Vocal audio effects : tuning, vocoders, interaction

Daniel Hernán Molina Villota

Sorbonne Université

L'École Doctorale Sciences Mécaniques, Acoustique, Électronique, et Robotique de Paris SMAER (ED391)

Vocal audio effects : tuning, vocoders, interaction

Thèse de Doctorat SMAER, specialité en Acoustique

Daniel Hernán MOLINA VILLOTA

Institut Jean Le Rond ∂'Alembert, CNRS (Centre National de la Recherche Scientifique), Sorbonne Université, Campus Pierre et Marie Curie, UMR7190

Soutenue le 4 Juillet 2024 à Paris devant un jury composé de :

Christophe D'ALESSANDRO Directeur de Recherche, Sorbonne Université - CNRS	Directeur de thèse
Maëva GARNIER HDR, Université de Grenoble - CNRS	Rapporteuse
Thierry DUTOIT Professeur, Université de Mons	Rapporteur
Nicolas OBIN HDR, Ircam - Sorbonne Université	Examinateur
Mitsuko ARAMAKI Directrice de Recherche, Aix-Marseille Université	Examinatrice Présidente du jury

This thesis explores how digital audio effects (DAFx), particularly pitch correction (PC), shape the voice in modern music. A taxonomic study thoroughly investigated the behind-the-scenes of vocal production in contemporary music. The goal was to find what makes effects such as PC, vocoders, and Autotune interesting and how they might be enhanced. The Dynamic Pitch Warping (DPW) method was revisited, proposed for vocal PC, and compared with Antares Autotune (ATA). Subsequently, a psycho-acoustic study of the PC methods was conducted, first by comparing four different vocoders and then the two PC methods. The study showed that each system imparts a unique coloration to the voice and provides directions for future improvements of the PC methods. Finally, the sound description of the vocoder for tuning and the interactive use of effects through real-time gestural control were examined.

Cette thèse s'intéresse à la manière dont les effets audio numériques, en particulier la correction de l'intonation (CI), façonnent la voix dans la musique moderne. À travers une étude taxonomique, les coulisses de la production vocale ont été dévoilées. L'objectif a été d'analyser ce qui rend intéressants des effets tels que la CI, le vocodeur et l'Autotune, et comment les améliorer. La méthode Dynamic Pitch Warping (DPW) a été revisitée, proposée pour la CI vocale et comparée à Autotune Antares (ATA). Par la suite, une analyse psychoacoustique a été menée, d'abord, entre différents vocodeurs, puis entre différentes méthodes de CI. L'étude a montré que chaque système donne une coloration particulière à la voix et fournit des pistes pour le développement futur de la CI. Enfin, la description sonore du vocodeur pour le tuning et l'utilisation interactive des effets par le contrôle gestuel en temps réel ont été examinées.

"And the stars will show where the waters flow,
where the gardens grow."

— Roxette 1999.

Remerciements

Tout d'abord, je voudrais exprimer ma plus profonde gratitude à ma maman et à ma tante Titina. Malgré la distance et le temps, nos sentiments restent intacts. Vous êtes toujours dans mon cœur. Les souvenirs partagés sont intemporels et infinis. Maman, ton soutien a été le vent qui m'a propulsé vers des idéaux extraordinaires. Merci.

Thomy, mon compagnon de folies, de paroles et d'amour, merci d'embrasser mes levés de soleil, merci d'avoir été à mes côtés lorsque la recherche scientifique devenait douce-amère.

Je remercie Nico et Titina encore une fois pour leurs rires contagieux, Soqui pour ses délicieux quimbolitos, mon grand-père Hernan pour son idée audacieuse de venir à la Sorbonne, et mon papa pour sa confiance en moi.

Je suis très heureux et ravi de partager avec vous tous le succès de ce projet. Cette thèse représente une expérience extrêmement enrichissante, marquée par des voyages captivants et des moments de réelle satisfaction. Chaque étape a été l'occasion d'élargir mes horizons et de m'épanouir tant sur le plan professionnel que personnel. Avec dévouement, j'ai sauvegardé l'originalité de mes idées, valorisé mon travail et maintenu mon intégrité.

Merci à Simona pour son aide, à Silvia pour sa sincérité piquante, et à Christophe, Pierre-Yves et Djimedo pour l'opportunité de venir à Sorbonne Université. Je tiens également à exprimer ma gratitude au jury — Maeva, Thierry, Mitsuko et Nicolas — pour leur lecture attentive du manuscrit, leur appréciation de mon travail, ainsi que pour les conseils enrichissants et pertinents qu'ils m'ont fournis afin d'améliorer le document final. Merci aussi à Sandrine, Antoine, Pablo, Krishan, Corto, Cecilia et Victor.

Enfin, un grand merci à l'Agence Nationale de la Recherche pour le financement de cette recherche doctorale à travers les projets : ARS ANR-19-CE38-0001 (Analysis and Transformation of Singing Style) et Gepeto ANR-19-CE28-0018 (Gesture and Pedagogy of Intonation), ainsi qu'à Colciencias - Convocatoria Doctorados en el Exterior 2021.

Abstract

This research focuses on the study of pitch correction, one of the most widely used digital vocal audio effects in modern music. The study aims to gain a deeper understanding of the perceptual coloration of this effect and to gather insights into potential improvements for pitch correction algorithms. The pitch correction method Dynamic Pitch Warping (DPW) is revisited with a focus on its vocal application and psycho-acoustical evaluation.

A perceptual taxonomic analysis of vocal digital audio effects is proposed, incorporating technical and contemporary musical examples. This perceptual taxonomy positions pitch correction as an effect that modifies pitch and vocal quality. Despite the widespread use of this effect, no descriptive and comparative scientific foundations exist outside of patents. Consequently, a compendium of technical-musical terms has been developed to distinguish the types of signals to be corrected and the relevant cases for study. As perspective, prototypes are proposed for the interactive use of vocal effects, capturing hand movements through wireless sensors.

A pitch correction system consists of three components: a pitch tracker, a pitch correction method, and a vocoder or vocal warper. Originally, Dynamic Pitch Warping (DPW) was a graphical pitch correction method designed for tablets. It has since been revisited and adapted for use in vocal correction. This method has been validated using theoretical pitch curves, accompanied by sound samples, and compared to Autotune Antares (ATA), the reference method in the field.

The vocoder plays a crucial role in implementing a given transposition or pitch curve. To psycho-acoustically study the coloration introduced by pitch correction, it is necessary to analyze both the effect of the vocoder and that of the pitch correction method. This study was approached through a psycho-acoustic comparative evaluation divided into two parts: (i) the vocoder techniques and (ii) the pitch correction methods. The psycho-acoustic evaluation of the vocoder aims to provide insights into the coloration produced by four systems (World, Circe, Retune, ATA). The evaluation of pitch correction methods seeks to determine whether perceptual differences exist between DPW and ATA. By comparing these two evaluations, insights can be gained into the relative contribution of coloration from pitch correction versus that from the vocoder, as well as potential avenues for developing new techniques of autotune effects. As perspective, a sonorous description of the vocoder is proposed, with an emphasis on its use for tuning.

Finally, a discussion on the integration of our work regarding the taxonomy of effects, pitch correction, and the psychoacoustic evaluation of this effect is presented. Within this discussion, perspectives for new research lines in the short and medium term are proposed.

Résumé

Cette recherche se concentre sur l'étude de la correction de l'intonation, l'un des effets audio numériques vocaux les plus utilisés dans la musique moderne. On vise à en savoir plus sur la coloration de cet effet d'un point de vue perceptif et à obtenir des indications sur ce qui pourrait rendre la modification de l'intonation plus intéressante à l'avenir. Pour mener cette étude, la méthode de correction par déformation mélodique dynamique, en anglais Dynamic Pitch Warping (DPW), est revisitée de manière intégrale, en incluant son application vocale et son évaluation psychoacoustique.

Une analyse taxonomique des effets numériques sur la voix a été proposée sur la base de la perception sonore, avec des exemples techniques et musicaux actuels. Cette taxonomie perceptive permet de placer la correction de l'intonation comme un effet qui modifie à la fois la hauteur et la qualité vocale. Malgré l'utilisation massive de cet effet, il n'existe pas de base scientifique descriptive ni comparative sur ce type d'effet en dehors des brevets. Par conséquent, on a développé un compendium de termes technico-musicaux pour distinguer les types de signaux à corriger et les cas d'intérêt à étudier. En perspective, on propose des prototypes pour l'utilisation interactive des effets vocaux, qui capturent les mouvements des mains grâce à des capteurs sans fil.

Un système de correction de l'intonation est composé de trois éléments : un suiveur de la courbe de hauteur, une méthode de correction de l'intonation, et un vocodeur ou système de transformation vocale. DPW est une méthode de correction de l'intonation graphique pour tablette. On l'a revisitée et adaptée pour son utilisation en correction vocale. Cette méthode a été validée par des courbes de hauteur théoriques (avec un support sonore) et comparée à Autotune Antares (ATA), qui est la méthode de référence.

Le vocodeur est très important, car il met en place une transposition donnée ou une courbe de hauteur. Ainsi, pour étudier la coloration psychoacoustique de la correction de l'intonation, il est nécessaire d'analyser l'effet du vocodeur et l'effet de la méthode de correction. On a abordé cette étude à travers une évaluation psychoacoustique divisée en deux étapes : la comparaison des techniques de vocodeur et la comparaison des méthodes de correction de l'intonation. La première est censée informer sur la coloration de quatre systèmes (World, Circe, Retune, ATA). La deuxième vise à déterminer s'il existe des différences perceptives entre DPW et ATA. En comparant les deux évaluations, on peut obtenir des indices sur le poids de la coloration due à la correction de la hauteur et celui dû au vocodeur. Par conséquent, on peut indiquer comment procéder pour le développement de nouveaux effets de type autotune plus intéressants. En perspective, on propose une description sonore du vocodeur en mettant l'accent sur son utilisation pour le tuning.

Enfin, une discussion sur l'intégration de notre travail concernant la taxonomie des effets, la correction de l'intonation et l'évaluation psychoacoustique de cet effet est présentée. Dans cette discussion, nous proposons également des perspectives pour de nouvelles lignes de recherche à court et moyen terme.

Contents

List of Figures

List of Tables

Publications

International Conferences

- D. H. Molina Villota, C. D'Alessandro, O. Perrotin, "Dynamic Pitch Warping for Vocal Correction" In : Digital Audio Effects, Copenhagen, Denmark, 2023, pp. 1–8.
- D. H. Molina Villota, C. D'Alessandro, "Comparing vocoders for automatic vocal tuning" In : 16th International Symposium on Computer Music Multidisciplinary Research, Tokyo, 2023, pp. 756–759.
- D. H. Molina Villota, C. D'Alessandro, G. Locqueville, T. Lucas, "A Singing Toolkit: Gestural Control of Voice Synthesis, Voice Samples and Live Voice" In : 16th International Symposium on Computer Music Multidisciplinary Research, Tokyo, 2023, pp. 704–707.

National Conferences

- Daniel Hernan Molina-Villota, Christophe d'Alessandro, Olivier Perrotin. Correction dynamique et adaptative de la justesse en voix chantée. CFA 2022 - 16ème Congrès Français d'Acoustique, Apr 2022, Marseille, France.

Introduction

It is an honor to present the culmination of this research thesis, entitled: "Digital Vocal Effects: Tuning, Vocoding, and Interaction." The main objective of this work has been to explore the various realms and nuances of vocal modification, particularly focusing on pitch correction methods and the vocoder. This research has been conducted from a transdisciplinary perspective, encompassing topics such as signal processing, experimental psychoacoustics, interactive interface control, and the use of digital effects in the music industry.

The first chapter aims to attain a general understanding of vocal effects and their musical application. A taxonomy of digital effects applied to the voice has been developed from a perceptual perspective, establishing a glossary of effects and examples of usage. Additionally, concepts such as preserving vocal quality are proposed, emphasizing the contemporary importance of modular effects usage, pitch modification, vocoder as a vocal modification algorithm, and autotuning. Vocal effects are increasingly being incorporated into music more creatively, akin to the historical utilization of guitar pedals. The transversal study proposed here can lead to the development of new technically and musically interesting vocal effects.

Although autotune has existed for over 25 years, there is a notable absence of a technical-musical glossary related to vocal pitch correction. That is likely due to the predominance of patents (such as Antares Autotune, Melodyne, and Retune, among others) rather than freely accessible and open studies on the subject. Therefore, this second chapter begins with the proposition of a compendium of terms and concepts related to pitch correction, aiming to define the types of pitch modification and the fundamental cases that are interesting both musically and technically (in music production). This glossary highlights potential improvements in pitch correction methods.

Chapter two also revisits the Dynamic Pitch Correction (DPW) method; DPW was initially developed as a graphical pitch correction method by [Perrotin and D'Alessandro, 2016]. This method has been adapted for use on vocal samples as an alternative autotuning method. A comparative study is conducted between our method and the reference standard (Antares Autotune, abbreviated ATA). This study is based on theoretical pitch curves, revealing substantial differences in pitch management. The analysis indicates that DPW differs from the reference method (Antares Autotune). Nevertheless, such a difference is subtle and does not precisely correspond to the bibliographic information. This finding underscores the need to re-formulate the comparison

between both methods and psycho-acoustic verification through a subjective test, which is addressed later. The need to psycho-acoustically compare pitch adjustment methods also involves a perceptual study of the coloring generated by the vocoder in pitch correction. The vocoder is a primary component of the pitch correction algorithms (composed of a pitch tracker, a pitch corrector, and a vocoder).

The process of vocal modification, whether for transposition or correction, necessitates the imposition of a pitch (tuning) or a pitch adjustment value and the use of a vocoder to effectuate such modification. Different vocoders can be employed for this purpose; each one imparts its own un-characterized coloration. Chapter three conducts a psycho-acoustical evaluation using three pitch correction scenarios (original pitch, extreme tuning, and soft tuning) and four systems (world, Circe, retune, and ATA). This study enables the comparison of each vocoder's transparency for the re-synthesis of the given pitch contour curve; in other words, the coloration of the vocoder through the different tuning cases can be evaluated by comparing them. The coloration evaluation is directly related to the timbre change and the singer's vocal quality modification when using each vocoder. Additional conclusions can be drawn regarding how the panel responds to the different tasks. For example, do musicians perceive the difference more easily? Are there consistent results through the different tuning cases and subdivisions?

The development of Chapter 3 also allowed us to consolidate a protocol evaluation structure, which was transferred to the subjective psycho-acoustic comparison of pitch correction methods (ATA and DPW) presented in Chapter 4. Such a study treats three tuning scenarios among five experiments (Tasks). The results give insights into the cases where the difference between ATA and DPW is more significant and musically interesting and where it is not. Additional conclusions can be drawn regarding how perceptually close the samples are to the original audio and how the auditor's panel subdivisions behave statistically. Both psycho-acoustical evaluation tests (vocoders and pitch correction methods) help to analyze if autotuning coloration depends more on the vocoder or the pitch correction method and give insight into future developments related to pitch correction methods.

Chapter Five is a compendium of research perspectives and conclusions. Section one concerns the sonorous description of the vocoder based on its tuning use, avoiding addressing parameters related to other perceptive characteristics different from pitch modification. Four principal parameters to describe the vocoder quality are defined. Perspectives on the psycho-acoustic evaluation of these parameters are proposed. Section two explores real-time motion-controlled interactive effects. The interactive use of vocal effects is relevant for disseminating new technologies and developing a new digital lutherie. Some prototypes are presented among with insights for their future psychoacoustical evaluation. The final section addresses the research contributions summary and conclusions, integrating all the work, from development to the results, demonstrating how the chapters are interconnected. Perspectives on

the advancement of each topic are addressed, as well as guidelines for future research.

I hope this thesis serves as a valuable contribution to the world of tuning vocals, offering a fresh and comprehensive view of vocal effects, pitch correction, and vocoder's coloration and giving exciting perspectives for research on this fascinating topic.

Chapter 1

Perceptual DAFx Taxonomy and Vocal Transformation

The musical creative process is inherently linked to artistic expression, connecting with the social environment in all its aspects. Indeed, the creative process varies according to the musical genre (orchestra, acoustic music, street rap, experimental music). Modern music, distinctive in its creative process, enables contemporary musicians to utilize Digital Audio Effects (DAFx) from their home studios. Consequently, DAFx play a crucial role in contemporary music, not only for their use in the creative process, but also because they contribute to the sonorous identity the artist wishes to convey. DAFx are employed to impart texture, intentionality, and poetic meaning to the musical message. The imminent use and widespread vocal effects underscore the need for various technical-musical supports.

This chapter corresponds to a review of vocal digital audio effects, showcasing the creative possibilities offered by vocal modification and the importance it holds in music today. There is a call for a taxonomic study on vocal effects from a perceptual standpoint, distinguishing it from other taxonomies not focused on voice. Also, this chapter highlights the significance that vocal transformation, autotuning, and vocoder (as a vocal transformer) have gained as stylistic and interpretative elements and why delving deeper into the study of these effects is interesting.

1.1 Historical Context

The 20th and 21st centuries have witnessed unprecedented technological advancements that have shaped and redefined music profoundly [Wilmering et al., 2020]. This metamorphosis has influenced not only sound generation and lutherie but has revolutionized musical creation, diffusion, staging, and catalyzed the emergence of new musical genres and styles.

At the heart of this revolution lie digital tools designed to generate and manipulate sounds. These tools allow artists to explore vast sonorous landscapes and craft auditory experiences with a distinct message and intention. These

tools not only enable the creation of new sounds but have also successfully replicated natural acoustic phenomena digitally, such as reverberation and the distortion of electromechanical signals.

Artificial reverberation, for instance, was a key 20th-century innovation, patented by the Radio Corporation of America (RCA). Its adoption in classical music recordings of the 1920s marked a turning point in the perception of recorded sound. Concurrently, avant-garde artists like Stefan Wolpe challenged conventions by introducing revolutionary sound manipulations in their Dada performances using phonographs.

Bell Laboratories, with their electromechanical delay developed in 1939, and groundbreaking artists like Les Paul, with guitar effects like chorus, echo, and flanger, opened the gateway to an era dominated by electric and electronic sound. This era reached its zenith in the 1960s and 1970s with the popularization of electric and electronic effects.

During the 1960s, Schaeffer provided a theoretical and analytical perspective on the world of sound, defining and exploring concepts that would form the backbone of contemporary musical techniques [Schaeffer, 1966]. These theoretical and practical explorations paved the way for unprecedented musical experimentation, bolstered by the democratization of computer technology and the development of powerful tools like the Digital Audio Workstation (DAW) [Wilmering et al., 2020]. The 4X synthesizer, developed by IRCAM, stands as a notable example of these advancements, laying the groundwork for revolutionary platforms like Pure Data and MAX/MSP [Boulez and Gerzso, 1998].

The 1990s became a breeding ground for innovative Digital Signal Processing (DSP) developments. This era marked a confluence of disciplines, including acoustic physics, computer science, and music. Tools emerged that could replicate and expand natural acoustic phenomena, broadening compositional and performative possibilities for artists.

The vocoder is the latest advancement in terms of vocal transformation, in which the concepts of transform, filtering, sampling, and convolution converge. Its origin dates back to 1930 by Bell Labs with the aim of vocal compression; however, it was not until the 1960s that it was musically used and popularized by artists like Kraftwerk, Pink Floyd, and Giorgio Moroder. Its use exploded after 1999 with the implementation of autotune by Cher and bands like Daft Punk. Today, the vocoder is in full swing, much like guitar pedals were at one point. With digital technology, it is now more accessible and versatile than ever. Producers use it not only to give that characteristic robotic tone to voices but also to merge instruments and voices, create harmonies, and add textures and depth to mixes. The vocoder has evolved from being a communication tool to an essential and versatile musical instrument that continues to evolve.

At the zenith of this development, Autotune ANTARES emerged in 1998, quickly becoming an iconic tool in contemporary music production. Despite its initial imperfections, it found its place at the heart of pop music (with tracks like Cher's "Believe") and other genres like rap and hyperpop. The impact of Autotune is undeniable, even leading it to be recognized in the industry with

awards like the Grammy given to Hildebrand in 2023.

Today, DAFx play a starring role in contemporary music, with advanced tools like adaptive effects leading the forefront [Verfaille and Arfib, 2002]. These effects, which evolve based on the relationship between sound and gesture, present a world of possibilities yet to be explored. In this rich and diverse musical landscape, it is essential to establish study and application frameworks for vocal effects, constantly seeking to enrich and improve the sonorous capacities and qualities that nourish the musical experience.

1.2 Review of the DAFx Technical Classification and its Vocal Use

The human voice is the most important musical instrument. There are vast regions of our brains that help to control and perceive sounds. Even people who have no musical training can perform imitations and vocal features[Cook, 1999]. Voice changes depending on our feelings, technical features, and training. Voice effects involve aspects such as dynamic, time, space, and timber. As with any other instrument, voice can be modified by the player (singer in this case) in numerous ways. However, unlike any other instrument, the voice system is part of our body and involves complex brain processes. In that way, interactive DAFx for singing can help to understand the generation and perception of sound better.

To introduce the use of Digital Audio Effects (DAFx) for voice, we consider the singer's vocal quality features and how some of those features are modeled or modified through signal processing algorithms. Vocal quality is a very complex term that, according to[Garnier et al., 2007] [Garnier, 2007], can be considered an evaluative judgment, adequately described as a set of acoustic, perceptual, and semantic properties of a voice. It gives cues to the singer's vocal quality, state, and intentions.

The voice is characterized by several resonant peaks in the spectrum called formants, whose positions and intensities crucially determine the unique quality of different vowel sounds. Various methods exist to modify formants. One common technique is the use of band filters, which can modify the voice spectrum by amplifying or attenuating certain frequency bands, and thus the formants. We must mention the source-filter model of speech production. In this model, the source (the vocal folds' vibration waveform or the turbulence caused by a constriction) is modeled by an excitation signal (an impulse train for the voiced component, or a random signal for the unvoiced component). Its spectrum is obtained by filtering the signal with the source filter, and the effect of the vocal tract shape and lips is also modeled as filters. Linear predictive coding (LPC) is a source-filter-based method often used by linguists as a formant extraction tool and is appropriate for modeling vowels that are periodic, except for nasalized vowels. In this model, a source signal passes through a synthesis filter that represents the spectral envelope [Zölzer, 2011].

A spectral envelope is "a curve in the frequency-amplitude plane, derived from the Fourier magnitude spectrum" [Schwarz and Rodet, 1999].

The importance of these studies in understanding the perception of vocal features cannot be overstated. By examining how technical attributes of the voice are perceived, researchers gain insights into how modifications affect listener interpretation, similarly to studies of vocal quality [Henrich Bernardoni et al., 2008] [Garnier et al., 2005]. For instance, synthesis and re-synthesis using these techniques can enhance musical creation. Musicians often seek to play with unnatural changes to the voice, as we will see by referencing many modern uses of digital effects on voice.

What we would like to point out is that a technical review of digital audio effects is helpful to understand the techniques but is not yet linked to the characteristics of vocal quality. Such a study is not undertaken in this thesis, but we would like to introduce the use of digital effects for voice in modern music and how autotuning is integrated within such use.

1.2.1 The Technical Classification

All stakeholders involved in modern music creation (engineers, composers, performers, producers) utilize digital equipment and tools. Their language and understanding of these tools may differ depending on their role in the creative process. Establishing a framework to study this language can enhance collaboration between artists and developers, aiding in developing new interactive tools. These tools might be digital musical instruments, stemming from an interface and a sound generation engine [Perrotin and D'Alessandro, 2016], or DAFx, originating from an interface and a sound transformation engine. DAFx emulate effects from acoustic, electric, and electronic sources, and their various combinations. They are crucial in crafting sonorous environments. Verfaille conducted a taxonomic study on DAFx [Verfaille et al., 2006a], suggesting an interdisciplinary classification that bridges understanding between artists and engineers. While his proposition is insightful from a musicological and research perspective, it may be intricate to apply in a studio session.

Various methods exist to analyze vocal DAFx taxonomically. Effects can be analog or digital. Within the analog domain, we find mechanical effects like reverberation and electromechanical ones, such as guitar pedals or changes in vinyl playback speed [Verfaille et al., 2006b]. Digital effects replicate analog ones, like reverberation, but also include purely digital effects, such as distortion or changes in vocal formants. Considering those existing general reviews on digital effects, we will focus exclusively on vocal effects. As we will elaborate, the perceptual approach is most suitable given its practicality in the creative process.

An initial classification of DAFx is based on the digital signal processing (DSP) technique, organizing effects by their implementation method, like filtering or delays. Subcategories can be formed based on the application domain and processing method. While this approach highlights technical similarities,

it might be too complex for a diverse audience in a varied creative context. A second classification revolves around how the performer modifies the effect during a performance: whether the control is constant or variable, modifiable by wave generators (LFO), adaptive to the signal, or through gestural control. A third classification focuses on vocal perception, entailing modifications in melody, dynamics, timing, and vocal timbre.

1.2.2 Main Techniques

There are two main techniques to deal with Vocal Effects: the time domain techniques and the frequency domain techniques. The vocal effects in the time domain offer a broad range of manipulations for the voice. **Time-shuffle** [Geslin, 1998] rearranges the order of words or phrases, while **Delay-Line** introduces echoes or repetitions, with notable effects [Dattoro, 1997] such as the *Chorus* and *Flanger*. Various **filters** allow for the modification of vocal timbre, adjusting its frequency characteristics. Techniques **SOLA** [Roucos and Wilgus, 1985] [Makhoul and El-Jaroudi, 1986] and **PSOLA** [Moulines and Charpentier, 1990] are essential for altering the rhythm and pitch of the voice, allowing for subtle or dramatic changes. **LPC** is used to transform voices and is particularly useful for preserving formants or doing cross-synthesis [Keiler et al., 2001] [Moorer, 1979a]. On the other hand, **Resampling** tailors the voice to different pitches while preserving its natural character. Within the **Gain** category [Verfaille et al., 2006a], we find a variety of modulations for the amplitude of the voice [Zölzer, 2011], from the *Tremolo* to dynamic tools like the *Compressor*. When applied appropriately, these effects can enhance or completely transform a vocal recording.

Vocal effects in the frequency domain (FD) offer a variety of techniques specifically designed to transform and enhance vocal characteristics. These are primarily categorized [Verfaille et al., 2006a] as Inverse Fast Fourier transform ***IFFT*** and ***Oscillations Bank***. Within ***IFFT***, techniques are differentiated based on their use of phase correction. They are identified as either the Phase Wrapping Technique or the Phase Unwrapping Technique [De Götzen et al., 2001]. Under the Phase Warping Technique, the *Time-Scaling* [Zölzer, 2011] stands out, adapting vocal duration with variations such as *Adaptive-Time-Scaling* [Verfaille et al., 2006c] with or without temporal synchronization. The *Time Domain-Resampling* utilizes *LPC* [Makhoul and El-Jaroudi, 1986] to modify vocal pitch while maintaining or altering timbral characteristics. Moreover, *Cross-Synthesis* combines spectral attributes from different sounds.

In the context of the Phase Unwrapping Technique, effects such as *Robotization* [Zölzer, 2011] can modify or harmonize voices for mechanical effects. *Whisperization* [Zölzer, 2011] transforms regular voices into whispers. Transformations targeting the spectral envelope are provided, as are adjustments in the amplitude spectrum and techniques like *Time-Shuffle with SE Modif* that merge temporal and spectral modifications. Lastly, ***Osc. Bank*** introduces methods such [Verfaille, 2003] as *Spectral Ring Modulation* [Kameoka

and Kuriyagawa, 1969] and *Spectral Tremolo* [Hoffman and Cook, 2008] to achieve distinctive vocal modulations.

Beyond Verfaille's study, artificial intelligence (AI) emerges as a revolutionary category that has driven advancements in vocal synthesis and transformation [Roche, 2016] [Martinez Ramírez, 2020]. What has been observed is an adaptation of deep learning techniques [Peeters and Richard, 2021], intitially designed for images, applied to information blocks in the frequency domain and, to a lesser extent, in the temporal domain. In the musical field, deep learning has been primarily employed for music information retrieval, such as melody extraction, source separation, instrument recognition, and tempo estimation. However, its application for sound synthesis and transformation is still in its beginnings. Autoencoder-based models have been used, which process data in the frequency domain and encode them into a latent space, subsequently decoding them. Likewise, models based on Generative Adversarial Networks (GAN) have been explored for sound synthesis [Engel et al., 2019] [Donahue et al., 2019], yielding promising outcomes.

The technical overview of the taxonomy of vocal effects is depicted in Figure 1.1, this is an analogy to the technical classification proposed of DAFx [Verfaille et al., 2006a]. While the chart does not detail every effect, variations can be identified, such as low-pass or high-pass filters in filtering, vocoders in PSOLA techniques, and band equalizers in gain effects. Two main observations can be made: firstly, different techniques can achieve the same sound effect, and secondly, varied techniques can influence the same perceptual aspect. For instance, filtering can be done using a digital filter, band filter, or delay-line, while pitch changes can be attained through resampling or techniques like PSOLA. Although the techniques are not equivalent, they impact the same perceptual element. However, this does not imply that other perceptual elements remain unaffected. For example, in the case of autotune, while the primary adjustment is to the pitch, the timbre is also altered.

1.2.3 Time-Frequency Vocoding Technique

The vocoder plays a pivotal role in facilitating vocal transformation while maintaining the inherent human qualities and distinctive timbre characteristic of individual voices. This section provides a brief overview of the primary vocoding techniques (not the musical instrument for cross-synthesis). A more detailed exploration of signal processing, time-frequency representations, and the signal reconstruction process using the vocoder is presented in Appendix A.

Among the various vocoding techniques, there is recurrent discussion surrounding time-frequency representations, which delineate the temporal evolution of a signal's frequency spectrum. These representations serve as a fundamental basis for vocal reconstruction. This process entails an analysis/synthesis schema known as the phase vocoder, which constitutes one of the vocoding techniques. The phase vocoder adheres to the structure out-

Figure 1.1: Analogy of the Technical Classification Proposed by Verfaille [Verfaille et al., 2006a], Applied to Vocal DAFx

lined in Figure 1.2, detailed further in Appendix A. Within the phase vocoder framework, an input signal, $x(n)$, undergoes windowing with a window of size N, generating a continuum of windowed segments. The Fast Fourier Transform (FFT) is computed for each successive segment, yielding a time-varying spectrum $\mid X(n,k) \mid e^{j\varphi(n,k)}$ where $k = 0, 1, .., N$. This spectrum can be manipulated for vocal modification purposes and reconstructed to retrieve the vocal signal. Reconstruction is accomplished using the Inverse FFT (IFFT), employing a window and overlapping segments, as illustrated in Figure 1.2

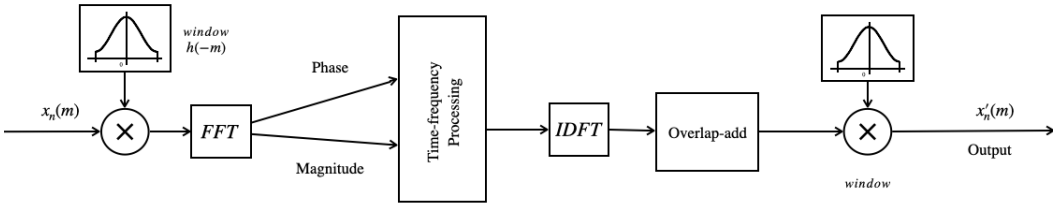


Figure 1.2: Time-frequency processing of the phase vocoder [Zölzer, 2011]

Another fundamental term in vocoding techniques is the spectral envelope. This envelope is extracted from time-frequency representations and is

closely related to the model of the vocal system. The spectral envelope is a smoothed version of the spectrum that disregards the spectral line structure while preserving the general shape of the spectrum. The problem of vocal signal reconstruction is based on extracting the spectral envelope through a model of the vocal system. In this model, vocal production originates from the vocal cords acting as the excitation source, while the mouth and nose act as a resonant or antiresonant system. This model of the vocal system is referred to as an excitation-resonance model, also known as the source-filter model in the literature.

