371/journal.pone.0089642>
- Nieuwenhuis, R., te Grotenhuis, M., & Pelzer, B. (2012). Influence.ME: Tools for detecting influential data in mixed effects models. *R Journal*, 4(2), 38–47. <https://doi.org/10.32614/rj-2012-011>
- Norris, R. P., & Harney, B. Y. (2014). Songlines and Navigation in Wardaman and other Australian Aboriginal Cultures, 17(2), 1–15. Retrieved from <http://arxiv.org/abs/1404.2361>
- Patel, A. D., & Morgan, E. (2017). Exploring Cognitive Relations Between Prediction in Language and Music. *Cognitive Science*, 41, 303–320. <https://doi.org/10.1111/cogs.12411>
- Paulmann, S. (2016). The neurocognition of prosody. In *Neurobiology of language* (pp. 1109–1120). Elsevier.
- Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice Modulation: A Window into the Origins of Human Vocal Control? *Trends in Cognitive Sciences*, 20(4), 304–318. <https://doi.org/10.1016/j.tics.2016.01.002>
- Poeppel, D. (2011). Genetics and language: A neurobiological perspective on the missing link (-ing hypotheses). *Journal of Neurodevelopmental Disorders*, 3(4), 381–387. <https://doi.org/10.1007/s11689-011-9097-0>
- R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., ... Rupp, A. (2005). Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nature Neuroscience*, 8(9), 1241–1247. <https://doi.org/10.1038/nn1530>
- Schulze, K., Koelsch, S., & Williamson, V. (2018). Auditory working memory. In *Springer handbook of systematic musicology* (pp. 461–472). Springer.
- Seifert, U., Verschure, P. F. M. J., Arbib, M. A., Cohen, A. J., Fogassi, L., Fritz, T. H., ... Rickard, N. (2013). Semantics of internal and external worlds. In *Language, music, and the brain: A mysterious relationship* (pp. 203–232). MIT Press.
- Semal, C., Demany, L., Ueda, K., & Hallé, P. (1996). Speech versus nonspeech in pitch memory. *The Journal of the Acoustical Society of America*, 100(2), 1132–1140. <https://doi.org/10.1121/1.416298>
- Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2020). afex: Analysis of Factorial Experiments.
- Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PLoS ONE*, 12(10), 1–21. <https://doi.org/10.1371/journal.pone.0186773>
- ten Cate, C., & Rowe, C. (2007). Biases in signal evolution: learning makes a difference. *Trends in Ecology and Evolution*, 22(7), 380–387. <https://doi.org/10.1016/j.tree.2007.03.006>
- Tierney, A., Dick, F., Deutsch, D., & Sereno, M. (2013). Speech versus song: Multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex*, 23(2), 249–254. <https://doi.org/10.1093/cercor/bhs003>
- Tillmann, B., Albouy, P., & Caclin, A. (2015). *Congenital amusias*. *Handbook of Clinical Neurology*

- (1st ed., Vol. 129). Elsevier B.V. <https://doi.org/10.1016/B978-0-444-62630-1.00033-0>
- Tong, Y., Gandour, J., Talavage, T., Wong, D., Dziedzic, M., Xu, Y., ... Lowe, M. (2005). Neural circuitry underlying sentence-level linguistic prosody. *NeuroImage*, 28(2), 417–428. <https://doi.org/10.1016/j.neuroimage.2005.06.002>
- Trehub, S. E. (2013). Communication, music and language in infancy. In *Language, Music, and the Brain: A Mysterious Relationship*.
- Ueda, K. (2004). Short-term auditory memory interference: The Deutsch demonstration revisited. *Acoustical Science and Technology*, 25(6), 457–467. <https://doi.org/10.1250/ast.25.457>
- Verhoef, T., Kirby, S., & Padden, C. (2011). Cultural emergence of combinatorial structure in an artificial whistled language. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, (2008), 483–488. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.208.4746>
- Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory and Cognition*, 38(2), 163–175. <https://doi.org/10.3758/MC.38.2.163>
- Winkler, I., Korzyukov, O., Gumenyuk, V., Cowan, N., Linkenkaer-Hansen, K., Ilmoniemi, R. J., ... Näätänen, R. (2002). Temporary and longer term retention of acoustic information. *Psychophysiology*, 39(4), 530–534. <https://doi.org/10.1017/S0048577201393186>
- Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory Distance Perception in Humans : A Summary of Past and Present Research. *Journal of the Acoustical Society of America*, 91, 409–420.
- Zarate, J. M. (2013). The neural control of singing. *Frontiers in Human Neuroscience*, 7(JUN), 1–12. <https://doi.org/10.3389/fnhum.2013.00237>
- Zatorre, R. J., & Baum, S. R. (2012). Musical melody and speech intonation: Singing a different tune. *PLoS Biology*, 10(7), 5. <https://doi.org/10.1371/journal.pbio.1001372>

CHAPTER 3

Spectrotemporal Cues and Attention jointly modulate fMRI Network Topology for Sentence and Melody Perception

Felix Haiduk^{1,2} and Philippe Albouy^{3,4}

¹ Department of Behavioral and Cognitive Biology, University of Vienna, Vienna, Austria

² Department of General Psychology, University of Padua, Padua, Italy

³ International Laboratory for Brain, Music and Sound Research (BRAMS) - CRBLM, Montreal, QC, Canada

⁴ CERVO Brain Research Centre, School of Psychology, Laval University, Quebec, QC, Canada

Submitted (12 September 2023)
to

SCIENTIFIC REPORTS

Abstract

Speech and music are two fundamental modes of human communication. Lateralisation of key processes underlying their perception has been related both to the distinct sensitivity to low-level spectrotemporal acoustic features and to top-down attention. However, the interplay between these bottom-up and top-down processes needs to be clarified. In the present study, we investigated the contribution of acoustics and attention to melodies or sentences to lateralisation in fMRI functional network topology. We used sung speech stimuli selectively filtered in temporal or spectral modulation domains with crossed and balanced verbal and melodic content. Perception of speech decreased only with degradation of temporal information, whereas perception of melodies decreased only with spectral degradation. Applying graph theoretical metrics on fMRI connectivity matrices, we found that local clustering, reflecting functional specialisation, linearly increased when spectral or temporal cues crucial for the task goal were incrementally degraded. These effects occurred in bilateral fronto-temporo-parietal network for processing temporally degraded sentences and in right auditory regions for processing spectrally degraded melodies. In contrast, global topology remained stable across conditions. These findings suggest that lateralisation for speech and music partially depends on an interplay of acoustic cues and task goals under increased attentional demands.

1 Introduction

Spoken language and singing are two fundamental modes of human communication. As vocalised instances of music and language, respectively, they share several structural features such as underlying syntactical operations, the exploitation of the same acoustic dimensions (pitch, time, formants), the vocal apparatus to produce them, or their cultural transmission (Fitch, 2006; Haiduk & Fitch, 2022; Jarvis, 2019; Kirby et al., 2015; Rohrmeier et al., 2015; Tomlinson, 2015). However, song and speech also differ in numerous ways. Notably, the formation of hierarchical tonal organisation (Krumhansl, 2001) in music is typically related to discrete rather than continuous pitch, such as typical in speech (Fitch, 2006; Ozaki et al., 2023).

Recent studies have offered a framework to explain such differences between speech and song by showing that complex sounds can be characterised by the distribution of their spectrotemporal modulation (STM) power (Elhilali, 2019). Over the last decades, it has been demonstrated that i) syllable rate of speech is associated with a high rate of temporal modulation (Poeppel & Assaneo, 2020) resulting in faster temporal modulations as compared to music (Ding et al., 2017; Poeppel & Assaneo, 2020), whereas syllable duration in singing is longer, and enables the production of stable pitch values for a better encoding of tonal relationships (Mantell & Pfordresher, 2013); ii) the complex physiology of phonation typical of singing (i.e. energy in the upper harmonics) is associated with higher spectral modulation rate (Kob et al., 2011; Sundberg, 1989) than speaking; iii) even when spectral modulations are degraded, temporal modulation cues are sufficient for perceiving speech (Shannon et al., 1995); iv) degradation of spectral modulations considerably reduces the perception of melodic content in song, but does not affect speech comprehension, while v) degradation of temporal modulations reduces speech comprehension but has little or no effect on melody perception (Albouy et al., 2020), and vi) spectrotemporal modulation differences between song and speech are found across several cultures (Albouy et al., 2023).

Recent studies have suggested that song and speech are processed in distinct auditory cortical regions because of hemispheric differences in sensitivity to spectral and temporal modulations (Albouy et al., 2020): the right auditory cortex has higher spectral resolution but lower temporal resolution whereas the left auditory cortex has the reverse specialisation (Flinker et al., 2019; Jamison et al., 2006; Schonwiesner et al., 2005; Zatorre & Belin, 2001). According to this framework, music and speech utilise opposite ends of the spectrotemporal continuum (Albouy et al., 2020), explaining their hemispheric lateralisation. However, speech and music also appear to occupy distinct cortical regions beyond mere acoustics (Norman-Haignere et al., 2015, 2022). Indeed, despite spectrotemporal differences being processed by distinct brain regions, top-down factors likely influence the perception of a vocalisation as song or as speech, as illustrated by the “speech-to-song” illusion wherein acoustics

are held constant (Deutsch et al., 2011). Left-hemispheric lateralisation has, for example, been observed when listeners attend to propositionally meaningful features of spoken language (e.g. prosodic intonation, (van der Burght et al., 2019). Auditory speech imagery without any acoustic input has been found to be left-lateralised, depending on emotional valence (Prete et al., 2016, 2020). Acoustically identical stimuli perceived as song rather than speech appear to focus on the right temporal cortex (Hymers et al., 2015). Auditory attention and prediction can interact differently depending on task demands, with partially lateralised effects (Bendixen, 2014; Morillon & Baillet, 2017). Therefore, top-down attention can be expected to interact with spectrotemporal acoustic cues regarding hemispheric lateralisation to perceive speech and song.

In the present study, we asked whether brain asymmetry for speech and music is modulated by both available spectral or temporal acoustic information and task goals (mediated by top-down attention) or by acoustical cues only. We considered two alternative hypotheses: (1) attention to melodies or sentences of sung speech interacts with spectral or temporal acoustic cues to contribute to hemispheric lateralisation (interaction hypothesis), or (2) hemispheric lateralisation depends solely on bottom-up acoustic information, that is spectral or temporal acoustic cues (acoustic hypothesis).

We, therefore, manipulated both the spectral or temporal acoustic information available in sung speech stimuli and the task goal to recognise either the sentence or the melody (for simplicity, we will refer to the sentence aspect as "sentence" or "speech" in this paper, despite lacking prosodic features; Angenstein et al., 2012). Using behavioural and fMRI data from Albouy et al. (2020), we examined the corresponding modulation of brain connectivity patterns manifested in network topology of functional connectivity networks.

Since attention-related effects can be expected both in distributed networks (Lee et al., 2014) and single hub regions that are integrated with surrounding networks (e.g. Chennu et al., 2013), we used graph theoretical metrics (Fornito et al., 2016; Watts & Strogatz, 1998). Graph theory investigates the topology of a network, that is, the way single regions are connected, revealing how networks subserve functional dynamics (e.g. integration or segregation). It can characterise several topological properties at global (the entire network), medium (subnetworks) and local (hub regions) levels of networks and is therefore well suited for functional connectivity analyses.

Topological network properties are modulated dynamically to cope with different listening situations (Alavash et al., 2019; Quante et al., 2018). For spatial attention to two competing speech streams, Alavash et al. (2019) observed a dynamical reconfiguration of brain network connectivity, altering functional integration and segregation relative to resting-state, in auditory, ventral attention, and cingulo-opercular regions. Integration and segregation were measured by the graph metrics "clustering

coefficient", "efficiency", and the network modularity. Specifically, increased modular segregation (as revealed by the graph theoretical modularity index Q) in an auditory fronto-temporal control network predicted individuals' listening success in this study. Along with modular segregation, Alavash et al. (2019) reported an increase in local but a decrease in global efficiency with more challenging speech comprehension relative to resting-state. These measures are related to localised (subnetwork) or global (entire network) integration of information, respectively (Fornito et al., 2016). Similarly, Quante et al. (2018) observed increased clustering and efficiency in auditory modules when attending to auditory (piano-like chords) instead of visual stimuli in an auditory-visual task. Alavash et al. (2019) suggested that functional segregation was a general principle of auditory attention. If attention was directed either to sentences or to melodies in sung speech stimuli, such segregation might thus be lateralised between song and speech.

We, therefore, predicted the following for the present study: If lateralisation for speech vs music depended on the interaction of task goal and affordances (interaction hypotheses), that is of attention to melody or sentence and acoustic cues present, we would expect to find a hemispheric difference in modular segregation, scaling under attentional demands. We also predicted to observe increased local clustering and local efficiency but decreased global efficiency when task-relevant acoustical cues are degraded. We expected these effects to be lateralised (left for attention to sentences and temporal degradation, right for attention to melodies and spectral degradation) in auditory, cingulo-opercular and/or ventral attention networks. If lateralisation for speech vs music depended only on acoustical cues (acoustic hypothesis), we would expect lateralisation to occur independent of task goals and attention, solely based on acoustical cues.

2 Results

The behavioural and fMRI data used in the present study are a subset of the data presented in Albouy et al. (2020) (15/27 French-speaking individuals who performed the behavioural task and the fMRI session). In this previous study, they created a unique stimulus set in which ten original sentences were crossed with ten original melodies, resulting in 100 naturalistic acappella songs (Fig. 1, stimuli available at: <https://osf.io/9ub78/>). This orthogonalisation of speech and melodic domains across stimuli allows for the dissociation of speech- (or melody-) specific from non-specific acoustic features, thereby controlling for any potential acoustic bias (Mcgettigan & Scott, 2012). They created two separate stimulus sets, with French and English sentences. In the present study, data recorded with the French set only have been used. They then parametrically degraded each stimulus selectively in either the temporal or spectral dimension using a manipulation that decomposes the acoustical signal using the Spectrotemporal Manipulation framework (Elliott & Theunissen, 2009).

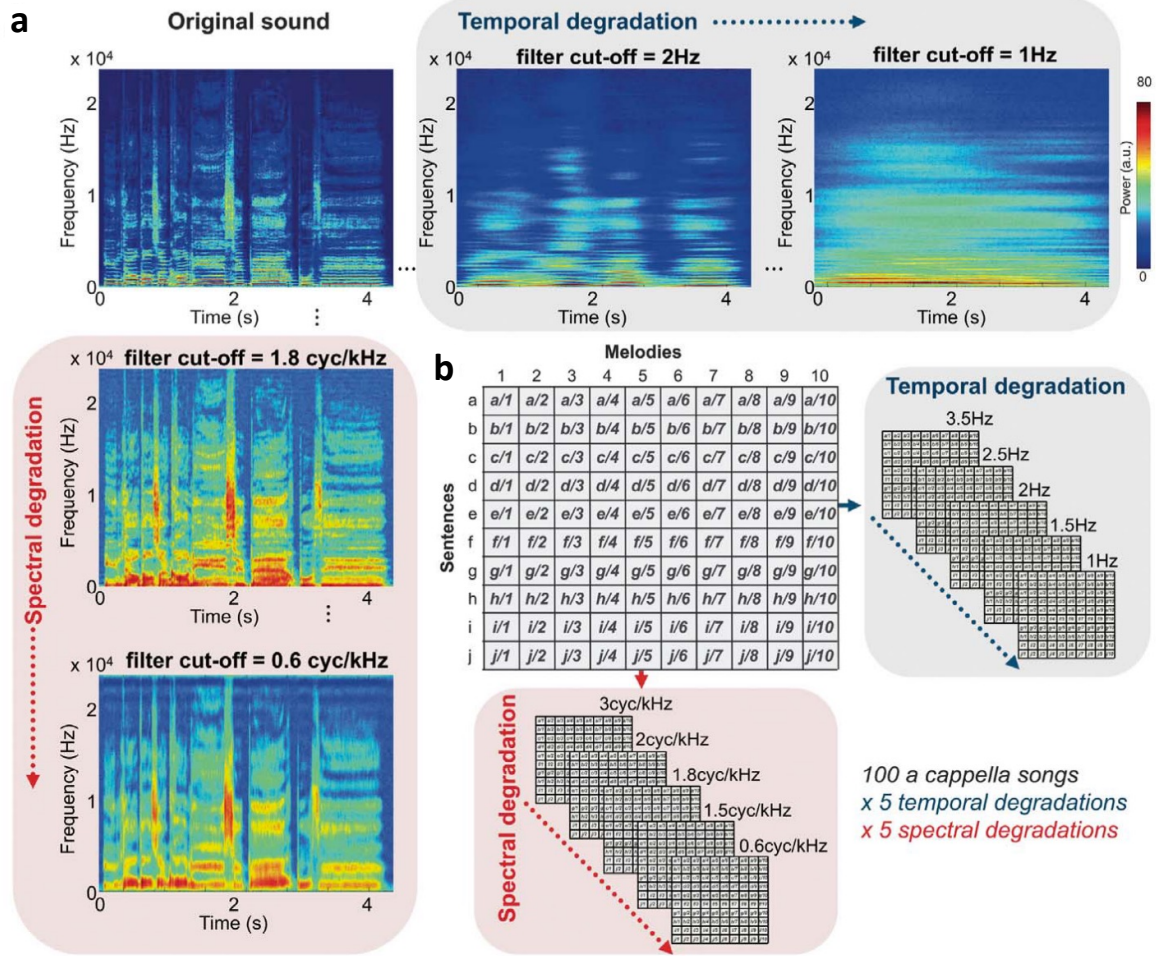


Figure 1: Spectrotemporal filtering and stimulus set. **a.** Spectral and temporal degradations applied on an original a cappella song. **b:** One hundred a cappella songs in each language were recorded following a 10x10 matrix with ten melodies (number code) and ten sentences (letter code). Stimuli were then filtered either in the spectral or in the temporal dimension, with 5 filter cut-offs, resulting in 1000 degraded stimuli for each language (taken with permission from Albouy et al., 2020).

2.1 Behavioural results

We first aimed to confirm the importance of spectrotemporal rates on sentence or melody recognition scores (Fig. 1A) in the subset of native French-speaking ($n = 15$) listeners from Albouy et al. (2020) (who performed both the behavioral and fMRI experiment). Participants were presented with pairs of stimuli and asked to discriminate either the speech or the melodic content. Thus, the stimulus set across the two tasks was identical, only the instructions differed. To manipulate the amount of spectral or temporal acoustic information, stimuli were degraded in the modulation power spectrum domain either temporally or spectrally with 5 cut-off values each (see Fig. 2 and Methods), leading to different amounts of acoustical information retained. For each participant, the modulation of performance between degraded and original (baseline, non-degraded) stimuli was computed as follows:

$$\text{normalized score} = \frac{(\text{raw score} - \text{chance})}{(\text{baseline} - \text{chance})} - 1$$

with chance corresponding to 50 % of performance. Baseline score corresponds to performance obtained with the original songs (without degradations) for each participant. A normalised score of 0 indicates no decrease in performance as compared to baseline, while positive and negative scores, respectively, indicate beneficial and deleterious effects on accuracy. This strategy was chosen to consider individual differences in overall melody or sentence recognition (without degradation). Behavioural results were computed based on these normalised deviation scores from the baseline, integrated over trials using a linear function for each participant (see methods).

We fitted a linear model with the interaction of "cutoff" (1 to 5), "task" (sentences or melodies), and "degradation" (temporal or spectral) as predictors. Comparing the full model with a reduced model lacking the predictors of interest revealed a significant interaction ($F_{13, 280}=22.154$, $p<.001$, $R^2_{\text{adj}}=0.555$) with behavioural performance decreasing as a function of degradation, task, and cutoff (see Fig. 1B and Supplementary Material for detailed statistics). More precisely, the degradation of information in the temporal dimension impaired sentence recognition but not melody recognition, while the degradation of information in the spectral dimension impaired melody recognition but not sentence recognition. No collinearity issues were found.