The implementation of the source-filter model can be achieved through three techniques, referred to as vocoding techniques. The first technique is known as the channel vocoder. It utilizes parallel bandpass filters and calculates the RMS value for each band. This process enables the estimation of the spectral envelope, as illustrated in Figure 1.3. The greater the number of channels (or filters), the more points of the spectral envelope are computed. The bank of filters can be modeled using either a linear or a logarithmic scale. The method depicted in Figure 1.3 is based on the time domain; however, it is feasible to obtain the spectral envelope in the time-frequency representation as well. In the frequency domain, a channel can be perceived as the summation of elementary energies of each bin weighted by the envelope of this channel filter. The resulting amplitude would then be the square root of these energies.

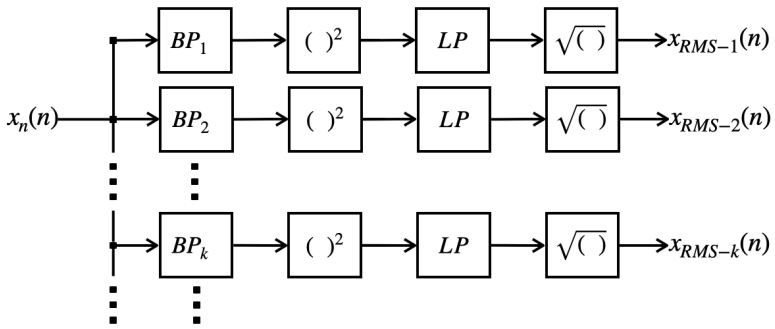


Figure 1.3: Channel Vocoder Technique, based on [Zölzer, 2011]. The RMS values correspond to the output of each band. The sum of all these signals generates the final output.

The second technique employed is Linear Prediction Coding (LPC), which directly stems from the source-filter model, as illustrated in Figure 1.4. In this model, the filter characterizes resonances with just poles, forming an all-pole filter that effectively matches the spectral content of a given sound. The estimation of such a filter $A(z)$ involves approximating the input signal $x(n)$ as a linear combination of its past samples, aided by a Finite Impulse Response (FIR) filter $P(z)$. This process entails configuring the prediction error filter, also referred to as the inverse filter $A(z)$, and subsequently generating the prediction error signal $\tilde{e}(n)$, which represents the disparity between the predicted signal denoted by $\hat{x}(n)$ and the actual input signal. For synthesis purposes,

the notation $\tilde{e}(n)$ is adopted, wherein this signal serves as an excitation signal. Synthesis is achieved through the inverse of the analysis filter $H(z) = \frac{1}{A(z)}$, known as the LPC filter, which embodies the spectral model of the input signal $x(n)$. The primary methods for obtaining the synthesis filter coefficients filter include the autocorrelation method [Makhoul, 1975], the covariance method [Orfanidis, 1990], and the Burg algorithm [Makhoul, 1977]. LPC is particularly effective for speech applications due to its ability to model both the vocal tract and the source, where the source typically comprises pulses and noise. The periodicity observed in voiced sounds determines the pitch, while unvoiced sounds exhibit a noise-like excitation pattern.

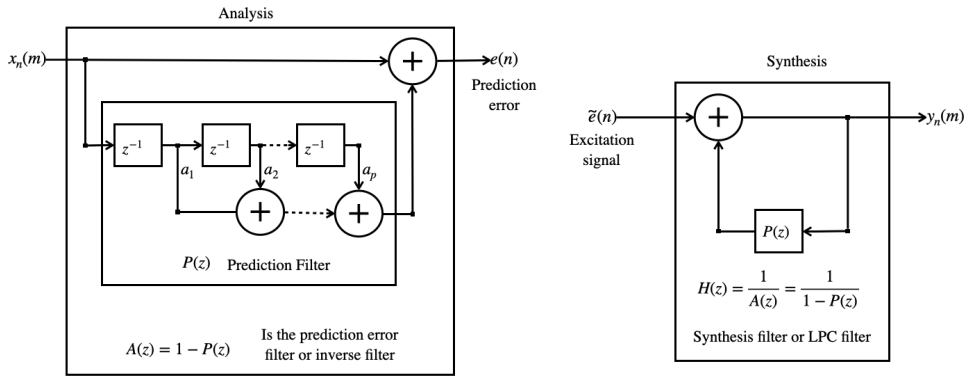


Figure 1.4: Linear Prediction Coding Technique [Zölzer, 2011]

The third technique is cepstrum, which involves smoothing the logarithm of the FFT spectrum to separate its slow-varying part (the spectral envelope) from its fast-varying part (the source signal), as illustrated in Figure 1.5. In this technique, we consider the signal $y(n)$ can be separated in a source signal $x(n)$ and the response to the impulse $h(n)$. The signal $y(n)$ is first passed through a window and its FFT is computed. Subsequently, the logarithm of the obtained spectrum, denoted as $\hat{Y}(k) = \log Y(k) + j\varphi_y(k)$, is taken, followed by an inverse FFT operation, resulting in the complex cepstrum $\hat{y}(n)$. The real cepstrum can be obtained by performing the IFFT of the real part of $\hat{Y}(k) = log|Y(k)|$, denoted as $c(n)$, which is also equal to $\frac{\hat{y}(n)+\hat{y}(-n)}{2}$. The cepstrum is then passed through a low-pass window and subjected to an FFT operation, yielding a smoothed version of the spectrum $Y(k)$ in dB scale, which is considered the spectral envelope and denoted as $Ch(k)$. The complementary high-pass filter can be applied to compute the source envelope, denoted as $Cx(k)$, through FFT. To retrieve the reconstruction $e(n)$ of the source signal $x(n)$ one can utilize the expression $e^{C_x(k)} = |X(k)|$ along with the initial phase of the signal $e^{i\varphi_y(k)}$. Subsequently, $e(n)$ can be obtained by applying the inverse fast Fourier transform IFFT$[X(k)^{i\varphi_x(k)}]$.

The importance of this review lies in the existence of various vocoding techniques primarily used to modify sound descriptors such as time, melody,

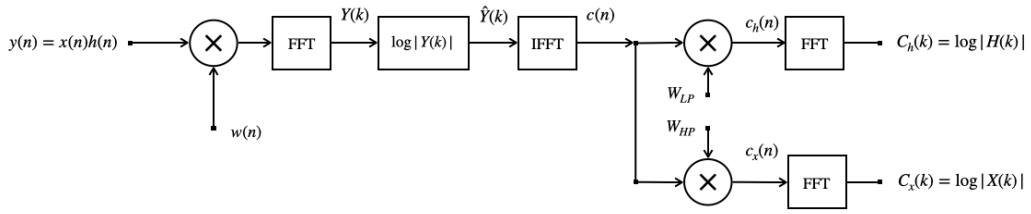


Figure 1.5: Cepstrum technique pour source-filter method [Zölzer, 2011]

and timbre. Perceptually, these changes can vary in their impact on vocal quality and perception, leading listeners to prioritize one sound descriptor over others. Consequently, the technical classification of vocal effects may lose significance from a vocal musical standpoint. Hence, it is crucial to undertake a perceptual taxonomic study that encompasses not only DAFX modification and the vocoder capacities to modify vocal quality but also the modular nature of contemporary music production.

1.2.4 Limits of the technical classification

In creative environments where modularity and fusion of techniques can lead to similar sounds, its perception becomes vitally important. When integrating various techniques and interfaces simultaneously, two fundamental perspectives emerge. The first highlights the ability to expand the range of vocal effects through innovative modular combinations. The second emphasizes how musicians perceive and interpret such sound, not just as an effect but as a unique and specific musical expression. While this modular vision broadens horizons, it also challenges traditional taxonomic classifications, which tend to focus on specific techniques.

Taxonomy, as Verfaille well articulates in his work, can adopt different lenses: from the technological nature of the effect (analog or digital), through its complexity, to the type of control employed, whether constant or variable. The latter, for example, can be adaptive, responding to a specific control signal or guided by a low-frequency oscillator. Faced with this multiplicity, the pressing need to value sound transformation beyond mere techniques stands out, placing sound perception at the core of our reflection.

In line with Verfaille's perspectives, it is clear that a classification based on perception is more coherent and relevant, especially when different approaches — whether analog, DPS, or AI — lead to similar auditory outcomes. We must add the modularity of effects within a processing chain and the complexities introduced by interactive control. Therefore, it is essential that our classification centers on vocal perception, which is the true heart of our auditory experience as we will see in 1.3.

1.3 Proposal for a Perception-Based Taxonomy of Vocal DAFx

An introduction to vocal effects has been conducted by Coralie Vincent in the chapter of the book "La voix Chantée" titled "From Antipop to Autotune". In her chapter, she explores different types of vocal sound effects, describing them generally without delving into the taxonomic classification of these effects, the modular approach, and the uses of the vocoder. We will expand on Vincent and Verfaille's work by applying it to the vocal context, taking into account current capabilities in music production.

Within the expansive musical landscape, a composer can explore and manipulate sound using various tools that modify vocal quality in numerous ways. These transformations may aim to enhance vocal qualities, emphasize certain voice characteristics, or introduce a specific texture or timbre with a distinct musical intention. Such actions play a pivotal role in crafting the musical narrative throughout the piece. The sonorous landscape that emerges from this composition is shaped by a series of vocal interventions which, regardless of the technique employed, are determined by the composer. Our aspiration is to provide a conceptual framework that helps understand the artistic goals underlying these sonorous transformations.

In the vastness of the sound universe, modifiable perceptual properties encompass acousmatic descriptors such as melody, dynamics, temporal and spatial aspects, and timbre. Such descriptors will be examined in this section. Under the melody category, there are alterations affecting the melody itself, the tone, and the harmonic and inharmonic components. Regarding dynamics, it considers the dynamic range, nuances, phrasing, and accents. Temporal aspects address duration and tempo. Spatiality refers to the size and location of the sound source, its movement in space, and the environment or setting where the sound propagates. Spatialization tools become essential in composition and performance, adding greater depth and meaning to human gestures. Timbre relates to vocal characteristics that distinguish an individual, such as age, emotion, vocal range, roughness, brightness, and vocal effort. As highlighted by Wilson and Fazenda [Wilson and Fazenda, 2013]: "Timbre can be adapted towards other nuances and, in this way, be steered to provide harmonic spectra, with the aim of better integrating the sound within a specific tonal context." Hence, we now seek to define a taxonomy of digital vocal effects from a sonorous perception standpoint, outlining uses and applications that are understandable to the composer, the engineer, and the listening audience. While it is possible to define a taxonomy in this way, it is important that effects are classified based on their primary objective. For instance, a robotization effect primarily affects the quality of the sound source, that is, its timbre. Of course, it involves changes in dynamics and tone, but its primary goal is the change in timbre.

In this section, our focus is on presenting concisely our perceptual classification of vocal effects with a technica summarized revision and audio support

example. Vocal effects could be divided into five primary types:

1. Effects on dynamics
2. Temporal modification effects
3. Spatial effects
4. Effects for pitch changes
5. Effects for timbre changes

The first four cases are effects that do not intentionally impact timbre; any resulting timbral alterations are a byproduct of the technique used. Regarding timbre effects, a special division is proposed:

1. Effects that preserve vocal quality
2. Effects that can distort vocal quality
3. Effects that change vocal quality

The resume of our classification can be found in Figure 1.6

1.4 Effects on Dynamics

Dynamics in music, and particularly in this case (voice), refers to the dynamic range or the variation in sound intensity over time. This variation in intensity is essential for giving character to a performance. Amplitude changes within the voice can be subtle or pronounced. When we talk about a small dynamic range, we refer to minor fluctuations in the audio signal over a short period of time. These subtle variations can be related to vocal techniques such as vibrato or slight nuances in intonation. On the other hand, large modulations refer to more noticeable changes in intensity, which can unfold over both short time spans (like an accented note) and long ones (like a gradual crescendo).

Several tools and techniques are used to manage or modify the dynamic range in post-production or recording. Linear effects, such as amplification and normalization, are used to adjust the overall level of a recording. On the other hand, adaptive effects respond to the changing characteristics of the audio signal, adjusting the dynamics as needed. Non-linear effects, in turn, can introduce more complex variations in the signal, like distortion or saturation, which can affect the perception of dynamics depending on the specific nature of the alteration applied.

Figure 1.6: Perceptual classification of Vocal DAFx

1.4.1 Amplification and Compression Effects

Amplification involves adjusting the amplitude of the samples $x[i]$ by multiplying them by a specific factor. The primary purpose of amplification is to adjust the recording level that might not align with other elements in the mix; for instance, if the vocal track sounds too loud or soft. This technique can enhance the presence of a vocal recording within a mix. Amplification in its simplest form can be expressed as:

$$y(n) = 10^{\frac{A}{20}} x(n) \tag{1.1}$$

Where A is the amplification factor. It is essential to be cautious when amplifying to avoid clipping or saturation, which could distort the sound; ***saturation*** itself can be considered as a timbral effect.

Normalization adjusts the maximum level of a recording to the desired

value while preserving its dynamic range. This technique is crucial when balancing vocal recordings that might be uneven in volume, especially before applying compression. it is essential to approach normalization with caution; pushing it to the extreme can lead to ***saturation*** and, consequently, signal distortion. Normalization is applied to a signal whose maximum level is 0dB. It is performed in deferred time. The maximum level of the signal is calculated, and then, sample by sample, it is divided by the maximum, as follows:

$$y(n) = \frac{x(n)}{\max_n |x(n)|} \tag{1.2}$$

The ***Expander*** is a non-linear treatment that increases the dynamics of the signal, attenuating low-level sounds without affecting high-level ones. The signal level controls the expander: if it is high, the expander maintains unit gain, but if it is low (below an adjustable threshold), it reduces the gain. ***A noise gate*** is a version that silences sounds below said threshold. The application of the expander is to reduce noise in parts of the recording when no instrument is being played. Typically, they are used in conjunction with compressors. ***Compressors*** are used to raise the level above a threshold, thereby compressing the dynamic range of the sound, hence their name. The subsequent use of an expander allows to restore the dynamics of the sample. Generally, expanders are used after filters to prevent amplifying noise generated by them, but before echo or reverb effects to prevent the replication of noise or abrupt cuts of these types of effects.

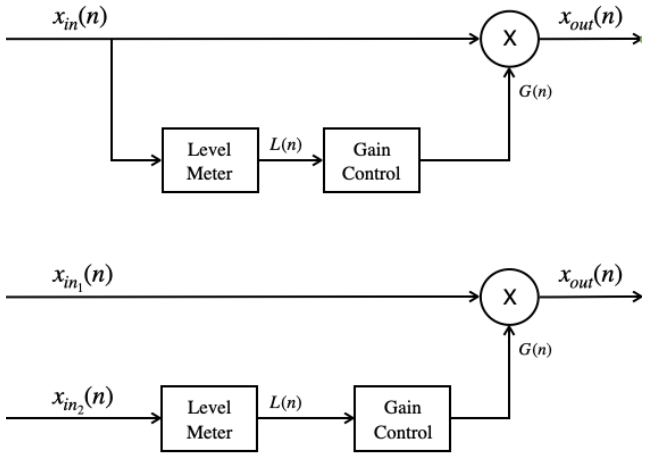


Figure 1.7: Schema for the expander/compressor classic technique and the sidechain technique, adapted from [Zölzer, 2011] and [Verfaille, 2003]

In Figure 1.7, the standard scheme of a compression or expansion system can be seen. The signal level is obtained concerning to a reference, and then it is defined whether the gain is 1 or varies according to the input value. A signal can be utilized to autonomously to manage another signal, enriching the auditory mix in a musical project to generate a unified, "full" auditory

experience. Techniques such as sidechain compression can be applied, where the music volume automatically lowers upon detecting vocals, thereby ensuring apparent prominence of vital elements within the mix.

The utilization of amplification and compression effects is highly beneficial for providing body, dynamics, and capturing the listener's attention, which accounts for their popularity from the late 1990s to the early 2000s. Applying such effects is commonplace in the vocal effects chain, akin to using equalizers and filters. Butch Vig, a founding member of the band Garbage and producer of significant albums such as Nirvana's "Nevermind" (1991) and Foo Fighters's "Wasting Light" (2011), which are characterized by typical compression for their respective eras, provides a tutorial on YouTube [1]. This tutorial shows the application of filters and the compressor in the song "Blood for Puppies" [Garbage, 2012]. This instance illustrates the moderate use of compression in alternative rock, yet its application is systematic; for instance, the utilization of the Teletronix LA2A compressor can be observed in divergent genres, for example, in "StoneMilker" [Björk, 2015] [2] (avant-garde), and also in the album 1989 by Taylor Swift (north-american pop) [3].

While the album "Wasting Light" has faced criticism for its high compression, it is widely regarded as a rock album and received the Grammy Award for Best Rock Album. The application of the compressor is exemplified in songs such as "White Limo" [Foo Fighters, 2011] from the 00:23 mark onwards. The heightened use of compression in the 2000s, known as the "Loudness War," is not necessarily detrimental. Examples of elevated compression usage without compromising dynamics can be found in "bad guy" [Billie Eilish, 2019] from the 00:16 timemark.

1.4.2 Modulation

There are several types of amplitude modulation, the most simple of which is ***tremolo***. Tremolo periodically modulates the signal's volume, creating a "pulse" or "beat" sensation in the voice. This amplitude modulation occurs at a specified rate and depth, where "rate" refers to the frequency of the modulation and "depth" indicates its intensity. This effect introduces an additional rhythmic dimension to the vocal sound. The general form for a tremolo is given by:

$$y(n) = [1 + \alpha m(n)] \cdot x(n) \tag{1.3}$$

Where $y(n)$ and $x(n)$ are the input and output, α is the quantity of modulation to apply (depth of modulation), and the internal frequency of $m(n)$ is the rate of tremolo.

[1] https://www.youtube.com/watch?v=YmBA4syh1dA

[2] https://www.soundonsound.com/techniques/inside-track-bjorks-vulnicura

[3] https://www.billboard.com/music/music-news/taylor-swift-1989-louder-acdc-back-in-black-6538870/

Another is the ***ring modulator***, which produces synthetic or "metalized" tones, then considered a ***timbral effect***. It is widely used to create robotic voices and sound effects in electronic music and science fiction. It takes two input signals, commonly referred to as the "carrier" and "modulating" signals, and multiplies them to generate an output:

$$y(n) = x(n) \cdot m(n) \tag{1.4}$$

A sonorous example of ring modulation in soft and hard configurations can be check in: https://www.youtube.com/shorts/t9Z2GELc-hw. Tremolo modulation effects are extensively explored in tracks such as "xanny" [Billie Eilish, 2019] from the segment at 00:36-00:50, "bad guy" [Billie Eilish, 2019] at 00:56, "ilomilo" [Billie Eilish, 2019] at 00:48-00:57 and in a 2023 remix of the 1998 song "Frozen" [Madonna and Sickick, 2021] in time segments 0:19-0:23, 0:54-0:58. Ring modulation could be the effect used in "CUUUUuuuuuute" [Rosalía, 2022] from the segment at 01:02-1:05.

1.5 Time perception effects

Before diving into the second classification category, we will tackle a delicate topic. Similarly to classification by technique, a single effect, from a perceptual viewpoint, can impact several descriptors upon which we are basing our classification. For example, an effect that is fundamentally temporal can influence tone perception since these two variables are intrinsically linked in signal terms. Nonetheless, each effect has a primary musical purpose, and it is this purpose that allows us to categorize it appropriately.

1.5.1 Time Stretching

1.5.1.1 Variable Speed Replay

Time, one of the critical sound corpus descriptors, is crucial for the voice which is recorded on a timeline. The signal, sampled at a specific frequency, is defined by this line, and altering the playback frequency modifies the temporal perception of the sample. If, for example, it is played back at half the sampling frequency, the reproduction will take twice as long. And, since any finite signal can be decomposed into sinusoids, each will also take twice as long to play back, thereby halving its frequency and consequently sounding an octave lower. This technique is called resampling, and this type of temporal stretching results in modifying other descriptive parameters of the voice, such as pitch and formants, as well as of the signal that holds the vocal recording, such as duration.

1.5.1.2 PSOLA

Temporal stretching, while preserving the formants and pitch, is more complex than it may seem at first glance. Although sonorously, it may appear simpler to someone unfamiliar with the technique, in reality, the temporal stretching method is quite complicated and is performed in the time domain using the PSOLA method. This technique involves taking sound segments that coincide with the glottal pulses, as it is shown in figure 1.8, and performing a synchronous overlap-add between the segments, adding a segment that replicates a glottal pulse, and then the ends of the grain segments are alined by superposition. PSOLA allows temporal stretching while preserving both pitch and formants.

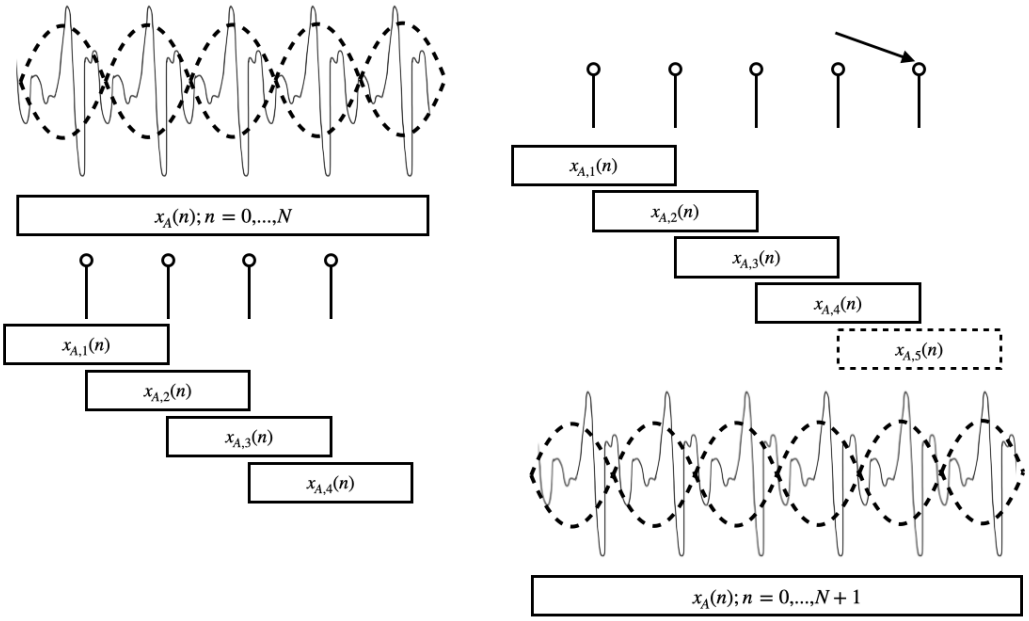


Figure 1.8: Diagram of the Psola Method, adapted from [Zölzer, 2011] and [Verfaille, 2003]

1.5.1.3 Phase Vocoder Approach

The time-stretching effect is achieved through phase vocoder techniques, using either the sum of sinusoids or the sliding FFT approach. Both methods require considering a phase condition, as outlined in equation A.35. This effect necessitates a reconstruction synthesis grid distinct from the analysis grid, and for simplicity, we consider here the case of integer hop size R_a and R_s. The core technique follows the sum of sinusoids approach, maintaining magnitude unchanged while adjusting the phase to preserve instantaneous frequency. This implies that the change in instantaneous frequency is determined by:

$$\Delta\psi(k) = \frac{R_s}{R_a}\Delta\varphi \tag{1.5}$$

that means that for the synthesis segment that have a change of phase equal to $\Delta\varphi$, each sample have a phase increment equal to:

$$d\psi(k) = \frac{\Delta\varphi}{R_a}, \tag{1.6}$$

For the output samples of the re-synthesis, we use:

$$\tilde{\psi}_k(n+1) = \tilde{\psi}_k(n) + d\psi_k, \tag{1.7}$$

and the sum of signals is given by:

$$y(n) = \sum_{N/2}^{k=0} A_k(n) cos(\tilde{\psi}_k(n)) \tag{1.8}$$

As previously mentioned, the sliding window (block-by-block approach) can also be employed for time-stretching. Output phase values calculation must adhere to the same condition as the sum of sinusoids approach, determining the unwrapped phase with a factor of $\frac{R_s}{R_a}$. It is imperative to consider specific details: the window size should equal the length of the analysis FFT hop size (R_a) and a submultiple of the synthesis IFFT (R_s). Notably, the sliding window implementation is faster than the sum of the sinusoids approach.

It is necessary to apply a circular shift (multiplying by $(-1)^k$) (The developed version is in the Appendix A.1, equations A.20 and A.26). Therefore, we do a zero-padded synthesis window, preferably truncated Gaussian windows, is recommended. This window should ensure correct re-synthesis for a ratio $\frac{R_s}{R_a} = 1$ corresponding to a no-time-stretching case.

A significant challenge with this approach is the unresolved phase unwrapping between different bins, which may vary from window to window. This issue, known as the dispersion of phase, is addressed by the proposed solution from [Laroche and Dolson, 1999] under the term "phase-locked vocoder." Assuming the processed sound comprises quasi-sinusoidal components, we can approximate the spectrum as the sum of these components. During time-stretching, phases must propagate accordingly, involving a constant phase rotation for each sinusoid, affecting all the spectral bins. In

1.5.1.4 Examples

The effects of time can exert a significant influence on vocal quality, thus qualifying as a timbral effect under appropriate circumstances. An example of varied tape playing can be observed in "Tomorrow Never Knows" [Beatles, 1966] at timemark 00:56-01:04, and at 1:26 with an experimental vocal timbral-time effect. Two additional examples of variable tape play are evident in "Third Stone From the Sun" [Jimi Hendrix, 1967] and "Habits (Stay High) - Hippie

Sabotage Remix" [Tove Lo, 2014], where vocals throughout the entire songs are accelerated, leading to the loss of formants.

Conversely, the utilization of recent time-stretching devices is notable in "Levitating (The Blessed Madonna Remix)" [Dua Lipa et al., 2020], where the vocal samples of Dua Lipa and Madonna are accelerated while still retaining their vocal identities. Another example of time stretching, including autotune and EQ-compressing chain, can be observed in "Ain't Me" [Kygo and Selena Gomez, 2017] at the timestamp 0:57-1:07, containing the phrase "the the the bowery, the the whiskey neat, grateful, I'm so, grateful," which is a refrain part of the song.

1.5.2 Temporal Inversion

The time inversion effect, also known as "reversed audio," involves playing recordings backward, resulting in a distinct and often ethereal sound while maintaining the original frequency and timbre. Utilized across both analog and digital platforms, this effect can create enigmatic voices and conceal messages, offering listeners a surreal auditory experience. Although explored in various media such as music, film, and advertising, it has also stirred controversy due to alleged subliminal messages, particularly within the realm of rock music. Artists and studios in the 1950s and 1960s, employing tape manipulation, delved into time inversion in audio to craft novel soundscapes. This technique is notably showcased in "Tomorrow Never Knows"[Beatles, 1966], where certain vowels were reversed during the chorus. Another more recent example can be found in "Tenochtitlan" [Mon Laferte, 2023] at the timestamp 3:47-3:51, featuring the phrase "Maria Madre de" reversed.

1.5.3 Granulation

The granular synthesis, originating from perceptibly "indivisible" audio segments, represents a form of controlled alteration of timbre and vocal characteristics. Alone or combined with other effects, it can become a highly useful experimental tool. A simple example of granulation involves generating segments that follow the expression:

$$g_k(i) = x(i + ik)w_k(i) \tag{1.9}$$

Where $i = 0, .., L_{k-1}$ represents the length of the audio segments. w_k represents a window with fade-in and fade-out. Longer grains may preserve more timbral content, while shorter ones may resemble pulses. Filtering or other effects can also be mixed before or after granulation to achieve more varied results.

Regarding vocal granulation, a prominent exemplar is Björk [Björk, 2004]. She achieves a heightened level of intricacy in her album "Medúlla" (2004), which comprises entirely vocal tracks and vocal effects. This approach is particularly evident in songs such as "The Pleasure Is All Mine" at timestamps

0:09-0:33, 1:02-1:10, and 1:13. Additionally, another more recent example is "Mycelia" [Björk, 2022], with timestamps at 0:13, 0:23, and 0:37, or in "claws" [Charli XCX, 2020a] at 2:18-2:29 in hyperpop music genre, openly explored in her album "how i'm feeling now" [Charli XCX, 2020a].

1.6 Spatial Effects

The spatial perception of sound, similar to that of vision, illustrates how we localize and process information across distinct cognitive levels and it is called auditory perspective. Factors such as intensity, spectrum, and timbral definition play crucial roles in this perception [Chowning, 1999], varying depending on the location and acoustic characteristics of the environment. Space perception can be modified though directionality modification, reverberation and adding spacial cues [Politis et al., 2012].

1.6.1 Amplitude Panning

Amplitude panning is a virtual source positioning technique primarily based on loudness control. It involves applying gain to the signal sent to each loudspeaker to create a virtual position from the listener's perspective. The output of the $g_i = 0, 1, ..., N$ speaker can be written:

$$x_i(t) = g_i x(t) \tag{1.10}$$

Where the gain factors g_i must to be normalized as $\sum g_i^2 = 1$. Various methods can be employed for amplitude panning. The most straightforward type of panning is the stereo panning. If two speakers and a listener form a triangular configuration where the listener is placed at the same distance from the two speakers. The standard two-channel stereophonic setup, popularized in the late 1950s, consists of two loudspeakers in front of the listener, often deviating from the shown 60-degree separation as shown in Figure 1.9, it was initially proposed by Clément Arder in 1881 [Henrich Bernardoni, 2014].

Despite variations in domestic or car audio setups, two-channel reproduction is preferred over monophonic configuration, it gives more richness to the music recording. Although panning was initially developed for loudspeakers configurations, it can also be utilized for headphones. The aim is to achieve a perception of movement or positioning of the sound source in the auditory space. The advantage of an headphones system is that the signals for each ear are isolated, thus avoiding losses due to phase differences of the signals upon entering the ear.

Some examples of songs with the use of vocal panning are as follows: "I Wanna Sex You Up" [Color Me Badd, 1991], where throughout the song, a vocal loop moves from one extreme to the other. In the track "Forbidden Love" [Madonna, 2005], at time marks 0:23-0:35, there is an echo that additionally employs panning, alternating from side to side, similarly in the segments 0:45-0:50. In the song "As Heaven is Wide" [Garbage, 1995], at time stamps 2:09,

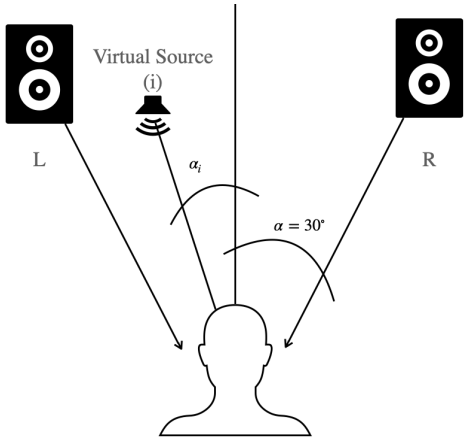


Figure 1.9: Configuration of loudspeakers in stereo, adapted from [Zölzer, 2011] and [Verfaille, 2003]

2:13, 2:17, 2:20, a phrase "I Wish" oscillates from side to side and repeats four times, also at 2:52 and 3:33. In the track "Who Is It" [Björk, 2004], from 0:00-0:21, there is a set of stereo-arranged tracks; then, from 0:32 onwards, the vocal tracks have a well-differentiated and fixed stereo arrangement that allows for cohesive localization of the sources.

1.6.2 Echo (Delay)

An echo is an acoustic phenomenon that occurs when sound reflects off one or more surfaces and returns to the place of origin after a perceptible period of time, typically at least 50ms. This phenomenon is commonly observed in open and mountainous areas or enclosed spaces, provided that the distance to the walls allows the reflected sound to maintain a sufficient level for perception. Musically, echo can be utilized to enhance the depth and spaciousness of sound. In its simplest form, the echo algorithm creates one or several replicas of the input sound, each with attenuation and a time delay, as shown in figure 1.10 (left side). More complex systems may employ a feedback system wherein the repetitions are attenuated with each cycle, as figure 1.10 (right side), and the input level also decreases when the sound ends. Thus the echo becomes inaudible when its level is comparable to the ambient noise level.

Examples of echo can be discerned in musical compositions such as "Forbidden Love" [Madonna, 2005], where echoes are notable at time intervals 0:23-0:35, employing panning techniques that alternate from one stereo channel to another, similarly observable in segments from 0:45 to 0:50. Additionally, in "Human Being" [Robyn and Zhala, 2018], echoes are evident at marks 0.30, 0.38, 1:06, 1:13, and 1:21, among others, emphasizing specific words like "being" and "body." Another illustrative instance is found in "Neon Lights" [Loreen, 2022], where echoes and automated panning are synchronized with the singer, who strategically utilizes them to enhance harmonies at 0:29, 0:41, 0:51, 1:39, 1:45, 1:57, and various other points. A live rendition of this song,

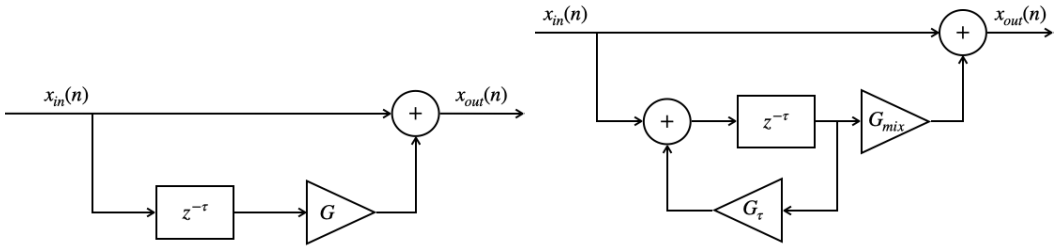


Figure 1.10: The simplest case of delay-echo. Right: Echo with feedback. Adapted from [Zölzer, 2011] and [Verfaille, 2003].

accessible via the following link [4], showcases a more pronounced utilization of echo and reverb effects, particularly noticeable at marks such as 2:39 and 3:42.

1.6.3 Reverberation

Reverberation and room acoustics are the first phenomena that humans used to express and transform sound, for example, historical public performances were perceived in a different way depending on the place and its acoustic properties [Bouty and Sabine, 1901]. Reverberation results from reflections of sound on the surfaces of an enclosed room. This phenomenon acts like a filter/coloring effect influenced by the room itself and it differs from an echo. Reverberation involves fast, multiple reflections that are close in time, while an echo has distinctive reflections that are well delayed in time and clearly distinguishable from the original sound. Moreover, reverberation reflections are so close in time that they are perceived as an ensemble, contributing to the perception of space, including the characteristics of the materials (environment) in the room. *Reverberation* is a robust perceptual cue that allows the identification of the properties of room surfaces (absorptive, reflective, diffusive) and can be divided into three components: direct sound, primary reflections (short echoes), and secondary reflections (later ones). There are three principal perceptual attributes of reverberation:

Reverberation modeling has four principal approaches: delay lines, sets of all-pass filters/comb filters, networks of delay lines with feedback, and convolution with a room's impulse response. The use of delay lines and filters, prominent in the 80s, was pioneered by Schroeder and Logan and extended by Chowning. The recursive comb filters and delay-based allpass filters for the inexpensive simulation of echoes were introduced by Schroeder at Bell Labs. The Schroeder allpass filter based on the recursive delay line is shown in Figures 1.11 and 1.11 [Zölzer, 2011] [Schroeder, 1962]:

$$x_{out}[n] = -gx_{in}[n] + x_{in}[n-m] + g[n-m] \tag{1.11}$$

[4] https://www.youtube.com/watch?v=advvAPiYQRw

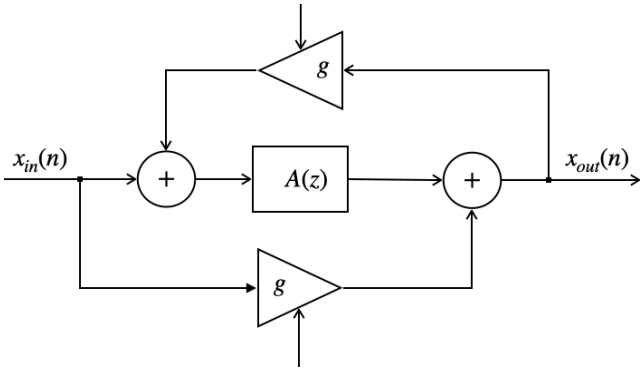


Figure 1.11: An all-pass Schroeder filter mixes the direct undelayed sound and the feedback delayed sound, where $A(z)$ corresponds to a delay and g to a gain. The use of these filters in cascade allows for an aperiodic echo response and increases echo density. Adapted from [Zölzer, 2011] [Verfaille, 2003] [Moorer, 1979b]

In the seventies, Michael Gerzon expanded the allpass filter to a multi-input multi-output structure, increasing the complexity of the impulse response. Later, Moorer incorporated additional elements such as other allpass filters, parallel comb filters, gains, and lowpass filters. Finally, in 1985, Julius Smith introduced digital waveguide networks based on:

$$x_(n) = x(n-m) + gy(n-m) \tag{1.12}$$

Where the m-samples delay is replaced by several delays in parallel m_i and the feedback gain is replaced by a matrix G, as shown in Figure 1.12.

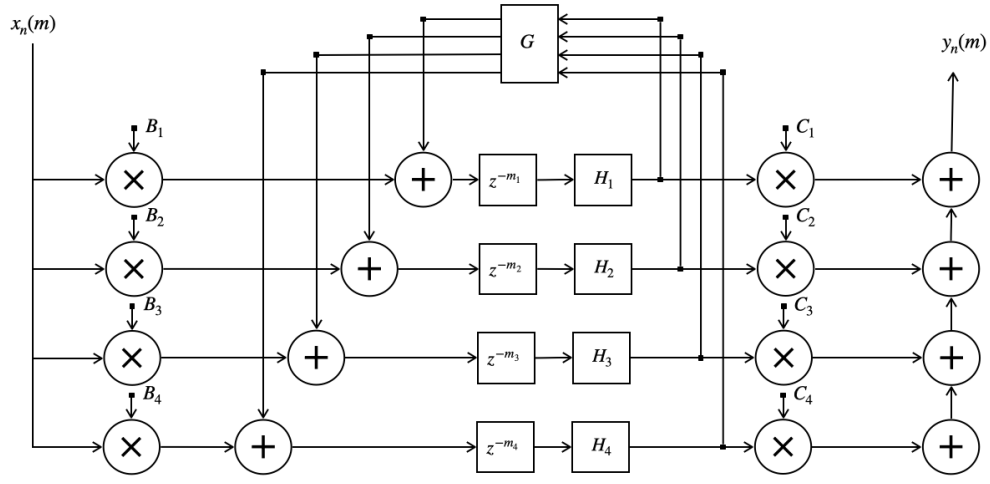


Figure 1.12: Delay Network Reverb, adapted from [Zölzer, 2011] and [Verfaille, 2003]

Reverberation can also be studied about the room's geometry, for which two main modeling approaches can be employed. Wave-based methods aim to

numerically solve wave equations using finite element and boundary element methods,. Ray-based methods, assuming a particle-like behavior similar to light, focusing on finding trajectories through image-source methods and ray tracing. This thesis will not address such models as they are quite extensive. We merely mention them to highlight their significance. However, for our perceptual taxonomic purpose, the key point is that any reverberation technique helps simulate speaking or singing within a space with certain reverberation characteristics, and that this serves a specific musical or stylistic intention.

In general, reverb, alongside compression and delay effects, plays a pivotal role in shaping the sonorous characteristics of vocal tracks [5] even for vocal song like "Rolling in The Deep" [Adele, 2011]. To elucidate this concept, we have elected to showcase instances where intensive reverb is utilized for artistic effect. In the track "Tattoo" [Loreen, 2023], by Loreen, a pronounced reverb effect coupled with a subtle echo is consistently integrated throughout the composition. This effect is particularly discernible at the conclusion of each phrase within the refrain and recurs in various sections of the song, notably following the 1:06 mark. Furthermore, it is noteworthy that the release of the reverb is modulated through automation, heightening its impact on certain input levels of the vocal track after 2:10.

Another artist known for stylistically employing reverb is Lana Del Rey, notably in tracks such as "Summertime Sadness" [Lana Del Rey, 2012b], where reverb envelops the entirety of the vocal performance, contributing significantly to the overall atmosphere of the song, same as "Doin' Time" [Lana Del Rey, 2012a]. Mon Laferte also utilizes reverb to create transitions in her album "Autopoietica", particularly evident in the song "NO+SAD" [Mon Laferte, 2023] from the 2:08 mark until the end. The use of reverb in that segment contrasts completely not only in musical style but also in spatial perception with the first part of the track. Other examples can be tracks such as "Team" and "Royals" [Lorde, 2013].

1.6.4 Binaural audio and 3D audio

Humans employ various mechanisms to analyze the position and direction of a sound source. Sound signals originating from different points in space reach each human ear canal differently based on the source's direction. This discrepancy is approximately described by head-related transfer functions (HRTFs), influenced by factors like arrival time at each ear position, skull-induced shadowing (relevant for $freq > 2kHz$, and filtering. HRTFs also depend on the relative distance to each ear and involve a more intricate neuronal learning process. There are mechanisms such as interaural time difference (ITD) and interaural level difference (ILD) that enable listeners to interpret the angle between the sound source and the median plane. Head movements also affect binaural signals, favoring the ear closer to the source. The combination of these mechanisms allows humans to perceive the spatial position and details

[5] https://www.soundonsound.com/techniques/tom-elmhirst-recording-adele-rolling-deep

of sound sources.

The binaural effect is achieved through precise binaural recordings or the use of head-related transfer functions (HRTF). Binaural recordings are made with the help of a mannequin with a structure similar to a person's ear canal. The results of binaural audio allow for a very efficient simulation of reality, with limitations such as physiological differences among individuals and a lack of dynamics (sound variation when moving the head).

Binaural perception depends on distance to the source, orientation (β_{plane}), and elevation ($\beta_{\text{elevation}}$) as shown in Figure 1.13 (on the left). The reconstruction of a binaural audio system with a stereophonic loudspeaker configuration can be seen on the right side. Some advanced 3D and binaural commercial audio tools are Ircam Spat[6] and Panoramix [7]. Examples of Dolby Atmos (based on filters) can be listened in Apple Music; in the section for spatial audio, audio is only compatible with AirPods or suggested by Apple devices.

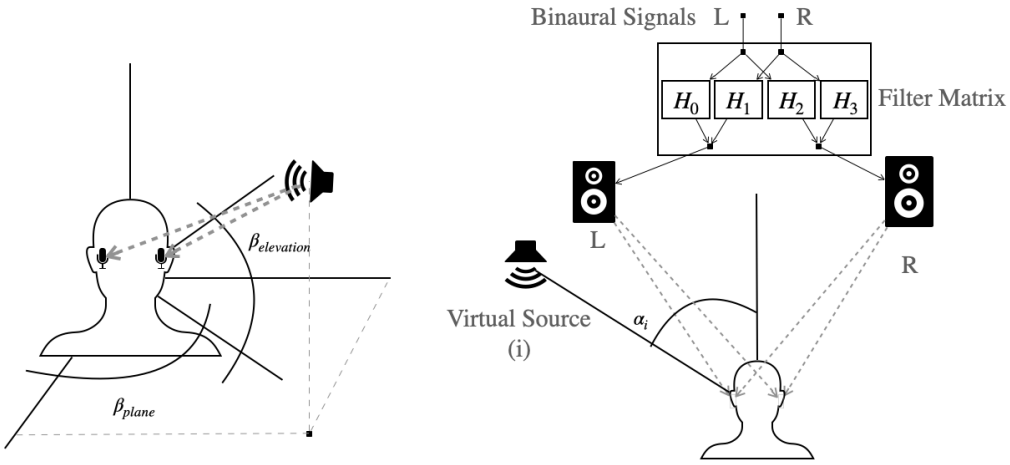


Figure 1.13: Distance, orientation and elevation perception, and binaural audio reconstruction. On the left, the subject is located at the origin, and a sound source is positioned at an arbitrary distance r in space. The position of the source is defined by the distance r, the orientation in the plane β_{plane}, and the elevation above the plane$\beta_{elevation}$. On the right is a binaural stereo system for a virtual source position sonorous reconstruction. It uses two signals, L and R, which pass through four filters representing the ear's responses to the signals directed to each ear from either the L or R side. The diagram also includes the speaker setup and the person's positioning. Adapted from [Zölzer, 2011] and [Verfaille, 2003]

[6]https://forum.ircam.fr/projects/detail/spat/
[7]https://forum.ircam.fr/projects/detail/panoramix/

1.7 Effects for Pitch Changes

The melody can be considered in its more simple form as is the succession of musical sounds [Apel, 1969] [Rehding and Rings, 2020] [Hijleh, 2012]. A melody can helps to express a musical idea, regardless of tonality, microtonality, tuning, or the poetic explanation cause they vary according the musical style. The melody of a composition contains various components, such as temporal, timbral, and dynamic elements. Pitch contour combined with silences (Rythm) define the melody.

Pitch refers to the perceived height of a sound and is related to the frequency of the sound wave. Pitch helps us identify sounds as high or low, facilitating the recognition of a melody when played by two different instruments or singers. It is also related to age and gender; for example, a child will generally sing higher than a woman, and a woman, in turn, will sing higher than a man (generally). According to [Hallam et al., 2016] and [Hijleh, 2012] we can conceptualize pitch patterns in a vertical dimension as harmony, and horizontal dimension as melody.

1.7.1 Pitch-shifting and Transposition

When a vocalist aims to transition to a higher pitch, the necessity for altering the tonal register is compounded by concurrent shifts in timbre, complicating the transposition process. In addition to timbre changes, there are relationships among the so-called formants of the voice. These formants manifest as peaks or maxima within the spectrum, with their respective positions influencing phoneme perception. Therefore, this transposition should stretch the spectrum, preserving the resonances of the formants. Here, we will discuss some methods for pitch-shifting.

The general method for creating pitch changes is called pitch-shifting. To achieve this, each frequency must be multiplied by a transposition factor. Pitch-shifting can be done in various ways; for example, one can resample a time-stretched signal and then return it to its initial duration. However, there are solutions that allow more direct calculations of the output signal.

1.7.1.1 Delay Line Modulation

Delay Line Modulation is a time-segment technique where the signal is divided into small chunks (small buffers of audio) that are reproduced faster or slower to produce higher or lower pitches. There are two delay lines that generates audio chunks (small buffers) using sawtooth-type function. To produce a continuous signal output, two chunks are read simultaneously with a time delay equal to one half of the cross-fade block length (two half are equal to 1). A cross-fade is made from one chunk to the other at each end of a chunk. The algorithm is shown in Figure 1.14.

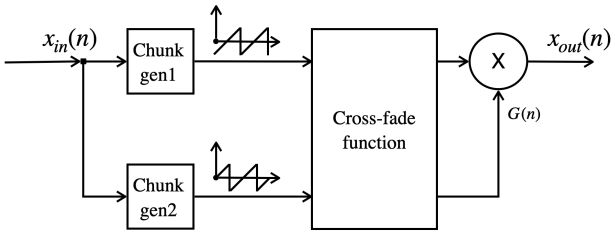


Figure 1.14: Delay-Line Modulation based pitch-shifter, adapted from [Zölzer, 2011]

1.7.1.2 Filter Bank - Phase Vocoder

According to the vocoder techniques discussed earlier and taking the reference [Zölzer, 2011], we start from the equations already established for time stretching. Taking these equations, we calculate the phase increment from:

$$d\varphi(k) = \Delta\varphi(k)/R_a \tag{1.13}$$

Then, we multiply the phase increment by the transpose factor t_f so:

$$d\psi(k) = t_f \Delta\varphi(k)/R_a \tag{1.14}$$

and add it to the previous phase to obtain:

$$\tilde{\psi}_k(n+1) = \tilde{\psi}_k(n) + d\psi_k, \tag{1.15}$$

$$\tilde{\psi}_k(n+1) = \tilde{\psi}_k(n) + t_f \Delta\varphi(k)/R_a, \tag{1.16}$$

Finally, we calculate the sum of sinusoids in the same way as in time-stretching. When the transpose factor is greater than one, we keep only the frequencies that satisfy the Nyquist theorem, i.e., taking N/t_f frequency bins.

1.7.1.3 Pitch-shifting Preserving Formants

The problem of formants is significant, as it can lead to intelligibility issues such as the "Donald Duck effect". The general approach involves calculating the spectral envelope of the sound (using techniques like cepstrum or linear predictive coding LPC), performing transposition (with methods like phase vocoder or additive model), and then correcting the peaks based on the original spectral envelope [Verfaille, 2003].

The frequency domain pitch shifting technique is based on a formant shift prior to the reconstruction of time-stretched grains, as shown in Figure 1.15. First, the log values of both the input and interpolated input signals' FFT are calculated. Next, the difference between these two generates a spectral correction factor, which is transformed to the cepstrum domain, low-pass weighted, and transformed back to the frequency domain. Then the correction is applied

to the time-stretched grain FFT, and finally, the grains are resampled and added.

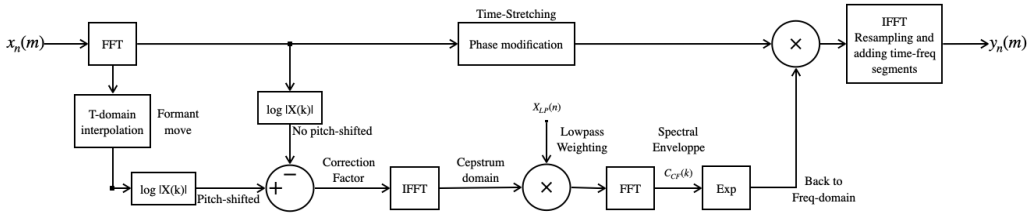


Figure 1.15: Example of pitch-shifting preserving formants

The general idea is to perform a source-filter separation and enforce the spectral envelope during reconstruction. This can also be achieved through two other techniques known as cepstrum and Linear Predictive Coding.

1.7.1.4 Examples

The use of modulation can be found in the track "Anything Could Happen" [Ellie Goulding, 2012] throughout the entire song and is audible at 0:12 and 0:23, according to [8]. Modulation can be achieved through devices prior to the vocoder, such as the Talkbox, where the signal is modulated by an instrument, as seen in "Digital Love", [Daft Punk, 1997], the segment 1:03-1:34 which is repeated over all the song.

Pitch shifting with formant loss as a stylistic hallmark can be found in some musical pieces such as "Gorgeous," [Taylor Swift, 2017] in segments 0:57 and 2:03. Additionally, some opt to use only formant deformation, as in "Team" [Lorde, 2013], from time 0:20 to 0:35. While pitch shifting is typically used as support in precise segments of songs, there are some tracks like "Good Love" that are entirely performed using pitch shifting without formant loss, or like "Te Juro Que Volveré,"[Mon Laferte, 2023] which is performed using pitch shifting with formant loss, and is even performed live with the same arrangement and without autotune [9], indicating that it is simply a stylistic choice. Other examples of pitch shifting with formant preservation include "Set Me Free,"[Robyn & la bagatelle magique, 2015], which uses a higher note for the segment 0:34-1:02 and a lower one for the segment 1:04-1:18. However, pitch correction is also used to provide greater support to the vocal sample, as in "Apeshit" [The Carters (Beyoncé, Jay-Z), 2018], especially in segment 0:47-0:52 [10], or systematically putting in stereo two tracks pitch-shifter in "What About Now" [11] [Bon Jovi, 2013].

[8] https://www.soundonsound.com/people/mix-review-46
[9] https://www.youtube.com/watch?v=A90nnXICJlA
[10] https://www.soundonsound.com/techniques/inside-track-beyonce-and-jay-z-apeshit
[11] https://www.soundonsound.com/techniques/inside-track-bon-jovis-what-about-now

1.7.2 Harmonization

Harmonization is an effect that can be applied with the help of transposition; it is possible to create a layer of harmonization as long as the pitch-shifting effect has adequate quality. In other words, through harmonization, a vocal layer is created with a defined transposition that follows the melodic progression of the original vocal. The presence of an additional vocal layer can create harmonies that enrich the audio. Therefore, if used in conjunction with a filter and multiple transpositions, more body can be given to the vocal sample, creating a choir-like effect or simply enhancing its presence. Examples can be "What About Now" [Bon Jovi, 2013], or in segments of "Cruel Summer"[Taylor Swift, 2019] using several layers of vocals and vocoded voices, especially in 1:06-1:08, 1:27-1-33 among others moments, in Me! in the segment 0:00-0:02 [Taylor Swift, 2019], or in "The Contorsionist" [Melanie Martinez, 2023a] in the second vocal layer at time marks: 0:39-0:50 or 1:04-1:08.

1.8 Effects for Percetive Timbral Changes

The spectrum of an instrument changes as the loudness changes, and the same occurs for voice. This happened because the force applied and its distribution to active the different modes of vibration change for louder sounds. However, it has been shown that listeners perceive loudness with a preference for spectral cues above acoustical intensity [Chowning, 1999].

So, all effects applied to a sound signal will affect its spectrum somehow; however, the degree of coloration will vary depending on the effect. Here lies a fundamental difference between the specificity of a vocal signal and an instrumental or other type of signal. A vocal signal has a rich prosodic content, for which our brain is trained to analyze. In other words, we can identify subtle changes much more efficiently than spectral changes or colorations in other non-vocal or instrumental sounds. Effects that may be interesting in musical production can be destructive to the voice. Therefore, the classification of what we propose as vocal spectral effects differs from traditional classifications of such effects.

We consider as spectral effects all those that aim to introduce coloration into the vocal signal. We classify them into three types: those that preserve vocal quality, those that destroy vocal quality, and those that change vocal quality. Consider that this classification is not about a side effect due to another effect with a different goal (dynamics, tone, time, space), but the main objective is to change spectral content. This classification can be summarized as it follows:

1. Effects that Preserve Vocal Quality: These effects subtly add coloration while maintaining the actual voice recognizable. These types of effects enhance the vocal sample within the musical mix. The use of these effects is often contextual and depends on the specific musical style in which they are employed.

2. Effects that Distort Vocal Quality: These effects apply strong coloration while keeping the original voice slightly recognizable. Some effects depend on parameters; if used in an "extreme" configuration, they end up distorting the vocal quality. These extreme configurations are included within this category.

3. Effects that Change Vocal Quality: These effects significantly alter the signal to the extent that the quality of the original voice is lost. This can include changes in size, vocal range, or other vocal qualities.

The dividing line between these effects lies in their intentional use; for example, a "whisperization" effect can be used to add a hint of whisper to the voice or to entirely remove vocal timbre. A "pitch-shifting" effect with formant preservation can be used to transpose within a small range or to move two octaves. While the technique is the same, the intentional use—determined by the configuration parameters—defines whether the effect is being applied to preserve, distort, or completely change the vocal qualities.

1.8.1 Effects that preserve vocal quality

This category includes effects specifically designed to enhance the spectral content of the vocal signal intentionally. The aim is to make adjustments that improve quality without significantly altering vocal quality, as well as the prosody and linguistic content of the voice. Within this category, effects that produce coloration as a side effect are not considered. This includes coloration resulting from specific characteristics of reverberation or the spectral modification induced by compression, which may impact different frequency bands.

The vocal audio signal has many details and characteristics that shape the vocal quality of the person singing. Audio effects, particularly those in the frequency and time-frequency domain, must allow for the preservation of vocal quality. This means avoiding distortions such as phasiness or the "Donald Duck" effect.

The category of effects that preserve vocal quality is based on subtle changes that maintain the prosodic-linguistic content as well as the primary qualities of the voice. For example with effects specifically designed to intentionally enhance the vocal signal's spectral content. This application is subjective and depends on the musical context. For instance, one may desire a vintage-style filter for the voice or want it to stand out above the instrumentation.

Within the spectral modification effects, there are various types of filters, such as low-pass, band-pass, high-pass, and notch filters, including different techniques for their implementation. This category also encompasses equalization effects, which involve various techniques such as parametric equalization, dynamic equalization, and equalization, among others.

One example of the use of formants modification and harmonization with the device SoundToys Little Alterboy is the song "Don't Start Now" [Dua Lipa,

2020] [12]. Billie Eilish also uses this Device in the previous referenced songs in this chapter. All the examples of compression, reverb, and modulation noted before also apply to this category when the vocal quality is preserved.

1.8.1.1 Filters

Filtering effects are very common for both instrumentals and vocals, as they are used to add a certain coloration to sound samples and enhance the cohesion of the musical composition. It is most common to use filtering effects to pass certain parts of the spectrum. As their name suggests, low-pass, high-pass, and band-pass effects allow the passage of specific spectrum regions and are typically parametrizable with elements such as quality factor and cutoff frequency. On the other hand, notch filters can attenuate a narrow band of frequencies, making them helpful in correcting unwanted resonances or interference. The use of filters serves as the foundation for parametric equalization, signal degradation (like telephonic line using a bandpass filter) and phasing effect.

Filters have an electric origin, and their digital pairs are intended to follow a similar transfer function. For example, for the given digital state variable filter in Figure 1.16, the corresponding transfer function is given by:

$$H(z) = \frac{r^2}{1 + (r^2 - q - 1)z^{-1} + qz^{-2}} \tag{1.17}$$

In this filter, $x_n(m)$ is the input and $y_n(m)$ is the output. $H(z)$ is the transfer function that depends on the values of $r = F_1$ and $q = 1 - F_1Q_1$, where F_1 and Q_1 are the tuning parameters of the filter. $H(z)$ represents the delay line version of three parallel analog filters: high-pass, band-pass, and low-pass.

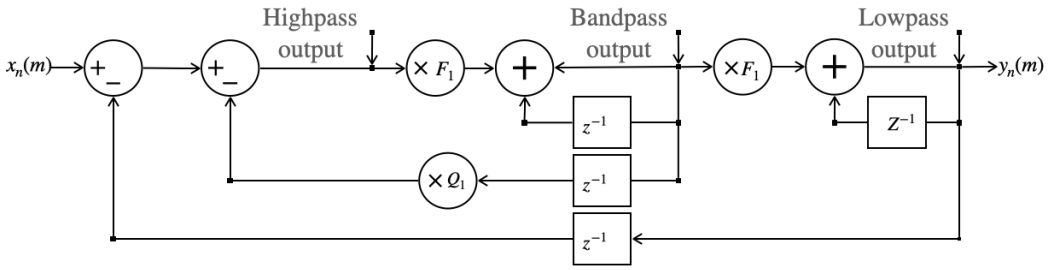


Figure 1.16: Digital state variable filter, adapted from [Dutilleux, 1998]. The digital version of the analog state variable filter containing three filters: high-pass, band-pass, and low-pass. The transfer function is given by $H(z) = \frac{r^2}{1+(r^2-q-1)z^{-1}+qz^{-2}}$ where $r = F_1$ and $q = 1 - F_1Q_1$, which depend on the tuning parameters F_1 and Q_1.

[12] https://www.soundonsound.com/techniques/inside-track-dua-lipa-dont-start-now

1.8.1.2 Equalizers

Unlike filters, which attenuate the spectrum concerning a cutoff frequency, equalizers shape the spectrum by enhancing certain frequency bands. They are typically constructed as a sequence of shelving and peak filters. The shelving filters boost or cut the bands with a cutoff frequency and a gain, and they have the form described by:

$$H_s(z) = 1 + \frac{H_0}{2}[1 + -A_s(z)] \tag{1.18}$$

Where the sign $+/-$ works for a low-pass and high-pass filter respectively, And from [Zölzer and Boltze, 1995] [Zölzer, 2008]:

$$A_s(z) = \frac{z^{-1} + a_{B/C}}{1 + a_{B/C}{}^{-1}} \tag{1.19}$$

Where the variables a_B and a_C refer to boost and cut respectively and depend on the cut-off frequency. The peak filters, in addition to using a cutoff frequency and a gain G, also employ a bandwidth f_b and are given by the transfer function:

$$H_p(z) = 1 + \frac{H_0}{2}[1 - +A_p(z)] \tag{1.20}$$

Where:

$$A_p(z) = \frac{-c + (d - dc)z^{-1} + z^{-2}}{1 + (d - dc)z^{-1} - cz^{-2}} \tag{1.21}$$

Where $-/+$ denotes band-pass and band-reject operations. Parameters d and c are related to the band and the cut-off frequency. Peak filters offer nearly independent control of their control parameters. When used in conjunction with shelving filters, they are useful for creating equalizers with a structure as Figure 1.17.