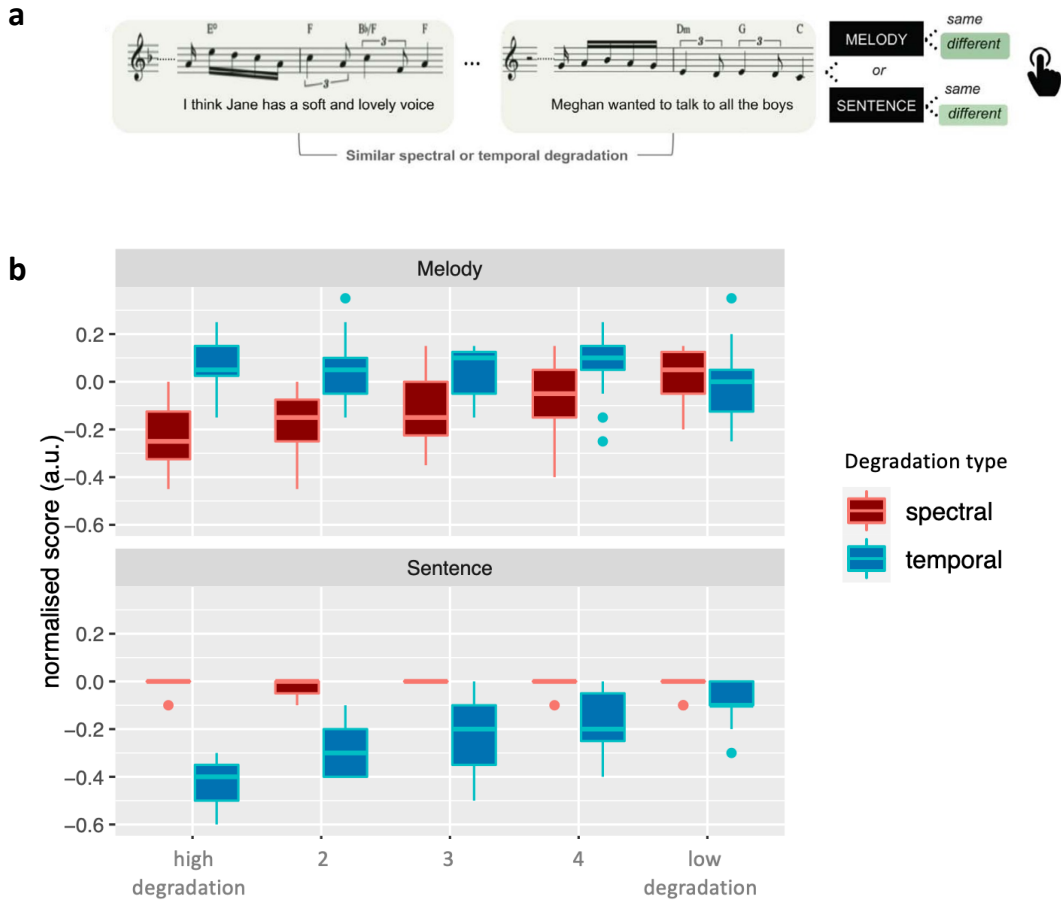


Figure 2: Behavioural task. **a.** 15 French-speaking participants listened to degraded (either in the spectral or temporal dimension) a cappella songs presented in pairs. After the second song, a visual instruction indicated the domain of interest (sentences or melodies). The lower panel shows example trials (adapted from Albouy et al., 2020, with permissions). **b.** Behavioural performance of 15 French-speaking listeners who participated in both behavioural and fMRI experiments presented as a function of task instructions (judging melodies or sentences) and degradation cut-off values. Blue bars: Temporal degradations; Red bars: Spectral degradations. Boxplot whiskers indicate first quartile - 1.5* interquartile range and third quartile + 1.5 interquartile range, respectively.

2.2 fMRI results

Blood-oxygenation level-dependent (BOLD) activity was recorded while the 15 French speakers who had participated in the behavioural experiment listened to blocks of five songs degraded either in the temporal or spectral dimension. Participants attended to either the speech or the melodic content (Fig. 3A). More precisely, participants were asked to attend either to the melody or the sentence of a stimulus, visually cued before each of 110 blocks of five degraded stimuli, resulting in a 2 (attention) x 2 (degradation type) x 5 (degradation strength) design. The participant's task was to detect repetitions of sentences or melodies (1-back task). We will abbreviate the four attention x degradation conditions

as follows throughout the paper: MS: attention to melody, spectral degradation (temporal acoustic cues retained), MT: attention to melody, temporal degradation (spectral acoustic cues retained), SS: attention to sentences, spectral degradation (temporal acoustic cues retained), ST: attention to sentences, temporal degradation (spectral acoustic cues retained).

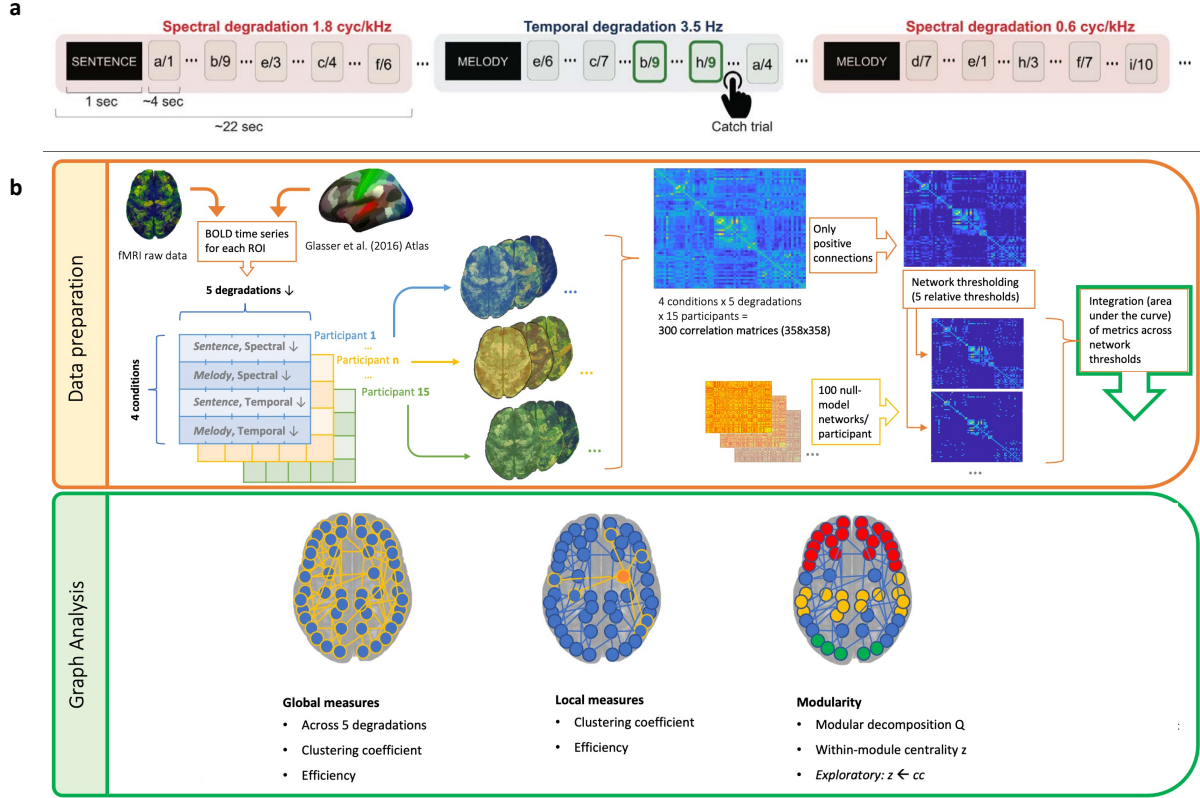


Figure 3: fMRI task and analysis. **a:** fMRI design: BOLD activity was collected while participants listened to blocks of 5 sung-speech stimuli (degraded in the temporal or spectral dimension). To modulate attention, participants were asked to detect two catch trials (with the high filter cut-off (low degradation): 3cyc/kHz, and 3.5 Hz) containing melody (or sentence) repetition (1-back task) (taken from Albouy et al., 2020, with permission). **b:** Overview of data preprocessing to enter the graph theoretical analyses (upper panel) and the different graph theoretical analyses (lower panel). We created ROI-to-ROI correlation matrices (Fisher Z-transformed) as a function of attention (sentences/melodies), degradation type (spectral/temporal) and cutoff value (5 steps per degradation type) for each of the 15 participants. This resulted in 300 358x358 functional connectivity matrices (20 per participant, see Table 1). We then applied several analyses using measures from Graph Theory, a field that investigates networks regarding their topological organisation. It models brain connectivity networks as "nodes" (ROI) connected by "links" (or "edges"), which in our case were the Fisher z-transformed correlations across ROIs. Graph measures can be calculated regarding the entire functional connectivity network (global measures) and subnetworks or single nodes (local measures).

In order to investigate how connectivity patterns changed depending on spectrotemporal acoustic cues and attention to speech or melody, we calculated complementary graph theoretical measures using the GraphVar toolbox (Kruschwitz et al., 2015; Waller et al., 2018) in Matlab. This allowed us to measure the entire network's behaviour, the differential emergence of subnetworks, and the functional role of single ROI in a statistically rigorous way. We thresholded the 358x358 connectivity matrices (see Fig. 3 and Methods, consisting of Fisher z scores derived from correlation coefficients between ROI) for each subject to exclude spurious or noisy weights by using several thresholds to prevent measures from being biased (see Methods). We conducted several confirmatory analyses based on Graph measures (see below) and an exploratory analysis (network-based statistic, see Supplementary Material).

We calculated the clustering coefficient and the efficiency both for the entire network (global) and for each ROI (local). Since we applied multiple thresholds to the connectivity matrices, we used the area under the curve across all thresholds to obtain an integrated measure of global and local clustering coefficient and efficiency, respectively (Bassett et al., 2006, 2012; Fornito et al., 2016). For simplification, we will further refer to these integrated measures as local and global clustering coefficient and efficiency, respectively. We also analysed network modularity as a further global metric.

We used Generalised Linear Mixed Models (GLMM), including the fixed factors "attention" (speech or melody), "degradation" (spectral or temporal degradation), "cutoff" (5 cutoffs from high to low degradation) and their interactions. We also included participants as random effects and random slopes of all fixed effects within participants. All statistical analyses were done in R (R Core Team, 2019).

2.2.1 Global graph measures

2.2.1.1 Global clustering coefficient

The clustering coefficient measures what fraction of the neighbours of a given node in a network are also connected among each other, forming closed triangular patterns (motifs). In our case, neighbours are ROI with connectivity retained after thresholding the entire connectivity matrices. The global clustering coefficient measures the mean clustering coefficient across all ROI. The clustering coefficient can be interpreted as integration within triangular node motifs, which could support functional specialisation (it does, however, not directly measure segregation; see Fornito et al., 2016).

We calculated Variance Inflation Factors (Field, 2005) to assess collinearity but did not observe any collinearity issues (R package "car"; version 3.0-2; Fox et al., 2011). Model stability was assessed

by comparing the full model estimates with estimates obtained from reduced models with levels of the random effects excluded one at a time. The model was stable regarding fixed effects estimates (regarding random effects estimates were not as stable; however, random effects were not the focus of these models). We obtained 0.95 confidence intervals using the function “bootMer” (package “lme4”, version 1.1-21; Bates et al., 2015) with 1,000 parametric bootstraps (see Supplementary Material).

The comparison of the full model with a null model devoid of fixed effects terms was significant ($\chi^2 = 41.358$, $df = 19$, $p = .002$, $R^2m = .01$, $R^2c = .297$). Thus, the predictors of interest explained the data significantly better than the null model.

We then tested the effects of the predictors using likelihood-ratio tests, comparing the full model with models reduced by the predictors of interest (R function “drop1” with argument “test” set to “Chisq”). The three-way interaction "attention:degradation:cutoff" was significant (LRT = 13.124, AIC = -722.523, $p = 0.01$, $R^2m = .01$, $R^2c = .297$). Thus, lower-order effects (two-way interactions and main effects) were conditional on each other and not meaningful to interpret independently.

To test which cutoffs drove the interaction "attention:degradation", we fitted GLMMs for each cutoff separately and compared the resulting full (fixed effects "attention" by "degradation", random effect "participant", random slopes of fixed effects within "participant") and null models as post-hoc tests. However, we did not obtain a clear pattern across cutoffs (see Supplementary Fig. S1).

2.2.1.2 Global efficiency

Global efficiency is the inverse of the average shortest path length of the network. Both path length/efficiency and clustering are two fundamental network characteristics (Watts & Strogatz, 1998). Efficiency can be interpreted as measuring communication facility between nodes and therefore as a measure of integration. We used the same model structure and diagnostic functions as for the global clustering coefficient. Model stability was acceptable in terms of fixed effects (regarding random effects estimates were not as stable, however, random effects were not the focus of these models). We did not obtain any collinearity issues. The full model explained the data significantly better than the null model ($\chi^2 = 39.173$, $df = 19$, $p = .004$, $R^2m = .109$, $R^2c = .231$). As for the global clustering coefficient, the three-way interaction "attention:degradation:cutoff" was significant (LRT = 17.514, AIC = -2121.311, $p = 0.002$, $R^2m = .109$, $R^2c = .231$). As for the global clustering coefficient we fitted separate models with the "attention:degradation" interaction as fixed effects (random effect "participant", random slopes of fixed effects within "participant") for each cutoff and ran full-null model comparisons as post-hoc tests. However, we did not obtain a clear pattern across cutoffs (see Supplementary Fig. S2).

2.2.1.3 Modularity

We investigated whether segregation in network modularity increased as a function of attention to melody or sentence and spectral or temporal degradation across cutoffs. Using GraphVar, we ran 1000 iterations of the Louvain algorithm for each condition, cutoff and participant, with a resolution factor gamma of 1.0 (standard value), to derive the subject-level community structure and the metrics Q. The modularity index Q designates the optimal modular decomposition of the network, and a higher Q value indicates more functional segregation (see Alavash et al., 2019).

We used the Q values as response variable in a GLMM with the fixed factors "attention", "degradation", and "cutoff". We included the three-way interaction of these factors, as well as "participant" as random factor and random slopes of all fixed factors within "participant" (i.e., the same model structure as for Clustering coefficient and Efficiency). The model diagnostics revealed no collinearity issue and acceptable model stability regarding fixed effects (the model was not very stable regarding random effects; however, random effects were not the focus of this model). The full-null model comparison remained not significant ($\chi^2 = 15.632$, $df = 19$, $p = .682$). Thus, although some ROI changed their affiliation with given modules by condition (see OSF repository <https://osf.io/merwk/>), the results indicate no significant difference in the overall modular segregation across conditions and cutoffs. We additionally analysed whether any of the ROI would change in their within-module centrality, however we did not observe any significant changes (see Supplementary Material).

2.2.2 Local graph measures

2.2.2.1 Local clustering coefficient

Next, we calculated the clustering coefficient for each ROI, indicating the integration of each ROI with its neighbourhood in closed-triangle motifs. We used the same model structure as for the global clustering coefficient, fitting a model for each of the 358 ROI. After FDR correction, full-null model comparisons of models for 5 ROI were significant (see Fig. 4, Supplementary Table S4). These ROI form a bilateral network comprising left medial prefrontal (l8BM), left hippocampal (lH), and right auditory associative regions (rA5, rSTSvp), and right superior parietal (rAIP) region. We then conducted likelihood ratio tests for the three- and two-way interactions (see Table 1).

Table 1: Likelihood-ratio tests for ROI with significant full-null model comparisons for the local clustering coefficient. Highest order effects that are significant are in bold.

ROI	effect	Sum Sq	Mean Sq	NumDF	DenDF	F	Pr(>F)
l8BM	attention:degradation:cutoff	0.003	0.001	4.000	263.361	3.753	0.005
IH	attention:degradation:cutoff	0.003	0.001	4.000	203.825	1.996	0.096
	cutoff:attention	0.004	0.001	4.000	226.328	2.564	0.039
	cutoff:degradation	0.007	0.002	4.000	226.328	4.808	0.001
	attention:degradation	0.000	0.000	1.000	226.328	1.012	0.316
rA5	attention:degradation:cutoff	0.005	0.001	4.000	239.211	6.075	0.000
rAIP	attention:degradation:cutoff	0.004	0.001	4.000	224.421	3.196	0.014
rSTSvp	attention:degradation:cutoff	0.003	0.001	4.000	240.218	3.620	0.007

To see whether for any of the conditions, there was an effect of degradation intensity (cutoff) at any of the five ROI, we conducted post-hoc full-null model comparisons for each ROI and condition separately, with "cutoff" as fixed and "participant" as random effect in the full model. FDR-corrected results revealed that for ST in all five ROI, thus bilaterally, the local clustering coefficient increased significantly with higher degradation, while for MS this increase occurred in the right-hemispheric ROI rA5 and rSTSvp. At rAIP and for MT, local clustering decreased across cutoffs (see Table 2 and Fig.'s 4 and 5).

Table 2: Results for post-hoc full-null model comparisons regarding the effect of degradation intensity (cutoff) for the five ROIs of the local clustering coefficient subnetwork.

ROI	condition	df	intercept	slope	χ^2	p (FDR corrected)	R ² m	R ² c
l8BM	ST	4	0.061	-4.97E-03	25.938	0.001	0.275	0.325
IH	ST	4	0.064	-5.77E-03	20.595	0.003	0.223	0.280
rA5	MS	4	0.055	-4.20E-03	13.110	0.031	0.136	0.265
rA5	ST	4	0.060	-4.07E-03	12.337	0.038	0.112	0.354
rAIP	MT	4	0.049	9.65E-04	17.196	0.007	0.162	0.356
rAIP	ST	4	0.066	-4.38E-03	16.623	0.008	0.171	0.293
rSTSvp	MS	4	0.059	-4.70E-03	17.540	0.007	0.197	0.228
rSTSvp	ST	4	0.060	-4.64E-03	22.670	0.001	0.250	0.250

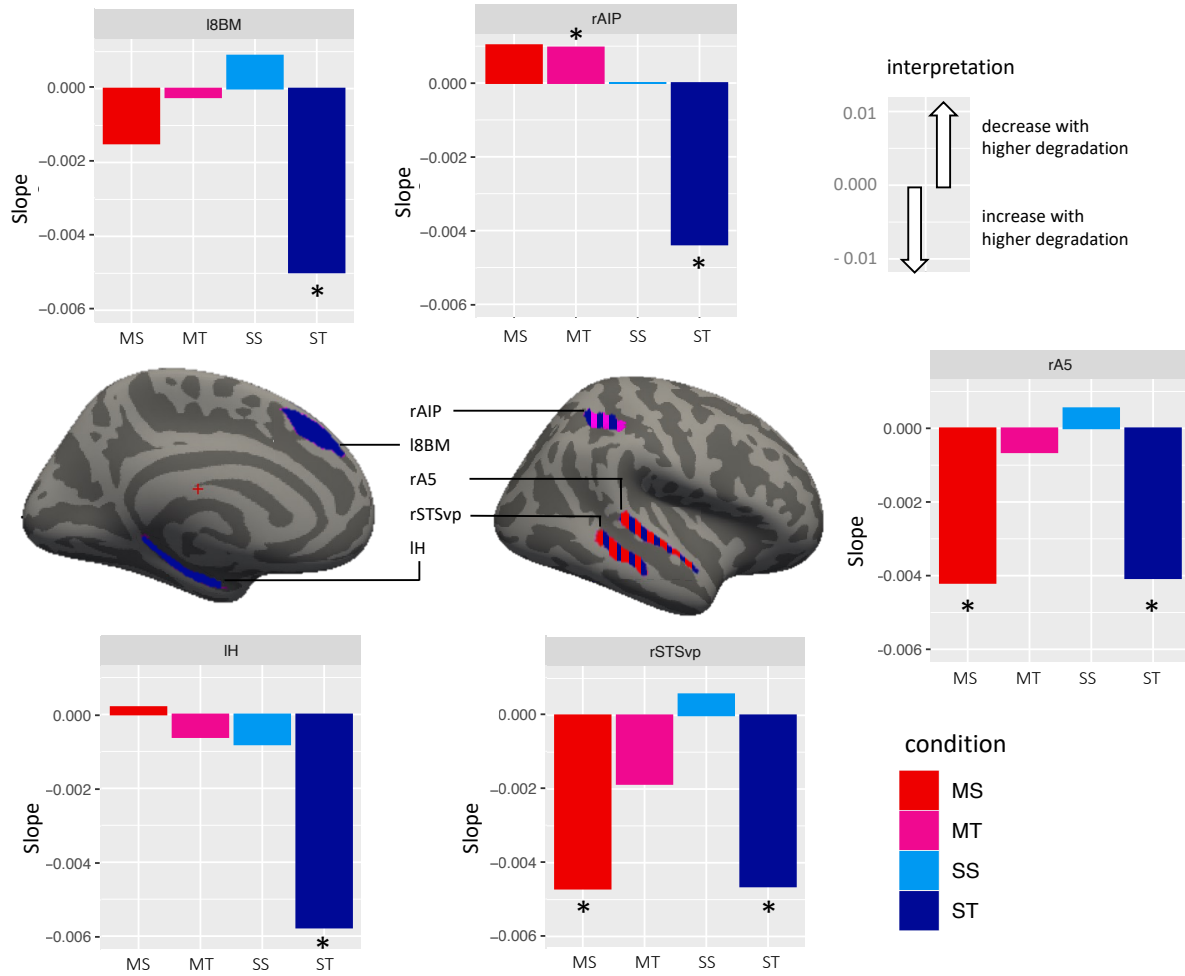


Figure 4: ROI with significant full-null model comparisons for the local clustering coefficient, and their respective model slope values (see Table 3) for each condition. Negative slopes indicate a decrease of local clustering with lower degradation intensity (cutoff), or likewise an increase with higher degradation intensity. Significant effects for degradation intensity occurred for ST bilaterally, thus in all the five ROI of this subnetwork, and for MS for right STSvp and right A5 (see also Fig. 5). Asterisks indicate significance. Brain regions are colored according to significant effects. Brain images throughout the paper have been done with FreeView (version 3.0) of the FreeSurfer package (Fischl, 2012).