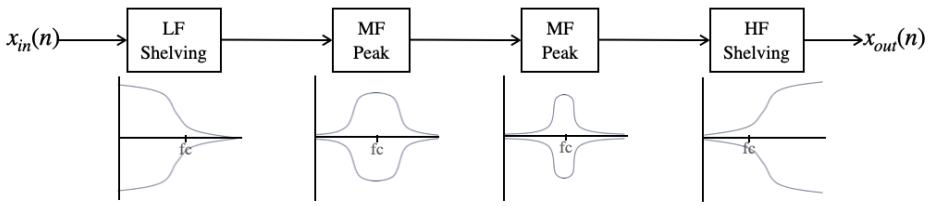


Figure 1.17: Equalizer based in shelving and peak filters, adapted from [Zölzer, 2011]

1.8.2 Effects that can distort the vocal quality

In this section, we address effects whose main objective is to distort the vocal sample in an extreme manner or even destroy its vocal quality. That is to say, there is a deliberate intention to apply distortion that makes it impossible to recognize the vocal quality of the person singing or speaking. The goal is also to make the signal lose its prosodic content to the extent that we cannot determine if it is a person speaking. Vocal distortion finds its utility, especially in experimental electronic music, and presents numerous approaches. However, in this section, we will first focus on some effects that have been intentionally developed to create distortion, where distortion is not an occasional side effect but a central goal.

1.8.2.1 Time varying filters

Variable time filters are filters that either use direct control over the parameters controlling the filter or an oscillator that automatically controls the filter. For example, flanging is an effect that uses a variable delay added to the input signal. The delay is periodically varying as $d(n) = \frac{D}{2}(1 - \cos(2\pi F_d n))$, with D between 0 and 10 ms and F_d around 1 Hz. The output of the delay is added to the input signal, resulting in a periodic shape $H(\omega)$ transfer function, as shown in Figure 1.18.

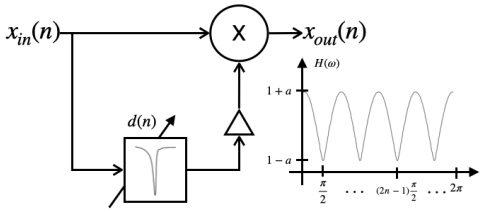


Figure 1.18: Time varying filters - flanger, adapted from [Strange, 1983]. Based in time varying delay generates a variable transfer function.

Phasing (phase shifting) is another time-varying effect achieved by passing the signal through a narrow notch filter (Band Reject Filter) and combining a portion of the signal with the direct sound. The time variation occurs in the notch filter, which is controlled by a low-frequency oscillator (LFO). The phases are combined so that cancellations or enhancements occur. A typical realization of this effect is shown in Figure 1.19, where the notch filters can be controlled independently.

For example a phaser is used in Music [Madonna, 2000] in the segment 0:14-0:30 with lost of vocal quality, and variable lost of quality in the segment 0:56-1:04.

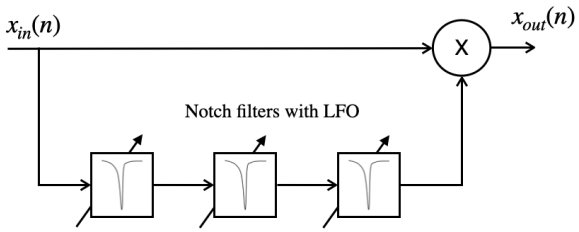


Figure 1.19: Time varying filters - phaser, adapted from [Zölzer, 2011] and [Strange, 1983]. Multiple notch filter controlled by LFO generates the phase shifting effect.

1.8.2.2 Distorsion and Overdrive

Distortion effects made their appearance in the 1940s with different techniques for capturing and amplifying acoustic and electric guitars. Distortion began as a flaw in amplification processes, which turned out to be musically useful, becoming an essential part of rock and influencing music immensely to this day. The exploitation of these distortion effects occurred when attempts were made to use them to create higher harmonics. There are mainly three terms known: overdrive, distortion, and fuzz. These effects can be applied to vocal samples, leading to a rapid degradation of the voice. An example of digital treatment for this kind of effect is given by [Zölzer, 2011].

$$f(x) = \begin{cases} 2x & for \quad 0 \leq x < 1/3 \\ \frac{3-(2-3x)^2}{3} & for \quad 1/3 \leq x \leq 2/3 \\ 1 & for \quad 2/3 < x \leq 1 \end{cases} \tag{1.22}$$

Through the piecewise-defined function, symmetric soft clipping is included at the edges, which also generates compression simultaneously on the edges. When a logarithmic scale is applied to this function, a linear response can be observed along with compression at the upper end. This effect is known as overdrive.

Overdrive is used in "Rolling in the Deep" [Adele, 2011] [13] without losing vocal quality, so this example shows how the preset is fundamental and can be used to distort or, on the contrary, to improve vocals, as in this case.

1.8.2.3 Reusing other effects for distortion

Effects, such as those discussed earlier for perceptually addressing dynamics, space, time, and timbre, have been carefully developed to generate those specific changes and avoid introducing distortion. In these effects, distortion can be considered an undesirable side effect. In other words, the goal is always to keep these effects as clean as possible. However, the parametrization that allows for filtering, equalization, and reverb can sometimes introduce distortion. This distortion may be quite pronounced, so these types of effects, even

[13] https://www.soundonsound.com/techniques/tom-elmhirst-recording-adele-rolling-deep

though distortion is not their initial purpose, can be used for vocal distortion. Now, we can consider these effects, with their timbre-shaping capability, within this category, as long as their aim is the alteration of timbre and a highly destructive coloration of the initial vocal qualities. In this context, distortion, while a side effect of the technique, becomes the primary intended usage, and the effects of equalization or reverb become less noticeable secondary effects. For example, the song "Sorry" [Madonna, 2005] uses vocoders, variable equalization automated with a pitchshifter, and some kind of over drive over the segment repeating "sorry" in the segment 2:56-3:26, while other vocal layers played simultaneously with other effects of presets. Time-stretching, formants deformation, and pitch-shifting are also effects that can be used with a distortion purpose, as we showed before.

1.8.3 Effects that change vocal quality

As we have seen, effects can be used to modify signal properties such as dynamics, space, and time in a way that preserves vocal quality. It is also possible to intentionally modify the timbre with the help of various types of filters or equalizers. On the other end, there is the destructive use of vocal quality, in which these same effects are employed either to partially or entirely destroy the prosodic content and/or vocal quality of the person singing. In between these two extremes are effects that allow intentional modification of the vocal quality and/or prosodic content of the person singing. Below, we will provide examples of these types of effects.

1.8.3.1 Robotization with vocoder

Robotization is a signal processing technique based on the structure of a phase vocoder. In this effect, the phases are set to zero before the audio reconstruction. This means that, for a given grain of length L_w, the phases would be set to zero for each FFT before the reconstruction. It is as if each cosine making up the FFT had the same phase. With all components in phase, there would be a peak in the center of the L_w segment in the time domain, much like a cosine function over time. When combining all the grains, the resulting sound will be a robotic voice with a pitch equal to f_s/N_w, where f_s is the sampling frequency, and N_w is the window length. The larger the audio grains, the more secondary peaks there will be, and the greater the influence of the initial pitch of the voice.

Some examples of this kind of effect are: "Harder, Better, Faster, Stronger" [Daft Punk, 2001] from time mark 1:45, "Got to Work it Out" [Robyn & la bagatelle magique, 2015] at time marks 0:21:023, 0:28-0:38 and all over the song, and Sorry [Madonna, 2005] at time mark 0:24-0:47. In "Death" [Melanie Martinez, 2023b] in the segment 2:01-2:21, or in "Monster" [Lady Gaga, 2009] at different levels for each phrase of the chorus in the segment 1:05-1-35, different pitch and different gender too.

1.8.3.2 Timbre Scaling

This category includes effects specifically designed to intentionally enhance the spectral content of the vocal signal. Audio effects, in general, make changes without deeply altering vocal quality and prosody; such an approach to coloration is not considered here. An example of timbre scaling is formant modification [Verfaille, 2003], it used in the song "Diablo" [Rosalía, 2022] in 0:22-0:42 with some autotune, with the same configuration over other parts of the song and mixing with the regular voice of Rosalia. In "Te Juro Que Volveré" [Mon Laferte, 2023] formant modification is done without using autotune, as a stylistic choice. In "Bury a Friend" [Billie Eilish, 2019], in the segment 1:12-1:26, a harmonization layer is done with original vocals and formants modified vocals. Or in "Frozen" [Madonna and Sickick, 2021] in segments 0:14-0:18, 0:23-0:26, 0:31-0:44 among others time marks.

1.8.3.3 Gender Change

The voice of a man and a woman is situated in different frequency ranges, and the gender change in the voice is possible through signal processing, which can be carried out in various stages. The first stage will involve performing a transposition with pitch shifting while preserving formants. On the other hand, it should be considered that formants (especially the first one) vary as the fundamental frequency evolves; this can be modeled with a shift of the spectral envelope after transposition. For instance, if transitioning from a male to a female voice, this would be done above frequencies of 100 Hz up to 500 Hz, linearly shifting from 0 to 50 Hz. To transition from a female to a male voice, the reverse shift is performed in the same range. The shift depends on the fundamental, making it adaptive. One example can be of gender change is "Good Love" by Prince which is done through a pitch shifting technique preserving formants [Prince, 1998]; other examples can be "Te Juro Que Volveré" [Mon Laferte, 2023] and "Bury a Friend" [Billie Eilish, 2019] in the segment 1:12-1:26.

1.8.4 Other perceptual possibilities

There are various techniques to modify the voice. Although the vocoder originated as a tool for telecommunications, it has evolved into an instrument for vocal analysis and resynthesis. Numerous transformation tools have emerged, perceived as part of perfecting the vocoder. From our perspective, it is crucial to musically understand what can be achieved with the perceptual effect of these diverse techniques.

For example, pitch-shifting without formants yields more striking perceptual results than pitch-shifting with formants. While we acknowledge that the technique behind the latter is more complex and provides a higher-quality result (for perfection purposes), the change is less noticeable from a perceptual standpoint. Additionally, countless techniques allow the modification of

perceptual characteristics of vocal timbre, such as breathiness, vocal fry, and softness, among others.

Thus, it is possible to change the gender, age, and timbre of a person's voice and, consequently, their vocal quality. These techniques also enable the alteration of other parameters linked to timbre, such as the person's size, intentionality, emotional expression, etc.

1.9 Tuning - Discussion

The melodic progressions in singing are composed of a succession of tones executed through a vocal technique that manages attacks and adjustments of dynamics and timbre. In this context, effects that affect the pitch entail, in most cases, modifications in the perceived execution of vocal technique.

Effects like autotune rely on forcing certain melodic curves onto the signal, notably to make it follow a melodic line more faithfully to a given tonality. There is also the possibility of manually forcing a well-established curve, as with Melodyne. These types of tools involve a more complex pitch change than a simple constant transposition on an audio file; we will refer to this as dynamic transposition or dynamic pitch modification.

Pitch correction as an effect is systematically used in modern music, primarily driven by tools like Autotune. However, the pitch correction as an algorithm modifying the pitch curve has changed minimally since the release of Autotune. Therefore, it is interesting to explore how new pitch correction algorithms can be developed and under what premises they could be designed. The next chapter proposes a compendium of terms and case studies, introducing a novel pitch correction method.

Autotuning does not necessarily cause a loss of vocal quality, but it affects0 coloration depending on the amount of correction and the transposition, potentially leading to extreme distortion. This kind of effect was popularized thanks to Cher in 1998 with the song "Believe" [Cher, 1998] and took the name "Cher-effect" for more than one decade. Autotune is now one of the most essential vocal effects in the DAFx vocal chain in music, with various preset variations and applications. Melodyne is also a pitch modification device used to autotune manually and off-line melodies. The use of pitch correction and pitch modification techniques with Autotune is widespread. They are included in the mainstream hit songs of various genres such as: "I'm Sprung" by T-Pain [T-Pain, 2005], "One More Time" [Daft Punk, 2000], "Somebody That I used to Know" [Gotye, 2011][14], "What Do You Mean" [Justin Bieber, 2015] [The Chainsmokers, 2016][15], "I Took A Pill in Ibiza" [Mike Posner, 2015][16]. Undoubtedly, Autotune has aided these songs in achieving interesting and more robust harmonies, and in being appreciated both for intentional use seeking

[14]https://www.soundonsound.com/techniques/mixing-gotyes-somebody-used-know-francois-tetaz

[15]https://www.soundonsound.com/techniques/inside-track-justin-bieber-purpose

[16]https://www.soundonsound.com/techniques/inside-track-mike-posner-took-pill-ibiza

extreme correction transitions, as well as for coloration purposes in a makeup Autotune, that is, with a gentle correction without extreme transitions.

The vocoder, regardless of the type of technique used (time domain, phase vocoder, channel vocoder, artificial intelligence), is employed on techniques such as autotune or melodyne for pitch transpositions leading to secondary changes in vocal timbre. This raises the question of to what extent these vocoder reconstruction techniques imply a change in vocal quality. In other words, does the vocoder, as a processing technique, add an inherent coloration to the signal?... Furthermore, beyond these potential changes in vocal quality, the question arises of whether the possible coloration of the vocoder carries more or less weight than the melodic changes that can be imposed on the vocoder. The use of vocoder layers and Autotune can be observed in songs such as "Instant Crush" [Daft Punk, 2013], where several vocoders and autotune are consistently present throughout the entire song. Is it the coloration (or possible change in vocal quality) due to the transposition? Or is it due to the inherent modeling of the vocoder? And... Can effects such as autotune be considered just pitch effects? Or also timbral effects? We will implicitly explore these questions in this thesis, particularly in chapters 3 and 4.

Chapter 2

Pitch and Tuning Adjustment Methods

As discussed in the previous chapter, pitch transposition produces perceptible changes that can be musically intriguing. This could be one of the reasons why effects such as Autotune and the use of the vocoder as an instrument have gained popularity over the years. In just two and a half decades, genres like hyperpop have emerged, and the utilization of the vocoder and Autotune has extended across numerous musical genres globally, including more regional genres and various languages. The stylistic interest in this effect is undeniable, and its widespread use signifies a general appreciation for tonal and timbral transformations by both artists and the audience.

The central issue of this chapter revolves around pitch correction and ways to enhance it. While several important references and patents related to this topic exist, there are no references that allow the classification of the problem conceptually and establish a connection with musical usage patterns and potential artist applications. The standardization of the autotuning concept in the musical realm overlooks nuances within its conceptual usage. This complicates proposing innovations to the pitch correction problem. Nevertheless, we have developed simple and beneficial concepts from a cross-disciplinary perspective—applicable to engineers, musicologists, and musicians alike. These concepts help us to clarify the problem and to propose new solutions.

Furthermore, in this section, we will explore using a graphical method for pitch correction. The development of this proposal will help us identify different types of usage and pitch corrections and understand the improvements that our method brings to the table.

2.1 Proposal for Terminology

Pitch transposition and correction may seem like simple terms, but there is no precise definition for these terms within the autotuning topic. This section proposes an accurate terminology for pitch transposition and pitch correction and its types. This terminology is essential for understanding the following

sections of this thesis, but it is also crucial for future autotuning studies.

2.1.1 Pitch Transposition

As observed, the pitch transposition issues encompass various challenges, including the preservation of instantaneous frequency, formant preservation, and timbre preservation. These issues are crucial in assessing a pitch transposition method for functionality. Once a functional pitch transposition method is available, we then explore potential applications. As implied by its name, the objective is to carry out a transposition, which can be accomplished in several manners.

The types of pitch transposition can be seen in Figure 2.1. There are two reasons for distinguishing between types of transposition. The first is that depending on the vocal transformation algorithm or vocoder, the input value given to the system is usually the transposition value. There are very few devices that allow the desired pitch value to be entered at the output. The second reason is that the quality evaluation to which vocoders are subjected includes only tests with constant transposition. What happens is that when the transposition is variable, there is a remaining coloration in the audio that is not present when the transposition is constant.

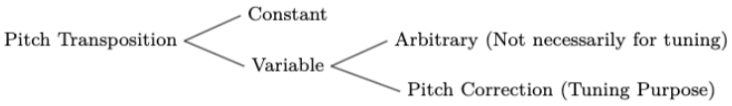


Figure 2.1: Pitch Transposition Cases.

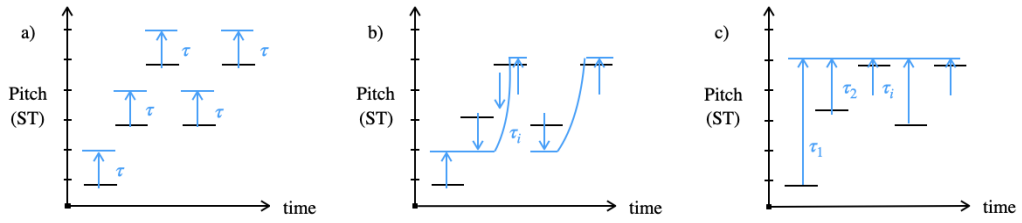


Figure 2.2: Pitch transposition possibilities: a) constant transposition, b) Variable transposition can follow any arbitrary curve imposed over the original audio signal. c) Tuning is a special case of transposition, but it requires that the imposed melody lies in a well-defined scale and has a well-defined sequence. A particular case could be, for example, a flat note.

In constant transposition, the entire signal is transposed by a constant number of semitones (and/or microtones), as shown in Figure 2.2 a). For a

given melody, a correction τ is uniformly applied to the entire audio. However, constant transposition is not always desired. Variable transposition throughout the sample involves imposing an arbitrary pitch curve τ_i, as illustrated in Figure 2.2 b). This is the case with systems like Melodyne™[Neubacker, 2011], which can use the curve to improve intonation. A special case of variable transposition is autotuning, which requires that the imposed melody be in a given scale and have a precise sequence. A particular case could be the use of extreme autotuning with a single active note, as shown in Figure 2.2 c). There, the transposition for each note is different: $\tau_1 \neq \tau_2 \neq \tau_i$. This example aligns with the use of Antares Autotune™(ATA) [Hildebrand, 1998] with a single note active and in an extreme configuration.

Transposition is performed using a vocoder-like re-synthesis technique. This process generates an output signal with the desired pitch and requires an adjustment factor. The adjustment factor represents the rate of change needed for the fundamental frequency. In the case of a constant transposition, the factor can be derived from the desired transposition value in semitones and the input fundamental frequency; from there, we can easily obtain the expected frequency and then the adjustment factor. The expected output frequency is calculated with the help of conversion between semitones and frequency. Consequently, the adjustment factor is determined as the division between the input and output frequencies. Sound examples of pitch transposition for constant transition can be found in "Te Juro Que Volveré" [Mon Laferte, 2023], which is performed using pitch shifting with formant loss and without autotune, or in "Set Me Free" [Robyn & la bagatelle magique, 2015], which employs a higher note for the segment 0:34-1:02 and a lower one for the segment 1:04-1:18.

For variable transposition, the adjustment factor must be applied point by point, representing the relationship between the input frequency and the expected output frequency, which is calculated. In both cases, knowledge of the output frequency is required, regardless of whether obtaining such a value is more or less direct, and the input frequency, which is determined using a pitch tracker. This way, the transposition process follows the scheme depicted in Figure 2.3. The pitch warping process involves applying an adjustment factor to a given signal for resynthesis, resulting in another signal with the expected pitch.

A blend of variable transposition and autotuning (with Autotune Antares) can be heard in "Starboy" [The Weeknd, 2016] (constant throughout the entire song) or in "Memories" [Maroon 5, 2019], probably using Melodyne, a software to impose a given pitch curve. Strong transition can be seen in examples like "Instant Crush" [Daft Punk, 2013] or in rap songs like "Headlines" [Drake, 2011], where jumps can be heard in segments like 1:24 in the word "know".

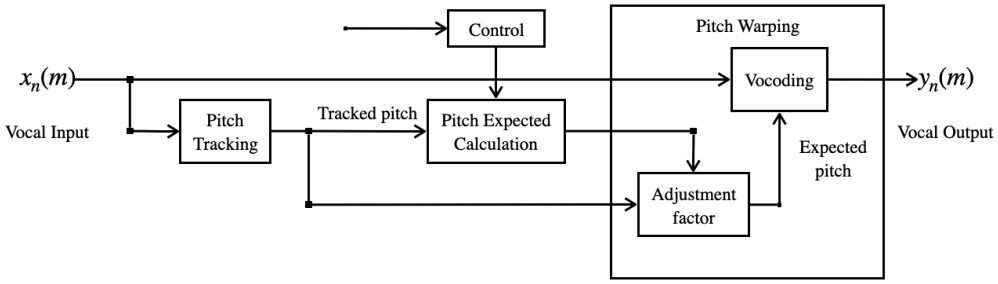


Figure 2.3: Pitch Transposition Schema. It consists of three stages: pitch tracking, which is the technique that identifies the frequency corresponding to the melody of the signal; transposition control, which is the amount of semitones to transpose; and the vocal transformation algorithm, which is generally a vocoder.

2.1.2 Pitch Correction

When addressing the transposition of a vocal signal, the concepts of expected pitch and adjustment factor naturally emerge. The process by which the expected pitch is obtained can be arbitrary, meaning without a corrective harmonic goal. However, it can also be leveraged to achieve an intonation-improved pitch. In this case, we refer to this stage as pitch correction. We define pitch correction as a specific case of variable transposition to improve intonation. The algorithm is based on the scheme in figure 2.3, where the pitch expected calculation is replaced by the pitch correction algorithm shown in 2.4.

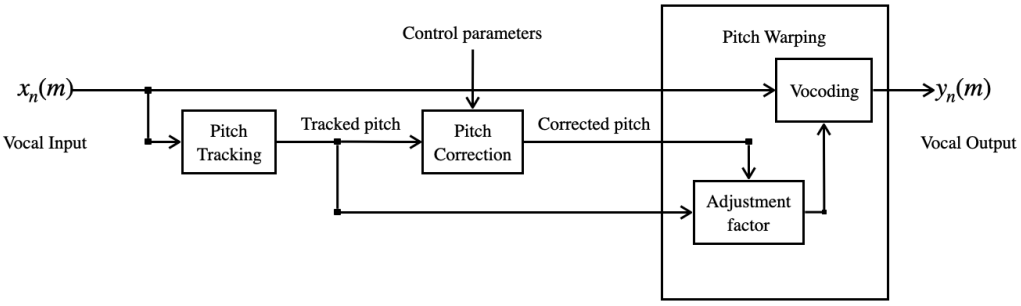


Figure 2.4: Pitch Correction Schema. This schema is similar to the pitch transition schema. The key difference is that pitch correction allows for a variable transposition for improving intonation.

A pitch correction system aims to generate a pitch curve better tuned within the tonal scale, for example, using a MIDI scale. This means pitch values (in semitones ST) are closer to integer values. To achieve this, we can employ mathematical algorithms that depend on a control parameter, enabling control over the deformation introduced to the pitch signal. In cases where

these algorithms rely on the signal's history as well as the control parameter, it must be referred as adaptive control.

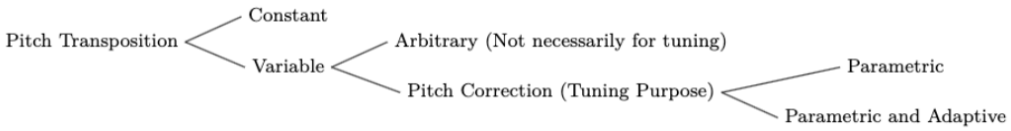


Figure 2.5: Pitch correction cases placed within pitch transposition. Parametric depends on parameters given by the user. Adaptive depends also in the pitch signal

The diagram (Figure 2.5) allows us to place pitch correction within the possibilities of pitch transposition, which can be classified as either constant or variable. In the case of variable transposition, it can be arbitrary or corrective (pitch correction). Pitch correction can be performed using parametric algorithms such as ATA or adaptive methods that employ adaptive functions. These adaptive functions are present in instruments like Cantor Digitalis [Feugère et al., 2017] [Feugère, 2013] [Perrotin and D'Alessandro, 2016], Continum Fingerboard [Haken, 2009] [Haken et al., 1998], TouchKeys [McPherson et al., 2013], and Seaboard [Lamb and Robertson, 2011]. The adaptive functions change according to the note's position in the scale. The algorithm used in Cantor Digitalis is particularly interesting. It is called Dynamic Pitch Warping [Perrotin and D'Alessandro, 2016], and we aim to study it in this doctoral thesis.

Melodyne™and ATA are cases of arbitrary and automatic pitch correction, respectively. In Melodyne™[Neubacker, 2011], a pitch curve is generated graphically (and assisted) and superimposed onto the audio sample. In ATA, the difference from a considered correct frequency (in semitones) is calculated, and the required difference is proportionally applied with the help of a smoothing filter (step 48 of the patent [Hildebrand, 1998]). Examples of autotuned songs with a improvement on correction purpose are "Happy" [Pharrell Williams, 2013] (with Melodyne) [1], "Memories" [Maroon 5, 2019] (probably ATA or melodyne) and Firework (Using a similar tool in Nuendo) [2].

2.2 Autotuning Review

In this research, we will consider the ATA method as a reference; therefore, we will study it in detail in this section. Antares Auto-Tune™ (ATA) is a vocal effect that emerged in 1998, and shortly after its release, gained massive popularity, primarily attributed to Cher's song "Believe" [Cher, 1998]. This

[1]https://www.soundonsound.com/techniques/inside-track-pharrell-williams-happy
[2]https://www.soundonsound.com/techniques/sandy-vee-recording-katy-perrys-firework

track reached number one in thirty countries, garnered over 25 accolades, and received multiple platinum certifications. However, beyond this singular song, ATA has had a profound impact on the music industry. The use of ATA grew exponentially over the years, becoming a global phenomenon.

The widespread popularity of the name "autotune" was initially caused by its extreme use, characterized by strong and distorted transitions between notes. However, the increased usage can be attributed to the parameterization capabilities of ATA. ATA allows for smooth automated parameterized use, enabling much smoother and minimal transitions if the parameters are chosen wisely. Later, a manual use was added, which meant the possibility of creating pitch curves manually, as we can do in Melodyne. Also, over time, the real-time response and overall quality of the device were improved. Thus, its use as a vocal enhancer to correct off-key notes significantly contributed to its widespread adoption.

As a result, ATA became a benchmark for vocal effects and is currently one of the most well-known and widely used effects. Determining when a song does not incorporate ATA or a similar effect is challenging. Its systematic use extends across most popular music genres worldwide, with some genres, such as pop and rap, embracing it. It is worth noting that musical genres like hyperpop have emerged, based on vocoders and autotuning, much like what happened in the past with rock music and the electric guitar. Furthermore, even today, ATA is used in genres that traditionally were less interested in using it, such as more acoustic and organic music (both for pitch improvement and for its extreme use). Without a doubt, nowadays, vocoder-type effects like ATA play an immense role in music.

Songs like "Starboy" [The Weeknd, 2016] maintain a consistent "soft" use of autotune throughout the entire song. Others like "Cardigan" [Taylor Swift, 2020] generate doubt; abrupt transition are evidenced at 2:17 and 2:08, even in a live session [Taylor Swift, 2020] at 2:54 and 3:05, nevertheless we cannot claim these particular abrupt transitions are due to Autotune Antares. It is important to note that autotune is not necessarily detrimental as part of the vocal chain, as singers must approach the note in live performances to avoid excessive transitions. It helps to enhance vocal consistency and presence across multiple layers of stereo and vocoded vocals. Furthermore, singers must be capable of reaching or approximating the correct notes sufficiently to properly replicate the song's studio version, as demonstrated in performances of "Believe" and "Cardigan", where live vocals are well replicated in live. The same applies to constant transposition on non-autotuned vocals, as seen in "Te Juro Que Volveré" [Mon Laferte, 2023]. If the performer cannot hit the correct notes to trigger the transition at the right moment, then vocal effects cannot be replicated exactly as in the studio.

Although the use of Autotune can be continuous and aimed at enhancing the performance of the vocal track, this effect can also serve a stylistic purpose. This is evident in the work of singers with inherently strong voices such as Cher (in Believe [Cher, 1998]), Rosalía and Mon Laferte. Some examples include

"De Aquí No Sales" [Rosalía, 2018], particularly at 0:06 and 1:12-1:45, in "Di Mi Nombre" [Rosalía, 2018] at 0:57-1:12 contrasting with untreated vocals, in "Como Un G" [Rosalía, 2022] from 3:05 onwards contrasting with the vocals at the beginning, in "Obra de Dios" [Mon Laferte, 2024] at 0:18-0:25 without Autotune between 0:24-0:29 and then with Autotune from 0:38 onwards, with the bridge being without Autotune, in "Casta Diva" [Mon Laferte, 2023] where vocals in parts like 3:01-3:28 contrast with 3:43-4:05, or in "NDA" in 0:57-1:18 [Billie Eilish, 2021]. Other artist have explored full use of autotune, in rap or in new genres based in extreme use of effects like hyperpop, an example is the album "how i'm feeling now" by Charli XCX [Charli XCX, 2020a], where autotune is explored in all the tracks with layers of vocoders, grains, autotune and distorsion, this album must be listened integrally. In these examples, the ATA effect is explored from different approches: improving vocals, intentional harmonies, or experimentally like it was done several decades ago with guitar amplification.

2.2.1 ATA - Pitch Tracking

Antares Autotune (ATA) is a digital audio effect developed by Harold Hildebrand, which is based on an algorithm for pitch correction. Hildebrand is the primary author and co-author of several patents related to geophysical exploration and the handling of digital signals for music. In fact, it can be assumed that the development of ATA is encompassed within several of his patents. ATA became commercially available in 1997. Being a pitch correction algorithm, it fullfill the structure shown in Figure 2.4. Pitch correction faces three problems, detailed at varying levels in Hildebrand's patents: pitch tracking, the pitch correction algorithm, and pitch warping. One of his more detailed patents is titled "Pitch detection and intonation correction apparatus and method" [Hildebrand, 1998]. In this document, Hildebrand primarily describes the mathematical foundations of his pitch tracking algorithm, but provides minor details about the pitch correction algorithms and the pitch warping vocoder.

As the author claims, pitch detection is virtually instantaneous and occurs before the sound has enough amplitude to be heard. To perform this, a sequence of data (signal) x_j with a repetition period L is used. The autocorrelation of a periodic signal is also periodic, so its value at n=L (if the period is L) is the same as its value at zero. This means that the autocorrelation of any signal could be compared with given reference values to estimate the fundamental frequency; such calculation is computationally expensive. Nevertheless, the patent describes that the computational cost can be reduced in several ways. As the analysis window changes over time, because the signal varies in real time, the autocorrelation can be written:

$$\Phi_{i,L}(n) = \sum_{j=i-L-1}^{i} x_j x_{j-n} \tag{2.1}$$

And two functions $E_i(L)$ and $H_i(L)$ are defined, in such a way that $E_i(L)$ is the cumulative energy of $H_i(L)$ over two periods 2L, with $j = 0, 1, 2, ..., i$:

$$E_i(L) = \Phi_{i,2L}(0) = \sum_{j=0}^{2L} x_j^2 \tag{2.2}$$

$$H_i(L) = \Phi_{i,L}(L) = \sum_{j=0}^{L} x_j x_{j-L} \tag{2.3}$$

The author indicates that the equations can be manipulated to obtain an inequality:

$$E_i(L) - 2H_i(L) <= epsE_i(L) \tag{2.4}$$

Remember that L is the period of the signal. This means that several values of L can satisfy equation 2.4 for several given eps values, meaning there are several "candidates". The smallest eps value is the one that minimizes the relationship 2.4. The detection operation mode works by down-sampling eight times (details of the reasons for this value are not given) to obtain a first value of L, which reduces the computational cost. Additionally, once an L that minimizes the "eps" equation is found, some verifications are performed before launching calculations within the neighborhood of L at full sampling.

2.2.2 ATA - Pitch Correction

Once pitch tracking has been performed, pitch correction is carried out, corresponding to step 67 of the patent. The difference between the actual pitch and the expected pitch (the integer part of the pitch in the MIDI scale) provides the correction factor. This correction factor is passed through a smoother filter that depends on a variable called Decay, which ranges from 0 to 1. For Decay=0, the correction is immediate, while for Decay=1, the correction takes 400ms. The user interface in ATA (as shown in Figure 2.6) allows control of pitch correction through two parameters, Flex Tune and Retune Speed. Retune Speed controls the Decay variable, and Flex Tune serves as a threshold for correction. This means that if the pitch curve is over given value of Flex Tune in the neighborhood of a note, then the pitch curve is corrected. The larger the Flextune value (maximum = 100), the less modification is applied. If Flextune is set to zero, all points of the pitch curve are corrected. In the ATA manual, it is quoted verbatim as follows: Retune Speed allows one to "set the rate at which the input audio is moved to target pitches," while increasing Flex Tune is used "to allow more pitch deviation, usually for expressive purposes."

2.2.3 ATA - Pitch Warping

The pitch warping process in ATA is very similar to the PSOLA process. Although Hildebrand has publicly denied using that algorithm, the technique is very similar to the OLA technique (predecessor of PSOLA). It is described in

Figure 2.6: ATA interface, protected by copyright ©. We can appretiate the retune-speed and flextune parameters control.

steps 46 to 50 and involves changing the playback rate according to the adjustment factor. If it is greater than one, it accelerates, so a cycle is continuously stored in memory that can be repeated when the rate needs to be increased (going sharper). Conversely, the excess cycle is deleted (going flatter).

2.2.4 ATA - Additional Patents

One of the issues not addressed in the main patent is how to guarantee a good resolution in low frequencies. It is discussed in another patent titled "Method and Apparatus for Digital Filtering of Audio Signals" [Hildebrand, 1996]. It is a high-definition and low-frequency fidelity equalization method. The method is based on frequency warping, which involves mapping the z-domain (of the z-transform). This domain is transformed through a mathematical process into another domain called warp-Z, which complies with certain equivalences with the z-domain and follows the filter laws that apply initially. This domain change allows high definition in low frequencies and low definition in high frequencies. On the one hand, the presumed use of this filter would improve the sound response through equalization, attenuating unwanted components of the spectrum, so it could be applicable during the pitch warping. On the other hand, this algorithm is compatible with the calculation of L (in the pitch tracking), which presumably could be used to achieve greater definition in the range of interest in low frecuencies.

Furthermore, there is a third patent titled "Virtual Tuning of a String Instrument" from 2012 that uses the information present in [Hildebrand, 2014], where the term "adjustment factor" is coined, as used by us in the previous

subsections.

The set of capabilities of ATA allowed its expansion in the music industry. However, possibly due to being patented, there have been very few advances in the science behind ATA, and the fact that it is patented has hindered the advancements that could have occurred. Some of the questions that could have been studied further are: If different pitch correction algorithms are used, what characteristics should such techniques have? What occurs musically if they are adaptive, meaning they depend on the present and past of the signal? What would be obtained musically, and what would be useful musically and technically (sound engineer approach)?

2.3 Autotuning Musical Analysis

In this section, we will briefly examine the changes implemented by Antares Autotune, as the primary effect of autotuning, exploring signal waveform, spectrum, pitch and formants. We take an audio file (realrt-yvesmontand) and apply extreme and minimal correction presets. We observe that the changes in the signal waveform are almost non-existent, and latence for ATA Artist (the actual real-time module) is 3 ms, as shown in figure 2.7.

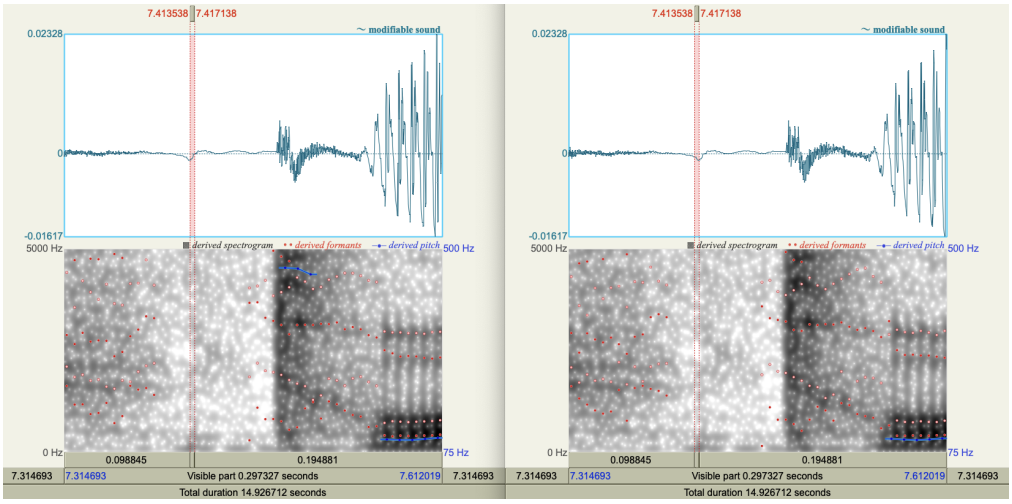


Figure 2.7: Signal shape is mostly preserved for the most transparent case and latence is equalt to 7.4 ms using ATA with extreme correction

Regarding the formants, as depicted in Figure 2.8, it can be observed that the transparent configuration (Retune Speed = 400ms, Flextune = 0) (center) of ATA preserves the formants of the original version (left). In contrast, the extreme correction (Retune Speed = 0, Flextune = 0) (right) introduces a consistent distortion across all formants. It can be noted that the variation of each formant is greater when the extreme correction is applied (at the beginning of a note when the singer is farther from the corrected note), and then decreases as the original pitch approaches the corrected pitch notes (when the singer is closer to the correct note).

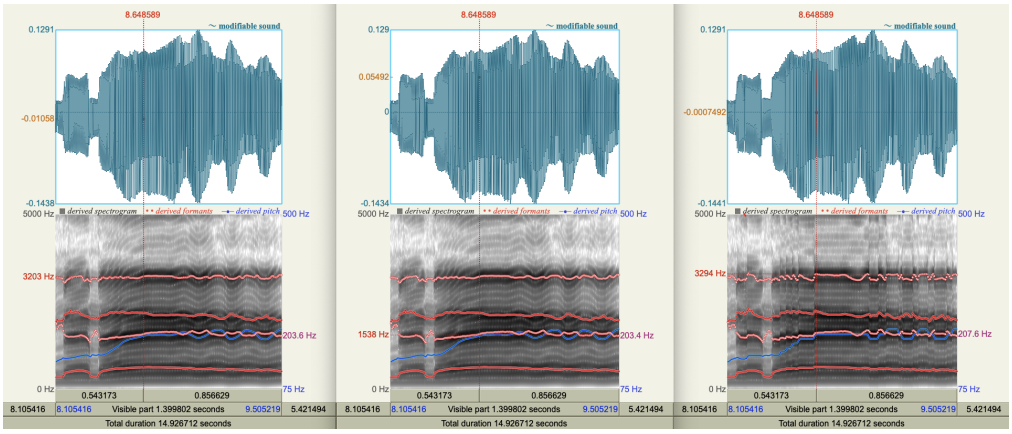


Figure 2.8: Original sound vs autotuning. Three cases are presented using a standard window on Praat (10 ms): original sound (left), sound treated with the most transparent configuration of ATA (center), and sound treated with an extreme autotuning configuration (right). We observe that the signal shape is preserved for both non-transforming and extreme-autotuning configurations. In the transparent case, F0 and formants are preserved. In the extreme tuning case, F0 is warped to have sharp transitions. As this new curve is imposed, formants are increasingly modified with more correction (F0 - int(F0)), resulting in variations in the formants for each note.

Regarding the transitions, according to Figure 2.9, it can be seen that in the most transparent configuration (center), all regions of the spectrum are well preserved relative to the original sound (left). Concerning the extreme correction (right), it can be observed that the extreme correction affects all harmonics. Furthermore, there is dispersion of f0 in all harmonics during the transition. We define f0-dispersion as the loss of the f0 value in the neighborhood of a given time mark, so f0 candidates are no longer concentrated in given values but smeared over all the frequencies, exhibiting as vertical segments at the time mark.

2.4 Pitch correction on non-vocal applications

Although pitch correction is primarily known thanks to ATA and Melodyne, there are other types of applications that use pitch correction algorithms. In fact, Hildebrand himself has a patent that employs the same algorithm for pitch correction in the electric guitar [Hildebrand, 2014]. But beyond that algorithm, there are others that are more innovative and interesting, being applied to human-computer interfaces to improve the response of transducers and the experience of playing digital instruments. These interfaces are mostly known as New Interfaces for Musical Expression (NIME), which have their own field of study and exploration, addressing issues such as the level of difficulty and learning with use, the capacity for exploration and control, among others [Orio et al., 2001]. Pitch correction methods applied to NIME are designed to

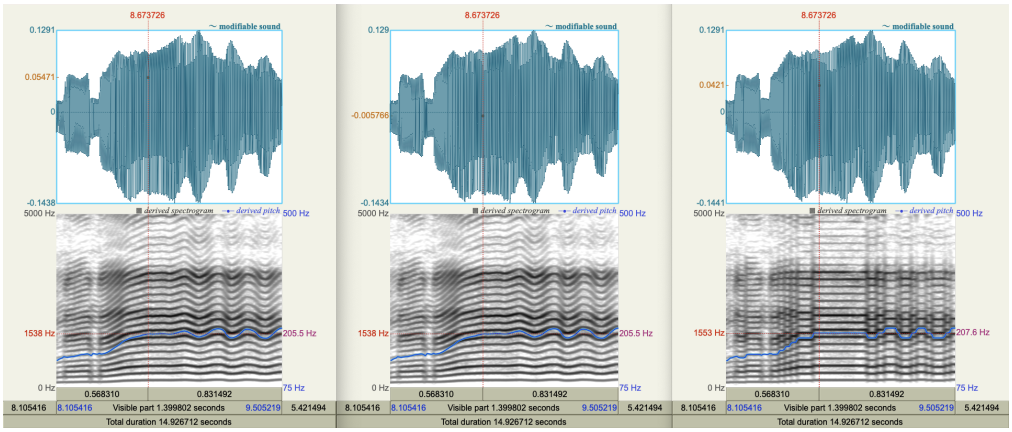


Figure 2.9: Original sound vs autotuning. The same three cases are presented using a larger window (100ms): original sound (left), sound treated with the most transparent configuration of ATA (center), and sound treated with an extreme autotuning configuration (right). No differences are observed for the non-transforming case. However, for the extreme-autotuning configuration, we can observe how the sharp transitions between generat a dispersion over F0 during the transition.

enhance the musical experience on those digital instruments with a physical interface.

Cantor Digitalis [Feugère et al., 2017] [Feugère, 2013] [Perrotin and D'Alessandro, 2016] is a digital instrument developed by the Lutherie-Acoustique-Musique group at the Institut Jean Le Rond D'Alembert; it is a vocal synthesizer that uses a graphical interface for a Wacom tablet. The data obtained on the tablet is corrected based on stability conditions and control parameters (set by the user) that modify the speed of the imposed correction. Additionally, the correction method employs an adaptive function, which is innovative compared to the other methods mentioned in this manuscript. This is particularly relevant because, unlike the other methods, it allows a transition based on the stability of the note and a variable transition time.

In the following section, we will address what was acclaimed as an innovation in pitch correction methods by its authors [Perrotin and D'Alessandro, 2016], the Dynamic Pitch Warping method, an algorithm for graphical correction of pitch. The novelty of this method lies in its three control variables, instead of the two (retune speed and flex tune) used in ATA. The use of systems like DPW can provide insights into the creation of new autotune-like effects.

2.5 Dynamic Warping Function for Pitch Correction

The pitch correction method based on its graphical position in a table, Dynamic Pitch Warping (DPW) [Perrotin and D'Alessandro, 2016], was developed to improve the intonation of the vocal synthesizer Cantor Digitalis [Feugère et al., 2017]. The name might be confusing because it contains the term "Pitch Warping," which generally refers to the process of vocal transformation (time-based techniques like ATA, time-frequency techniques like the vocoder WORLD, or others). From now on, we will use only the abbreviation DPW; however, we recommend renaming the method to Dynamic Warping Function for Pitch Correction.

Now we provide a general explanation before going into detailed explanations of the method. The method follows the same pitch transition format explained in Figure 2.4. We assume we have a pitch tracker that provides a pitch signal and a vocoder to impose the expected pitch. The DPW method acts over the pitch signal, it first checks if the pitch signal is stable. The stability condition is that the pitch remains within a specific interval during a critical time, as shown in Figure 2.10 a). If the condition is fulfilled, then a correction is applied progressively, as shown in Figure 2.10 b). The pitch curve varies, contouring a melody, so the adaptive function warps the curve to ensure the initial and final points are in-tune notes (in MIDI scale, integers), as shown in Figure 2.10 c). The stability condition and the transition time are set by the user. However, the DPW function depends on the pitch curve and is calculated at the place where the note was considered stable.

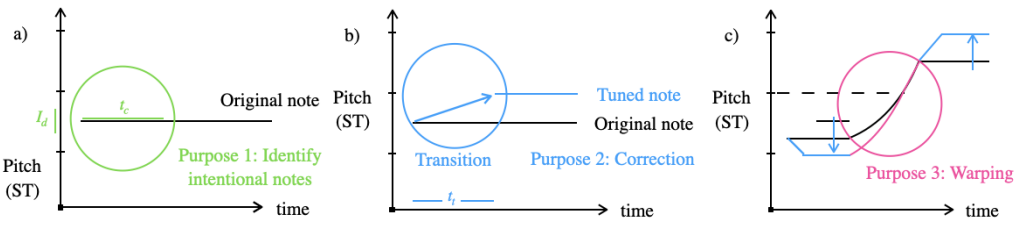


Figure 2.10: Schema for the purposes of DPW. The DPW function involves a stability condition, a correction, and a warping for convergence on in-tune notes

2.5.1 Review of the DPW adaptive function

The correction with DPW [Perrotin and D'Alessandro, 2016] was applied initially to the position data from a Wacom tablet; this data is tracked (in microtonal midi scale) in MAX to synthesize a voice with the corresponding fundamental frequency. The position data is calibrated and mapped within

a linear scale of semitones in the chromatic scale across a range of several octaves.

Intonation correction is applied to an f_0 (pitch) signal obtained from a pitch tracker (a topic we will discuss later). This out-of-tune f_0 curve, denoted as x, is the input to the DPW system. DPW improves x by adjusting it within a MIDI scale, where the numbers represent the correct notes.

First, let us consider that we are in the neighborhood of any note; for simplicity, we will take a relative position of 0 to denote any note in the MIDI space. The next and previous notes will be δ and $-\delta$, respectively, as shown in Figure 2.11, where δ separates the tuned notes. In the MIDI scale, $\delta = 1$, meaning the next and previous notes are one semitone above and below, respectively. Let x be the input and $y(x)$ the output. We trace an arc that goes from $-\delta$ to δ, passing through a given point (the out-of-tune intentional note) placed in the neighborhood of 0. The curvature of the arc is γ. This is the adaptive function described by [Perrotin and D'Alessandro, 2016] in Figure 2.11.

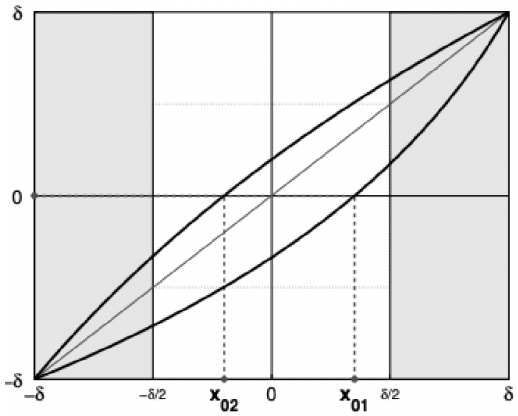


Figure 2.11: The arc of curvature for the dynamic pitch correction method, took from [Perrotin and D'Alessandro, 2016]

The correction is applied only if a stability condition is met, ensuring that only intentionally sung notes are corrected. A value of x is considered genuinely intentional if it remains within a microtonal interval (I_d) for a given period (critical time T_c), as shown in the pink interval in Figure 2.12. The correction is triggered by modifying the curvature γ of the adaptive function, changing it from a straight line to a curve that passes through the considered stable point, during a transition time T_t, as shown in the blue interval in Figure 2.12. Once the adaptive function is achieved, it ensures convergence to tuned notes by moving up or down the MIDI scale, as shown in the green interval in Figure 2.12.

Now, we will perform the mathematical calculation. According to [Perrotin and D'Alessandro, 2016], such an curvature arc $x = g(y)$ would have a curvature γ and could be written as :

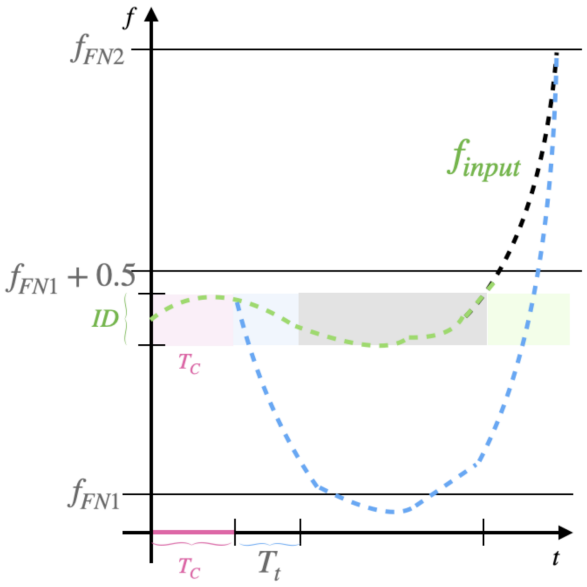


Figure 2.12: Illustration of the dynamics of DPW. The input f_0 curve (green curve), x, is stable within a detection interval I_d during the critical time T_c (pink region). The correction is triggered during the transition time T_t (blue region). The input f_0 can vary continuously during the transition time. It stays there (grey region time interval) until it moves to reach the next semitone (integer) on the pitch scale (black curve) where input and output converge (green region).

$$g(y) = Ae^{\gamma(y+B)+C} \tag{2.5}$$

The limit condition can be rewritten: $g(\pm\delta) = \pm\delta$, so we obtain a system of equations that leads to:

$$C = -\delta\Big(1 + \frac{2}{e^{2\gamma\delta}}\Big) \tag{2.6}$$

$$A = 2\delta\frac{e^{\gamma(\delta-B)}}{e^{2\gamma\delta}-1} \tag{2.7}$$

When replaced in 2.5, this makes the dependence on B disappear. If the input is in the neighborhood of an in-tune value (integer), then it is considered tuned, and the output should be equal to the input, that is, $x = g(y) = y$. In that way, the discontinuity for γ is avoided and we write:

$$g(y) = \begin{cases} \delta\Big[2\frac{e^{\gamma(\delta+y)}-1}{e^{2\gamma\delta}-1} - 1\Big] & si \quad \gamma \neq 0 \\ y & si \quad \gamma = 0 \end{cases} \tag{2.8}$$

Let us remember that $x = g(y)$ is the input as a function of the output y; therefore, the DPW function will be the inverse function of $g(y)$, denoted as $y_E(x)$:

$$y_E(x) = \begin{cases} 1/\gamma\Big[log\Big[(e^{2\gamma}-1)(\frac{x}{\delta}+1)\frac{1}{2}+1\Big]\Big] - \delta & si \quad \gamma \neq 0 \\ x & si \quad \gamma = 0 \end{cases} \tag{2.9}$$

The purpose is that once applied, the function y_E makes the output is in tune for an out-of-tune input. In our scale where 0 represents the in-tune note, this means that $y_E(x_0, \gamma_0) = 0$, so:

$$\gamma_o\Big[log\Big[(e^{2\gamma_o}-1)(\frac{x_o}{\delta}+1\Big]\Big])\frac{1}{2}+1] - \delta = 0 \tag{2.10}$$

$$x_o = \delta\Big[2\frac{e^{\gamma_o\delta}-1}{e^{2\gamma_o\delta}-1} - 1\Big] \tag{2.11}$$

Doing a change of variable $u = e^{\gamma_o\delta}$:

$$x_o = \delta\Big[2\frac{u-1}{u^2-1} - 1\Big] \tag{2.12}$$

That we can write as a second order system, and solve it:

$$\frac{x_o}{\delta}u^2 - 2u + (1 - \frac{x_o}{\delta}) = 0 \quad then \quad u = \frac{\delta \pm x_o}{\delta + x_o} \tag{2.13}$$

But for this part of the equation $\gamma_o \neq 0$ always, and so $u = e^{\gamma_o\delta} \neq 1$. Then, the only solution possible is $u = \frac{\delta - x_o}{\delta + x_o}$. Using again $u = e^{\gamma_o\delta}$, we obtain γ_o:

$$\gamma_o = \frac{1}{\delta}log\Big(\frac{\delta - x_o}{\delta + x_o}\Big) \tag{2.14}$$

Where δ is the value that separates the tuned notes, this means that $\delta = 1$ for pitch in the midi scale. γ_o is the value of γ that makes an output y_E in tune for a given input x out-of-tune, so γ can be considered the factor of correction. The value of γ can be calculated each time that the pitch stays in one interval (I_d) during a critical time (t_c). The value of γ is calculated from the inverse function, that is:

$$\gamma = \frac{1}{\delta}log\Big(\frac{\delta - x_o}{\delta + x_o}\Big) \tag{2.15}$$

With the value from 2.15, the output can be calculated through y_E, generating a tuned output. When the input moves to the previous or next note ($\pm\delta$), it perfectly converges to an integer value (tuned note), satisfying the boundary condition.

Once the correction is triggered, the value of γ varies, changing from a straight line to a curve defined by y_E during a transition time T_t at constant time steps. In the case of MAX, these time steps correspond to the inverse of the control rate of MAX, which is 1 ms.

Finally, we outline a summary of the key concepts to consider for DPW:

- The adaptive function is always localized with respect to a any reference note that corresponds to zero in our mathematical calculation.

- Only stable notes (within an interval I_d for more than the T_c interval) activate the correction algorithm.

- The transition between an out-of-tune note and a tuned one is done by modifying the γ value during the transition time interval (t_t) in constant time steps.

- We didn't mention before, but if no correction is needed, the input passes directly to the output, so DPW remains unnoticed.

2.5.2 Differences between DPW and ATA

The parameterization of the pitch correction method differs between ATA and DPW. While in both cases, it depends on triggering conditions, the transition to a tuned note, and the potential deformation of the curve, the ways of executing such actions are different in these two methods. ATA's patents do not mention the triggering method; however, by checking ATA's online documentation, it is found that triggering is done with a proximity condition to the tuned note (integer number in the semitone scale). If the input note is within a specific microtonal interval, called flextune, correction is initiated; otherwise, nothing is done. In the case of DPW, triggering is based on a stability condition. If a note is within a certain detection interval I_d for a time interval t_c, then correction is initiated.

The correction in ATA is done by modifying the adjustment factor with a smoother filter that transitions for the desired duration. In DPW, the correction is made by varying the value of γ at regular steps within the imposed time interval. While in ATA, only the adjustment factor is added, in DPW, the pitch change is passed through the adaptive DPW function, y_E, which warps, the pitch transitioning it to the next note (whether higher or lower), a tuned note is encountered.

2.6 Proposal for Defining Pitch Correction Cases

While the use of autotuning is widespread, and it is undeniable that almost any modern production includes it in vocals, there is no, whether due to appreciative reasons or questions about its use, a musical-technical language that tells us the type of autotuning we are using. Defining concepts related to the uses of autotuning will allow us to open the door to more varied and even more musically interesting uses.

2.6.1 Extreme Case

When using an autotuning-type effect, two main objectives can be pursued: either one wants to hear the distortion, or, on the contrary, vocal deficiencies are to be masked to improve the melody's tuning. The use that exploits the distortion generated by the "instantaneous" transition will henceforth be referred to as extreme autotuning. This configuration refers to parameterization with a minimum transition time possible, whether by the software used as it is schematized in Figure 2.13. The goal of this type of configuration is to hear the distortion. In fact, its definition already opens up new questions about where to place autotuning within our perceptual characterization of vocal effects. That is, does vocal quality get affected by the melody or by vocoder technique—a question that will be explored in Chapter 3.

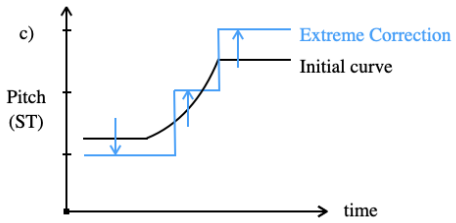


Figure 2.13: Schema for extreme correction

2.6.2 Transparent Case - Expressive Correction

The other type of use is when one aims to improve tuning. In this case, the goal is to enhance the singer's tuning without necessarily hearing the distortions caused by melodic deformation or the vocoder. This type of use is called transparent autotuning. In this case, the same question arises: iIs the pitch-warping technique (vocoder type) affecting vocal quality, or is it just not enough for the autotuning effect to be classified as an effect that changes vocal quality? we will explore this question in chapter 3. Transparent autotuning does not have a defined configuration; it depends on what the user is looking for. However, it can be said that it depends mainly on the transition time.

Now, what is expected to be done in the transparent case? The goal is to correct the note without hearing distortions. However, this depends on the input pitch signal. Therefore, we have gone for the most general classification possible. Basically, we find three types of melodies that we have defined as: staircases, vibratos, and free-paths (for example a glissando). They are schematized in Figure 2.14, where we illustrate the out-of-tune input and the expected improved in-tune output.

We can describe them as it follows:

- Staircases are intentional notes that are stable over time and clearly define a melody in a song. These notes can be more or less in tune

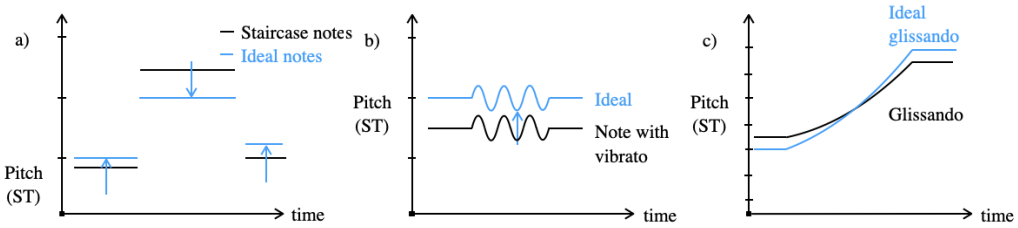


Figure 2.14: Schema for transparent correction cases: staircases, vibratos, and free-paths (for example a glissando). Input in black, and expected improvement in blue

(sharper or flatter than an integer semitone), and the goal of transparent correction will be to position them in the correct tuned notes (integer value).

- Vibrato is a common ornament in Western classical music [D'Alessandro and Castellengo, 1994], and it can be defined as a periodic modulation of f0 [Sundberg, 1994]. It typically occurs within intervals smaller than two semitones and at frequencies below 5.5 Hz. Since a periodic signal can be seen as a sum of sinusoids, we can use the vibrato category to include any pitch oscillation type, completing a full period covering all possible vibratos. Vibratos can be centered on an out-of-tune note, so our goal with a pitch correction is to shift the vibrato and align it to make it centered on a correct note. [Seashore, 1931] quoted: "a good vibrato is a pulsation of pitch, usually accompanied by synchronous pulsations of loudness and timbre, of such extent and rate as to give a pleasing flexibility, tenderness, and richness to the tone."

- Free-paths refer to free patterns that may be present as longer ornamentation in the transition between notes or may correspond to a glissando that is part of the melody and should pass without significant modification.

If the transparent correction is capable to deal with these cases with just one configuration, then we can refer to it like expressive correction.

2.7 Testing over pitch signals

In this section, we will go through different pitch correction scenarios. For this purpose, we will use pitch signals to test both ATA and DPW methods and make the differences between the two methods visible. Initially, we aim to illustrate the distinctions between the methods using the most straightforward case: correction for constant note out of tune with a constant pitch shift. After, we will cover several cases for extreme and transparent correction with ATA,

and extreme and expressive correction with DPW, varying the corresponding parameters to each method.

We have utilized pitch capture through a linear scale using a Wacom tablet compatible with the Cantor Digitalis digital instrument. Thus, we have internally modified the information flow of Cantor Digitalis, making the synthesizer receive a direct pitch signal from the tablet in the semitone scale, without going through the correction process included in the original version of Cantor Digitalis. In this way, a detuned voice is synthesized as is its shown in Figure 2.15.

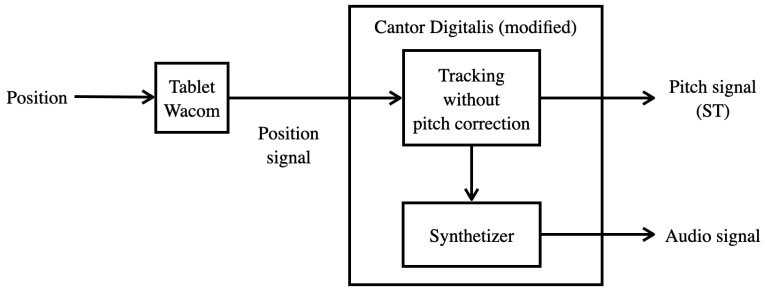


Figure 2.15: Generating sounds

This signal can be analyzed in Praat, and it is possible, as explained in Chapter 3, to generate a file with pitch information in various formats from Praat [3]. Praat has already been the subject of study among other pitch trackers [Babacan et al., 2013], such study shows that Praat is very precise.

To use Praat, we apply the function "Analyse Periodicity/Pitch" over an audio file, it generates the main pitch curve, including all f_0 candidates. Then, we use the function "Analyse/Down to PitchTier" to generate a Praat file with the most suitable f_0 candidates at specific time marks and we save this PitchTier file. Next, in a Python script, using the "interp1d" function of the "scipy.interpolate" package, we interpolate the PitchTier's time scale to match the actual audio's time scale. As a result, we generate a pitch signal in WAV format with the same time scale as the actual audio file.

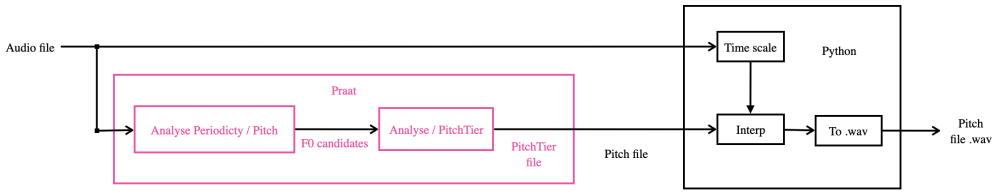


Figure 2.16: Tracking pitch from an audio file to generate a pitch .wav file

To use ATA, we first import the audio file into a digital audio workstation (DAW) such as Ableton or Reaper, where we implement ATA with a custom

[3] https://www.fon.hum.uva.nl/praat/

configuration and save the autotuned audio file. Next, we use Praat to perform a pitch analysis on the autotuned file, as shown in Figure 2.16, and we save the PitchTier file.