Explaining Spectrotemporal Differences between Music and Language

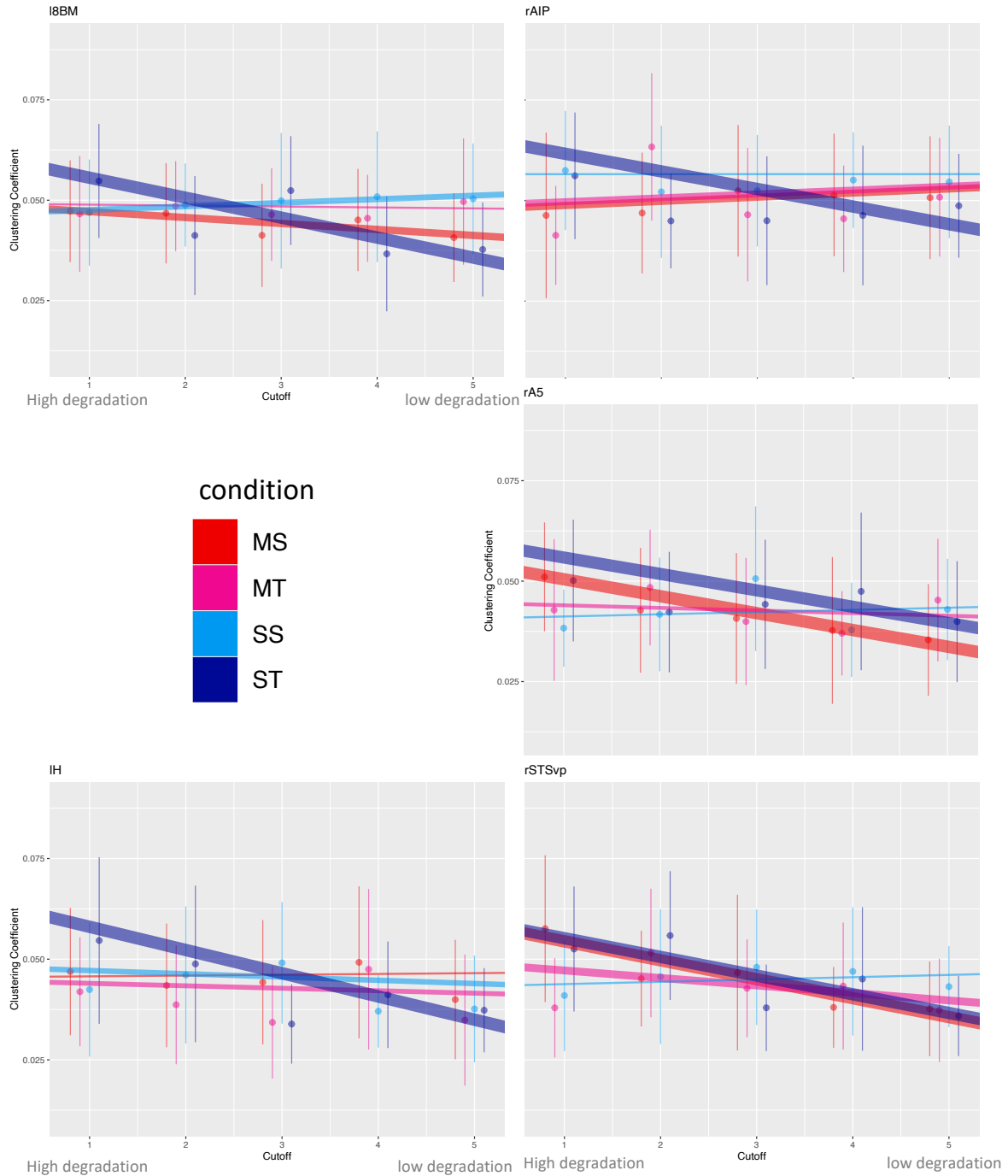


Figure 5: Mean, SD and model slopes for each significant region for the local clustering coefficient network, across cutoffs (high to low degradation). More extreme slope values are depicted with higher line thickness.

We investigated whether ROI could be specifically associated with sentence, melody, temporal or spectral processing by conducting post-hoc likelihood-ratio tests at each cutoff but did not obtain clear results regarding our hypotheses (see Supplementary Material).

2.2.2.2 Local efficiency

For the local efficiency, indicating the inverse of the shortest path between two ROI, we used the same model structure as for the global efficiency, fitting one model for each of the 358 ROI. For 264 ROI the full-null model comparison was significant (FDR corrected, see Supplementary Material). All the ROI in the fronto-parietal-temporal subnetwork found for the local clustering coefficient were also significant for the local efficiency. Equivalently to the analysis for the local clustering coefficient subnetwork, we investigated the effect of degradation intensity by conducting post-hoc full-null model comparisons for each ROI and condition separately, with "cutoff" as fixed and "participant" as random effect in the full model. FDR-corrected results revealed that local efficiency increased with higher degradation across various ROI in both hemispheres for all conditions but predominantly for MS and ST (see Fig. 6). A decrease with higher degradation was obtained only for a few ROI (and not for ST) (see Supplementary Material).

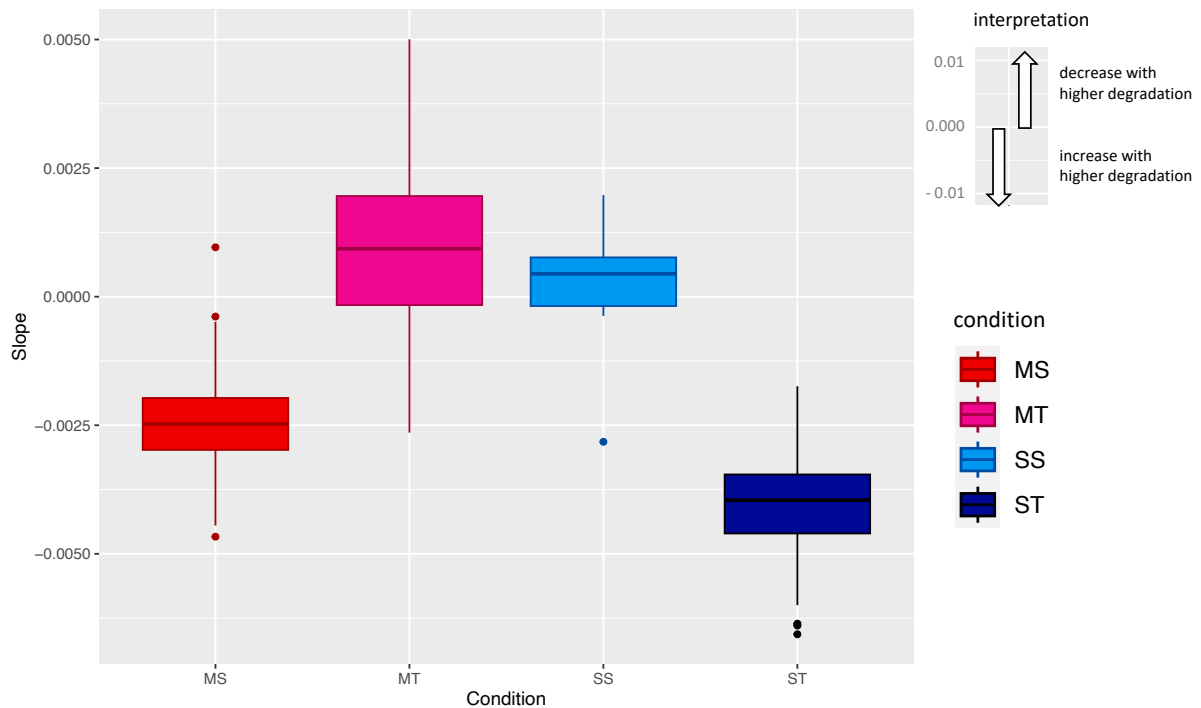


Figure 6: Model slope values across significant ROI for local efficiency. Negative values indicate a decrease of efficiency with lower degradation intensity (cutoff) or, equivalently, an increase of efficiency with higher degradation intensity, as can be obtained clearly for MS and ST. Of all ROI with significant changes in local efficiency across cutoffs, most showed effects for MS (107 ROI increasing with higher degradation, 1 decreasing with higher degradation) and ST (213, 0). In contrast, only few ROI showed significant effects for SS (2, 4) and MT (3, 8). Boxplot whiskers indicate first quartile - 1.5* interquartile range and third quartile + 1.5 interquartile range, respectively.

To see which ROI could be associated with processing of melodies, sentences, spectral or temporal acoustic information, we conducted post-hoc likelihood-ratio tests at each cutoff. Overall, we did not find a clearly localisable pattern (see Supplementary Material).

3 Discussion

We investigated whether task goals mediated by top-down attentional effects (i.e., attending to melodies or sentences of sung speech stimuli), complement bottom-up acoustic effects to contribute to hemispheric lateralisation for song or speech perception. To do so, we applied graph theoretical metrics on fMRI connectivity matrices extracted when participants attended to melodies or sentences of sung speech stimuli degraded in spectral or temporal dimensions. We had two alternative hypotheses: (1) that attention to melodies or sentences of sung speech would interact with spectral or temporal acoustic cues to contribute to hemispheric lateralisation (interaction hypothesis), and (2) that hemispheric lateralisation depends solely on bottom-up acoustic information, that is spectral or temporal acoustic cues (acoustic hypothesis). Regarding our predictions, the results in our planned analyses partially supported the interaction hypothesis, namely at the subnetwork level.

Using the behavioural data from a subset of participants of (Albouy et al., 2020), we first confirmed that perception of melodic content is most affected by degradation in the spectral dimension, while perception of speech content is most affected by degradation of information in the temporal dimension. These results align well with seminal studies (Flinker et al., 2019; Shannon et al., 1995) on the robustness of speech comprehension to spectral degradation and vice-versa.

For fMRI connectivity matrices, global graph theoretical metrics did not indicate clear evidence for lateralised functional segregation either under the form of an interaction of attention to melody or sentence and acoustic cues (interaction hypothesis) or as a function of acoustic cues alone (acoustic hypothesis). We predicted to observe lateralised segregation differences with different task goals and thus attentional focus. We did not observe a clear pattern regarding our hypotheses in global efficiency or global clustering coefficient. Also, the overall modular organisation into a motor, an occipital, and one or two fronto-temporal-parietal modules remained across conditions (see Supplementary Table S7). Changes in module affiliation between cutoffs and conditions did not significantly affect the modularity index Q . Thus, we cannot conclude that segregation in modularity increases with enhanced attentional demands when compared among different listening tasks (rather than compared with resting state as in Alavash et al. (2019)). We also did not obtain any significant change in the within-module centrality of any ROI, suggesting that their roles as hubs within their modules stayed the same across conditions. This suggests that differential reconfiguration of functional connectivity might instead be found below the level of modules.

Indeed, we found evidence for differential segregation regarding the local clustering coefficient. We obtained a network of five regions comprising left medial prefrontal (l8BM), left hippocampal (lH), right auditory association (rA5, rSTSvp), and right superior parietal (rAIP) regions, roughly reflecting the whole-brain modular organisation we observed (except for the occipital module). In contrast to our expectations, we obtained significant ROI only in auditory but not in cingulo-opercular or ventral attention regions in this analysis. Within the obtained subnetwork, all ROI showed an increase of local clustering with higher degradations for speech with temporal degradation (ST), thus in the two hemispheres, while an increase for MS occurred for right lateralised auditory regions (rA5 and rSTSvp). Thus, we obtained partially lateralised effects for the conditions where the crucial acoustic information was degraded, partially in line with the interaction hypothesis. Additionally, segregation, as indexed by the local clustering coefficient, increased with higher degradation, possibly reflecting higher attentional demands, as predicted. This means that neighbouring nodes of these ROI in the network increased in their closed-triangle motif structure with the ROI of these networks, leading to tight integration with these ROI, supporting functional specialisation for the task at hand (Fornito et al., 2016).

The left hemispheric regions lH and l8BM showed higher clustering with increased degradation of temporal cues and attention to sentences (ST). lH has previously (without using Glasser et al.'s [2016] parcellation) been associated with semantic speech processing, but also emotional valence (Bartha et al., 2003; Koelsch et al., 2006; Mueller et al., 2011; Rodd et al., 2005; Trost et al., 2012). l8BM has previously (without using Glasser et al.'s, [2016] parcellation) been associated with speech perception and production and musical familiarity and expectation (Foster et al., 2013; Geranmayeh et al., 2014; Kung et al., 2013; McNealy et al., 2006; Rothermich & Kotz, 2013; Seger et al., 2013; Wallmark et al., 2018). Despite associations with both music and speech processing in these areas, the topological reconfiguration we observed regarding clustering occurred with attention to sentences only. However, in Briggs et al.'s (2018) assignment of Glasser et al.'s (2016) parcellation to large-scale brain networks, l8BM has been assigned to the semantic language network, which is more in line with our results. Interestingly, this semantic language network also comprises the left hemispheric homologues of our obtained rSTSvp and rAIP. The left AIP is additionally also part of the dorsal attention network, while the right AIP has previously been associated with motor imagery, co-speech gesture and finger movement (Garbarini et al., 2014; Holle et al., 2008; Jonas et al., 2007; Péran et al., 2010; Sadato et al., 1997). Although music perception is often associated with motor processes (Zatorre et al., 2007), we found clustering effects in rAIP to be significant for ST but not MS. In contrast, clustering effects for rSTSvp were significant for both MS and ST. Along with rA5 (part of the auditory network in Baker et al., 2018), rSTSvp has previously been associated with music and speech syntax, perception and prosodic emotion (Angulo-Perkins et al., 2014; Bianco et al., 2016; Hesling et al., 2005; Humphries et al., 2014; Kyong et al., 2014; Park et al., 2015; Rogalsky et al., 2011; Sammler et al., 2013). Given that homologue areas in brain hemispheres can show similar, but not identical functions (e.g., Albouy et al.,

2020; Sammler et al., 2015), we propose that rAIP and rSTSvp can be considered as homologues to parts of the left-hemispheric semantic language network are related to a judgement of non-semantic meaningfulness, that in speech might be related to prosodic perception and in music to melody perception. In contrast, the left hemispheric l8BM might be related to processing meaningful propositional information. Since the task goal affords meaningfulness to aspects of the stimuli, these regions may change topographically when the task goal and, thus the meaning (semantic or not) is hard to obtain due to relevant acoustic cues being degraded.

Complementary, we observed an increase in local efficiency with higher degradation in distributed ROI bilaterally, but again mainly for MS and ST (see Fig. 6), thus, when spectrotemporal cues were crucial to solving the task and with increasing attentional demands. These changes included the ROI of the clustering coefficient subnetwork. If we interpret changes of efficiency resulting from topological reconfiguration in clustering, these results could mean that ROI that reconfigure their clustering act as a relay station from primary sensory areas to a wider functional network, distributing and integrating this information and recruiting attentional resources. These findings at the subnetwork level reflect the behavioural results.

Thus, we did find significant topographical reconfiguration partially supporting the interaction hypothesis at the subnetwork level but not at the global level, nor regarding the role of ROIs as hubs within their modules. It is possible that the different measures we used index not only different granularity levels but also different processing stages. Previous literature suggests lateralisation based on spectrotemporal acoustical cues (Albouy et al., 2020; Norman-Haignere et al., 2015) (sentences with temporal acoustical cues left, melodies with spectral acoustical cues right). However, Norman-Haignere et al. (2015) also obtained dedicated areas for speech and music beyond acoustic processing but in bilateral secondary auditory areas. The authors suggest that the speech-specific areas reflect early phoneme and syllable processing before the stage of speech intelligibility. Our results suggest that subsequently, when recognising melodies or sentences (thus beyond early phoneme processing), differential attentional demands partially recover lateralisation through reconfiguration of motif structure when task-relevant acoustic information is degraded. This reconfiguration influences efficiency in a wide bilateral network. Finally, the general modular organisation of the entire network and global clustering and efficiency were not significantly changed, and their stability might reflect late task-related processing remaining stable across conditions, such as decision-making for meaningful auditory stimuli.

Our study takes on a particular perspective on the problem of lateralisation for song and speech due to stimulus and task design, which should be complemented in future studies. First, our participants were presented with sung speech stimuli, a common but specific hybrid form of music and language.

These stimuli have the advantage of carrying all necessary features to understand melodies and sentences, thus keeping the acoustics constant and changing only the attentional focus. However, they lack features typical for spoken language, especially regarding prosodic features like intonation, the usual variability in syllable frequency and overall tempo, or the pragmatic conversational exchange in turn-taking. Sung music with lyrics is likely perceived as a typical instance of music rather than speech, along with invoking possible expectations and modes of perception (see e.g. Weidema et al., 2016), even if the attentional focus is on sentences. Even when spectral information is filtered out from the stimuli, the resulting percept likely reminds of harsh whispering, which is also no typical mode of spoken conversation, although not uncommon. The specificity of melodic vocalisations as a proxy for musical perception is less severe since sung music with lyrics is cross-culturally familiar from early childhood throughout the lifespan. Moreover, vocalisations likely were a focal trait for evolutionary niche constructive effects (Nishimura et al., 2022; Tomlinson, 2015). Thus, even if sung speech is common, it might not be generalisable to spoken language. We therefore recommend future studies to use effects like the song-speech illusion. However, a possible drawback would be the need to repeat a stimulus to sound like song, which might introduce other unwanted effects qua repetition (see Tierney et al., 2018).

Second, the fact that melodies or sentences need to be recognised, given that the crucial acoustic information is hard or impossible to perceive, is unusual. However, it might be compared to speech- or music-in-noise (see McDermott, 2009), and be, in that sense, a specific case of a general problem the auditory system faces. Thus, although the current tasks were quite specific, the lateralised changes in topology with increasing attentional demands might be generalisable, somewhat in line with the findings of Alavash et al. (2019). We recommend that future studies follow this line of research further.

Third, it might be objected that there are differences in how sentences and melodies are memorised, especially concerning the presence of non-crucial acoustic information (i.e., sentences without melodies and melodies without words). Spoken language is likely more abundant in everyday life than sung music for our participants (Western non-musicians), thus, language might be easier to remember for our participants. This is visible in the behavioural data, which, although remaining constant in both cases, are more variable for melodies with temporal degradation than for sentences with spectral degradation. On the other hand, discrete pitch, as in melodies, might enhance memory for sound sequences (Haiduk et al., 2020). Also, even if underlying cognitive processes differ between working memory for sentences and melodies (Albouy et al., 2019; Schulze et al., 2018), they appear to get integrated to yield comparable behavioural outcomes between different task goals in our study. Future studies should investigate the variation in everyday exposure to speech and song cross-culturally.

In summary, we found partial evidence for lateralisation regarding the interaction hypothesis in functional connectivity and no clear evidence for the acoustic hypothesis. While the overall modular organisation of functional connectivity remained relatively stable and the overall topological configuration in clustering and efficiency did not indicate a clear pattern regarding specific attentional demands, reconfiguration occurred at submodular stages in clustering in a specific bilateral subnetwork when task-relevant acoustic cues were degraded. This subnetwork might act as a relay station from spectrotemporally specialised auditory regions to recruit attentional resources when attentional demands are high, altering the efficiency of information transfer to remote brain regions as a function of task goal and available acoustical cues. We suggest that lateralisation is reduced from early to later processing stages on behalf of more distributed processing but recovered under higher attentional demands when task-relevant acoustic cues are hard to obtain.

4 Materials and Methods

The data used for the present study were the same as in Albouy et al. (2020). Please refer to their Supplementary Materials for further details.

4.1 Stimuli

Stimuli were based on ten professionally composed melodies with identical rhythms and consisting of 10 tones each. These were crossed with 10 French sentences, consisting of 10 syllables each, resulting in 100 different stimuli (duration of around 4.5s). These were recorded by a professional, bilingual (English, French) singer at McGill University, Montréal, Canada (48kHz sampling rate, 24-bit depth) and ramped. Stimuli were degraded in spectral or temporal acoustic components in the modulation power spectrum domain using a method by Elliott & Theunissen (2009). The modulation power spectrum represents the energy modulation across the spectral and temporal dimensions and is created by applying a 2D Fast Fourier Transform on the autocorrelation matrix of the spectrogram of a sound. Degradation in the modulation power spectrum domain reduces the spectrogram's resolution across spectral or temporal dimensions above a certain cut off-value. Via an iterative procedure (see Elliott & Theunissen, 2009), the waveform can then be recreated, resulting in a sound deprived of spectral or temporal acoustic information. Based on pilot testing, we used five cut-offs for each spectral (0.6, 1.5, 1.8, 2, 3 cyc/kHz) and temporal (1, 1.5, 2, 2.5, 3.5 Hz) modulations. These represented a wide range of degradation, from almost none to severe. The resulting stimuli were root-mean-square normalised in amplitude.

4.2 Participants

15 French native-speaking participants (8 Females, mean age = 22.2 ± 1.2 years, education 17.33 ± 1.11 years) without musical training participated in the fMRI and behavioural experiment. All participants were right-handed. Participants gave their written informed consent and received monetary compensation for their participation. The study was approved by and all experiments were carried out in accordance with guidelines approved by the Ethics Committee on Human Research of McGill University (MUCH-REB – 2017-332).

4.3 Procedure

4.3.1 Behavioural Experiment

Participants were seated in a sound-attenuated booth. Stimuli were presented with Sennheiser HD 280 pro headphones at ~ 60 dB SPL, using Presentation Software (Neurobehavioral Systems, Berkeley, CA, USA). The same software was used to record participants' responses. The behavioural experiment consisted of a same-different task in which participants had to judge whether pairs of stimuli, separated by a silent gap of 1 second, were the same or different in terms of the sentence or melody they consisted of (see Fig. 2A). Sentences or melodies were always different in the non-target domain and similar in the target domain for half of the trials. After each stimulus pair, a visual cue appeared, indicating whether the judgement was to refer to the sentence or the melody. Participants could decide between six possible responses: "Same", "Maybe Same", "Same not sure", "Different not sure", "Maybe Different", "Different". These responses were collapsed to "same" or "different" for the analysis. There was no time limit for participants to respond, and no feedback was given about the correctness of their responses. The experiment consisted of 6 blocks. The first was a practice block with ten trials, using familiar acapella songs derived from the internet that were not part of the stimulus pool. Four blocks of 50 trials each followed this block. Stimulus pairs were presented in pseudo-random order. Each possible combination of factors (same/different, attention to sentence/melody, temporal/spectral degradation) was distributed uniformly across the four blocks, and each of the five degradation cutoffs of both temporal and spectral degradations occurred 20 times. All 100 sung speech stimuli were presented twice as the first song in the stimulus pair but with a different type of degradation at each presentation. The last block consisted of 40 trials of non-degraded (original) stimuli, which were used to derive a baseline measure to normalise each participant's behavioural performance. The total duration of the experiment was approximately 90 minutes.

4.3.2 Analysis of behavioural data

To measure performance deviation for degraded stimuli from baseline comparable between participants, we computed a normalised score by $(\text{raw score} - \text{chance}) / (\text{baseline score} - \text{chance}) - 1$. The baseline score corresponded to performance with the non-degraded (original) stimuli. Chance level corresponded to 50% correct performance. Thus, a normalised score of 0 indicated no change in performance compared to non-degraded stimuli, while positive and negative values represented an increase or decrease in accuracy, respectively. We then modelled individual data with a linear function to obtain one integrated value per participant, degradation type (spectral/temporal) and attentional focus (sentence/melody). We used these values as a response variable in a simple linear model with cutoff (1 to 5) and task (sentences or melodies) as predictors, using the anova function of the R package "car" (version 3.0-2, Fox et al., 2011). Collinearity was tested using the function "vif" from the same package.

4.3.3 fMRI task

Participants' blood-oxygen level (BOLD) was recorded while they listened to 110 blocks of five degraded stimuli (either in the spectral or temporal domain; see Fig.'s 1 and 3A). A visual cue at the beginning of each block (~1 second duration) instructed participants to either attend to the melody or sentence aspect of the respective stimulus. Participants were asked to detect two catch trials where a sentence or melody was repeated (1-back task) to control for attention. All other sentences/melodies were not repeated in each block. Catch trial stimuli had low temporal (3.5 Hz) and low spectral (3 Cyc/kHz) degradation. All participants detected the catch trials. We again used Presentation Software (Neurobehavioral Systems, Albany, CA, USA) to present the stimuli and record participants' responses. Stimuli were presented via Sensimetrics MRI-compatible insert earphones at 70 dB SPL. The experiment consisted of two runs (about 20 minutes duration per run), and each degradation cutoff of both temporal and spectral degradations was presented five times, resulting in 50 blocks, complemented by four blocks of silence and one block of white noise (thus, 55 blocks per run). Between the runs, there was a break of 2 to 3 minutes. Degradation type (temporal/spectral) and respective cutoffs were distributed equally across each run. The order of presentation was pseudo-randomised, and the same degradation type could be presented a maximum of three times in a row.

4.4 fMRI acquisition parameters and pre-processing

We acquired high-resolution MPRAGE T1-weighted three-dimensional anatomical images using a gradient-echo sequence [192 sagittal slices; time to repetition (TR) = 2300 ms; time to echo (TE) = 2.98 ms; flip angle = 9°; matrix size = 256 × 256; field of view = 256 × 256 mm²; voxel size = 1 × 1 × 1 mm³]. To measure the whole-brain BOLD signal, we used a gradient-echo EPI pulse sequence (48

axial slices with interleaved, descending, multi-band acquisition (acceleration factor 6); TR, 570 ms; volume acquisition, TE, 300 ms; FA, 50°; 2.5 mm slice thickness; no gap; matrix size, 84 x 84, FOV 210 x 210mm²; voxel size, 2.5 x 2.5 x 2.5 mm³): All image pre-processing was performed using the Conn toolbox (version 19c, Whitfield-Gabrieli & Nieto-Castanon, 2012) in MATLAB (2017a) with the standard preprocessing pipeline (see <https://web.conn-toolbox.org/fmri-methods/preprocessing-pipeline>): coregistration and resampling to the first image of the first session (b-spline interpolation), slice-timing correction, outlier identification, normalisation into standard MNI space and segmentation into grey matter, white matter and CSF tissue, and functional smoothing (spatial convolution with a Gaussian kernel of 8mm full-width half maximum). Denoising was done using Conn's standard pipeline (linear regression of potential confounding effects in the BOLD signal and temporal band-pass filtering; see <https://web.conn-toolbox.org/fmri-methods/denoising-pipeline>).

4.5 ROI-to-ROI analysis and network generation

Functional connectivity analysis was performed using the CONN-fMRI toolbox for SPM (<http://www.nitrc.org/projects/conn>). Temporal correlations were computed between the BOLD signals of 358 ROI from Glasser et al. (2016) (see Fig. 7). To do so, a general linear model was fitted to analyse the BOLD activity of each participant for each condition. Data were band-pass filtered (0.008–0.09 Hz), and nuisance covariates were included to control for fluctuations in BOLD signal resulting from cerebrospinal fluid, white matter, and their derivatives. It resulted in ROI-to-ROI correlation matrices representing the level of functional connectivity between each pair of ROIs. Each element in the ROI-to-ROI correlation matrix is defined as the Fisher-transformed bivariate correlation coefficient between a pair of ROI BOLD timeseries. This analysis was performed as a function of attention (sentences/melodies), degradation type (spectral/temporal) and cutoff value (5 steps per degradation type) for each of the 15 participants. This resulted in 300 358x358 functional connectivity matrices (20 per participant, see Table 3).

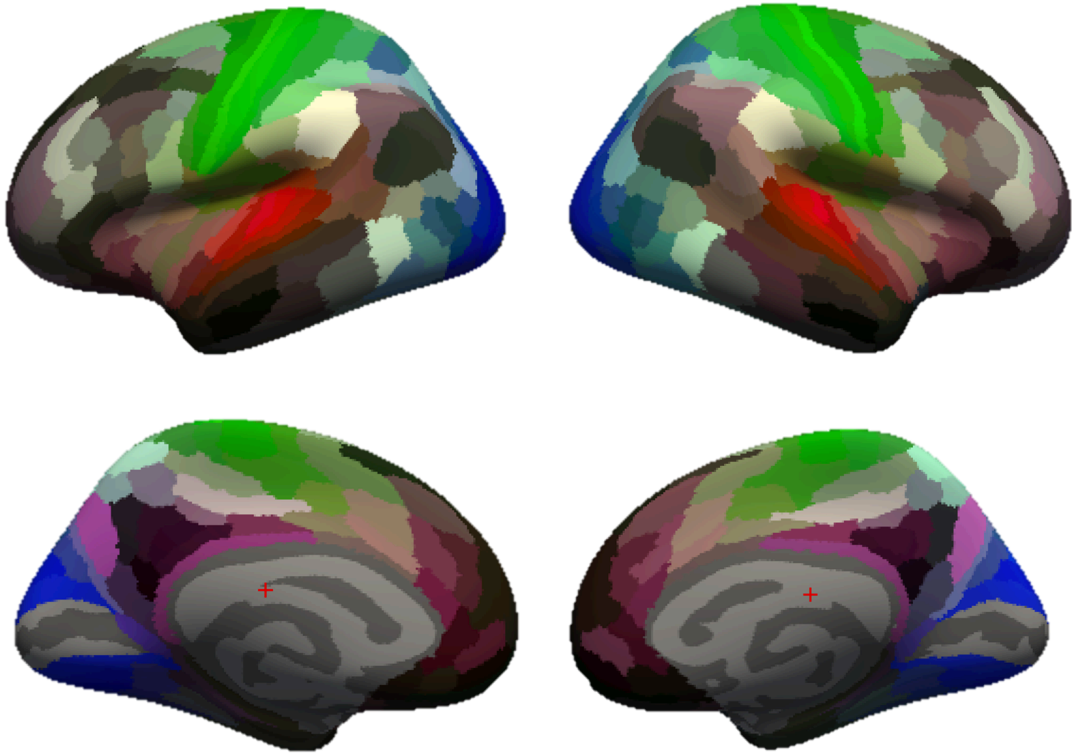


Figure 7: Illustration of the 179 ROI per hemisphere resulting from the conjugation with the Atlas by Glasser et al. (2016). Annotation file created by Mills (2016).

Table 3: Conditions investigated in our analyses.

	Degradation	spectral	temporal
Attention			
Melody		<i>MS</i>	<i>MT</i>
Sentence		<i>SS</i>	<i>ST</i>

Each condition has five degradation steps. MS and ST are the conditions where crucial acoustical information is missing.

4.6 Graph Theoretical Analyses

We applied several analyses using measures from Graph Theory (GraphVar toolbox version 2.0, Kruschwitz et al., 2015; Waller et al., 2018), a field investigating networks regarding their topological organisation. It models brain connectivity networks as "nodes" (ROI) connected by "links" (or "edges"), which in our case were the Fisher z-transformed correlations between ROI. Graph measures can be

calculated regarding the entire functional connectivity network (global measures) and subnetworks or single nodes (local measures, see Fornito et al., 2016).

4.6.1 Global graph measures

Using the GraphVar toolbox, we derived two global measures, the global clustering coefficient and global efficiency, to see whether there are connectivity differences across the entire network between conditions according to our hypotheses. The global clustering coefficient measures, across all nodes, whether the neighbours of a given node are also connected to each other, making it robust against perturbations and providing an indirect measure of functional specialisation if specialisation utilises tight integration in closed triangle motifs (see Fornito et al., 2016). Global efficiency, in contrast, measures the efficiency of information flow via the shortest paths across the network and provides a measure of integration. We used only positive correlations ("negative weights to zero"), a common approach to avoid biasing of graph measures by polarity. Since graph measures depend on the number of connections in the analysis, we used relative network thresholds across a broad range, retaining the strongest 10%, 20%, 30%, 40% or 50% of the edges (cells of the ROI-to-ROI connectivity matrices) ("relative thresholds" 0.1, 0.2, 0.3, 0.4, 0.5). We also created random networks preserving the degree, weight, and strength distribution of the original networks ("c_null_model_und_sign") to normalise our graph measures (100 per participant with ten iterations each). We calculated graph measures for weighted, undirected networks.

4.6.2 Statistical Analysis of clustering coefficient and efficiency

According to our predictions, we aimed to quantify whether the interaction of attention to melody or sentences with spectral or temporal degradation steps influenced the topological connectivity across the entire network. First, we integrated the derived measures across the different network thresholds (area under the curve, Bassett et al., 2006, 2012) to reduce multiple comparisons and increase model stability. All statistical analyses were done in R (R Core Team, 2019). We fitted the models for local and global efficiency and clustering coefficient using the function "lmer" (R package "lme4"; version 1.1-21; Bates et al., 2015). We used GLMMs including the fixed factors "attention" (speech or melody), "degradation" (spectral or temporal degradation), "cutoff" (5 cutoffs from high to low degradation) and their interactions, as well as "participants" as random effect and random slopes of all fixed effects within "participants".

We assessed collinearity by fitting a linear model without random effects and interactions between fixed effects and applying the function "vif" (R package "car"; version 3.0-2; Fox et al., 2011). We assessed model stability using a custom-built function to assess how robust the model estimates

would be given changes in the predictors. This function excludes random effects one at a time from the data. Model estimates of the reduced and the full datasets are then compared (please see Nieuwenhuis et al., 2012) for a comparable approach).

To test the effect of all predictors as a whole, we compared the full model with a null model lacking the predictors of interest, i.e., including only the random intercept and slopes. Models were compared using a likelihood ratio test, applying the function "anova" (argument test set to "Chisq"; Dobson, 2002; Forstmeier & Schielzeth, 2011). Effect sizes were calculated using the function `r.squaredGLMM` (R package "MuMIn"; version 1.47.1; Barton, 2009).

4.6.3 Modularity analysis

To investigate modularity using GraphVar, we again used only positive weights ("negative weights to zero") and normalised the matrices with 100 random networks per subject ("null_model_und_sign", ten iterations). We used a Louvain algorithm with 1000 iterations and a resolution (gamma) of 1.0, first at the individual participant level and then at the group level. The algorithm reaches a consensus classification by executing a modular decomposition on a co-classification matrix, which itself is based on iterative classifications of nodes (see GraphVar manual and Blondel et al., 2008, for details). For each condition, participant and cutoff, we calculated an agreement matrix across five relative network thresholds (0.1, 0.2, 0.3, 0.4, 0.5), which became the basis for calculating the modularity metrics, i.e. the modularity index Q and the classification consistency z (Fornito et al., 2012).

Additional Information

Acknowledgements: F.H. wants to thank Cliodhna Quigley, Daniel C. Mann, Narly Golestani, Olga Kepinska, Sevil Maghsadghagh, and Johann Kruschwitz for helpful advice. F.H. thanks Roger Mundry for providing the following R scripts: `overdisp.test` to test overdispersion, `glmm.model.stab` to test model stability, `xx.fe.re` to do dummy coding. We thank Christopher Soden for creating the melodies, Audrey-Kristel Barbeau for singing the stimuli, and Mariane Generale and Martha de Francisco for their expertise with recording. Parts of the computational results of this work have been achieved using the Life Science Compute Cluster (LiSC) of the University of Vienna. We thank Benjamin Morillon, Robert J. Zatorre and Lucas Benjamin for their help in data acquisition.

Funding: P.A. is funded by the FRQS, Brain Canada and NSERC. F.H. was supported by the Austrian Science Fund (FWF) number W1262-B29 (DK Grant Cognition & Communication). F.H. was also

supported by a VDS CoBeNe final fellowship. The funders had no role in decision on the article content, the decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no competing interests.

Data Availability: Raw data, additional tables and plots are available in the OSF repository (<https://osf.io/merwk/>), code is available from the corresponding authors upon reasonable request.

Supplementary Material: The Supplementary Material for this article can be found online at: <https://osf.io/merwk/>

Author contributions: Conceptualization, F.H., P.A.; Methodology, F.H. and P.A.; Analysis, F.H., P.A.; Investigation, P.A.; Resources, P.A.; Writing – Original Draft, F.H.; Writing – Review & Editing, F.H. and P.A.; Visualization, F.H.; Supervision, P.A.

References

- Alavash, M., Tune, S., & Obleser, J. (2019). Modular reconfiguration of an auditory control brain network supports adaptive listening behavior. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 116, Issue 2, pp. 660–669). <https://doi.org/10.1073/pnas.1815321116>
- Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, 367(6481), 1043–1047. <https://doi.org/10.1126/science.aaz3468>
- Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., & Zatorre, R. J. (2023). Spectro-temporal acoustical markers differentiate speech from song across cultures. *BioRxiv*.
- Albouy, P., Peretz, I., Bermudez, P., Zatorre, R. J., Tillmann, B., & Caclin, A. (2019). Specialized neural dynamics for verbal and tonal memory: fMRI evidence in congenital amusia. *Human Brain Mapping*, 40(3), 855–867. <https://doi.org/10.1002/hbm.24416>
- Angenstein, N., Scheich, H., & Brechmann, A. (2012). Interaction between bottom-up and top-down effects during the processing of pitch intervals in sequences of spoken and sung syllables. *NeuroImage*, 61(3), 715–722. <https://doi.org/10.1016/j.neuroimage.2012.03.086>
- Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F. A., Armony, J. L., & Concha, L. (2014). Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and non-musicians. *Cortex*, 59, 126–137. <https://doi.org/10.1016/j.cortex.2014.07.013>
- Baker, C. M., Burks, J. D., Briggs, R. G., Stafford, J., Conner, A. K., Glenn, C. A., Sali, G., McCoy, T. M., Battiste, J. D., O'Donoghue, D. L., & Sughrue, M. E. (2018). A Connectomic Atlas of the Human Cerebrum-Chapter 4: The Medial Frontal Lobe, Anterior Cingulate Gyrus, and Orbitofrontal Cortex. *Operative Neurosurgery*, 15(6), S122–S174. <https://doi.org/10.1093/ons/opy257>

- Bartha, L., Brenneis, C., Schocke, M., Trinka, E., Köylü, B., Trieb, T., Kremser, C., Jaschke, W., Bauer, G., Poewe, W., & Benke, T. (2003). Medial temporal lobe activation during semantic language processing: fMRI findings in healthy left- and right-handers. *Cognitive Brain Research*, 17(2), 339–346. [https://doi.org/10.1016/S0926-6410\(03\)00135-6](https://doi.org/10.1016/S0926-6410(03)00135-6)
- Barton, K. (2009). MuMIn: multi-model inference. In <http://r-forge.r-project.org/projects/mumin/>.
- Bassett, D. S., Meyer-Lindenberg, A., Achard, S., Duke, T., & Bullmore, E. (2006). Adaptive reconfiguration of fractal small-world human brain functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51), 19518–19523. <https://doi.org/10.1073/pnas.0606005103>
- Bassett, D. S., Nelson, B. G., Mueller, B. A., Camchong, J., & Lim, K. O. (2012). Altered resting state complexity in schizophrenia. *NeuroImage*, 59(3), 2196–2207. <https://doi.org/10.1016/j.neuroimage.2011.10.002>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. (R package version 1.1–21).
- Bendixen, A. (2014). Predictability effects in auditory scene analysis: A review. *Frontiers in Neuroscience*, 8(8 MAR), 1–16. <https://doi.org/10.3389/fnins.2014.00060>
- Bianco, R., Novembre, G., Keller, P. E., Kim, S. G., Scharf, F., Friederici, A. D., Villringer, A., & Sammler, D. (2016). Neural networks for harmonic structure in music perception and action. *NeuroImage*, 142, 454–464. <https://doi.org/10.1016/j.neuroimage.2016.08.025>
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10). <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Briggs, R. G., Conner, A. K., Baker, C. M., Burks, J. D., Glenn, C. A., Sali, G., Battiste, J. D., O'Donoghue, D. L., & Sughrue, M. E. (2018). A Connectomic Atlas of the Human Cerebrum-Chapter 18: The Connectional Anatomy of Human Brain Networks. *Operative Neurosurgery*, 15(00), S470–S480. <https://doi.org/10.1093/ons/opy272>
- Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., Owen, A. M., & Bekinshtein, T. A. (2013). Expectation and attention in hierarchical auditory prediction. *Journal of Neuroscience*, 33(27), 11194–11205. <https://doi.org/10.1523/JNEUROSCI.0114-13.2013>
- Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *Journal of the Acoustical Society of America*, 129(4), 2245–2252. <https://doi.org/10.1121/1.3562174>
- Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, 81, 181–187. <https://doi.org/10.1016/j.neubiorev.2017.02.011>
- Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC.
- Elhilali, M. (2019). *Modulation Representations for Speech and Music*. 335–359. https://doi.org/10.1007/978-3-030-14832-4_12
- Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, 5(3). <https://doi.org/10.1371/journal.pcbi.1000302>