The ATA documentation hints at how the pitch tracking and vocal transformation are performed, but there is very little information about how the pitch correction is carried out. Extreme correction is the only use case explicitly described in the patent: the transition time is zero, making the output pitch correspond to the integer part of the input pitch. The parameter flextune is not mentioned in the patents, and we do not have access to the ATA source code. Therefore, the path shown in Figure 2.16 is the most practical and precise way to obtain information about the resulting pitch correction when using ATA (the only limitation being the analysis window length, which is 10 ms in Praat).

To use DPW, We take the pitch signal file (from praat) and implement the DPW correction in MAX, thus obtaining an output pitch file. Then, we use the World [4] [5] [Morise et al., 2016] [Morise, 2016] [Morise, 2015] vocoder to generate the output audio (Figure 2.17).

We can compare the DPW-corrected pitch obtained in MAX with the pitch analyzed again by Praat over the World DPW-corrected audio. There is no difference, which verifies that Praat is precise and does not distort the pitch values.

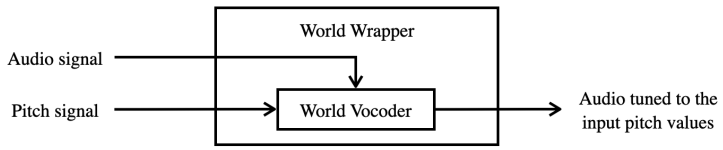


Figure 2.17: Imposing a given pitch on an audio file

2.7.1 DPW full implementation and issues comparing ATA and DPW

We have summarized the process to go directly to the testing phase, but such a process is not as straightforward to deduce. Here, we outline the challenge related to (i) implementing a pitch correction method and (ii) the comparison between pitch correction methods.

2.7.1.1 DPW real-time implementation

For real-time implementation, a pitch-tracker and a vocoder in real-time are required. The limitation lies in the precision and speed required. DPW is implemented by adapting the existing MAX code for Cantor Digitalis. DPW

[4]https://github.com/mmorise/World
[5]https://www.isc.meiji.ac.jp/ mmorise/world/english/

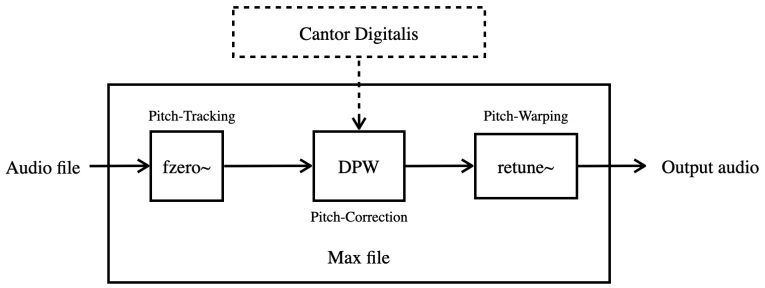


Figure 2.18: Real-time implementation of DPW

is isolated from the base software and adapted to a loop with the pitch-tracker and the vocoder, as shown in Figure 2.18.

Real-time implementation of DPW faces several issues:

1. In MAX, pitch trackers and vocoders are black boxes. We need to try all the available systems mixed between them. We used the following four pitch trackers: yin~ [de Cheveigné and Kawahara, 2002] [6], vb.pitch~ [Bohm, 2022] [7], sigmund~ [8], and fzero~ [9] [Zbyszynski et al., 2013]; and the following seven vocoders: pitchshift~ [10], freqshift~ [11], fbinshift~ [12], gizmo~ [13], hilbert~ [14], psych~ [15], supervp.trans~ [16], and retune~ [17][Bernsee and Gökdag, 2016]. This results in 28 different possible combinations of pitch tracker + DPW + vocoder. Considering that each "black box" works differently and receives different types of data, this process is very time-consuming.

2. Pitch tracking in the MAX architecture is not "instantaneous" as in ATA. We can improve latency, but we lost rapidly quality. Only the fzero pitch tracker can work with latency values below 1024 samples, but it results in poor quality. The latency problem keeps us from having a high-quality real-time implementation.

3. Vocoding in the MAX is not as simple. On the one hand, latency varies according to the device used. In fact, for each combination of pitch tacker+vocoder, we have to receive pitch and audio signals differently

[6] https://forum.ircam.fr/projects/detail/max-sound-box/
[7] https://vboehm.net/downloads
[8] https://github.com/v7b1/sigmund_64bit-version/releases
[9] https://docs.cycling74.com/max8/refpages/fzero~
[10] https://docs.cycling74.com/max8/refpages/pitchshift~
[11] https://docs.cycling74.com/max8/refpages/freqshift~
[12] https://docs.cycling74.com/max8/refpages/fbinshift~
[13] https://docs.cycling74.com/max8/refpages/gizmo~
[14] https://docs.cycling74.com/max8/refpages/hilbert~
[15] https://ismm.ircam.fr/maxmsp-externals/
[16] https://forum.ircam.fr/projects/detail/supervp-for-max/
[17] https://docs.cycling74.com/max8/refpages/retune~

(adjustment factor, rate between input/output pitch, difference in semitones, difference in Hertz); even some cases, the pitch signal must be delayed. On the other hand, some vocoders are slow or unresponsive to a variable transposition and we cannot know it without testing them because MAX does not provide full scientific support. Again, to ensure the problem lies with the vocoder, all the possible combinations (pitch tracker+vocoder) were done, so this part of the thesis took a lot of time. We hypothesized that some can have some kind of smoothing filter because they are slow-responsive even for the extreme correction pitch signal.

In the end, we did not succeed in developing a high-quality real-time implementation. However, we provided a **online implementation** that applies the DPW correction over a signal with a 2048-sample delay. This means it is not possible to use in real time, but at least it works online. This implementation uses the most accurate pitch tracker of the ones tested: fzero~ (set with @onsetamp 0.0001 and @onsetpitch 0.001) and retune~ which does not present discontinuities nor octave jumps. The best result was obtained with the vocoder retune~ a Max MSP object based on ZTX software (Precision Time Stretching and Pitch Shifting) [Bernsee and Gökdag, 2016] a patented method as ANTARES, whose implementation requires 1024 samples. Although this real-time MAX implementation shows that the DPW method works in real-time, faster and more efficient pitch trackers and vocoders, like those of ATA, would be needed to implement a real-time DPW system suitable for live singing.

2.7.1.2 DPW non-real-time implementation

Also, we worked in a non real-time implementation; for this purpose, we used various vocoders: world, Circe, retune . A wrapper for each vocoder was treated to implement the DPW corrected pitch file. The better quality was obtained with the vocoder World, so all the examples in this chapter have sound support generated using World [Morise et al., 2016] [Morise, 2016] [Morise, 2015] as shown previously in figure 2.17.

2.7.1.3 Issues comparing ATA and DPW

Comparing ATA and DPW is not straightforward because ATA's pitch tracking and ATA-pitch warping methods cannot be used isolately. The related documentation (patents and manuals) is insufficient for replicating ATA. Only the extreme correction case is replicable, as explained in detail in [Hildebrand, 1998]; this case happens when the time transition equals zero, making the output equal to the integer part of the input. This problem has been solved partially using the high-quality pitch tracker: praat. Nevertheless, a sonorous comparison using the ATA-pitch warper with both ATA and DPW pitch correction methods is impossible. Only this is possible when using an alternative high-quality vocoder compatible with variable transposition, such as World.

2.7.2 Constant out-of-tune note with constant pitch shift

With these tests, we will be able to observe several differences between the ATA and DPW methods. We use an out-of-tune note with a constant pitch shift for this test. Both DPW and ATA allow correcting the signal within a transition time. In the case of ATA, it is referred to as the retune speed, and in the case of DPW, it is denoted as t_t, but for practical purposes, they are very similar. The difference lies in the fact that DPW has a critical time, t_c, to assess the stability of the signal, so the correction is not triggered until after the time t_c has elapsed.

For our example, we use a constant and out-of-tune input signal with a constant pitch shift of 0.15 ST, as shown in Figure 2.19, where the green line represents the input, and the red and blue lines represent the corrections with ATA and DPW, respectively. The choice of these colors will continue throughout this subsection. The value of retune speed (for ATA) and t_t (for DPW) is the same and is set to 0.5 s. DPW is configured with $I_d = 0.1$ ST and $T_c = 0.5$ s. The transition time and critical time values have been chosen large for visibility purposes. Critical time stands out as the primary distinction between the two methods. Although it introduces a triggering delay in DPW, we observe comparable outcomes for both corrections following that initial trigger.

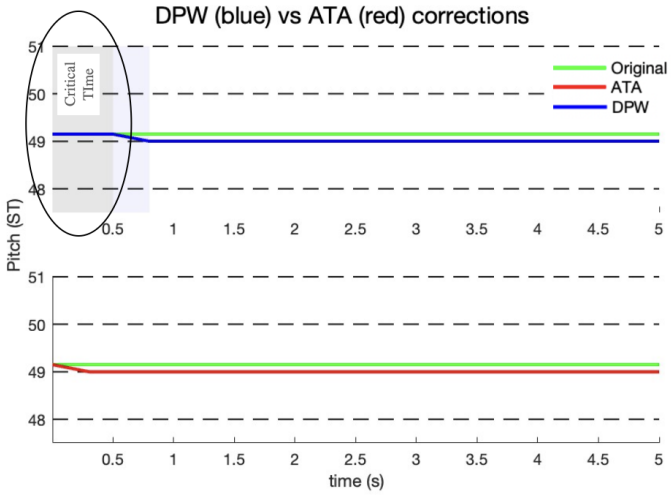


Figure 2.19: DPW correction (blue curve) and ATA correction (red curve) of a constant input pitch (green curve). The difference between both methods in this case is due to the DPW's critical time (different from zero).

2.7.3 Extreme correction with zero transition time parameter

As mentioned before, extreme correction is the configuration with an instantaneous transition, where the transition time must be the minimum allowed

by the algorithm. To achieve this, we need to set the configuration to a retune speed = 0 and a $t_t = 0.001s$, which are the minimum values allowed in ATA and DPW, respectively. Regarding triggering, in ATA, we choose the extreme value of flextune = 0 so that all notes are corrected. In DPW, the parameters are a critical time equal to $t_c = 0$ with a interval of detection $I_d = 0.01$, so the correction is triggered for any note, regardless of its stability. This configuration makes the two methods virtually equivalent. Two test signals are chosen for this test: the first one is a glissando, and the second is a melody taken from [Perrotin and D'Alessandro, 2016]. We can see the results in Figure 2.20 and 2.21. The f_o-signal treated with DPW is drawn in blue, the one treated with ATA is in red, and the original is in green.

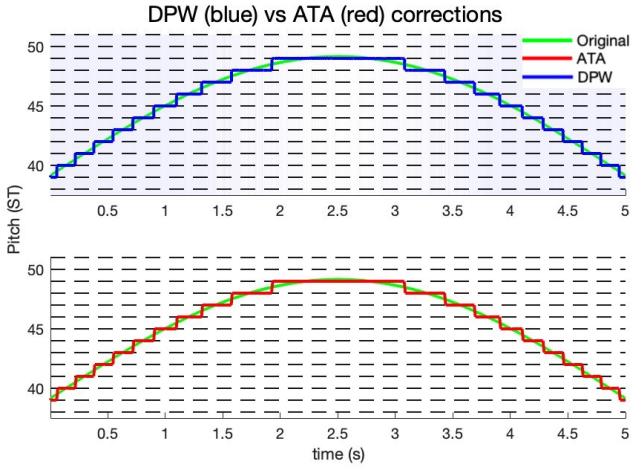


Figure 2.20: Extreme correction for a glissando. By choosing the minimal critical time in DPW, we can achieve an extreme DPW correction (blue) that closely resembles the extreme ATA correction (red).

2.7.4 Transparent and Expressive Correction

In the following sections, we will compare the parameters between ATA and DPW to understand how it is possible to achieve transparent correction in both cases and determine whether achieving expressive correction is feasible. We use the term "expressive correction" to denote that oscillatory ornaments and the free path are transposed and preserved while staircase notes are corrected. The idea is to explore the parameters that control activation, transition, and pitch curve deformation both individually and in combination. This will help us understand the changes generated in the test signal of Figure 2.21, as it contains the three basic types of vocal pitch signals (staircases, vibrato-like ornaments and freepath). Our goal is to identify optimal configurations with both methods and then compare these configurations to determine which one is better.

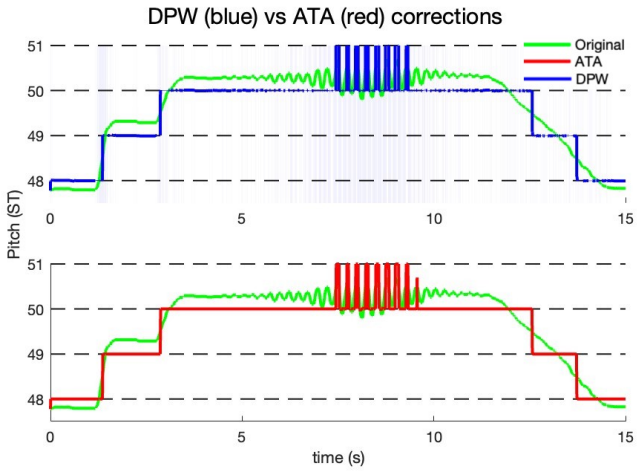


Figure 2.21: Extreme correction for an expressive melody. The expressive melody contains staircase notes, vibrato, and free-path notes, helping to show how extreme correction functions in various scenarios with both DPW (blue) and ATA (red) systems.

2.7.5 Pitch Correction with ATA

ATA comprises two parameters that control its pitch correction algorithm: retune speed and flextune. As we have mentioned, in principle, retune speed should be equivalent to the transition time t_t of DPW. We will apply different values of retune speed to observe how the performance of pitch correction changes on the test signal. The retune speed values we use are 0, 15, 50, 100, and 200 ms. The correction results can be observed in Figure 2.22 and following the given order from up to down.

When examining the correction for staircase notes, it becomes apparent that the optimal settings are 50 and 100 ms. For values of 0 and 15 ms, the correction is too fast, resulting in the loss of the original transition shape, while for 200 ms, it is too slow.

Regarding small vibratos, it is evident that these are eliminated in configurations with values of 15 ms and 0 ms, while they are preserved for values of 50 ms, 100 ms, and 200 ms. For large vibratos, it can be observed that they are not treated correctly in any case, presumably because they depart from the neighborhood of integer values. For values of 0 and 15 ms, the vibratos are shifted towards the two integer values, causing the vibrato to widen. Since the perceived frequency of the vibrato is the median between the notes, what is perceived is a note halfway between the two semitones 50 and 51. In other cases (50, 100, 200 ms), it can be seen that the correction trigger is slightly different, but the vibrato fails to be adequately corrected.

For the free path part, the best result is obtained with 200 ms; otherwise, a gradual transition from extreme correction to a smoother correction can be observed, ranging from 0 to 200 ms. This is a proof of the trade-off between free-paths preservation and overall correction with the retune speed parameter.

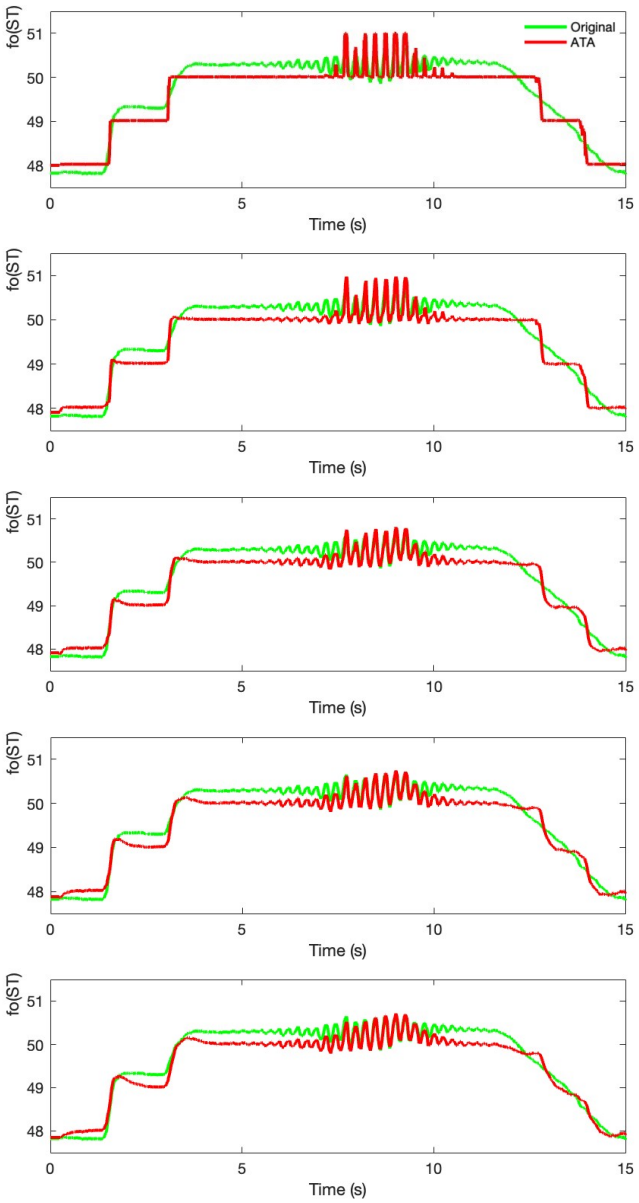


Figure 2.22: Correction using different values of retune speed on ATA: RS = 0, 15, 50, 100, and 200 ms (from top to bottom). This parameter acts as a smoother of the curve.

Now, we will analyze a broader scenario examining the functionality of the flextune parameter. In the first case, we keep the retune speed value constant and equal to zero, varying the flextune value between zero and 40, as shown in the results in Figure 2.23. The results are depicted in red for flextune = 0 and violet for flextune = 40. As observed, the flextune parameter allows for movement within the range defined by its value, in this case $(+20, -20)$ cents. This results in preserving smaller ornaments, those within the range of $(+20, -20)$ cents relative to any given integer note. Flextune significantly enhances the preservation of the expressive gesture.

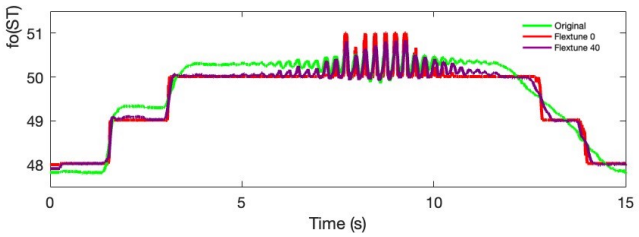


Figure 2.23: Correction with ATA, at zero retune speed and varying flextune: 0 cents (red) and 40 cents (violet). Notes within the flextune threshold are not corrected, so the larger the flextune value, the less precise the correction.

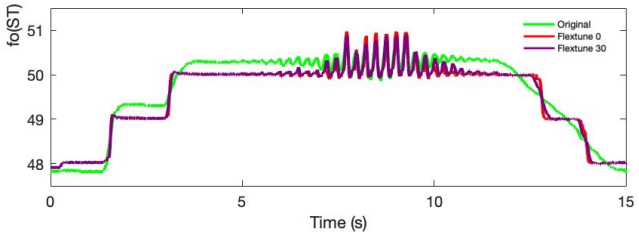


Figure 2.24: Correction with ATA, at retune speed equal to 15 ms and varying flextune: 0 (red) and 30 cents (violet)). Notes within the flextune threshold are not corrected, so the larger the flextune value, the less precise the correction.

In the upcoming example, we employ a non-zero retune speed value of 15ms along with flextune values of zero (depicted in red) and 30 cents (depicted in violet), as illustrated in Figure 2.24. As observed, akin to the previous example, ornaments smaller than the flex-tune range $(+15, -15)$ cents can be retained in the output. In this regard, we can observe how the previous case (flextune $= 40$ and retune speed $= 0$) exhibits greater preservation of the gesture than this case, showing that flextune improves ornamentation preservation and correlates with less reactivity.

Now, we proceed to test a bigger retune speed than the precedent two cases, set at 50 ms; and flextune set at 30 and 60. As illustrated in Figure 2.25, the larger flextune value results in reduced reactivity, and consequently, the vibrato remains uncorrected. At the same time the free-path is better preserved. On the other hand, the smaller retune speed in the given range shows more precision correcting ornamentation and more rapidity correcting staircases but is worse for the free-path because it gets more distorted. This is proof of the trade-off between free-paths preservation and vibrato correction precision with the flextune parameter.

Also, we examine the implications of varying retune speed while keeping flextune parameter fixed. We use a moderate flextune $= 40$, and vary retune speed $= 0, 50, 100, 200$ ms. The corresponding outcomes are depicted in Figure 2.26. For the given value of flextune, it can be observed that, in general, there is a better ornaments preservation compared to when this parameter is not used. However, the trade-off between the preservation of free paths and overall

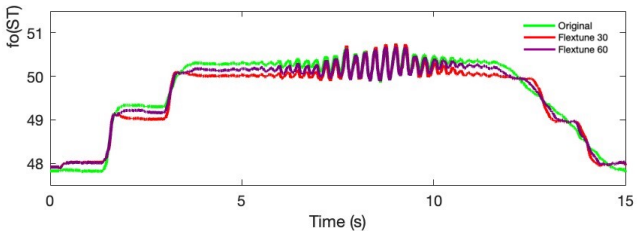


Figure 2.25: Correction with ATA, at retune speed equal to 50 ms and varying flextune: 30 (red) and 60 cents (violet)). Correction with ATA, at retune speed equal to 15 ms and varying flextune

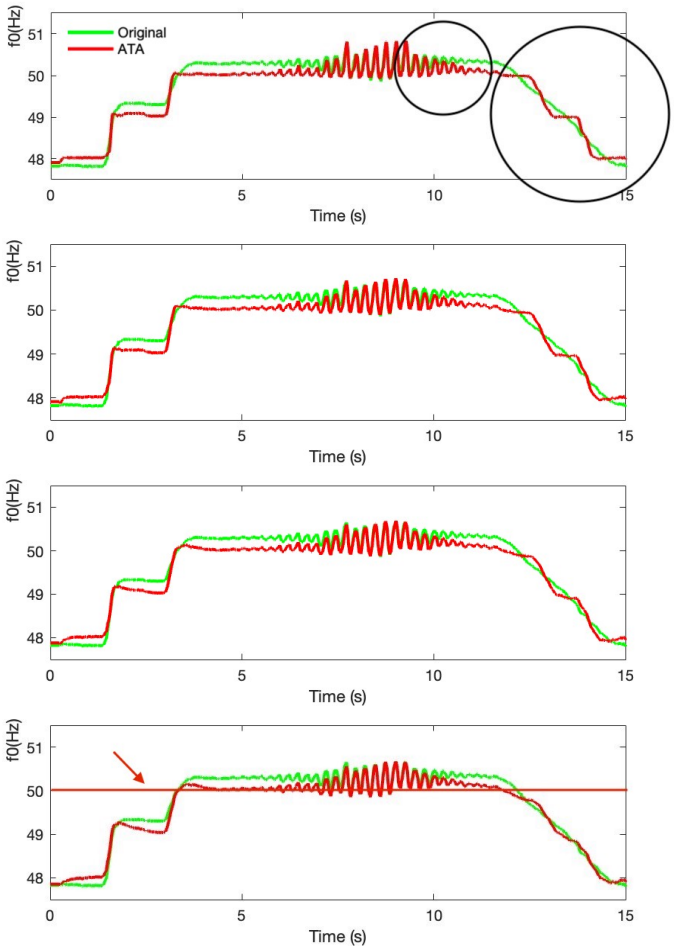


Figure 2.26: Correction using different values of retune speed on ATA, RS= 0, 50, 100, 200 ms (up to down) for the same flextune value (40 cents).

correction with the retune speed parameter still persists. In extreme cases, for retune speed = 0, it can be noted that the free path is lost, while for retune speed = 200, the correction on the staircase notes is excessively slow, resulting in more than half of the duration of each note being perceived as undefined. Visually, the best overall results are achieved for a retune speed = 50.

2.7.6 Pitch Correction with DPW

We have already seen in section 2.4.1 how the critical parameter t_c works, so for our first test, we use the transition time parameter t_t, varying it between $100, 200, 400$ ms for a $t_c = 100$. This gives us the results shown in Figure 2.27. As can be observed, varying the parameter t_t allows us to smooth the pitch correction, in the same way, the retune speed parameter does. However, it can already be appreciated that both small and large vibratos are well preserved and centered in 50. For DPW, the trade-off of the t_t parameter lies between the free path and the staircase notes (vibratos are preserved sufficiently well), whereas for ATA, the trade-off was between the free path and overall correction.

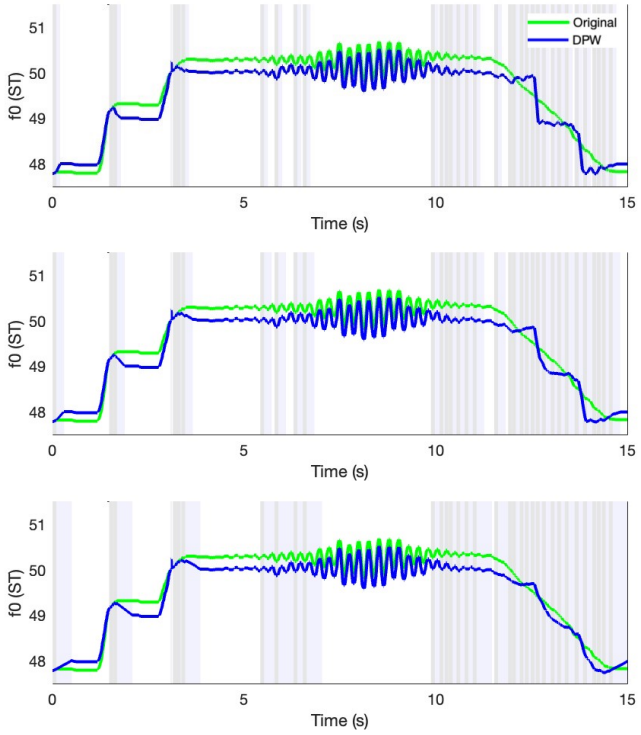


Figure 2.27: Correction using different values of transition time in DPW (from up to down: 100,200,400 ms), for the same critical time (100 ms)

Additionally, we conducted an experiment using a larger critical time of 250 ms (optimal according to [Perrotin and D'Alessandro, 2016]). We kept the critical time fixed while varying the transition time between 25 ms and 200 ms, yielding the results shown in Figure 2.7. As observed, the critical time acts as a trigger for the correction because the correction does not start until the critical time has passed. The transition time acts as a smoother, similar to how retune speed functions for ATA. Furthermore, the critical time parameter adds an ornament at the beginning of each note step in the staircase region. The slope is controlled by the transition time, as depicted in Figure 2.7; thus, a smaller transition time results in a fast change, while a larger transition time results in a slower pitch variation.

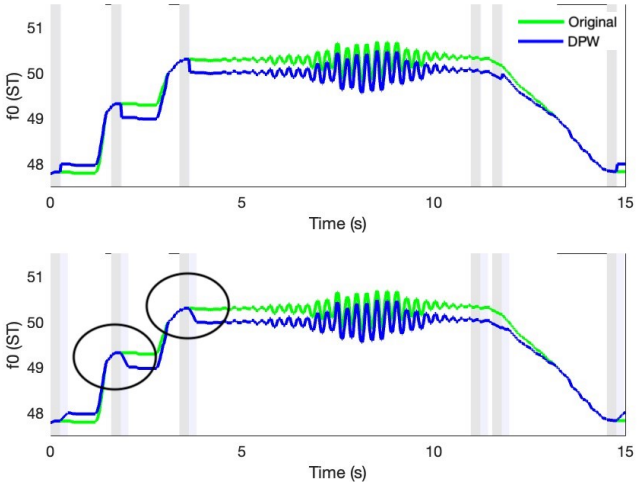


Figure 2.28: Correction using different values of T_t in DPW (from up to down: 25,200 ms), for the same T_c (250 ms)

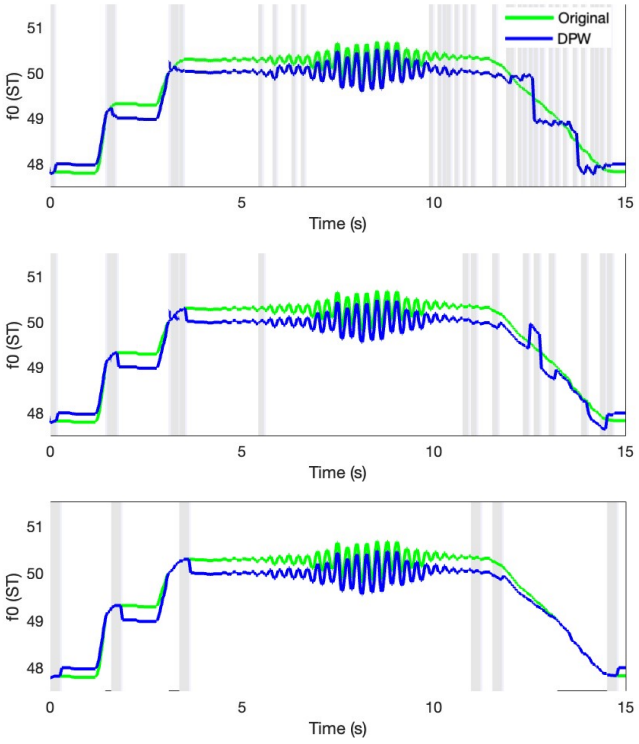


Figure 2.29: Correction using different values of T_c in DPW (from up to down: 100,150,250 ms), for the same T_t (50 ms)

Now we use a smaller transition time $t_t = 50$ ms while varying the critical time parameter $t_c = 100, 150, 250$ ms. The results are shown in Figure 2.29. We can see that the ornamentation due to the critical time changes significantly with different critical time values. However, the slope depends on the transition time, so it remains the same in all three cases. Additionally, the critical time parameter allows for the correction of vibrato sections, making it a valuable

tool for preserving vibrato ornaments. The configuration that achieves the best trade-off between the different parts of the expressive melody is $t_c = 250$ ms and $t_t = 50$ ms. From these results, we can now define optimal configurations for ATA and DPW for theoretical pitch curves and proceed to compare these configurations as done in Section 2.7.7.

Visualizing the changes due to the transition time and critical time variables is not straightforward. Therefore, we have created 3D examples to help understand how these variables work, showcasing the most practical cases. Regarding the transition time variable, we vary it while keeping the critical time constant at $ct = 100$ ms. This variation generates the surface shown in Figure 2.30, where the free-path region (between 10 and 15 seconds) can be visualized. It can be seen that for small t_t values, there is significant distortion. As the t_t value increases, the transition becomes slower, resulting in a smoother curve. We can examine the staircase notes part. As shown in Figure 2.31, just after the correction is triggered, there is a transition to the correct pitch; the t_t value controls the slope. As the t_t value increases, the slope becomes less steep. Finally, we can view the surface from a frontal cut in Figure 2.32, which allows us to confirm our observations for the staircase and free-path regions and to see that vibrato preservation is maintained while varying the transition time.