- Field, A. (2005). Andy Field - Discovering Statistics Using SPSS, Second Edition.pdf. In *Journal of Advanced Nursing* (Vol. 58, Issue 3, pp. 303–303). https://doi.org/10.1111/j.1365-2648.2007.04270_1.x
- Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
- Fitch, W. T. (2006). The biology and evolution of music : A comparative perspective. *Cognition*, 100, 173–215. <https://doi.org/10.1016/j.cognition.2005.11.009>
- Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature Human Behaviour*, 3(4), 393–405. <https://doi.org/10.1038/s41562-019-0548-z>
- Fornito, A., Harrison, B. J., Zalesky, A., & Simons, J. S. (2012). Competitive and cooperative dynamics of large-scale brain functional networks supporting recollection. *Proceedings of the National Academy of Sciences of the United States of America*, 109(31), 12788–12793. <https://doi.org/10.1073/pnas.1204185109>
- Fornito, A., Zalesky, A., & Bullmore, E. T. (2016). Fundamentals of Brain Network Analysis. In *Fundamentals of Brain Network Analysis*. <https://doi.org/10.1016/C2012-0-06036-X>
- Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner’s curse. *Behavioral Ecology and Sociobiology*, 65(1), 47–55. <https://doi.org/10.1007/s00265-010-1038-5>
- Foster, N. E. V., Halpern, A. R., & Zatorre, R. J. (2013). Common parietal activation in musical mental transformations across pitch and time. *NeuroImage*, 75, 27–35. <https://doi.org/10.1016/j.neuroimage.2013.02.044>
- Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., & Heilberger, R. (2011). *Package “car”: Companion to applied regression* (version 3.0-2). Thousand Oaks CA: Sage.
- Garbarini, F., D’Agata, F., Piedimonte, A., Sacco, K., Rabuffetti, M., Tam, F., Cauda, F., Pia, L., Geminiani, G., Duca, S., Graham, S. J., & Berti, A. (2014). Drawing lines while imagining circles: Neural basis of the bimanual coupling effect during motor execution and motor imagery. *NeuroImage*, 88, 100–112. <https://doi.org/10.1016/j.neuroimage.2013.10.061>
- Geranmayeh, F., Wise, R. J. S., Mehta, A., & Leech, R. (2014). Overlapping networks engaged during spoken language production and its cognitive control. *Journal of Neuroscience*, 34(26), 8728–8740. <https://doi.org/10.1523/JNEUROSCI.0428-14.2014>
- Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615), 171–178. <https://doi.org/10.1038/nature18933>
- Haiduk, F., & Fitch, W. T. (2022). Understanding Design Features of Music and Language : The Choric / Dialogic Distinction. *Frontiers in Psychology*, 13(April). <https://doi.org/10.3389/fpsyg.2022.786899>
- Haiduk, F., Quigley, C., & Fitch, W. T. (2020). Song Is More Memorable Than Speech Prosody: Discrete Pitches Aid Auditory Working Memory. *Frontiers in Psychology*, 11(December), 1–22. <https://doi.org/10.3389/fpsyg.2020.586723>

- Hesling, I., Dilharreguy, B., Clément, S., Bordessoules, M., & Allard, M. (2005). Cerebral mechanisms of prosodic sensory integration using low-frequency bands of connected speech. *Human Brain Mapping*, 26(3), 157–169. <https://doi.org/10.1002/hbm.20147>
- Holle, H., Gunter, T. C., Rüschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *NeuroImage*, 39(4), 2010–2024. <https://doi.org/10.1016/j.neuroimage.2007.10.055>
- Humphries, C., Sabri, M., Lewis, K., & Liebenthal, E. (2014). Hierarchical organization of speech perception in human auditory cortex. *Frontiers in Neuroscience*, 8(DEC), 1–12. <https://doi.org/10.3389/fnins.2014.00406>
- Hymers, M., Prendergast, G., Liu, C., Schulze, A., Young, M. L., Wastling, S. J., Barker, G. J., & Millman, R. E. (2015). Neural mechanisms underlying song and speech perception can be differentiated using an illusory percept. *NeuroImage*, 108, 225–233. <https://doi.org/10.1016/j.neuroimage.2014.12.010>
- Jamison, H. L., Watkins, K. E., Bishop, D. V. M., & Matthews, P. M. (2006). Hemispheric specialization for processing auditory nonspeech stimuli. *Cerebral Cortex*, 16(9), 1266–1275. <https://doi.org/10.1093/cercor/bhj068>
- Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science*, 366(6461), 50–54. <https://doi.org/10.1126/science.aax0287>
- Jonas, M., Siebner, H. R., Biermann-Ruben, K., Kessler, K., Bäumer, T., Büchel, C., Schnitzler, A., & Münchau, A. (2007). Do simple intransitive finger movements consistently activate frontoparietal mirror neuron areas in humans? *NeuroImage*, 36(SUPPL. 2), 44–53. <https://doi.org/10.1016/j.neuroimage.2007.03.028>
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102. <https://doi.org/10.1016/j.cognition.2015.03.016>
- Kob, M., Bernardoni, N. H., Herzel, H., Howard, D., Tokuda, I., Wolfe, J., Kob, M., Bernardoni, N. H., Herzel, H., Howard, D., Tokuda, I., & Wolfe, J. (2011). Analysing and Understanding the Singing Voice : Recent Progress and Open Questions. *Current Bioinformatics*, 6(3), 362–374.
- Koelsch, S., Fritz, T., Cramon, D. Y. V., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: An fMRI study. *Human Brain Mapping*, 27(3), 239–250. <https://doi.org/10.1002/hbm.20180>
- Krumhansl, C. L. (2001). *Cognitive foundations of musical pitch* (Vol. 17). Oxford University Press.
- Kruschwitz, J. D., List, D., Waller, L., Rubinov, M., & Walter, H. (2015). GraphVar: A user-friendly toolbox for comprehensive graph analyses of functional brain connectivity. *Journal of Neuroscience Methods*, 245, 107–115. <https://doi.org/10.1016/j.jneumeth.2015.02.021>
- Kung, S. J., Chen, J. L., Zatorre, R. J., & Penhune, V. B. (2013). Interacting cortical and basal ganglia networks underlying finding and tapping to the musical beat. *Journal of Cognitive Neuroscience*, 25(3), 401–420. https://doi.org/10.1162/jocn_a_00325
- Kyong, J. S., Scott, S. K., Rosen, S., Howe, T. B., Agnew, Z. K., & McGettigan, C. (2014). Exploring the Roles of Spectral Detail and Intonation Contour in Speech Intelligibility: An fMRI Study. *Journal of Cognitive Neuroscience*, 26(8), 1748–1763. <https://doi.org/10.1162/jocn>

- Lee, A. K. C., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, 307, 111–120. <https://doi.org/10.1016/j.heares.2013.06.010>
- Mantell, J. T., & Pfordresher, P. Q. (2013). Vocal imitation of song and speech. *Cognition*, 127(2), 177–202. <https://doi.org/10.1016/j.cognition.2012.12.008>
- McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, 19(22), 1024–1027.
- Mcgettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: what’s wrong, what’s right, and what’s left? *Trends in Cognitive Sciences*, 16(5), 269–276. <https://doi.org/10.1016/j.tics.2012.04.006>.Cortical
- McNealy, K., Mazziotta, J. C., & Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *Journal of Neuroscience*, 26(29), 7629–7639. <https://doi.org/10.1523/JNEUROSCI.5501-05.2006>
- Mills, K. (2016). *HCP-MMP1.0 projected on fsaverage. figshare. Dataset.* <https://doi.org/https://doi.org/10.6084/m9.figshare.3498446.v2>
- Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, 114(42), E8913–E8921. <https://doi.org/10.1073/pnas.1705373114>
- Mueller, K., Mildner, T., Fritz, T., Lepsien, J., Schwarzbauer, C., Schroeter, M. L., & Möller, H. E. (2011). Investigating brain response to music: A comparison of different fMRI acquisition schemes. *NeuroImage*, 54(1), 337–343. <https://doi.org/10.1016/j.neuroimage.2010.08.029>
- Nieuwenhuis, R., te Grotenhuis, M., & Pelzer, B. (2012). Influence.ME: Tools for detecting influential data in mixed effects models. *R Journal*, 4(2), 38–47. <https://doi.org/10.32614/rj-2012-011>
- Nishimura, T., Tokuda, I. T., Miyachi, S., Dunn, J. C., Herbst, C. T., Ishimura, K., Kaneko, A., Kinoshita, Y., Koda, H., Saers, J. P. P., Imai, H., Matsuda, T., Larsen, O. N., Jürgens, U., Hirabayashi, H., Kojima, S., & Fitch, W. T. (2022). Evolutionary loss of complexity in human vocal anatomy as an adaptation for speech. *Science*, 377(6607), 760–763. <https://doi.org/10.1126/science.abm1574>
- Norman-Haignere, S., Kanwisher, N. G., McDermott, J. H., 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*, 32(7), 1470-1484.e12. <https://doi.org/10.1016/j.cub.2022.01.069>
- Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J., Benetos, E., Proutskova, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., & Savage, P. E. (2023). Globally songs are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]. *Peer Community In Registered Reports, Preprint:*
- Park, M., Gutyrchik, E., Welker, L., Carl, P., Pöppel, E., Zaytseva, Y., Meindl, T., Blautzik, J., Reiser, M., & Bao, Y. (2015). Sadness is unique: Neural processing of emotions in speech prosody in musicians and non-musicians. *Frontiers in Human Neuroscience*, 8(JAN), 1–8. <https://doi.org/10.3389/fnhum.2014.01049>

- Péran, P., Démonet, J. F., Cherubini, A., Carbebat, D., Caltagirone, C., & Sabatini, U. (2010). Mental representations of action: The neural correlates of the verbal and motor components. *Brain Research*, 1328, 89–103. <https://doi.org/10.1016/j.brainres.2010.02.082>
- Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21(6), 322–334. <https://doi.org/10.1038/s41583-020-0304-4>
- Prete, G., Marzoli, D., Brancucci, A., & Tommasi, L. (2016). Hearing it right : Evidence of hemispheric lateralization in auditory imagery. *Hearing Research*, 332, 80–86. <https://doi.org/https://doi.org/10.1016/j.heares.2015.12.011>
- Prete, G., Tommasi, V., & Tommasi, L. (2020). Right news, good news! The valence hypothesis and hemispheric asymmetries in auditory imagery. *Language, Cognition and Neuroscience*, 35(4), 409–419. <https://doi.org/10.1080/23273798.2019.1659990>
- Quante, L., Kluger, D. S., Bürkner, P. C., Ekman, M., & Schubotz, R. I. (2018). Graph measures in task-based fMRI: Functional integration during read-out of visual and auditory information. *PLoS ONE*, 13(11), 1–18. <https://doi.org/10.1371/journal.pone.0207119>
- R Core Team. (2019). *R: A Language and Environment for Statistical Computing* (Version 3.5. 2). R Foundation for Statistical Computing.
- Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, 15(8), 1261–1269. <https://doi.org/10.1093/cercor/bhi009>
- Rogalsky, C., Rong, F., Saberi, K., & Hickok, G. (2011). Functional anatomy of language and music perception: Temporal and structural factors investigated using functional magnetic resonance imaging. *Journal of Neuroscience*, 31(10), 3843–3852. <https://doi.org/10.1523/JNEUROSCI.4515-10.2011>
- Rohrmeier, M., Zuidema, W., Wiggins, G. A., & Scharff, C. (2015). Principles of structure building in music , language and animal song. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(20140097). <https://doi.org/http://dx.doi.org/10.1098/rstb.2014.0097>
- Rothermich, K., & Kotz, S. A. (2013). Predictions in speech comprehension: FMRI evidence on the meter-semantic interface. *NeuroImage*, 70, 89–100. <https://doi.org/10.1016/j.neuroimage.2012.12.013>
- Sadato, N., Yonekura, Y., Waki, A., Yamada, H., & Ishii, Y. (1997). Role of the supplementary motor area and the right premotor cortex in the coordination of bimanual finger movements. *Journal of Neuroscience*, 17(24), 9667–9674. <https://doi.org/10.1523/jneurosci.17-24-09667.1997>
- Sammler, D., Grosbras, M. H., Anwander, A., Bestelmeyer, P. E. G., & Belin, P. (2015). Dorsal and ventral pathways for prosody. *Current Biology*, 25(23), 3079–3085. <https://doi.org/10.1016/j.cub.2015.10.009>
- Sammler, D., Koelsch, S., Ball, T., Brandt, A., Grigutsch, M., Huppertz, H. J., Knösche, T. R., Wellmer, J., Widman, G., Elger, C. E., Friederici, A. D., & Schulze-Bonhage, A. (2013). Co-localizing linguistic and musical syntax with intracranial EEG. *NeuroImage*, 64(1), 134–146. <https://doi.org/10.1016/j.neuroimage.2012.09.035>
- Schonwiesner, M., Rübsamen, R., & Von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *European Journal of Neuroscience*, 22(6), 1521–1528. <https://doi.org/10.1111/j.1460-9568.2005.04315.x>

- Schulze, K., Koelsch, S., & Williamson, V. (2018). Auditory working memory. In *Springer handbook of systematic musicology* (pp. 461–472). Springer.
- Seger, C. A., Spiering, B. J., Sares, A. G., Quraini, S. I., Alpeter, C., David, J., & Thaut, M. H. (2013). Clinical Practice Guideline for the Treatment of Posttraumatic Stress Disorder (PTSD). *Journal of Cognitive Neuroscience*, 25(7), 1062–1077. <https://doi.org/10.1162/jocn>
- Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, 270(5234). <https://doi.org/10.1126/science.270.5234.303>
- Sundberg, J. (1989). *The Science of the Singing Voice*. Northern Illinois University Press.
- Tierney, A. T., Patel, A. D., & Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, 147(6), 888–904.
- Tomlinson, G. (2015). *A million years of music: The emergence of human modernity*. MIT Press.
- Trost, W., Ethofer, T., Zentner, M., & Vuilleumier, P. (2012). Mapping aesthetic musical emotions in the brain. *Cerebral Cortex*, 22(12), 2769–2783. <https://doi.org/10.1093/cercor/bhr353>
- van der Burght, C. L., Goucha, T., Friederici, A. D., Kreitewolf, J., & Hartwigsen, G. (2019). Intonation guides sentence processing in the left inferior frontal gyrus. *Cortex*, 117, 122–134. <https://doi.org/10.1016/j.cortex.2019.02.011>
- Waller, L., Brovkin, A., Dorfschmidt, L., Bzdok, D., Walter, H., & Kruschwitz, J. D. (2018). GraphVar 2.0: A user-friendly toolbox for machine learning on functional connectivity measures. *Journal of Neuroscience Methods*, 308(July), 21–33. <https://doi.org/10.1016/j.jneumeth.2018.07.001>
- Wallmark, Z., Deblieck, C., & Iacoboni, M. (2018). Neurophysiological effects of trait empathy in music listening. *Frontiers in Behavioral Neuroscience*, 12(April), 1–19. <https://doi.org/10.3389/fnbeh.2018.00066>
- Watts, D. J., & Strogatz, S. H. (1998). Strogatz - small world network Nature. *Nature*, 393(June), 440–442. <https://www.ncbi.nlm.nih.gov/pubmed/9623998>
- Weidema, J. L., Roncaglia-Denissen, M. P., & Honing, H. (2016). Top-Down modulation on the Perception and categorization of identical pitch contours in speech and music. *Frontiers in Psychology*, 7(JUN), 1–11. <https://doi.org/10.3389/fpsyg.2016.00817>
- Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. *Brain Connectivity*, 2(3), 125–141. <https://doi.org/10.1089/brain.2012.0073>
- Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, 11, 946–953. <https://doi.org/10.1093/cercor/12.2.140>
- Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience*, 8(7), 547–558. <https://doi.org/10.1038/nrn2152>

DISCUSSION

In this thesis, I have attempted to understand why music and language differ in certain spectrotemporal features. To illustrate how the three chapters relate to each other, I will follow four steps: First, I will describe the three chapters and the different perspectives and assumptions they make to see how their conceptual differences limit their interpretation. Second, I will argue that two additional conceptual frameworks, active inference and biosemiotics in Terrence Deacon's interpretation, can provide a shared conceptual grounding to the three chapters. Third, I will interpret the three chapters based on the active inference framework and biosemiotics. Finally, I will give some future research directions from the three chapters.

1 Three chapters in three perspectives

Considering Tinbergen's and Marr's frameworks that I heuristically used to lay out the three chapters in the introduction, it is already clear that tackling complex traits or systems such as language and music requires multiple perspectives (Massimi, 2022; Wimsatt, 2007) and levels of investigation. In cognitive biology, phenomena and explanatory mechanisms are often not linked straightforwardly, thus not in one-to-one but many-to-many mappings (B. Hallgrimsson, personal communication, June 22, 2023; Wimsatt, 2007). Also, there are multiple approaches to evolutionary questions, and the content of evolutionary theory and the relevance of different evolutionary processes are debated (Laland et al., 2015; Laland & Brown, 2011; Wagner & Tomlinson, 2022). From an epistemological point of view, multiple perspectives such as those articulated regarding evolutionary theory are considered valuable in explaining natural phenomena (Massimi, 2022; Wimsatt, 2007). To guide the interpretation of the three Chapters in the light of these different perspectives it is therefore useful to clarify their conceptual assumptions regarding evolution. One helpful reference frame concerning conceptual assumptions in evolutionary theory has been proposed by Hallgrimsson (personal communication, June 22, 2023): evolutionary explanations can be conceptualised in two dimensions: the first reaches from functionalistic explanations, with an emphasis on factors external to organisms, to structural explanations, emphasising factors internal to organisms like exploration and exploitation of niches constrained by selection. The other dimension reaches from reductionist approaches which attempt low-level mechanistic explanations to holistic approaches which emphasise emergent properties and systems level explanations (B. Hallgrimsson, personal communication, June 22, 2023). Thus, functional, reductionist approaches rely on concepts like adaptation and natural selection, while structuralist, holistic approaches explain evolution as exploring ecological niches constrained by selection. I will use this proposal below.

Chapter 1:

Study 1 laid out a framework about spectrotemporal differences in terms of design features of language and music that are present in typical spoken language and music despite both systems operating along the same acoustic continua: pitch, formant transitions and duration. We suggested a framework of three dimensions to account for the emergence of spectrotemporal features in music and spoken language and linked it to general cognitive processes of multilevel prediction. We operationalised predictability with concepts from information theory - uncertainty (or entropy) and surprisal (or information) - and made predictions concerning how the deployment along these dimensions should change these information trajectories at multiple levels. The general nature of the dimensions proposed (especially the social interaction dimension), the notion of prediction of sequences, and the fact that information and entropy can be measured empirically make the framework suitable for applications in music, language, and animal communication. Thus, it invites comparisons of language and music with animal communication regarding predictability and social interaction. In doing so, it situates both language and music in a biocultural evolutionary context.

Chapter 1 is the most process-based of the three chapters, thus, in Halgrimsson's framework, it represents a structuralist and systemic attempt since it does rely on explaining features by emergence without invoking natural selection, at least not directly. In fact, should it turn out empirically that features of music and language would emerge by social interaction (cf. Miton & Charbonneau, 2018; Thompson et al., 2016), this would shift the explanatory burden towards motivational questions. Thus, rather than attempting to explain what adaptive function a given spectrotemporal feature would reflect, questions to address would be why humans engage in choric or dialogic modes of interaction in the first place, or how the capacity for vocal imitation relates to engaging in repetition. The notion of feature emergence by social interaction illustrates that Marr's computational account is not ideal as a guiding heuristic here. Marr's account assumes fixed algorithms implemented neurally to fulfil specific goals (French & Thomas, 2015; Samuelson et al., 2015), while our framework suggests that goals are dynamically negotiated and shared between interlocutors, and that neural physiology is exploited dynamically and flexibly to fulfil these goals.

Being a theoretical account, our framework from Chapter 1 leaves some issues unexplained that require clarification by empirical or computational modelling work. One issue relates to the geometry of the three dimensions we proposed, which are not orthogonal. For example, propositionality and novelty are closely related since, in conversational language, most of what is semantically communicated must be informative and thus novel to the receiver. Both repetition and choric interaction suggest the emergence of discrete pitch and metrical regularity, which suggest a close relation between these two poles. The non-orthogonality suggests a historical dependence of the

three dimensions, that is, their geometry potentially has changed over time by differentiation of dimensions or by locking into codependence because of shared causal relations by constraints on predictability. This potentially indicates evolutionary processes like differentiation into "kinds" (see Wagner & Tomlinson, 2022) and relaxed selection (Deacon, 2010, 2022a), which have been proposed to apply to both biological and cultural traits. I will address this point below. It also suggests that such dimensions, if we take them as a working model for communicative interaction, may differ between different species and across individual development. This would invite research projects investigating the differentiation and ontogeny of these dimensions comparatively.

Another limitation of the framework that could be addressed by empirical comparative work is that we do not explain enculturation. Instead, we rely on findings of encultured humans who develop in an already musical and linguistic environment (Boyd & Richerson, 2005). An evolutionary account would need to address this by showing how modes of social interaction lead to feature emergence, in turn providing the basis for cultural transmission before music and language existed (Tomlinson, 2015; Tomlinson, 2022). For empirical studies based on our framework, these concerns must be considered in the design, for example, by complementing human experiments with comparative analyses.

Finally, Chapter 1 is limited by relying on the "machine metaphor" of life in using the terms "reverse engineering" and "design features". This metaphor has provided helpful guidance for evolutionary thinking (Tooby & Cosmides, 2005), but it has been criticised by philosophers and evolutionary biologists for falling short when used theoretically instead of heuristically. Specifically, these critics point out that the analytical reductionism and search for deterministic explanations reach their limits when explaining evolutionary processes on a systemic level (Nicholson, 2013, 2014, 2019). For example, an engineering perspective invites us to look for optimisation (in fact, this is one of the associations that one reviewer pointed out for the article forming Chapter 1). Evolution, however, is not necessarily about optimisation, given historical constraints and multiple non-adaptive processes (Gould & Lewontin, 1979; Magnon & Corbara, 2022; Pigliucci & Kaplan, 2000). Thus, although a reverse engineering approach helped us heuristically in coming up with this framework, in further development we better refrain from using the metaphors of "reverse engineering" and "design" and instead think in terms of "characteristic" or "typical" features of music and language.