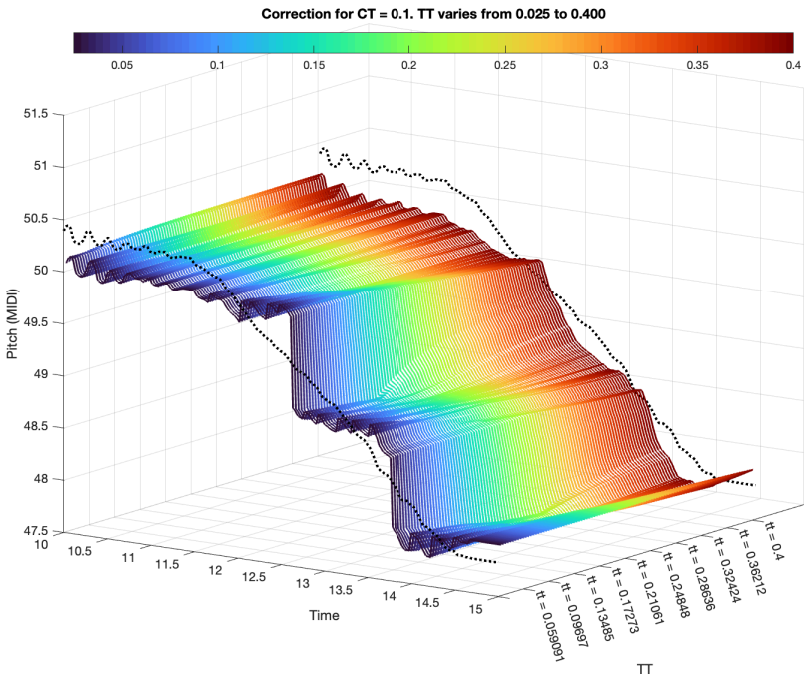


Figure 2.30: Effect of varying transition time t_t with a fixed critical time $ct = 100$ ms with DPW. In the free-path region (between 10 and 15 seconds), smaller t_t values result in significant distortion. As t_t increases, the transition becomes slower, leading to a smoother curve.

Now we will vary the critical time while keeping the transition time constant at $t_t = 50$ ms. To begin, let us examine Figure 2.33. This graph shows the

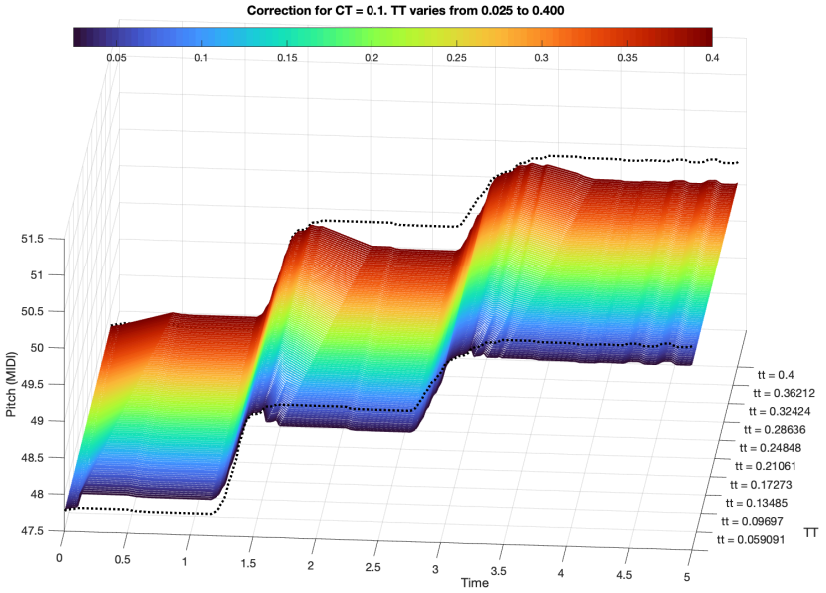


Figure 2.31: Effect of varying transition time t_t with a fixed critical time $ct = 100$ ms with DPW in the staircase notes part. The scope after releasing each correction changes according to the transition time values.

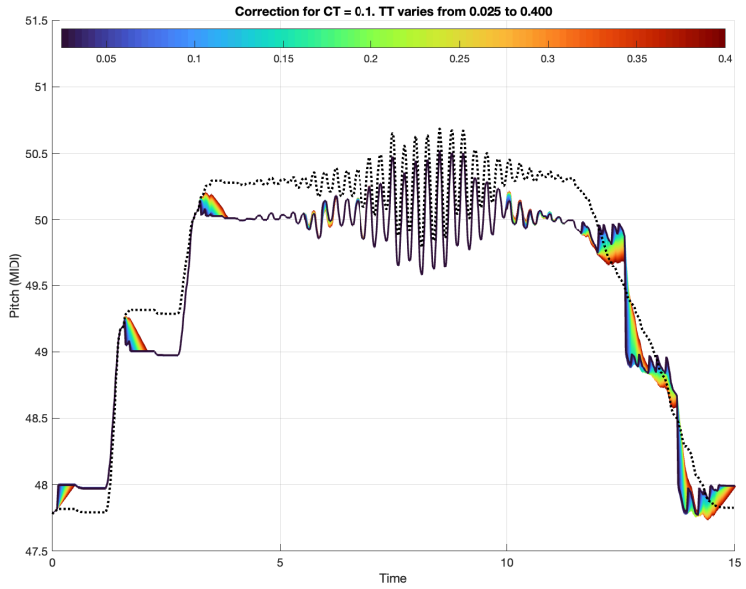


Figure 2.32: Frontal view of the surface generated by varying transition time for fixed critical time $ct = 100$ with DPW. While there is not much change in the vibrato region, the staircase and free-path regions exhibit the changes mentioned above

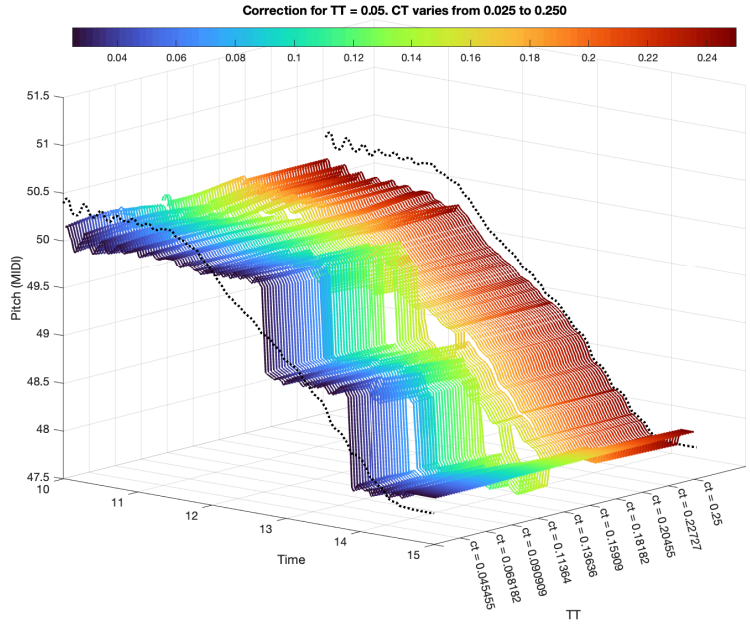


Figure 2.33: Correction of the free-path using DPW as t_c varies from 50 ms to 250 ms, with $t_t = 50$ ms. The changes in the pitch curve are discontinuous when t_c is varied. The free-path is better preserved for larger values of t_c due to the pitch curve in the free path not being stable enough to trigger a new γ

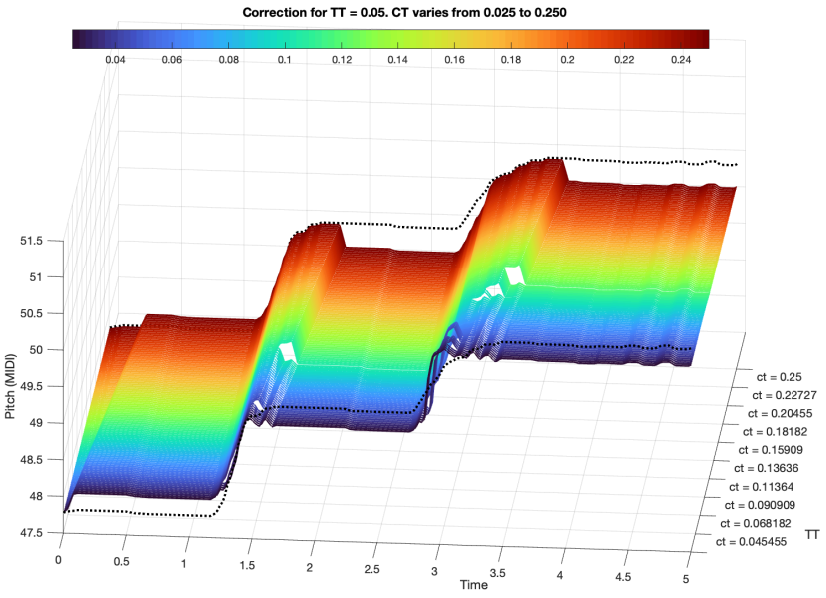


Figure 2.34: Correction of the staircase notes using DPW as t_c varies from 50 ms to 250 ms, with $t_t = 50$ ms. Similar to the free-path region, discontinuous changes are observed. The critical time creates a lobe at the beginning of each step, and the size of this lobe increases as t_c increases.

correction of the free-path as t_c varies from 50 ms to 250 ms. The pitch curve changes when t_c is varied are discontinuous, unlike when t_t is varied, which results in gradual changes as seen in Figure 2.30. In Figure 2.33, it is also evident that the free-path is better preserved for larger values of t_c. This is because the the free path pitch curve is not sufficiently stable to trigger a correction, and therefore, the gamma calculation is not updated but remains constant from the last region that triggered a correction.

Next, let us look at the graph corresponding to the staircase notes (Figure 2.34). Similar to the free-path region, a discontinuous change is observed here as well. Additionally, it can be seen that the critical time creates a lobe at the beginning of each step, and the size of this lobe increases as t_c increases.

Finally, examining the frontal view (Figure 2.35) and posterior view (Figure 2.36 with time scale inversed), it can be verified that the highest number of jumps and positioning errors of the vibrato occur at low t_c values, while for higher t_c values, the vibrato correction varies less with t_c.

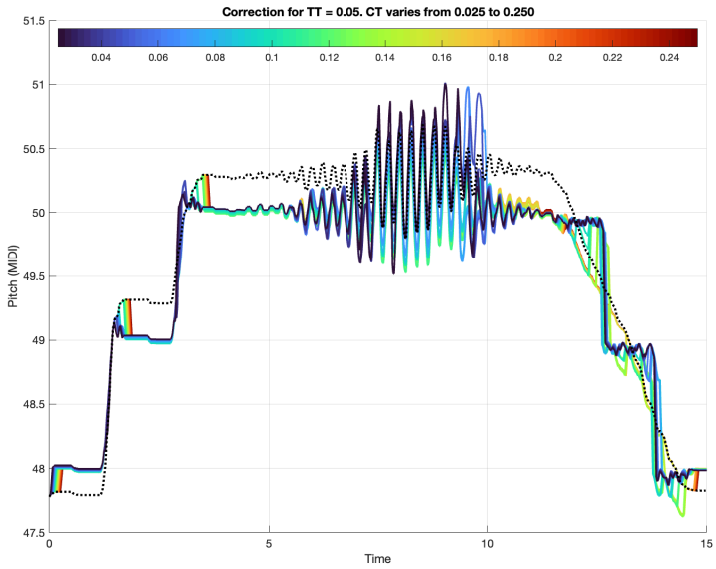


Figure 2.35: Frontal view of vibrato positioning errors using DPW as t_c varies, with $t_t = 50$ ms.

Remember that the parameter I_d represents the detection intervals, defined in code 2.11, where this parameter specifies a constant grid size that divides the pitch axis but is not evaluated on the pitch curve itself. Reducing I_d to a smaller value will result in a stability condition that is difficult to satisfy, thus decreasing the likelihood of note correction. Conversely, selecting a larger I_d will result in less precise correction, as it will be triggered depending on the grid position and the note's neighborhood. Varying I_d causes the correction shape to vary discontinuously and spontaneously, making it impossible to use the detection interval value to control the correction. The impact of varying I_d can be observed from various perspectives in the following graphs: lateral

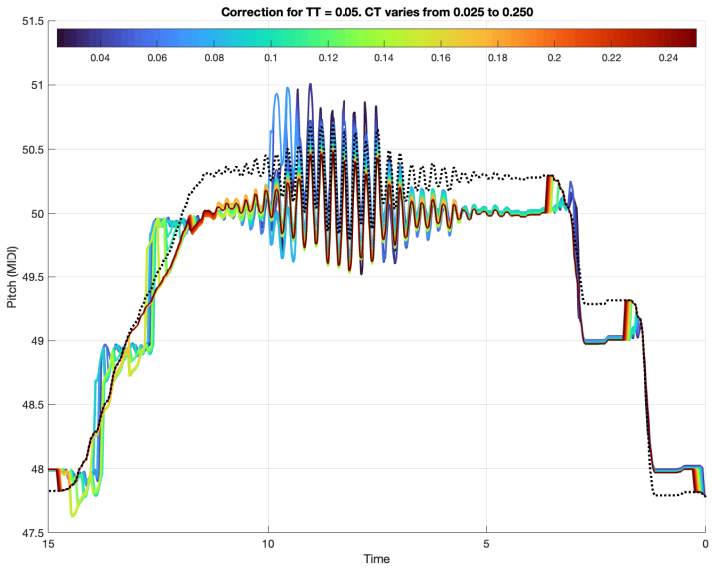


Figure 2.36: Posterior view of vibrato positioning errors using DPW as t_c varies, with $t_t = 50$ ms.

(Figure 2.37), frontal and posterior(both in figure 2.38), where we have set the values of t_c and t_t as recommended by 2.11 and then varied the parameter I_d using the reference code.

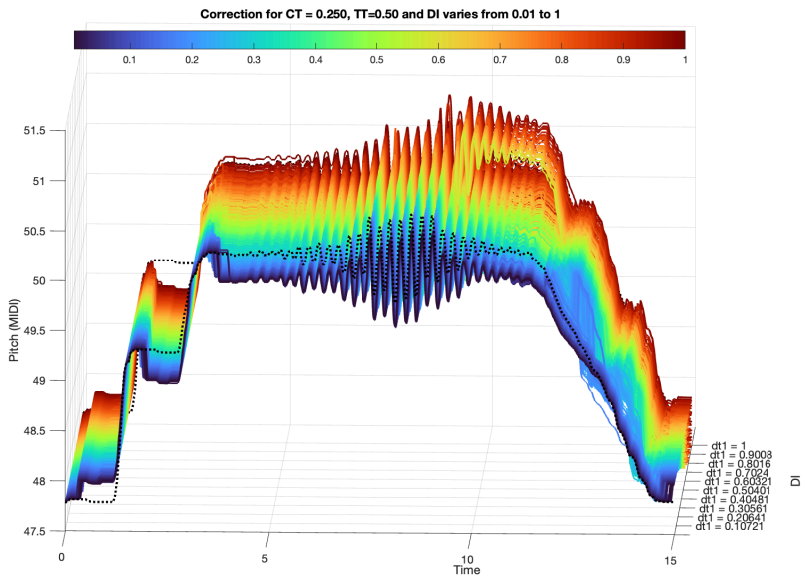


Figure 2.37: Lateral view when varying I_d with constant t_c and t_t.

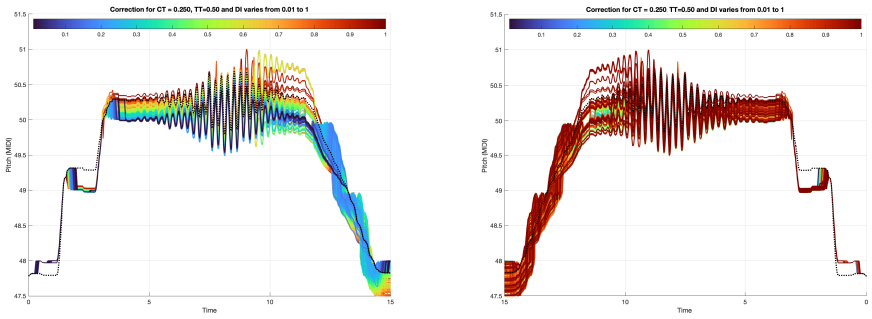


Figure 2.38: Frontal (left) and Posterior (right, with inversed time scale) view when varying I_d with constant t_c, t_t

2.7.7 Comparison between ATA and DPW

The optimal configurations for ATA (retune speed = 100 and flextune = 40) and DPW (t_c = 200 ms, t_t = 50 ms, I_d = 0.1 ST) are shown in Figure 2.39, along with the error concerning the ideal signal that should be achievable. It is evident that, in ATA, simultaneously achieving optimal correction for vibrato and free path is not possible. The optimal result is a moderate configuration of ATA that performs both tasks reasonably well. In the case of DPW, it can be observed that the trade-off lies between the distortion in the free path and the total treatment time of the staircases. Conversely, vibratos are effectively corrected. Next, we will estimate the error that these configurations present concerning the ideal signal that should be obtained with them according to Figure 2.39.

We can quantify the disparity between two curves using different metrics and have introduced two here. Firstly, the Mean Squared Error (MSE) gauges sensitivity to quadratic errors by computing the difference of squares. This method assigns more significance to larger errors, offering a measure of variance between the curves. Secondly, the Mean Absolute Error (MAE) provides an average measure of the magnitude difference between the curves. In contrast to MSE, MAE does not magnify larger errors. We express these concepts using the following equations:

$$\text{Mean of MSE} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ij}-\hat{y}_{ij})^2\right) \tag{2.16}$$

$$\text{Mean of MAE} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}|y_{ij}-\hat{y}_{ij}|\right) \tag{2.17}$$

Where N represents the number of samples, n_j is equal to 1, because there is always a comparison of one curve with the reference, j represents the curve to compare (ATA or DPW), y_{ij} are the values of the original curve j, and $\hat{y}_{ij}$ are the values of the comparison curve j.

Our example serves to illustrate three types of pitch modification. The initial segment in the $0 < t < 5$ time range exhibits a signal resembling a stair-

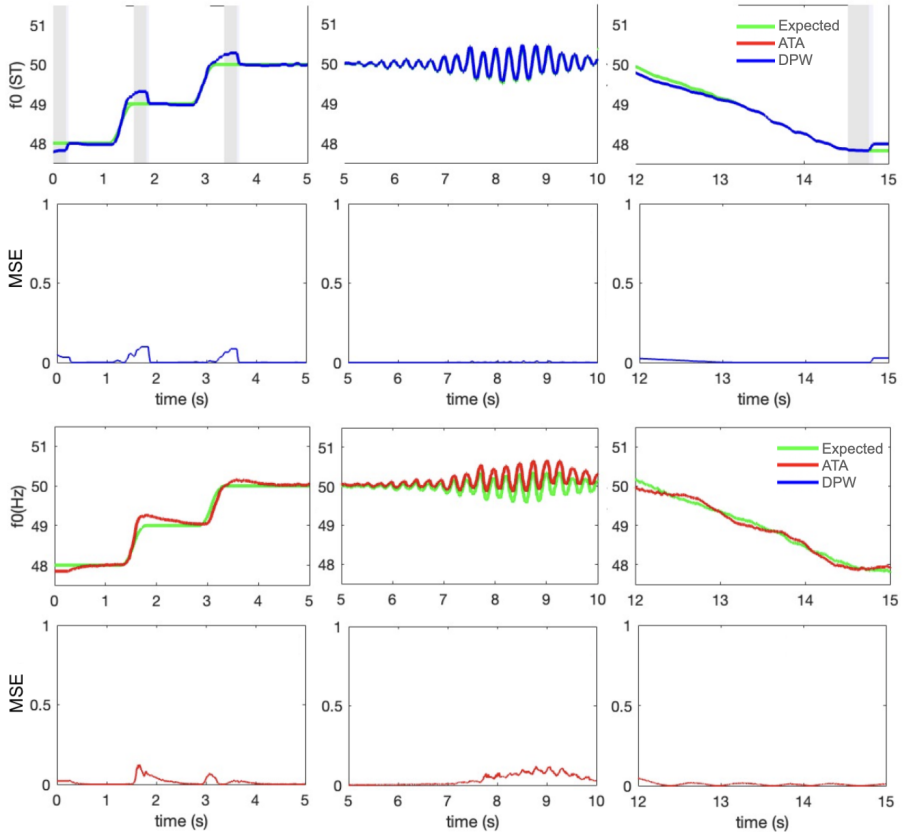


Figure 2.39: Correction for the same T_t (50 ms) using flextune at 40 cents for ATA and T_c at 200 ms for DPW and the difference with the ideal signal.

case between notes 48, 49, and 50. The subsequent segment in the time range of $5 < t < 10$ represents the correction of a poorly intonated frequency modulation akin to the human vibrato. The final part depicts a gentle trajectory of the f_o that should not be corrected; the free path signifies the scenario where the singer does not intend to produce any specific note. Each of these segments needs to be compared to the desired pitch curve, which varies for each region. For instance, the desired signal for the staircase segment is a staircase itself. In the vibratory segment, the ideal pitch would be the same vibration but well-centered. As for the third segment, the original signal serves as the ideal pitch; here, the goal is not correction but preservation. These assumptions are depicted in Figure 2.16, and the MSE calculation is performed point by point. The mean over each region is summarized in Table 2.1. As previously mentioned, DPW demonstrates superior correction of vibratos while preserving the free path of the note. In contrast, ATA performs better for the staircase segment but sacrifices more regarding vibrato and free path components.

It is important to clarify that all comparisons focus on pitch correction curves. The sound implementation of DPW involves adding pitch tracking and pitch warping methods that differ from those of ATA. However, despite

Table 2.1: MSE and MAE between input and corrected f_0 for the different regions

	MSE		MAE	
Region	DPW	ATA	DPW	ATA
1	0.0146	0.0146	0.0747	0.0914
2	0.0415	0.0642	0.1304	0.2103
3	0.0539	0.0280	0.2015	0.1463

these limitations, it is crucial to emphasize the valuable insights obtained from this comparison, providing a nuanced understanding of the distinct strengths and weaknesses inherent in each method.

2.8 Summary

Through our research, we studied the DPW algorithm for audio pitch correction. It is possible to control and trigger a pitch correction thanks to three degrees of freedom that preserve low-amplitude vibratos and ornaments in the neighborhood of the target note. We have also shown how the pitch correction methods are composed of two stages (triggering and warping) and how modifying the control parameters can lead to equivalent configurations for different systems. We have identified a scenario where ATA and DPW exhibit similarity: extreme correction. Moreover, we have identified three types of correction: staircases, vibratos, and free paths, and have illustrated that DPW performs better for vibratos and free paths while also being adequate for staircase correction. DPW also exhibits a different trade-off between its parameters compared to ATA, also it provides a better response for ornamentation.

We have developed an audio support that includes the DPW and ATA methods use over the testing signals from figures 2.22 to 2.39. DPW compared to ATA presents a smoother pitch trajectory transition towards the nearest notes on a defined scale, minimizing distortion of melodic ornaments between the notes. However, it is important to note that the vocoder used in our application may not provide the same level of quality, precision, and accuracy as the ATA vocoder; and that we cannot replicate pitch tracking and warping of ATA vst.

A comprehensive perceptual evaluation of the two systems in a formal setting later in this manuscript. This evaluation aims to assess the perceptual salience of the pitch effects introduced by the DPW method and their potential musical relevance.

Chapter 3

Vocoders and Tuning

The vocal production techniques can involve many digital audio effects (DAFx), of which the vast majority are not designed specifically for the voice, even though they are used for it. The voice is the richest musical instrument in terms of style and technique. However, it is also the most difficult to study. This difficulty is due to the number of timbral subtleties that a phonatory system prototype must be capable of reproducing. Those subtleties can change for different people, even for the same person of different ages. The vocoder as effect is widely expanded today in studio and live performances techniques in an extensive range of possible configurations. The term vocoder describes the numerous vocal transformation and reconstruction techniques, an elemental component of vocal research. It is mainly used in music but is also relevant for other fields such as health, communication, and computer-machine interaction.

The vocoder as a tool for vocal signal reconstruction is closely related to pitch correction. We must use a vocal transformation algorithm such as a vocoder to impose a pitch curve or a pitch transposition over a signal. Our research is interested in studying the pitch correction perception. Indeed, a characterization of the vocoder's sonorous perception is needed before perceptually studying the pitch correction methods. Such characterization can be done through a psycho-acoustical evaluation considering specific vocoders and particular cases of pitch correction for a given pitch correction algorithm (ATA as the primary reference of pitch correction). In this chapter, we will address the basics of the vocoder study. Therefore, we will prepare a sound library implementing specific vocoders and use cases (presets, such as extreme autotuning or transparent autotuning). Then, we select samples and explain how to integrate them into a vocoder perceptual evaluation. The results of the psychoacoustical assessment will allow us to define the weight of the vocoder's contribution to the vocal signal coloration when the signal has been tuned. Then, in the next chapter, we will be able to study how to characterize different pitch correction methods (ATA or DPW) and their corresponding contribution to the vocal signal coloration.

3.1 Vocoder Evolution Context

Nowadays, the term vocoder agglomerates the numerous vocal transformation techniques for high-quality modification of natural vocal sounds. Among them, we can mention the spectral techniques such as the channel vocoder, the time domain techniques such as PSOLA, the phase vocoder, the source-models, and the use of neural networks.

Initially, the term comes from channel vocoder, which used a contraction of "voice coder" [Dudley, 1937] [Dudley, 1939] [Flanagan and Golden, 1966] [Moulines and Laroche, 1995]. The channel vocoder breaks down the spectrum into sections called subbands, which are analyzed and manipulated through parametric data, resulting in a sound transformation [Cook, 1998]. Later, other techniques appeared to make it suitable for musical application; the phase vocoder was first described by Flanagan in 1966 [Flanagan and Golden, 1966]. The phase vocoder [De Götzen et al., 2001] [Dolson, 1986] [Moulines and Laroche, 1995] calculates and maintains both instantaneous magnitude and phase using the Fast Discrete Fourier Transform. However, using the source-filter model and the parametric modeling of the source has been fundamental for vocal reconstruction when performing tasks beyond simple time stretching. In the following section, we mention some of the main improvements of the vocoder and vocal models.

One of the inherent vocoder sonorous artifacts is phasiness. It appears when slowing down a sound; it makes it muffled, reverberant and/or moving away from the microphone. Phasiness is due to the loss of coherence between the phases across the bins of the Short-Term Fourier Transform STFT over time, so phases must be regularly reset in order to keep them coherent [Moinet and Dutoit, 2011]. Phasiness in the vocoder is not immediately apparent but takes a few frames to become noticeable. Particularly for speech, the phasiness effect sounds strangely reverberant or with a lack of presence of the speaker. This problem comes from the unpredictable relationship of the attacks with previous frames of the signals. The PVSOLA (Phase-Vocoder with Synchronized OverLap-Add) is a method intended to improve the presence of phasiness. Roebel has treated it [Roebel, 2003] on the level of spectral bins by reinitializing the phase spectrum.

Various other sonorous artifacts of the phase vocoder that are present in several of its implementations. We can mention frequency smearing, reverberation [Favreau, 2001], and transient softening [Roebel, 2003]. Some researchers focus on improving naturalness by addressing transient irregularities [Loscos and Bonada, 2004] and vocal pulses[Bonada, 2004]. Much vocoder research focuses on naturalness, principally treating transient irregularities [Loscos and Bonada, 2004] and vocal pulses [Bonada, 2004]. For example, there are improvements in the growl phonation based on a time-domain pitch-synchronous overlap-add (TD-PSOLA) [Bonada, 2004] that controls pitch by the frame reading speed and distance between pulses.

This research approach also allows for modeling vocal disorders. Vocal dis-

orders, intentional or not, are related to irregularities in the excitation glottal pulse in time (jitter) and amplitude (shimmer). They affect the subharmonics spectrum and vary over time.

The vocal timbre can be modified or improved by scaling, warping, and equalizing the estimated spectral amplitude of each vocal pulse. Phase vocoder improvements like the pre-warping function for frequencies [Roebel and Rodet, 2005] [Roebel, 2010] help to rebuild the signal without requiring pitch mark. The standard implementation of the phase vocoder uses instantaneous frequency estimation and phase unwrapping. One disadvantage of the STFT is its rigid time-frequency resolution trade-off and its constant absolute frequency resolution. Another improvement is the constant-Q transform (CQT) [Schörkhuber et al., 2012] to create a multi-resolution frequency scale that eases the detection of harmonic structures and reduces interference in lower-frequency areas.

There are applications well documented as the ones of IRCAM, for both non-real-time (SuperVP) [1] and for real-time (TRAX) [2] [3], which can transform gender, age, vocal quality, etc, rather than trying to attain a specific target voice, as noted by Farner [Lanchantin et al., 2011]. These applications have improvements to deal with the transients [Roebel, 2003], waveform preservation [Roebel, 2010], spectral-envelope estimation [Roebel and Rodet, 2005], and dynamic voicing with spectral-peak triage. Concerning vocal parameters, the major enhancement outside of those already discussed is the parametrization of vocal tension by the parameter R_d, which characterizes the slope of the glottal spectrum [Lanchantin et al., 2011]. Voice quality (breathy, harsh voice) can be transformed via the Rd parameter; glottal closure instants marks allow for adding jitter (e.g. creaky voice). YIN [de Cheveigné and Kawahara, 2002] is a robust f0-estimator used to define the voice-unvoiced regions and a base for several devices developed at IRCAM (34), including SuperVP and TRAX.

3.1.1 Vocal Research Approach

The voice is not only defined by its natural pitch range but also by its timbre, which is influenced by physiological and phonatory factors. Timbre, concerning the identity and qualities of a sound, has been the subject of numerous musicological studies aiming to define and characterize sound events [Schaeffer, 1966]. In the realm of voice, terms such as dark, bright, soft, rich, noisy, pure, rough, etc., are commonly employed by musiciens as shown in [Garnier et al., 2007]. Determining the gender and age of an individual can provide insights into their voice characteristics [Lanchantin et al., 2011], for example, age plays a significant role in the frequency range of the vocal folds (pitch measured as f0), the spectral distribution of the glottal source (measured as spectral

[1] https://forum.ircam.fr/media/uploads/software/SuperVP20for20Max/supervp-for-max.pdf

[2] https://www.flux.audio/project/ircam-trax/

[3] http://anasynth.ircam.fr/home/category/logiciel-associ%C3%A9/supervp-trax

tilt), and the vocal tract acoustics (formants and anti-formants). Younger voices generally have a smaller vocal tract. In contrast, aged voices exhibit characteristics such as decreased intensity, breathiness, relatively high pitch (primarily in men), lower flexibility, and perhaps trembling [Lanchantin et al., 2011, Klatt and Klatt, 1990]. However, beyond these general trends, each person may have a distinct vocal timbre at different ages.