Chapter 2:

Chapter 2 was designed with Tinbergen's four questions in mind, focussing on explaining a currently observable trait with a possible adaptive function. The chapter thus takes an adaptationist stance, emphasising functionalist and reductionist explanations (B. Hallgrimsson, personal

communication, June 22, 2023). In that sense, it is also influenced by evolutionary psychology, one of several approaches to the evolution of human behaviour (Laland & Brown, 2011). Evolutionary psychology emphasises the search for adaptative functions of psychological traits, assuming a modular organisation of cognition (Laland & Brown, 2011). In this chapter, I tried to find a possible adaptive function for a feature at the level of the audible phenomenon that is abundant in instances of music cross-culturally - discrete pitch - assuming that such a statistically universal feature is pointing to an evolved trait in humans. In other words, I assumed that it indicates an underlying purpose that is measurable and, in turn, explains the feature's existence. In this case, the purpose would be enhancing auditory working memory for vocal phrases. Indeed, we observed that vocal phrases consisting of discrete pitches are remembered better than phrases of the same contours with gliding pitches, confirming the prediction.

However, in creating the hypothesis, I conflated current utility with adaptive value (see Bateson & Laland, 2013). A trait's current function may differ from that for which it was selected (exaptation), and that it is reliably observed across the life span and across cultures does not mean this robustness is maintained by selection. In fact, several processes have been described that explain the emergence of traits without selection, for example by means of pleiotropy, ontogenetic processes (evodevo, see Laland & Brown, 2011), cultural inheritance (Boyd & Richerson, 2005), relaxed selection (Deacon, 2010, 2022a), drift etc. It could also be that an effect of pitch on memory is an epiphenomenon of some other cognitive process; this interpretation is, in fact, suggested regarding Chapter 1 (see below). Investigating current utility does not distinguish between these possible origins. I also assumed that a statistically universal feature difference, like the use of discrete versus gliding pitch, has to represent an optimal solution to a given problem or, following Marr (1982), the outcome of an algorithmic solution for a given goal and thus an adaptive function, neurally implemented. However, as pointed out by Patel (2023), statistically universal traits provide no evidence for evolved neural specialisations. My thinking in this chapter, influenced by Marr's and Tinbergen's frameworks, thus assumes evolution to optimise, the same assumption as criticised above. It is not unlikely that statistical universals indicate similar causes. However, they need not necessarily be adaptations but could, for example, be "rediscovered" by exploitation under similar social, environmental, and physiological constraints. The relationships between these constraints might themselves be subject to various biocultural evolutionary processes (see, e.g., Deacon, 2010, 2022a; Wagner & Tomlinson, 2022). The study design in this chapter cannot discriminate between these possibilities.

The study is also naturally limited by having tested encultured participants, that is, participants had musical and linguistic knowledge. We observed that higher musical sophistication correlated with remembering stimuli better, independently of whether pitches were gliding or discrete. This suggests that general musical enculturation is sufficient for this effect, at least in participants with Western

musical enculturation. Even though we tried to minimise these influences by using both an unfamiliar scale and an unfamiliar prosodic contour to create our stimuli, and even though the study can be implemented cross-culturally, this means we could not disentangle bottom-up acoustic influences from top-down interpretations as musical or as speech-like. In other words, we cannot conclude whether it is discrete pitches per se that convey the memory advantage or whether it is a perception of the stimuli as musical or as speech-like that is crucial. Interestingly, though, we also observed that ambiguous pitches, partially flat and partially gliding, showed the same enhancement effects as discrete pitches, possibly reflecting the strong tolerance for mistuning in singing (Pfordresher & Brown, 2017). This suggests that acoustic cues alone are not sufficient explanations and also that they indicate but do not determine a top-down interpretation. The interaction of acoustical cues and top-down interpretation was investigated in Chapter 3.

Chapter 3:

Chapter 3 examined spectrotemporal differences in the narrower sense of temporal and spectral acoustical modulations. Using sung speech stimuli, we investigated the differences in neural functional connectivity networks when relevant spectral or temporal acoustic cues were degraded, but recognition of sentence or melody was required. It thus can be seen as an attempt to address both Marr's implementational and algorithmic levels. We hypothesised that attention to melodies or sentences of sung speech would interact with spectral or temporal acoustic cues to contribute to hemispheric lateralisation (interaction hypothesis), or to the contrary that hemispheric lateralisation depends solely on bottom-up acoustic information, that is, spectral or temporal acoustic cues (acoustic hypothesis). At different levels of investigation - global, at the subnetwork level, and concerning single regions - and using different measures, the analyses overall yielded only partial evidence for the interaction hypothesis. However, we observed topological reconfiguration in local clustering, indicating segregation as the proposed attentional mechanism, along with higher efficiency of information transmission, when degraded acoustical cues were task-relevant. This effect partially contributed to lateralisation. These topological changes in subnetworks contrast with the overall topological architecture, measured by modularity, which stayed relatively constant. Insofar as our different graph measures indicated different processing stages, our results also suggested that lateralisation dynamically changes depending on processing stage and requirements: lateralisation, present in early acoustical processing, was partially recovered for later processing stages under high attentional demands, namely when intelligibility for sentence or melody was required, but relevant acoustic cues were degraded. Later, more integrated processing stages making use of the entire network seem not to be lateralised in this study. We used the same stimuli in all conditions, altering the cognitive tasks for the participants, which controlled acoustic input and disentangled its influences from top-down processing (see Albouy et al., 2020). However, in our case, this meant using stimuli

that represent only certain aspects of speech (lacking prosody) and a task that required several cognitive processes, including decision-making, auditory memory, and attention to only one stimulus aspect.

Study 3 aimed to investigate lateralisation and the interaction of acoustic spectrotemporal features and attention to melodies or sentences. In that respect, it drew on the classic distinction between top-down and bottom-up processing (Awh et al., 2013; Rauss & Pourtois, 2013). Relevant regarding evolutionary questions is that the modulation power spectrum framework assumes that language and music inherit their feature differences from exploiting auditory physiology that is not specific to music or language. This is in line with systemic evolutionary thinking insofar as this proposes reliance on existing, physiologically grounded traits to be evolutionarily refined or to interact in novel ways with other traits, in contrast to assuming the evolution of language-specific or music-specific processing modules *de novo* (Jacob, 1977; see also O'Malley et al., 2015). Comparative evidence indicates that the spectrotemporal hemispheric divide indeed occurs in several species (see Flinker et al., 2019). However, one needs to consider that lateralisation due to semantic meaning might interact with lateralisation to spectrotemporal differences, at least in humans (see, e.g., McGettigan & Scott, 2012; van der Burght et al., 2019). In that sense, the perspective of this chapter is also a more structuralist, holistic one *sensu* Hallgrímsson (exploration of ecology, constrained by selection; personal communication, June 22, 2023), even though it also investigated current utility (Bateson & Laland, 2013). Also, this study has methodological limitations. The focus on brain regions and their interactions with fMRI as a method with low temporal resolution somewhat enforces reductionism regarding cognitive processes insofar as functional connectivity represents time averages. Graph theoretical measures at least allow for a more integrative and multilevel network-based analysis (e.g., Chang & Chen, 2021). However, even if two networks are structurally the same, they can yield rather different dynamical regimes (Jaeger & Monk, 2021b), adding another layer of dynamicity on top of topological reconfiguration that could not be tracked with our study. Also, whether undifferentiated signals would exploit networks and merge on certain spectrotemporal features or whether these networks are a response to spectrotemporal features in the first place cannot be decided with this study design. The mutual interaction and consolidation of these two processes requires a more process-based framework that allows us to model and measure these dynamics. This calls for studies investigating the process of signal emergence during interactions of individual agents, possibly under different modes of interaction (see Chapter 1). FMRI, with its low temporal resolution - yielding an observation of instances cut relatively coarsely from the temporal flux of cognitive processing - thus comes to its limits (see, e.g., Chang & Chen, 2021).

As we have seen, the three chapters represent different perspectives regarding their focus on and implementation of structuralistic or functionalistic and reductionist or holistic concepts in

Hallgrimsson's framework. Chapter 1 takes an emergentist stance without making use of the notion of adaptation, while Chapter 2 was designed with a focus on adaptive functions. Study 3 is only indirectly concerned with evolution by focussing on the exploitation of hemispheric spectrotemporal lateralisation. We also notice that although Marr's and Tinbergen's heuristics provided guidance in designing the studies, they come to their limits as interpretative guidance, especially regarding Chapter 1. Thus, frameworks that can extend Marr's and Tinbergen's levels with a dynamic perspective comprising systemic concepts like emergence would be useful to interpret the three chapters. These frameworks should also be able to address questions on communication to account for music and language and add a subjective, agential component to account for the role of task goals and social interaction (Chapters 1 and 3). Processes of attention and memory, which were addressed in this thesis, should also be part of their conceptualisation. Of the wide range of possible frameworks in cognition and evolution, two seem to be particularly helpful to this end: active inference (Parr et al., 2022) and biosemiotics as conceptualised by Deacon (Deacon, 2022b; see also Tomlinson, 2015; García & Ibáñez, 2022). Both frameworks attempt to link structural and functional perspectives sensu Hallgrimsson (personal communication, June 22, 2023) in emphasising links between factors both internal and external to an organism, using a systems theoretical perspective. Both frameworks have also been used to explain aspects of communication, cognition, and evolution due to their formalised nature (e.g. Badcock et al., 2019; Deacon, 2010, 2022b; Ramstead et al., 2019; Vasil et al., 2020) and thus appear useful to address the questions in this thesis. We, therefore, can use them to explore how they can inform the interpretation of the three thesis Chapters.

2 Additional helpful perspectives: active inference and semiotics

A framework that influenced Chapter 1 and has the potential to integrate the three chapters is the active inference framework (Parr et al., 2022). In a nutshell, it states that organisms reduce uncertainty and make inferences about the world by either selecting the most appropriate prediction fitting the data (in a Bayesian way by prediction error minimisation) or by selecting actions that lead to the world fitting the prediction (active inference). This is modelled as Bayesian inference. The framework is rooted in information theory and thermodynamics and, although originating in an attempt to explain cognitive processes, has been extended in recent years to comprise communication (Friston & Frith, 2015; Vasil et al., 2020), evolutionary processes like niche construction (Constant et al., 2018; see also Badcock et al., 2019) and even the dynamics of living beings (Ramstead et al., 2021, but see Raja et al., 2021). This is done by applying a formalised account, the free energy principle, that models the mutual externalisation and internalisation of dynamics between two entities, for example, between individuals or between organisms and their environments. These entities are separated by boundaries that can be formalised statistically as Markov blankets (but see recent criticisms of shortcomings of this account in Bruineberg et al., 2022 and related commentaries).

Attention and memory are fundamental cognitive processes in this account. Prior beliefs about what to encounter in the world are either evolved (Badcock et al., 2019) or learned, and moment-to-moment predictions are thus based on working memory framed as evidence accumulation (Parr & Friston, 2017). Attention is conceived of as modulating the precision of prediction errors, thus the signal-to-noise ratio, while salience is understood as the planned sampling of data by action, which is based on working memory (Parr & Friston, 2017). In the auditory domain, this could relate to task-dependent modulation of early auditory processing (see Bendixen, 2014). The scope of this framework is broad and provides us with concepts that can account for both reduction in uncertainty and variances in the linkage between cognition and environment over time.

Biosemiotics is another framework with a similar explanatory scope that is especially relevant for language and music (Barbieri, 2008; Deacon, 2022b; Tomlinson, 2016; Vehkavaara & Sharov, 2017). Here I will use Terrence Deacon's (2022b) conceptualisation of this framework and for simplicity will use the term "biosemiotics" to refer to this conceptualisation. Biosemiotics is generally concerned with the generation of meaning, or the way organisms interpret their environment (Barbieri, 2008). In other words, it focuses on what states or objects in an organism's environment "are about" for the organism at a given moment, what they stand for or represent to the organism. Such interpretative relationships between objects in the world and their representations for the organism are called "signs". The representational forms, such as words or pictures, are termed "sign vehicles", and the cognitive operations creating the relationship between them and the objects they represent are termed "semiosis" or "sign-making". Deacon, in following Peirce (1994), distinguishes three types of signs: icons, indices, and symbols. Note that in his conceptualisation, these terms are used in an abstract sense and are not confined to linguistics.

The three types of signs are hierarchically related by cognitively building onto each other (Deacon, 2022b). Fundamental for iconic signs are shared properties between an object and its representation. One example from the visual domain would be a drawing of an elephant representing an actual elephant. Multiple such drawings are possible, some may depict only the elephant's boundaries, and some may rely on colour or emphasise the trunk. This illustrates that there are multiple possible attributes of sign vehicles (the drawing) that bear a resemblance to an object (the actual elephant) and that an organism needs to establish a relationship between the two that is relevant to it. This can vary between individuals and across time. To establish an iconic relationship, a way of abstracting certain features from a given object is therefore required. This is done by attention and memory formation processes such that these features come to stand for the object they refer to (Tomlinson, 2022). Iconic interpretation thus provides the basis for categorical perception. An example might be that multiple occasions of encountering sounds of footsteps bear similarity to each other. Sounds of footsteps can thus become iconic and form a category.

The second type of sign, an index, is based on correlations between entities and can cognitively be formed by recognising shared attributes between remembered iconic interpretations (Deacon, 2022b). One example would be hearing footsteps and interpreting this as someone approaching, based on previous experience that such categorical events (sounds of footsteps and someone walking) frequently occur together. The sound of footsteps would, therefore, be an index for a walking person. Such an indexical interpretation can even be invoked without ever having heard a person approaching, based on the experience that objects hitting other objects often create a sound (iconic interpretation), that feet touching ground fall into this category, and that walking involves feet regularly touching the ground (also iconic). This illustrates that indexical interpretations are based on iconic interpretations.

The third type, symbols, are culturally conventionalised both in the way they refer to objects and in their sign vehicle attributes (e.g. words in different languages referring to the same object). Because of this cultural conventionalisation, the sign vehicle attributes of symbols typically do not relate to their referents in iconic or indexical ways. Instead, their real-world grounding must be learned based on indexical and iconic processes. This would, for instance, happen in language learning, with the help of iconic gestures or indexical pointing in preverbal children during caregiver-infant interaction (e.g. Behne et al., 2014; see also Cooperrider, 2020). This grounding then stabilises by learned interpretative habits and by learning indexical relationships between the sign vehicles themselves, for example in learning relationships between words (Deacon, 2022b). Thus, the same cognitive processes of interpretation can apply recursively to create different kinds of signs.

Active inference and biosemiotics appear to share some conceptual ideas: both are concerned with mapping relationships between organisms and environment, draw on information theoretical concepts, and argue from a system theoretical perspective. Indeed, there have been attempts to link the two frameworks. The three kinds of signs can be expressed in active inference terms, linking the two frameworks (Milette-Gagnon et al., 2022): Iconic signs have been interpreted as multiple instances based on which an organism can infer the same hidden causes. According to Deacon (2022b), such iconic interpretations can be adapted on a moment-to-moment basis depending on sensory input and behavioural goals, and several interpretations can exist in parallel, contributing to the organism's percept. This resembles Bayesian belief updating and attentional precision gain in the active inference framework. Turning to indexical relationships, Milette-Gagnon et al. (2022) interpret those as beliefs about moment-to-moment state transitions, that is, how one state transitions into another and what the hidden causes of the future state may be. If an organism hears particular tapping sounds, it can infer iconically, due to previous experience with similar sounds, that they most likely are sounds of footsteps. This belief then reduces uncertainty about the future state that the sound of footsteps is most

probably caused by a person approaching. This would provide an example of an indexical inference. Finally, symbols are conventionalised signs understood as shared generative models, that is, shared beliefs about states of the world. These beliefs are culturally transmitted. Symbols are also higher-order indices insofar as they point to other indices. For example, symbols like words can indicate other words with given transition probabilities. These insights will become relevant below in interpreting spectrotemporal differences between music and spoken language.

Two essential concepts utilised in active inference and biosemiotics are "affordances" and "constraints". Affordances (Gibson, 1966) are opportunities for an organism that its environmental niche provides, depending on its current needs and goals. In active inference, affordances are expectations of action possibilities, thus expectations about possible agent-environment couplings (Ramstead et al., 2016). In biosemiotics, attributes of sign vehicles do not determine but provide affordances for a given interpretation (Deacon, 2022b). The term "constraints" in system theories refers to a limitation on the possible states of a system, thus a limitation of the exploitable state space. Affordances and constraints are closely linked, and this relationship will become important in interpreting the three chapters of this thesis.

Another fundamental insight from these two frameworks is that there exists a close in-principle correspondence between cognition and communication and the general relation between organisms and their environmental niche (which for some species comprises vocalised social interaction). Both biosemiotics and active inference share an agential perspective on organisms, in the sense that organisms act upon their Umwelt (Mildenberger & Herrmann, 2014), that is, the aspects of their environment that are relevant to them and thus correspond to their specific ecological niche (Vehkavaara & Sharov, 2017). This relevance means, in semiotic terms, that there is aboutness of the niche to an organism and that organisms and their niches correspond to each other. In that sense, they are complementary and in physical correlation, thus stand in higher-order iconic and indexical relationship to each other (in active inference, this correspondence has been modelled as mutual inference between organism and environment; Constant et al., 2018). In that sense, organisms' physical structure and dynamic relationships are about their world (Deacon, 2011). This aboutness *is* information; it reduces uncertainty by providing constraints to the agent-environment system.

Such agent-environment couplings can become culturally transmitted affordances (Ramstead et al., 2016) if their component relationships are interdependent and functionally integrated (Wagner & Tomlinson, 2022). One example would be a sequence of gestures resulting in the creation of a tool, which due to its function persists across generations, such as Acheulean tools. Other examples may include certain expressions and gestures individuating rituals from each other, birdsongs consisting of certain learned vocalisation patterns used in mating contexts, or gene expression patterns

differentiating cell types, each with their own lineage of descent (Wagner & Tomlinson, 2022). Wagner & Tomlinson and Deacon would see no in-principle differences between component relationships in biological or cultural entities (see also Laubichler & Renn, 2015; cf. Badcock et al., 2019; Deacon, 2010, 2022b; Ramstead et al., 2019; Vasil et al., 2020). Relevant for our purposes is that the components of vocalised interactions, in that sense, are sign relationships between social interaction and corresponding acoustics, that is, they are cognitive interpretations, and thus constraints. Therefore, they provide relations between their constituting behavioural actions that are shared between the interlocutors of the social interaction (see Sebanz & Knoblich, 2021). Thus, the constraints on the agent-environment system, such as the sign relationships in vocalised interaction, can become the subject of cultural transmission (Deacon, 2022a; Tomlinson, 2015). Such cultural "kinds" (Wagner & Tomlinson, 2022) can have evolutionary relevance: qua their component interdependence they can develop relatively autonomously, for example by becoming ritualised, and feedback into niche construction by providing new affordances (Tomlinson, 2015). One example of such a new affordance would be remembering vocal sequences if they consist of discrete pitch (see below).

Both frameworks can be summarised as dealing with mapping relations between environment and organism (including acoustic signals) by interpretation (biosemiotics) and inference (active inference). The mutual influence between environment and organism and the changes in these relations over time are inherent to both frameworks. Thus, they could guide our interpretation of spectrotemporal differences in speech and song regarding evolution. Both approaches can be empirically addressed using various methods, including dynamical systems methods (cf. Poincaré, 1890; Strogatz, 2018; Winfree, 1980). This will become relevant in section 4 of this discussion. In my view, both biosemiotics and active inference attempt to cover a greater range within the structuralist-functional/reductionist-holistic space proposed by Hallgrímsson and thus have integrative potential for the three chapters.

The current thesis has three important takeaways which I will outline below: First, the spectrotemporal structure that emerges from different requirements of predictability by social interaction (mutual inference) is itself about this social interaction. Thus, it inherits an indexical relationship to its dynamic origin. This is the fundamental assumption behind our framework from Chapter 1, but this also means that measurements of vocalisations can be used to infer what brought them about. Second, the emergent spectrotemporal structures themselves bear new affordances, both as opportunities for action and perceptively, for example by enabling indexical relationships and thus predictability, leading to effects on working memory. And third, when these kinds of vocalisation are shared among individuals of a group, they can become cultural practices that can feed back into the biological make-up of organisms and become evolutionarily relevant.

3 Interpreting the three chapters in the light of active inference and biosemiotics

Equipped with these concepts from biosemiotics and active inference, we can now return to the three thesis chapters. Our proposal in Chapter 1 suggests that spectrotemporal differences between song and speech result from a differentiation of vocalisations depending on the deployment along the proposed three dimensions. According to Deacon (2011), such a differentiation occurs by adding new constraints to the system's dynamics. These constraints are introduced by aboutness, that is, by semiosis. This means that the resulting spectrotemporal features of different modes and goals of vocal interaction (that is, the deployment along the three axes) are, in turn, indicative of the modes and goals that brought them about. In other words, they are themselves indices of their social interaction origin and can, therefore, cue their own top-down interpretation in this respect. This is interesting in two ways: first, the introduction of constraints is exactly what we would predict in Chapter 1 to happen through alphabet reduction, which allows us to reinterpret the results of Chapter 2, that is, enhanced memory for discrete pitches. Second, it describes a circular relationship between the spectrotemporal features as results of constraints (Chapter 1) and their role as acoustical cues for top-down interpretation (Chapter 3).

Regarding the reinterpretation of the results of Chapter 2, that is, enhanced memory for discrete pitches, the underlying proposed mechanism would be the following: if during vocalised social interaction shared and repeated attributes of mutual vocalisations are abstracted away and their likeness recognised (by attentional selection, the recognition of similar inferences and the learning of these likenesses), this would lead to the formation of icons that bear similarity to each other, that is of categories. In the case of choric interaction, this would result in discrete pitches, constraining the number of possible sequence elements (the "alphabet" in information-theoretic terms). These afford indexical relationships among each other, thereby introducing the possibility for prediction and learning of transition probabilities or, in active inference terms, beliefs about moment-to-moment state transitions. The effect of discrete pitch on memory would thus be an epiphenomenon on the social interaction goals, mediated by predictability, rather than being a direct result of selection. In this respect, it entails a shift from an adaptationist to an emergentist perspective. The enhanced similarity and the recognition of these indexical relations between contributions from different individuals as successful predictions (and aesthetically pleasing surprises) would be perceived as fulfilling the choric goal. That is, the respective interaction goal is constraining the space of possible vocalisations through successful predictions, reducing mutual uncertainty for interlocutors in interaction. In this respect, spectrotemporal features like discrete pitch directly reflect their predictability for the sake of social goals. Thus, they inherit the constraints that brought them about in their spectrotemporal structure. In other words, they are indices of their constraints (i.e., their choric/dialogic, propositional/aesthetic and repetition/novelty deployment) and thus represent affordances for their interpretation as such indices.

During joint action, these spectrotemporal affordances would be attentionally selected by precision-weighting of prediction errors and thereby, in a Gricean way, marked as relevant to the interlocutor (Ramstead et al., 2016). This marking would, in turn, enhance the likelihood of attentional selection and thus result in a self-sustained loop of mutual convergence on these affordances, incrementally contributing to shared expectations and learning of indexical relationships. Thus, during joint action, there is a constant cycle between constraints on the system and affordances provided by the resulting elements.

This idea reflects Wagner & Tomlinson's (2022) proposal of differentiation into "kinds" mentioned above that recursively leads to self-sustaining cultural systems that can be transmitted. For example, transition probabilities between discretised elements can form second-order icons as intervallic relationships through similarity, and those can again establish second-order indexical relationships. The same could happen with motifs and later melodies, incrementally and recursively applying the same cognitive operations, leading to what Tomlinson (2015) calls hyperindexicality, a hierarchical network of stable indexical relations that come to be shared among a community. This, in turn, provides new affordances, for example by exploiting the enhanced memory for cultural applications like ritual (see Ramstead et al., 2016) or to fulfil aesthetic goals by balancing surprise and uncertainty. The inherent stability of such a differentiated spectrotemporal cultural kind also enables vocalisations to differentiate again, for example, into hybrid forms like poetry under certain deployments along the proposed three axes. This adds a processual view to the analytical decomposition approach brought forward by notions of musicality (Honing, 2018b) and the faculty of language (Fitch et al., 2005).

This relation between cultural inheritance, affordances and constraints is also reflected in Chapter 3, which informs us in three ways. First, spectrotemporal lateralisation itself provides constraints on spectrotemporal feature emergence. Suppose we see the modulation power spectrum framework as a state space of acoustical outputs that cross-culturally differentiates between song and speech (Albouy et al., 2023). In that case, we can ask how constraints on dynamics of vocal interaction emerge to limit certain kinds of vocal interaction to certain spaces in the modulation power spectrum domain, or in other words, why different deployments along our proposed dimensions in Chapter 1 exploit spectrotemporal lateralisation in different ways. We can assume exploitation (at least incrementally over evolutionary time) because evidence for spectrotemporal lateralisation has been found in species of several clades, including bats, dogs, mice and zebra finches (Neophytou et al., 2022; Ratcliffe & Reby, 2014; Ruijssevelt et al., 2017; Rybalko et al., 2010; Washington & Tillinghast, 2015; see also Flinker et al., 2019). This suggests spectrotemporal lateralisation is not domain-specific for human speech and song (cf. Norman-Haignere et al., 2015). Instead, it could be a general response of the auditory system related to the uncertainty principle inherent to wave-like

systems, which leads to imprecision in either the temporal or spectral acoustical resolution (cf. Oppenheim & Magnasco, 2013).

Secondly, we can understand why choric and dialogic interaction exploit this lateralisation differently based on this insight. Why would, for example, a reduction in the alphabet, which we would predict to occur from choric interaction, capitalise on spectral processing, resulting in discrete pitches, and on regularity in the form of the establishing of a metrical grid rather than on fast temporal modulations in the signal? One possibility for converging on discrete pitches and slow regularised temporal structure might be a general slowing down of vocal phrases in choric modes for the sake of mutual predictability, naturally coming into the focus of right-lateralised processing. This enables further exploitation of general auditory physiology and auditory scene analysis, as Trainor (2018) suggested in her review. A second aspect is the moment-to-moment prediction of time in choric interaction, which leads to regularity. Similarly, the famous speech-to-song illusion, in which by repetition (thereby introducing regularity) spoken phrases shift from being perceived as speech to being perceived as song, seems to capitalise on right hemispheric processing (see Hymers et al., 2015). A third aspect might be the absence of a propositional goal. As we have seen in Chapter 1, propositional goals lead to the subordination of layers of the speech stream to enhance the probability of the correct semantic inference (e.g., by intonation changes or word choice), and slight deviations can lead to different inferences (as in the case of "Happy Birthday" sung in minor key). This suggests that inference on referential meaning must be precise, which requires a sufficient signal-to-noise ratio, suggesting the avoidance of masking effects (Yost, 1991; see also Ravignani et al., 2019). If this pressure is absent under a choric goal, the precision of inference only needs to go so far as to whether an element of a vocalisation relates to another event iconically or indexically, and more degrees of freedom are possible within this constrained domain.

A propositional goal may also explain a discretisation into phonemes. According to Ladd (2014) discretisation into phonemes would result from discretising a continuum to keep signal elements discriminable when the call repertoire size increases. This suggests that less differentiated holistic calls are propositional in the first place (through indexical and iconic world grounding, Deacon, 2022b). Recombination of such calls to indicate different referents (indexically) seems to occur across different clades of vocalising animals (Engesser & Townsend, 2019). The propositional referentiality itself might add a constraint: recalling semantic content when hearing multiple concurrent speech streams seems to require attending to one stream and is impaired for unattended streams (e.g. Mackay, 1973). Thus, it seems there is limited possibility for more than one trajectory through the semantic conceptual space in cognitive processing, thus for correct inferences. Therefore, there seem to be cognitive constraints limiting the processing of multiple semantic streams in parallel, leading to pressure to engage in turn-taking. If such a dialogic mode is established, there is limited

masking; thus, the differentiation into phonemes is not selected against (cognitively). Therefore, dialogic interaction is required if the goal is to find a common trajectory through the semantic space mediated by phoneme-based symbols (words). In choric contexts, finding a common trajectory would mean that interlocutors perceive the joint action outcome as coherent (which requires a shared and culturally acquired concept of what constitutes a joint acoustic outcome, e.g., singing in tune and in time). Since elements are primarily indexically related rather than symbolically, there is less constraint on this outcome, for example, in musical improvisation when several outcomes are possible without compromising the coherence goal. This leads to the interesting observation that with propositional goals, a higher constraint on the correct inference is achieved by less constraint at the spectrotemporal level. In contrast, in the indexical (non-propositional) space, less constrained inference of coherence is achieved by higher spectrotemporal constraint. We can, therefore, see that discretisation resulting in phonemes would have a different origin than discretisation resulting in discrete pitch or metrical regularity, again indicating that spectrotemporal features can show traces of their origin (and thus be indexical in that sense) rather than being arbitrary conventions.

We can learn a third aspect from study 3 in the light of semiotics: spectral and temporal acoustic information can be regarded as affordances to interpret the sounds iconically in a song-like or speech-like way (that is similar to other instances of song and speech), depending on whether the task goal is to recognise melodies or sentences. Temporal cues are needed to understand speech, and spectral cues are required to understand melodies (Albouy et al., 2020). Thus, the relationship between spectrotemporal structure and the ability to guide a specific interpretation as musical or speech-related appears to be not ambiguous conventionalisation (a symbol), but a culturally inherited (Boyd & Richerson, 2005) indexical relationship. In other words, the spectrotemporal structure is an index both of the social interaction and lateralisation constraints that brought them about, and of the culturally inherited iconic interpretation as song-like or speech-like. This indexicality, however, does not mean that interpretation is deterministic. In Chapter 3, we observed topological neural changes in clustering only when those acoustical cues were degraded that were crucial for the current task. The topological changes may indicate enhanced attentional precision to cope with a worse signal-to-noise ratio (due to higher acoustic degradation) when the acoustical cues are task-relevant. This illustrates that even when spectrotemporal cues index the mode of social interaction which brought them about, their inference as musical or language-like still is probabilistic and not deterministic, in line with the active inference framework (see also Graber et al., 2017). The logic of probabilistic inference on spectrotemporal cues is also consistent with the results of Chapter 2, where we observed enhanced memory also for intermediate, ambiguous stimuli (see also Pfordresher & Brown, 2017), and where some participants reported tapping rhythmically to aid their memory although we did not introduce any regularity in the stimuli.

To summarise, we can interpret spectrotemporal differences between song and speech as results of differentiations generated through predictability, attentional abstraction, and memory in creating interpretative relations (semiosis) and selection by (active) inference. Spectrotemporal differences exploit general auditory traits like spectrotemporal lateralisation under different deployments of social interaction, such as choric and dialogic goals. Resulting spectrotemporal features like discrete pitch are both indices of the constraints that led to their emergence and affordances, for example to remember vocal phrases or for aesthetic exploitation. They also afford their culturally learned interpretation as song-like or speech-like. This interpretation as song or speech is still probabilistic. This reading of the three chapters represents a hypothesis that can be empirically addressed (see next section).

Since social interaction is not confined to the audio-acoustic domain, a resulting question is how such a relationship between spectrotemporal structure and its origin in social interaction is realised in other modalities, like in body and gestural movement. According to Deacon (2011), constraints on system dynamics can become displaced and thus be realised in different modalities. This would predict that either similarities in body movement between music and language (dance and sign, for example) emerge in similar ways as song and speech in the vocal domain, or that choric and dialogic dynamics first emerge in one of the two domains and become displaced to the other. This can be empirically tested if such dynamics could be measured with similar metrics across different modalities. One way would be to directly track the predictive dynamics by using information-theoretic models (see below). In line with this, the process of differentiation, understood as dynamics that individuate (e.g. into song-like and speech-like vocalisations) while staying self-similar over time (iconic, to aid their inference as song-like and speech-like), has been proposed as representing a formal principle applying to development, biological and cultural evolution, and cognition in very similar ways (Wagner & Tomlinson, 2022; Deacon, 2013). This opens the possibility to measure and analyse such dynamics and derive ontological, evolutionary and cognitive insights simultaneously (see e.g. Badcock et al., 2019; Constant et al., 2018; Veissière et al., 2019). Below I will suggest how this could be done regarding spectrotemporal differences in music and language.

4 Future research: Complementing analytical decomposition with recomposition

The strategy to analytically decompose music and language into underlying traits (Fitch, 2015; Honing, 2018a) and investigate those has been quite successful (e.g. Kotz et al., 2018). A possible next step could be complementing such an analytical decomposition with functional recomposition. Recomposition means to analyse and understand how a system's different components and levels interact in their activities over time, within and between levels. This interaction is often nonlinear. Such an approach yields a model of the system's overall behaviour and how it changes over time in its

configuration. This supports a causal mechanistic understanding of the system (Strogatz, 2018; see also DiFrisco & Jaeger, 2020). This can be done by utilising methods from dynamical systems theory, also known as nonlinear dynamics or popularised as "Chaos theory" (Gleick, 1987; Poincaré, 1890; Strogatz, 2018; Winfree, 1980). In our case, this would mean investigating how traits interacted over different time scales to yield the capacities for music and language. Equipped with the biosemiotic and active inference frameworks, we can sketch a possible research agenda that follows up on the three chapters and takes a step towards recomposition.

First, Chapter 1 can be implemented empirically by investigating whether different modes of interaction, according to the dimensions we suggested in Chapter 1, would lead to an emergence of spectrotemporal structure. Participant dyads could be asked to imitate ambiguous sounds (for example, based on near-universal acoustic features that have been found cross-culturally for song and speech, see e.g. Ozaki et al., 2023) in either choric or dialogic modes, and the emerging spectrotemporal changes could be measured acoustically. This can be complemented by adding constraints regarding propositional semantics (i.e. conveying a message), repetition (i.e. imitating phrases repeatedly by themselves rather than in a dyad) or cultural transmission (cf. Mesoudi, 2021). This way, the framework can be tested and revised, and the geometrical relations of the framework's dimensions (and their changes over time) can be tracked by their correlative relationships. Such a setup could be transferred cross-culturally, cross-modally, and to children. Constraints on predictability should for example also apply to body movement, concerning dance or gestures. Under choric movement, we would therefore expect high moment-to-moment predictability through slower and more regularly timed gestures, while a dialogic mode would lead to marking of turn-inflection points through higher predictability of gestures at phrase endings. If many spectrotemporal features of music and language can be explained by emergence through different modes of social interaction, this shifts the explanatory requirements regarding evolution to the motivation to engage in these modes in the first place. At such a point, it might be valuable to investigate non-human species. Many species naturally engage in duetting and chorusing, such as various birds (e.g., wrens), primates (e.g., gibbons) or canids (e.g., howling wolf packs). By making use of annotated field recordings, the relation of spectrotemporal features and modes of interaction can be analysed, and we would expect that choric and dialogic modes would yield spectrotemporal features similar as in humans. In an experimental approach, vocal production learners like parrots, who also entrain body movements spontaneously to sound (Patel et al., 2009), could be trained by conditioning procedures to engage in choric or dialogic vocal interaction. Analysing their vocalisations can reveal to what extent spectrotemporal feature emergence depends on human brains and culture or represents general solutions and whether the proposed underlying mechanism of predictability can account for both.

How could the predictability of vocalisations be measured across species and linked to evolution? We can rely on dynamical systems methods (Poincaré, 1890; Strogatz, 2018; Winfree, 1980). If spectrotemporal changes inherit their dynamical structure, that is, their constraints, and if it is the constraints that get passed on over ontological and evolutionary times (Deacon, 2011), it seems natural to attempt to measure these dynamics. In other words, we can try to measure how and under which predictive constraints spectrotemporal features of music and language emerge and change over different time scales. Dynamical systems theory has long drawn on phase space analysis to this end. A phase portrait of a system can be constructed by tracking variations of system variables over time. Each variable forms a dimension of the phase space such that a trajectory through phase space represents the system's dynamic behaviour (its flow). Phase space analysis then allows for the investigation of regularities, for example regarding areas (and thus states) that are revisited by the system or towards which the system tends to converge (attractors). Information and entropy as metrics of predictability and uncertainty reduction in language and music can be measured using computational models like IDyOM and IDyOT (Forth et al., 2016; Pearce & Wiggins, 2012). The resulting trajectories of information and uncertainty across unfolding sequences can then be used to construct phase portraits and measure system dynamics change over time in dyadic interaction. Since such models are based on learning transition probabilities between elements of sequences, they are, in principle, not confined to specific acoustics of human song and speech. Entropy and surprisal could thus be used as cross-modal state variables of joint action to construct phase portraits comparatively between different signalling systems (like instances of music, language, or of hybrid forms) within and across species.

Another exciting insight from phase space analysis regarding evolution is that it can potentially track evolutionary transitions. Applying phase space analysis to gene regulatory networks, Jaeger & Crombach (2012) found that evolutionary change can only occur in four ways (Jaeger et al., 2012). Similarly, phase space analyses of predictive dynamics in communication can be analysed regarding topographical changes related to these transitions. Jaeger & Monk (2021a) also describe a method to functionally decompose the behaviour of a system, thus its dynamics, into constituting modules of activity. Each dynamical module is characterised by a particular dynamical regime with a specific phase space topology. Therefore, dynamical modules can vary independently during evolution, influence evolvability through their differential sensitivity to system parameters, and have, for this reason, been suggested as fundamental units both for ontogenesis and evolution (Jaeger & Monk, 2021a). This is in line with Deacon's idea of passing on constraints (Deacon, 2011).

The tacit assumption in the argumentation above is that music and language differentiated from a common precursor, at least regarding their spectrotemporal structure. This assumption, in turn, provides a working hypothesis for empirical investigation. In this sense, spectrotemporal

differentiation would be the outcome of social interaction dynamics in the audiovocal modality, suggesting that comparable outcomes may occur in other modalities like movement. A phase space analysis based on surprisal and uncertainty, complemented by analytical decomposition into modules, could clarify which components of the dynamical interaction are shared among species and modalities and which ones are derived (see Tomlinson & Wagner, 2022) and, therefore, be a means to reconstruct biocultural evolutionary processes. This approach might reveal links between communicative systems that are not captured by comparing only acoustics. For example, some of such dynamical modules characteristic for language and/or music in humans could be conserved and found in other animal communication systems, while contexts like social behaviour and related cognitive capacities might lead to different overall outcomes in the resulting communication systems. It may be possible, for example, to find certain dynamical modules basic for both choric and dialogic interaction in movement interaction among great apes, suggesting these provided a scaffold for vocalisation in human ancestors. Other modules comparable to some in human musical interaction may also be found in vocal production learning birds, suggesting convergent interaction dynamics. Traits like vocal production learning, entrainment or social group structure would represent parameter values on the phase portraits that can be manipulated by investigating different model species. Such a comparative research program can comprise many species engaging in duetting or chorusing behaviour. Thus, such an approach could help to falsify or refine the differentiation hypothesis.

5 Conclusion

In this thesis, I attempted to gain some insights into why music and language, although exploiting the same acoustic dimensions in the auditory-acoustic modality, show major differences in spectrotemporal features. Conceptualising music and language as two instances of bioculturally evolved communication systems that structure social interaction, I heuristically followed Marr's and Tinbergen's levels and an analytical decomposition approach based on design features (Chapters 1 and 2) and spectrotemporal modulations (Chapter 3). I thereby focused on comparing the vocal outcomes of music and language - song and speech, respectively.

The three chapters of the thesis revealed the usefulness of a reverse-engineering approach to gain insights into the emergence of spectrotemporal structure and the differential social and cognitive constraints that might be involved. They also showed how social behaviour, cognition and acoustic structure can be linked via information theory and how such design features like discrete pitch can fulfil social requirements by exploiting cognitive functions like auditory memory. Finally, they demonstrated how functional connectivity networks in the brain show a differential architecture at different levels in response to interactions of spectrotemporal processing and attention to musical or linguistic aspects of sung speech. The three chapters, however, also showed conceptual differences

and indicated the limitations of Marr's and Tinbergen's levels as guiding heuristics for their interpretation. Therefore, in this discussion, I used the active inference and biosemiotics frameworks to come to an integrated perspective on the three thesis chapters.

This interpretation suggested spectrotemporal differences between song and speech to be the results of differentiations generated by establishing interpretative relations (semiosis) and by inference depending on task goals. Predictability, attentional abstraction, and memory are underlying mechanisms to this end. Spectrotemporal differences exploit general auditory traits like spectrotemporal lateralisation under different deployments of social interaction, such as choric and dialogic goals. Spectrotemporal features like discrete pitch inherit the constraints that brought them about and, therefore, are indices of these conditions. At the same time, they represent affordances for exploiting epiphenomenal effects like enhanced memory for vocal phrases and are fundamental indices for a probabilistic and task-dependent interpretation as song-like or speech-like. Constraints and affordances in spectrotemporal features thus stand in an interdependent relationship that enables cultural transmission.

In this sense, the three chapters suggest empirical future work based on Chapter 1, both within humans and the audio-acoustic domain, as well as comparatively across species and in other domains like movement. This is aided by utilising metrics for information/surprisal and entropy/uncertainty to create phase portraits of social interaction dynamics. This abstract space will support the search for fundamental dynamic modules and a better understanding of evolutionary transitions. This synergistic perspective can complement the analytical decomposition of music and language with a recomposition to yield causal mechanistic understanding.

References

- Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, 367(6481), 1043–1047. <https://doi.org/10.1126/science.aaz3468>
- Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., & Zatorre, R. J. (2023). Spectro-temporal acoustical markers differentiate speech from song across cultures. *BioRxiv*.
- Awh, E., Belopolsky, A. V., & Theeuwes, J. (2013). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*, 16(8), 437–443. <https://doi.org/10.1016/j.tics.2012.06.010>.Top-down
- Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., & Hohwy, J. (2019). *The hierarchically mechanistic mind : an evolutionary systems theory of the human brain , cognition , and behavior*. 1319–1351.
- Barbieri, M. (2008). What is Biosemiotics? *Biosemiotics*, 1(1), 1–3. <https://doi.org/10.1007/s12304-008-9009-1>

- Bateson, P., & Laland, K. N. (2013). Tinbergen's four questions: An appreciation and an update. *Trends in Ecology and Evolution*, 28(12), 712–718. <https://doi.org/10.1016/j.tree.2013.09.013>
- Behne, T., Carpenter, M., & Tomasello, M. (2014). Young children create iconic gestures to inform others. *Developmental Psychology*, 50(8), 2049–2060. <https://doi.org/10.1037/a0037224>
- Bendixen, A. (2014). Predictability effects in auditory scene analysis: A review. *Frontiers in Neuroscience*, 8(8 MAR), 1–16. <https://doi.org/10.3389/fnins.2014.00060>
- Boyd, R., & Richerson, P. J. (2005). *The Origin and Evolution of Cultures*. Oxford University Press.
- Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The Emperor Is Naked: Replies to commentaries on the target article. *Behavioral and Brain Sciences*, 45, 1–38. <https://doi.org/10.1017/S0140525X22000656>
- Chang, C., & Chen, J. E. (2021). Multimodal EEG-fMRI: Advancing insight into large-scale human brain dynamics. *Current Opinion in Biomedical Engineering*, 18, 100279. <https://doi.org/10.1016/j.cobme.2021.100279>
- Constant, A., Ramstead, M. J. D., Veissie, S. P. L., Campbell, J. O., Friston, K. J., & Constant, A. (2018). A variational approach to niche construction. *Journal of the Royal Society Interface*, 15(20170685).
- Cooperrider, K. (2020). Fifteen ways of looking at a pointing gesture. *PsyArXiv*. <http://psyarxiv.com/2vxft/>
- Deacon, T. W. (2010). A role for relaxed selection in the evolution of the language capacity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(SUPPL. 2), 9000–9006. <https://doi.org/10.1073/pnas.0914624107>
- Deacon, T. W. (2011). *Incomplete nature: How mind emerged from matter*. WW Norton & Company.
- Deacon, T. W. (2022a). A degenerative process underlying hierarchic transitions in evolution. *BioSystems*, 222(August), 104770. <https://doi.org/10.1016/j.biosystems.2022.104770>
- Deacon, T. W. (2022b). Steps to a Semiotic Cognitive Neuroscience. In *The Routledge Handbook of Semiosis and the Brain*.
- DiFrisco, J., & Jaeger, J. (2020). Genetic Causation in Complex Regulatory Systems: An Integrative Dynamic Perspective. *BioEssays*, 42(6), 1–10. <https://doi.org/10.1002/bies.201900226>
- Engesser, S., & Townsend, S. W. (2019). Combinatoricity in the vocal systems of nonhuman animals. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(4), e1493. <https://doi.org/https://doi.org/10.1002/wcs.1493>
- Fitch, W. T. (2015). *Four principles of bio-musicology*. 10(3), 197–202.
- Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition*, 97(2), 179–210. <https://doi.org/10.1016/j.cognition.2005.02.005>
- Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature Human Behaviour*, 3(4), 393–405. <https://doi.org/10.1038/s41562-019-0548-z>

- Forth, J., Agres, K., Purver, M., & Wiggins, G. A. (2016). Entraining IDyOT : Timing in the Information Dynamics of Thinking. *Frontiers in Psychology*, 7(1575), 1–19. <https://doi.org/10.3389/fpsyg.2016.01575>
- French, R. M., & Thomas, E. (2015). Interactive effects of explicit emergent structure: A major challenge for cognitive computational modeling. *Topics in Cognitive Science*, 7(2), 206–216. <https://doi.org/10.1111/tops.12135>
- Friston, K. J., & Frith, C. D. (2015). Active inference, Communication and hermeneutics. *Cortex*, 68(Kelso 2012), 129–143. <https://doi.org/10.1016/j.cortex.2015.03.025>
- García, A. M., & Ibáñez, A. (2022). *The Routledge Handbook of Semiosis and the Brain* (A. M. García & A. Ibáñez (eds.)). Routledge.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*.
- Gleick, J. (1987). *Chaos: Making a new science*. Viking.
- Gould, S. J., & Lewontin, R. C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society B: Biological Sciences*, 205(1161), 581–598.
- Graber, E., Simchy-Gross, R., & Margulis, E. H. (2017). Musical and linguistic listening modes in the speech-to-song illusion bias timing perception and absolute pitch memory. *The Journal of the Acoustical Society of America*, 142(6), 3593–3602. <https://doi.org/10.1121/1.5016806>
- Honing, H. (2018a). *Musicality as an upbeat to music : Introduction and research agenda*.
- Honing, H. (2018b). *The origins of musicality*. The MIT Press. [https://doi.org/10.1016/s0163-6383\(84\)80425-7](https://doi.org/10.1016/s0163-6383(84)80425-7)
- Hymers, M., Prendergast, G., Liu, C., Schulze, A., Young, M. L., Wastling, S. J., Barker, G. J., & Millman, R. E. (2015). Neural mechanisms underlying song and speech perception can be differentiated using an illusory percept. *NeuroImage*, 108, 225–233. <https://doi.org/10.1016/j.neuroimage.2014.12.010>
- Jacob, F. (1977). Evolution and tinkering. *Science*, 196(4295), 1161–1166. <https://doi.org/10.1126/science.860134>
- Jaeger, J., & Crombach, A. (2012). Life's attractors understanding developmental systems through reverse engineering and in silico evolution. In *Advances in Experimental Medicine and Biology* (Vol. 751, pp. 93–119). https://doi.org/10.1007/978-1-4614-3567-9_5
- Jaeger, J., Irons, D., & Monk, N. (2012). The Inheritance of Process: A Dynamical Systems Approach. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 318(8), 591–612. <https://doi.org/10.1002/jez.b.22468>
- Jaeger, J., & Monk, N. (2021a). Dynamical Modularity of the Genotype-Phenotype Map. In A. Crombach (Ed.), *Evolutionary Systems Biology* (pp. 245–280). Springer. https://doi.org/10.1007/978-3-030-71737-7_11
- Jaeger, J., & Monk, N. (2021b). Dynamical modules in metabolism, cell and developmental biology. *Interface Focus*, 11(3). <https://doi.org/10.1098/rsfs.2021.0011>
- Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The Evolution of Rhythm Processing. *Trends in Cognitive Sciences*, 22(10), 896–910. <https://doi.org/10.1016/j.tics.2018.08.002>

- Ladd, D. R. (2014). *Simultaneous structure in phonology*. Oxford University Press.
- Laland, K. N., & Brown, G. R. (2011). *Sense and nonsense: Evolutionary perspectives on human behaviour*. Oxford University Press, USA.
- Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., Jablonka, E., & Odling-Smee, J. (2015). The extended evolutionary synthesis: Its structure, assumptions and predictions. *Proceedings of the Royal Society B: Biological Sciences*, 282(1813). <https://doi.org/10.1098/rspb.2015.1019>
- Laubichler, M. D., & Renn, J. (2015). Extended evolution: A conceptual framework for integrating regulatory networks and niche construction. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 324(7), 565–577. <https://doi.org/10.1002/jez.b.22631>
- Mackay, D. G. (1973). ASPECTS OF THE THEORY OF COMPREHENSION, MEMORY AND ATTENTION. *Quarterly Journal of Experimental Psychology*, 25, 22–40.
- Magnon, V., & Corbara, B. (2022). When the “satisficing” is the new “fittest”: how a proscriptive definition of adaptation can change our view of cognition and culture. *Science of Nature*, 109(5). <https://doi.org/10.1007/s00114-022-01814-9>
- Massimi, M. (2022). Perspectival Realism. In *Encyclopedia of Educational Philosophy and Theory*. Oxford University Press. <https://doi.org/10.1093/oso/9780197555620.001.0001>
- Mcgettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: what’s wrong, what’s right, and what’s left? *Trends in Cognitive Sciences*, 16(5), 269–276. <https://doi.org/10.1016/j.tics.2012.04.006>Cortical
- Mesoudi, A. (2021). Cultural selection and biased transformation: Two dynamics of cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1828). <https://doi.org/10.1098/rstb.2020.0053>
- Mildenberger, F., & Herrmann, B. (2014). *Uexküll: Umwelt und Innenwelt der Tiere*. Springer-Verlag.
- Milette-Gagnon, A., Veissière, S. P. L., Friston, K. J., & Ramstead, M. J. D. (2022). An active inference approach to semiotics: A variational theory of signs. In *The Routledge Handbook of Semiosis and the Brain* (pp. 49–65). Routledge.
- Miton, H., & Charbonneau, M. (2018). Cumulative culture in the laboratory: Methodological and theoretical challenges. *Proceedings of the Royal Society B: Biological Sciences*, 285(1879). <https://doi.org/10.1098/rspb.2018.0677>
- Neophytou, D., Arribas, D. M., Arora, T., Levy, R. B., Park, I. M., & Oviedo, H. V. (2022). Differences in temporal processing speeds between the right and left auditory cortex reflect the strength of recurrent synaptic connectivity. *PLoS Biology*, 20(10), 1–20. <https://doi.org/10.1371/journal.pbio.3001803>
- Norman-Haignere, S., Kanwisher, N. G., Mcdermott, J. H., 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>
- O’Malley, M. A., Soyer, O. S., & Siegal, M. L. (2015). A Philosophical Perspective on Evolutionary Systems Biology. *Biological Theory*, 10(1), 6–17. <https://doi.org/10.1007/s13752-015-0202-6>

- Oppenheim, J. N., & Magnasco, M. O. (2013). Human time-frequency acuity beats the fourier uncertainty principle. *Physical Review Letters*, 110(4), 1–5. <https://doi.org/10.1103/PhysRevLett.110.044301>
- Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J., Benetos, E., Proutskova, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., & Savage, P. E. (2023). Globally songs are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]. *Peer Community In Registered Reports*, Preprint:
- Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, 7(1), 1–21. <https://doi.org/10.1038/s41598-017-15249-0>
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference. The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press.
- Patel, A. D. (2023). Human Musicality and Gene- Culture Coevolution: Ten Concepts to Guide Productive Exploration. In E. H. Margulis, P. Loui, & D. Loughridge (Eds.), *The Science-Music Borderlands* (pp. 15–37). The MIT Press. <https://doi.org/10.7551/mitpress/14186.001.0001>
- Patel, A. D., Iversen, J. R., Bregman, M. R., & Schulz, I. (2009). Experimental Evidence for Synchronization to a Musical Beat in a Nonhuman Animal. *Current Biology*, 19(10), 827–830. <https://doi.org/10.1016/j.cub.2009.03.038>
- Pearce, M. T., & Wiggins, G. A. (2012). Auditory Expectation: The Information Dynamics of Music Perception and Cognition. *Topics in Cognitive Science*, 4(4), 625–652. <https://doi.org/10.1111/j.1756-8765.2012.01214.x>
- Peirce, C. S. (1994). *Collected Papers of Charles Sanders Peirce* (J. Deely (ed.)). Harvard University Press. <https://doi.org/10.1038/131639b0>
- Pfordresher, P. Q., & Brown, S. (2017). Vocal mistuning reveals the origin of musical scales. *Journal of Cognitive Psychology*, 29(1), 35–52. <https://doi.org/10.1080/20445911.2015.1132024>
- Pigliucci, M., & Kaplan, J. (2000). The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later. *Trends in Ecology and Evolution*, 15(2), 66–70.
- Poincaré, H. (1890). SUR LE PROBLEME DES TROIS CORPS ET LES EQUATIONS DE LA DYNAMIQUE. *Acta Mathematica*, 13, 1–270.
- Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews*, 39, 49–72. <https://doi.org/10.1016/j.plrev.2021.09.001>
- Ramstead, M. J. D., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, 31, 188–205. <https://doi.org/10.1016/j.plrev.2018.12.002>
- Ramstead, M. J. D., Hesp, C., Tschantz, A., Smith, R., Constant, A., & Friston, K. (2021). Neural and phenotypic representation under the free-energy principle. *Neuroscience and Biobehavioral Reviews*, 120, 109–122. <https://doi.org/10.1016/j.neubiorev.2020.11.024>
- Ramstead, M. J. D., Veissière, S. P. L., & Kirmayer, L. J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in Psychology*, 7(JUL), 1–21. <https://doi.org/10.3389/fpsyg.2016.01090>

- Ratcliffe, V. F., & Reby, D. (2014). Orienting asymmetries in dogs' responses to different communicatory components of human speech. *Current Biology*, 24(24), 2908–2912. <https://doi.org/10.1016/j.cub.2014.10.030>
- Rauss, K., & Pourtois, G. (2013). What is bottom-up and what is top-down in predictive coding. *Frontiers in Psychology*, 4(MAY), 1–8. <https://doi.org/10.3389/fpsyg.2013.00276>
- Ravignani, A., Verga, L., & Greenfield, M. D. (2019). *Interactive rhythms across species: the evolutionary biology of animal chorusing and turn-taking*. 1453, 12–21. <https://doi.org/10.1111/nyas.14230>
- Ruijssevelt, L. Van, Washington, S. D., Hamaide, J., Verhoye, M., Keliris, G. A., & Van der Linden, A. (2017). Song processing in the zebra finch auditory forebrain reflects asymmetric sensitivity to temporal and spectral structure. *Frontiers in Neuroscience*, 11(OCT), 1–10. <https://doi.org/10.3389/fnins.2017.00549>
- Rybalko, N., Šuta, D., Popelář, J., & Syka, J. (2010). Inactivation of the left auditory cortex impairs temporal discrimination in the rat. *Behavioural Brain Research*, 209(1), 123–130. <https://doi.org/10.1016/j.bbr.2010.01.028>
- Samuelson, L. K., Jenkins, G. W., & Spencer, J. P. (2015). Grounding cognitive-level processes in behavior: The view from dynamic systems theory. *Topics in Cognitive Science*, 7(2), 191–205. <https://doi.org/10.1111/tops.12129>
- Sebanz, N., & Knoblich, G. (2021). Progress in Joint Action Research. *Current Directions in Psychological Science*, 30(2).
- Strogatz, S. H. (2018). *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press.
- Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences of the United States of America*, 113(16), 4530–4535. <https://doi.org/10.1073/pnas.1523631113>
- Tomlinson, G. (2015). *A million years of music: The emergence of human modernity*. MIT Press.
- Tomlinson, G. (2016). Sign, affect, and musicking before the human. *Boundary 2*, 43(1), 143–172. <https://doi.org/10.1215/01903659-3340673>
- Tomlinson, G. (2022, February 25). *Biocultural Evolution of Music and the Illusion of “Selection-For.”* [Video]. YouTube. <https://www.youtube.com/watch?v=0wsyDAx8xMo>
- Tooby, J., & Cosmides, L. (2005). Conceptual Foundations of Evolutionary Psychology. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (1st ed.). John Wiley & Sons, Inc. <https://doi.org/10.1007/s12041-017-0789-4>
- Trainor, L. J. (2018). The origins of music: Auditory scene analysis, evolution, and culture in musical creation. In H. Honing (Ed.), *The origins of musicality* (pp. 81–112). MIT Press.
- van der Burght, C. L., Goucha, T., Friederici, A. D., Kreitewolf, J., & Hartwigsen, G. (2019). Intonation guides sentence processing in the left inferior frontal gyrus. *Cortex*, 117, 122–134. <https://doi.org/10.1016/j.cortex.2019.02.011>
- Vasil, J., Badcock, P. B., Constant, A., Friston, K., & Ramstead, M. J. D. (2020). A World Unto Itself: Human Communication as Active Inference. *Frontiers in Psychology*, 11(March), 1–26. <https://doi.org/10.3389/fpsyg.2020.00417>

- Vehkavaara, T., & Sharov, A. (2017). Constructive Aspects of Biosemiotics. *Biosemiotics*, 10(2), 145–156. <https://doi.org/10.1007/s12304-017-9304-9>
- Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2019). Thinking Through Other Minds: A Variational Approach to Cognition and Culture. *Behavioral and Brain Sciences*, 1–97. <https://doi.org/10.1017/S0140525X19001213>
- Wagner, G. P., & Tomlinson, G. (2022). Extending the Explanatory Scope of Evolutionary Theory: The Origination of Historical Kinds in Biology and Culture. *Philosophy, Theory, and Practice in Biology*, 14(0), 1–25. <https://doi.org/10.3998/ptpbio.2095>
- Washington, S. D., & Tillinghast, J. S. (2015). Conjugating time and frequency: Hemispheric specialization, acoustic uncertainty, and the mustached bat. *Frontiers in Neuroscience*, 9(APR), 1–14. <https://doi.org/10.3389/fnins.2015.00143>
- Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings*. Harvard University Press.
- Winfrey, A. T. (1980). *The geometry of biological time* (Vol. 2). Springer.
- Yost, W. A. (1991). Auditory image perception and analysis: The basis for hearing. *Hearing Research*, 56(1–2), 8–18. [https://doi.org/10.1016/0378-5955\(91\)90148-3](https://doi.org/10.1016/0378-5955(91)90148-3)

ABSTRACT

Language and music are two of the most remarkable human phenomena, abundant across cultures and combining characteristics rare within the realm of animal communication. Although both music and language share certain characteristics in structure, cognitive processing, and significance for human social interaction, they differ in others, particularly in spectrotemporal features apparent in their vocal forms as song and speech. In my thesis, I investigated the origin of these differences in three studies which form three Chapters. Following Marr's and Tinbergen's frameworks as guiding heuristics, I used an analytical decomposition approach based on design features (Chapters 1 and 2) and spectrotemporal modulations (Chapter 3). I investigated predictability, attention and memory in a theoretical framework, a behavioural study, and an fMRI experiment. I conclude that spectrotemporal differences can be explained as emerging from constraints on predictability under different goals in social interaction. Resulting features such as discrete pitch then afford enhanced fulfilment of task goals such as remembering vocal phrases. They also provide probabilistic cues that differentially modulate functional connectivity networks for the sake of task-dependent interpretation. Drawing on the frameworks of active inference and biosemiotics, I suggest that spectrotemporal differences between song and speech are indices of the predictive constraints that brought them about, entailing affordances which enable cultural transmission. Measuring these constraints across species and modalities with methods from evolutionary systems biology can complement the analytical decomposition of music and language with a recomposition approach, enabling understanding of the interplay of cognition, communication, and evolution.

ZUSAMMENFASSUNG

Jede menschliche Kultur besitzt Sprache und Musik, und beide Phänomene vereinen Merkmale, die im Kommunikationsspektrum von Tieren selten anzutreffen sind, etwa in Struktur, kognitiver Verarbeitung und Bedeutung für zwischenmenschliche soziale Interaktion. Zugleich unterscheiden sich Musik und Sprache insbesondere in spektrotemporalen Merkmalen, wie vorallem in Gesang und gesprochener Sprache deutlich wird. In meiner Dissertation habe ich den Ursprung dieser Unterschiede in drei Studien untersucht, die drei Kapitel bilden. Basierend auf den Heuristiken von Marr und Tinbergen habe ich mich auf grundlegende Bestandteile von Sprache und Musik fokussiert: Designmerkmale (Kapitel 1 und 2) und spektrotemporale Modulationen (Kapitel 3). Ich habe kognitive Vorhersagbarkeit, Aufmerksamkeit und Gedächtnis in einer theoretischen Studie, einer Verhaltensstudie und einem fMRT-Experiment untersucht. Meine Arbeit legt den Schluss nahe, dass spektrotemporale Unterschiede durch Einschränkungen der kognitiven Vorhersagbarkeit erklärbar sind, die abhängig von unterschiedlichen Zielen sozialer Interaktion variieren. Daraus resultierende spektrotemporale Merkmale wie diskrete Töne ermöglichen dann verbesserte kognitive Verarbeitung, z.B. erhöhte Merkbarkeit von vokalen Phrasen. Die entstandenen spektrotemporalen Merkmale können für das kognitive System auch als probabilistische Hinweise fungieren, welche neuronale Konnektivität modulieren, um bestimmte aufgabenabhängige Interpretationen zu ermöglichen. Im Hinblick auf Theorien aktiver Inferenz und der Biosemiotik legt meine Arbeit nahe, dass spektrotemporale Unterschiede zwischen Gesang und Sprache Indizes für die prädiktiven Einschränkungen sind, durch die sie hervorgerufen wurden, und dass sie Affordanzen darstellen, die eine kulturelle Weitergabe ermöglichen. Die spezies- und modalitätsübergreifende Messung dieser Einschränkungen mit Methoden der evolutionären Systembiologie ermöglicht uns, die Wechselwirkungen von Kognition, Kommunikation und Evolution mechanistisch zu erklären.