Voice classification by genre, age, and vocal range often prompts the use of additional sound descriptors, as mentioned earlier. This is because vocal modes and styles contribute to the diversity of voice characteristics. For instance, whispering involves the separation of vocal folds; spectrally, it is similar to a speaking voice at high frequencies but differs at low frequencies. From other hand, roughness can result from various pathologies, but not only, it can be modulated also in healthy voices and may combine with other characteristics such as hoarseness or creakiness. Mathematically, roughness refers to variations in the fundamental frequency and period amplitude (jitter and shimmer) [Loscos and Bonada, 2004]. Standard techniques for reproduction include source-filter model-generated aperiodicities in the time domain, statistical models, or the use of vocoders. The technique introduced by Loscos [Loscos and Bonada, 2004] involves adding sub-harmonics in the frequency domain with a phase-locked vocoder. The growl effect, a vocal technique involving simultaneous vibrations of the vocal folds and supra-glottal structures of the larynx, produces sub-harmonics. The growl algorithm adds these sub-harmonics to the original voice spectrum to emulate growl phonation, using magnitude-phase patterns from real growl recordings.

As shown constantly through the vocoder research, and particularly by Abe et al.[Abe et al., 2008], the phase vocoder distorts not only the harmonic part of the spectrum but also the inharmonic part, making timbral analysis challenging from a musical point of view. The non-parametric modification of the spectrum hinders analyzing timbral features as explicit parameters or sound descriptors. Moreover, decomposing audio signals into perceptually meaningful modulable components is desirable [Disch and Edler, 2010] for developing new effects and efficient audio compression. Artifacts and secondary effects are not necessarily to be avoided. A proper approach could make them controllable so we could enrich the musical creation environment with those defaults.

Within the research on vocal perception, we cannot be left out of the work of Michelle Castellengo [Castellengo, 2014], who, through her research, has found that it is "illusory" to see an "absolute" description of sonorous qualities in the singing voice. Because the descriptors only have a sense when organized by type of voice and singing style. Nevertheless, she mentions that it is possible to develop a transversal vocabulary that allows communication between physiologists, musicians, and researchers, as did in [Henrich Bernardoni et al., 2008]. So, it is possible to have some voice descriptors that illustrate some of the vocal characteristics. These descriptors with a more semantic purpose can be used to identify modelable and controllable parameters over the vocoding technologies. Searching for a detailed and complete description of the sound of

the vocoder is complex. This path is even more complicated if we consider that the vocoder technology advances to a transparent system capable of modifying the inner human vocal descriptors. As it happens with the guitar amplification, the musical use of the vocoder is founded through the sonorous defaults (the unique artificial sound produced by each vocoder). These artifacts are the ones that have applications in music and that are captivating both musicians and audiences. Then, we could look at these particularities through different vocoders and release a first trace about what defines the coloration of a vocoder and if it has similarities with other musical events on voice or instruments.

3.1.2 Vocoder relation with voice and pitch

As we have seen, the vocoder has involved two evolution lines, the improvement of vocal transformation and the parametric control of vocal characteristics. (S. Farner in [36]). This for sure has enormous impacts, notably for aid devices for vocal disabilities, compression, reconstruction of signal data in communication, and music applications. However, this research line contrasts with the contemporaneous musical use of the vocoder. Such an approach dismisses the use of the vocoder as a musical effect profiting from its artifacts for a musical application that is today the principal musical interest of the vocoder in music. Facts on modern popular music can support this premise. An enormous quantity of musical pieces contains backing vocals with vocoded vocal layers and autotuned voices. This is a typical course of action on mastering and production that seems systematic in both studio and live productions. It seems that vocoders are often employed more for their sonorous defaults than for their similarity with a real voice; for example, the robotic-artificial sound is sought with a musical purpose. Some artists, in fact, refer to vocoders as "just one more tint" of their own voices color (translated from spanish) [4] [5], so we could check effectively for that coloration inherent to the vocoding technique.

Apart from the parameterized descriptors of voice and the vocoder coloration, let us remember that pitch transformation is one of the foundations of the vocoder application. Pitch is the first element appreciable concerning vocal perception, and it can give the listener and idea of the singer's age, genre, size, etc. These elements define the melody and the expression. This means f0 in singing not only carries the melody but also the singer's expression. Moreover, F0 conveys the singer's identity, the musical style, and the emotion required by a musical interpretation. In speech, the f0 carries essential information such as mood, intent, and identity. The modification of f0 can result in changing or obfuscating the speaker's gender, as mentioned in the previous subsection.

It is fundamental to understand further than the research advances on the vocoder as a technique of speech transformation, there is a musical use. The musical use is based on giving the musicians new coloration to their voices.

[4] https://youtu.be/8xGgFmoLRAE?si=NonuPIQQ7Ftp9cAv&t=1587

[5] Rosalia interview about the production of her album Motomami

They can keep their identity and the vocal intention of their vocal performance using the vocoder as a layer to improve some aspects technically and use them musically as the pedals for the guitars.

The musical use of the vocoder in modern music began with artists like Daft Punk, Cher, and T-Pain, as previously mentioned, this usage has evolved over the years, becoming systematic across all musical genres. Initially, with autotune in rap and trip-hop, for example, with Drake, Future, Kanye West, and Post Malone, and more recently with hyperpop. While the coloring provided by the vocoder and autotune is reproducible live in terms of sound, it is not in terms of performance, as specific autotune transitions cannot be replicated exactly live.

Live applications has been explored by Taylor Swift, Dua Lipa and Loreen, who use the effects as supports aiming to deliver a consistent performance with studio versions; in fact, vocoder is part of the chain of effects today as seen in several sources online [6] [7] . Meanwhile, artists and producers experimenting vocal effects have emerged such as Sophie, Arca, Raye, Mon Laferte, Rosalía, and Charli XCX.

An example of vocal production with layered vocals can be seen in the video provided by Charlie Puth [8], where he explains how to use Antares Autotune software to create vocoded or pitch-shifted vocal layers. This approach is evident in musical pieces like "Made You Look", where variable segments are used to give body and chorus to the song [9], in "Delicate" [Taylor Swift, 2017] in segments like 0:00-0:19. Moreover, exploratory effects can be appreciated in tracks like "c2.0" [Charli XCX, 2020b], or in "Black Mascara" at 0:00-0:32 [Raye, 2024], but with several layers throughout the song, creating a complete atmosphere with vocal replicas of the vocalist. Additionally, we recall many more examples have been provided in the preceding sections regarding timbre modification and autotune.

3.2 Review of Vocoders Techniques

The comparison between vocoding algorithms is extensive, requiring access to different vocoding systems and methodologies of use, which may vary with each algorithm. These methodologies must adapt and be reactive enough to make pitch changes. Given the multitude of usable vocoders, we have narrowed down our search to 4 vocoders: Retune, which is a frequency-time technique; ATA, which is a temporal technique; and Circe and World, both of which are artificial intelligence-based approaches.

[6] https://www.soundonsound.com/techniques/inside-track-dua-lipa-dont-start-now
[7] https://www.youtube.com/watch?v=7Y6aFCS8evg&t=416s
[8] https://www.youtube.com/watch?v=Ja_emre9Wwc
[9] https://www.soundonsound.com/techniques/inside-track-meghan-trainor-made-you-look

3.2.1 Autotune Antares

Autotune Antares (ATA) is a software developed by H. Hildebrand that employs time domain techniques and advanced filtering techniques with high resolution at low frequencies. The details of Autotune have already been addressed in the previous chapter of this thesis, including details of Hildebrand's patents so we don't need to provide additional information here.

3.2.2 World

World is an open-source vocoder developed to improve the sound quality and processing speed of real-time speech applications. According to the developer, a comparative evaluation regarding conventional systems shows that World provides better sound quality for natural speech, is over ten times faster than conventional systems, and is suitable for real-time processing. In Figure 1, the adapted information provided by Morise in his paper [Morise et al., 2016] is illustrated. The World's processing relies on the three following processes:

The first step involves multiple low-pass filtering at different cutoff frequencies. For each output, fundamental frequency candidates are estimated. Over two complete cycles, the signal consisting solely of the fundamental component should exhibit the same positive zero-crossing, negative zero-crossing, and peak-to-peak interval values. Therefore, these values' standard deviation is considered as the fiability parameter for each F0 candidate, with the average being the F0 candidate. The candidate with the highest fiability is selected.

The second step involves calculating the spectral envelope with CheapTrick [Morise, 2015], using the original signal and the fundamental frequency. Traditionally, the spectral envelope is computed using Cepstrum and linear predictive coding, which suffers from temporal position dependence issues. CheapTrick is a method for computing the spectral envelope based on the synchronous pitch analysis idea, with a lifting function to smooth the logarithmic power spectrum (cepstrum) that effectively removes the time-dependent component and parameterizes it in terms of two values that the authors have studied and optimally determined.

The third process involves estimating aperiodicity using the D4C method (Definitive Decomposition Derived Dirt-Cheap) [Morise, 2016]. The algorithm begins with the group delay, describing it as a function of the spectrum and its derivative. The time-dependent spectrum is written as a sum of components for each band, and a parameter C is introduced into the time-dependent component. This parameter corresponds to a temporal shift. When calculating aperiodicity, the term C is used to eliminate the temporal dependence of aperiodicity.

The synthesis is performed based on Figure 3.1. Essentially, the signal is considered to be composed of a train of pulses and a noise signal, which allows for defining a spectral envelope and aperiodicity. In World, the output signal can be seen as a function that can be calculated from the aperiodicity and spectral envelope values obtained from the original signal, by applying a new

fundamental frequency.

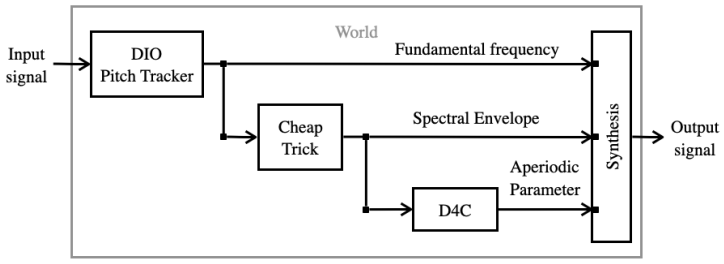


Figure 3.1: Schema of the World vocoder process

3.2.3 Circe - Neural Vocoder

The Crazy IRcam neural auto-encoder [10] [Bous, 2023] for voiCE is a neural vocoder that allows for pitch transposition and time stretching. Neural vocoders employ various approaches, such as signal pitch and energy, utilizing the mel-spectrogram as a parametric space, and models of acoustic dependencies in the source-filter model. The source-filter model enables the parametrization of a vocal signal through the source frequency (glottal source signal) and the noise excitation signal. These properties are not actually independent. Deep learning methods such as WaveNet can generate the glottal source and noise excitation interdependently, as described in [Bous, 2023]; however, these processes are slow and can only be used for a specific voice. Through their work, Bous demonstrates how the use of the mel-spectrogram preserves the spectral dependency concerning fundamental frequency and noise. The mel-spectrogram, due to its frequency axis, is more efficient than other methods and serves as part of their parametric space combined with the use of pitch.

The voice transformation with the CIRCE neural vocoder consists of two components. The first one is the bottleneck autoencoder [Bous and Roebel, 2022], which allows for pitch extraction from the original mel-spectrogram and the generation of a latent code from the mel-spectrogram and the fundamental frequency. Additionally, the autoencoder also serves to estimate the scaling factor (voice volume of the sample), thus ensuring good sound quality in the output. Bous also verified the bottleneck size to obtain the best possible results. The second component is the mel-spectrogram inverter, the Multi-Band Excited WaveNet (MBExWN), which allows for obtaining raw audio from a modified mel-spectrogram [Roebel and Bous, 2022].

The CIRCE vocoder is based on the scheme shown in Figure 3.2. In this algorithm, the mel-spectrogram calculation is performed first, followed by the estimation of the pitch corresponding to that mel-spectrogram using the autoencoder. Together, the mel-spectrogram and the pitch are used to create the latent code also with the help of the autoencoder. A new pitch value is

[10] https://forum.ircam.fr/projects/detail/circe/

defined by the user to perform resynthesis (pitch transposition) using a decoder that generates a new mel-spectrogram. The new mel-spectrogram is then transformed into raw audio using a mel-inverter.

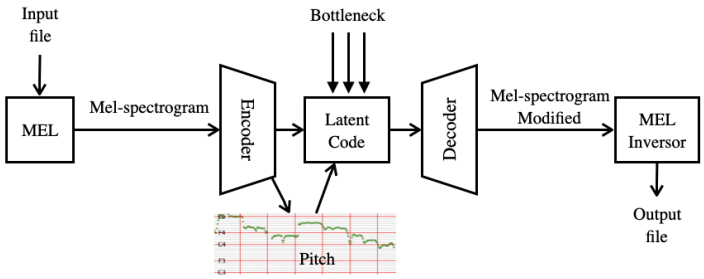


Figure 3.2: Schema of the Circe vocoder process

3.2.4 Retune

Retune is a vocoder based on time-frequency techniques, which is available in MAX MSP. The manufacturer of this technique is Zynaptiq GmbH (Hannover, Germany). On their website, Retune is referenced as utilizing a technique called ZTX, which enables time stretching, pitch shifting, formant shifting, and pitch correction. The only documentation found regarding these aspects is a patent concerning the employed vocoder method [Bernsee and Gökdag, 2016]. The technique in its main structure resembles the time-frequency approach for vocoder deduction but includes several peculiarities.

The general structure of the system is described in Figure 6 of the patent and steps 602 to 620, and the equations are described in columns 6 to 18. The first step involves selecting an audio grain and computing the discrete Fourier transform, obtaining the magnitude and phase components. With these components, a time-frequency representation is constructed. This time-frequency representation undergoes smoothing, which has a coupling effect between segments of the time-frequency representation, increasing the presence of artifacts in adjacent grains. The time-frequency matrix plus smoothing is called Cross-Frequency Phase Coupling (CFPC). Additionally, the CFPC is made dependent on a smoothing parameter, which is parameterizable and helps to better define according to the location in the time-frequency space; for example, it allows for better resolution at low frequencies. Additionally, smoothing aids in mitigating the effects of the uncertainty principle. Consequently, the resulting representation is called a reduced uncertainty transform representation.

The resulting time-frequency representation is then passed back to the time domain using an Inverse Discrete Fourier Transform, resulting in audio identical to the original. The significance of the method lies in the fact that the resulting time-frequency representation can be modified (step 614). However, despite being mentioned in claims 5, 9, and 11, no further details are provided regarding such modifications.

The methods used for retune are applicable on the commercial versions of ZTX software, the objects retune, pitchshift, freqshift on MAX, and the commercial retuning devices on Digital Performer and MOTU software.

3.3 Cases

To evaluate the sound of the mentioned vocoders, it is necessary to understand what is possible and useful from a preliminary approach. As mentioned earlier, the description of the voice is complex and varies according to the musical style. Therefore, the differentiation between vocoders should be based on what is intrinsic to a vocoding process and not on descriptors for which a particular vocoding technique may eventually be used. A counterexample would be the simulation of the hoarse effect by a vocoder; in principle, that is not the general use of a vocoder. Such a descriptor would only be comparable if several vocoders were compared, all used to simulate a hoarse effect on the voice. The comparison must be based on a parameter present in all vocoding techniques. The melody is the sound element more important when vocoding, so we can take a melodic approach.

The vocoder has two approaches: musical and scientific. The musical approach is highly influenced by the melody imposed on the output voice after vocal reconstruction, which is evidenced by both: the use of the channel vocoder and the use of autotuning systems. The scientific process seeks to resynthesize the signal in the most natural way possible preserving the original melody.

3.3.1 Natural pitch resynthesis

The primary characteristic fulfilled by a vocoder is resynthesis; just in the previous example, if a vocoder algorithm serves to create a hoarse voice, in principle, that algorithm should also serve to perform resynthesis or bypass. Vocoder techniques have very diverse foundations. Resynthesis and/or bypass are interesting evaluation cases, as they allow the analysis of the coloration intrinsic to the vocoding technique. Moreover, it can help us understand which techniques are more prone to color the signal through resynthesis. The main parameter of vocal melody is pitch; therefore, resynthesis can be controlled by preserving the melody and, thus, by imposing the original pitch curve on the audio file using various vocoders.

3.3.2 Extreme autotuning (Integer-part) pitch re-synthesis

The use of the vocoder is closely linked to the perception of melody, which is a primary goal in any vocoder technique. The target pitch can be the same as the original in the case of resynthesis, or it can be different when employing pitch correction techniques or when using the vocoder as a musical instrument, as is the case with the musical channel vocoder. This leads to countless possible

ways of modifying the target pitch. Target pitches for the vocoder are akin to singing styles for the voice. Given a melody, be it the style in the case of the voice or the target pitch curve for the vocoder, they would provide a different impression of the same melody.

The most general melodic case for target pitch is the integer part of the signal in semitones, i.e., autotuned pitch in the chromatic scale. This is applicable to both autotuning techniques and musical channel vocoders. Additionally, it is equally applicable to any melody; that is, it does not create dissonances that could introduce additional coloration but rather affects any melodic line similarly by using a chromatic scale.

3.3.3 Soft autotuning

Exploring other scenarios is useful for comparing how a pitch curve modification is executed in different vocoders and for seeing if there is a predominance between melody and vocoder for cases other than resynthesis and extreme autotuning. This type of application is valuable as it represents the primary use in studio vocal correction applications, aiming to enhance tuning without the autotuning transient effect.

3.4 Sound Catalog Generation

The next subsections outline the steps for generating the files used to compare vocoders. It involves pitch tracking stages, choosing a target pitch depending on the two study cases (natural or whole-tone), and pitch warping (vocoders).

3.4.1 Pitch Tracking

The pitch tracking stage has been explored using various Max objects, such as yin , fzero , vb.pitch, and sigmund , as well as the vocoder world software. However, the pitch tracking results obtained with Praat software are significantly superior. Unlike the other tools, Praat does not exhibit discontinuities or octave errors, which are common issues with the other objects used. Additionally, Praat allows testing melodic curves with test sounds, making it easier to adjust values that differ from the reference in the PichTier files or even to eliminate values taken in error. This feature gives preference to pitch data obtained with Praat. Furthermore, these data can be processed in Python to generate pitch curves in the same time scale as the original audio file.

The f_0 curve generated by Praat may not have the same time scale as the audio file. To align the time scales, we perform a resampling by interpolation using Python. Also, Praat extracts f_0 using a Hertz scale, so we converted them to MIDI data. We utilized the *wave* package to obtain the time array of the audio signal and the *path* package to read the *PitchTier* file generated by Praat. This file was reshaped to form two arrays - one for the time and another for f_0. Praat requires a delay of 40ms to obtain f_0 within an interval of time,

so we subtracted this time from the time array. We then obtained three arrays to work with: the time frames of the original audio, the time array of the *PitchTier* file, and the f_o array of the *PitchTier* file. To interpolate between these arrays, we use the *numpy.interpolation* method. This function performs a linear interpolation between the gaps and a constant interpolation on the borders, yielding the same f_o curve as the one obtained using *PitchTier* on Praat but framed on a time scale identical to the original audio. We can then use this information on Max/MSP by writing the f_o tracking over a wave file.

The pitch tracking is carried out in the following way The audio file is taken and processed through Praat, where it is verified that there are no pitch errors. Subsequently, a Praat-type pitch file is generated. This file can be read as text in Python, allowing for a Hz to semitone scale conversion. The Praat file is used in conjunction with the original sound file to generate a WAV output file containing pitch information on the same time scale as the audio file. In Python, it is also possible to calculate the integer part of the pitch curve, thus generating the pitch file in WAV format for the autotuning case. From now on, we will refer to these files as f0-wav, f0-natural, or f0-autotuned files.

As result, we have f_0 files on MIDI and Hz scales, coded on a 0-1 scale. We can use python to have the Hz or MIDI versions of f_o and use it conveniently according to re-synthesis protocol. As we will see the information required for the different vocoders is not always the same. The wav 0-1 decoding will always be the same, but the scaling part will differ. For the MIDI files, we multiply by 128. Furthermore, for Hz files, we use the equivalent Hz scale of midi (0 cents to 12800 cents) with the *expr* object on Max/MSP.

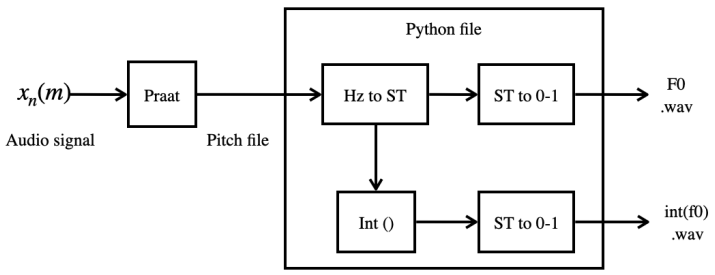


Figure 3.3: Pitch-Tracking procedure

3.4.2 Resynthesis

Resynthesis is done by imposing the pitch curve or the transposition value on the different audio files. Next, we will describe the process that must be carried out with each of the vocoders used to reproduce the resynthesis cases with the natural pitch and resynthesis cases with the autotuned pitch.

3.4.2.1 Resynthesis with Retune

The correction with retune is done through a patch that allows retrieving the pitch curve and imposing it on the audio file in real-time, but with a latency of 1024 samples at a sample rate of 44.1 KHz. To achieve this, two buffers are available where the audio and F0 WAV files are placed and then read in parallel (in MAX) and sent to the retune object. The retune object is designed to receive transposition values as an adjustment factor, but it can also receive the pitch curve through an alternative mode.

3.4.2.2 Resynthesis with World

The vocoder World, similar to Retune, operates with transposition values in Hz. To use it, we receive the expected frequency and then internally instruct it on how to calculate the difference between the current frequency and the frequency we wish to impose. Subsequently, we perform a detailed time-scale re-synthesis to obtain retuned output audio with the given pitch.

3.4.2.3 Resynthesis with Circe

This vocoder is installable on Mac and comes with a visual interface. It allows us to perform constant transpositions by adjusting parameters and dynamic transpositions using a WAV file containing pitch information. The pitch file should have information on a scale from 0 to 1 under the MIDI semitone protocol, preserving microtones. To use Circe, the original audio file is loaded, followed by the pitch file. The transposition is then generated, and a retuned audio file with the desired pitch is saved.

3.4.2.4 Resynthesis with ATA

The re-synthesis with ATA cannot be used with the pitch file obtained from Praat, as ATA does not allow the use of information from an external pitch tracker. Therefore, we are compelled to compare the audio tracks with the files modified by ATA for two configurations that represent our use cases. One, the mildest possible, with $retune - speed = 400$ and $flex - tune = 100$, and another extreme with $retune - speed = 0$ and $flextune = 0$. The gentle configuration will exhibit a sound very similar to the natural sound, while the extreme configuration (autotuning),according to the information from the patent, corresponds to calculating $int(pitch)$ and forcing it into the output.

3.4.2.5 Re-synthesis for soft-ATA-autotuning

The final case study utilizes a pitch-modified curve with ATA, for which a specific smooth value of $retune - speed = 50$ is chosen, and an audio file is generated. Pitch tracking is then performed on this file using Praat, resulting in a .Wav file. This file is taken through each of the previous protocols to generate a replica of the ATA correction with the other vocoders.

3.4.3 Summary of audio files

Here, we present a description of the main files of the sound corpus. The purpose of using these audio files is to check how using different vocoders affects the general perception of the sound. The library is composed of 7 original files summarized in table 3.1. The samples come from the PhD thesis of [Henrich Bernardoni, 2001] and the projects VOQUAL [D'Alessandro, 2003] and CHANTER (Chant numérique avec contrôle temps Réel) [Feugère et al., 2016].

Table 3.1: Catalog of Original Samples

Genre	**Style**	**File (.wav)**	**abbv. name**
Male	Intervals	real3Maleintervals	a
Male	Phrase on Legatto	real19Malevoicelegatto	b
Female	Legatto and Virtuoso	real23Femalelegattovirtuoso	c
Male	Belting	realJF-mem-6-a-male2	d
Female	Belting	realLP-mem-6-a-fem2	e
Female	Pop Style	realms-celinedion	f
Male	Variété Française	realrt-yvesmontand	g

The samples have been treated with the previous explanations to impose three pitch curves: original pitch (f_o), extreme pitch correction ($int(f_o)$), and a soft correction (through ATA and tracked with praat). The systems used, and cases are summarized in Table 3.2.

Table 3.2: Systems Used and Correction Cases

	ATA			**World**			**Circe**			**Retune**			
Sample	Original	Soft Corr.	Ext. Corr.	Original	Soft Corr.	Ext. Corr.	Original	Soft Corr.	Ext. Corr.	Original	Soft Corr.	Ext. Corr.	Original File
a	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
b	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
c	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
e	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
f	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
g	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The samples have been created to be used in a psycho-acoustical comparison we conducted; such a test will be explained in the next section. The

samples have been cut into 16 files, which duration and content (staircase, vibratos, free-path) are summarized in table 3.3. The following section will explain how these samples are integrated into the psycho-acoustic test.

Table 3.3: Samples for the Psychoacoustic Test of the Vocoder (abbreviated as v)

	Content					**Content**			
Sample	Staircase	Vibratos	Freepath	Duration [ms]	Sample	Staircase	Vibratos	Freepath	Duration [ms]
va part 1		✓		2000	ve part 1	✓	✓	✓	3000
va part 2		✓	✓	2000	ve part 3	✓	✓	✓	5124
vb part 1	✓	✓		2937	ve part 4	✓	✓	✓	4500
vb part 2	✓	✓		4500	vf part 2	✓	✓	✓	3625
vc part 1	✓	✓	✓	3171	vf part 3	✓	✓	✓	3000
vc part 2	✓	✓		2500	vg part 4	✓	✓	✓	2500
vc part 3	✓	✓		3500	vg part 1	✓	✓		4250
vd part 2	✓		✓	4750	vg part 2	✓	✓	✓	8250

3.5 Subjective Evaluation of Vocoder for pitch tuning

The primary psychoacoustic attributes, such as intensity, pitch, duration, and timbre, play an essential role in auditory organization. Pitch orders sounds from the lowest to the highest, contributing to the definition of melody in conjunction with rhythm. Vocal transformation through the vocoder predominantly involves melody. Depending on the use case, it is necessary to decide between introducing a change or preserving the existing melodic line, leading to perceptible alterations in the signal. Our objective is to observe how the use of different vocoders influences the variation of such sonorous changes according to the technique, as well as to analyze the preponderance or absence of it when imposing extreme melodic alterations.

The comparison of vocoders through sound evaluation should be carried out using a subjective psychoacoustic test, also known as an affective test. The coloration produced by each vocoder can have varying effects on the sound, and due to this variability, it does not make it a parameter (perceptible attribute) easily quantifiable. This characteristic leads to our test being a subjective assessment rather than a perceptual testing. However, since it involves com-

paring multiple vocoders simultaneously, a multi-stimulus technique must be employed. Typically, such comparative evaluations are conducted using a integrative tests like MUSHRA [ITU-R-BS.1534-3, 2015] (Multiple Stimuli with Hidden Reference and Anchor); nevertheless, this test is designed to compare (in terms of quality or preference) a single original sound with a reference and its degraded versions. On the one hand, using a reference without any sound treatments outright prevents the evaluation of resynthesis when the imposed pitch differs from the original pitch. This type of test, as its standard version, would only allow the comparison of a vocoder with respect to the reference sound, meaning that multiple pairs of vocoders cannot be compared. This limitation, present in integrative tests like MUSHRA, is mitigated in discriminative methods [Zacharov et al., 2018].

Discriminative methods such as pairwise comparison, the ABX method, n-forced comparison, allow, with certain levels of complexity and precision, discrimination solely between pairs of sounds. In other words, we can compare vocoded sound with vocoded sound, even with an imposed pitch different from the original. However, using this type of test involves a pairwise testing approach, making the differences between vocoders more challenging to identify and requiring extended testing times for a small number of sound examples [Zacharov et al., 2018]. Considering the number of vocoder and cases of tuning, this approach is not practical.

One of the additional challenges in test design is linked to the type of question; it is crucial that the formulation be as straightforward as possible. In this regard, the discriminative and multiple sense (MUSHRA) of the test must be considered, which is compatible with the DFC (Difference from Control) test, in which the MUSHRA reference is called the "control", which is not necessarily a sound without treatments but rather a reference sound. We will continue to refer to our test as "subjective" instead of incorrectly using the term MUSHRA [ITU-R-BS.1534-3, 2015, ITU-R-BS.1116-3, 2015], as it is not a proper MUSHRA test. The discriminative question for our subjective DFC test would be: "Please evaluate the degree of similarity between the reference and the different conditions on the scale", as shown in Figure 3.4. The choice of reference and comparables (vocoded sounds) will influence the objective of each test, as detailed below in each of the stages that comprise our test. The package used for the DFC test was webMUSHRA as referenced in [Schoeffler et al., 2018].

3.6 Tasks

There are 4 different tasks, each containing 16 questions. Throughout all 64 questions in the test, the same query is consistently posed. Participants are asked to provide a score indicating the level of similarity between the comparables and the reference. The questions are fully randomized, and each task serves a distinct purpose. For simplicity, the same question is employed, but the hypotheses and potential conclusions differ for each task, as outlined

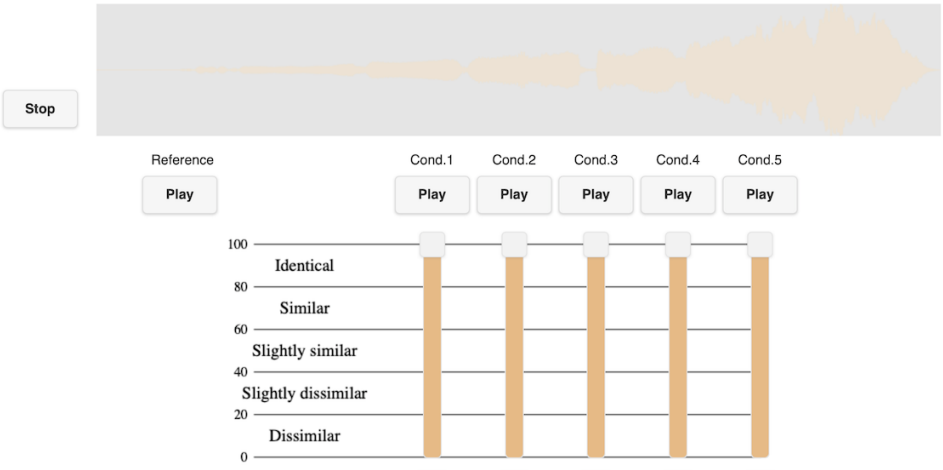


Figure 3.4: Example training trial question. The reference is on the left, and five comparable sounds are presented in random order, including four vocoder-processed versions and one hidden reference. The listener must classify the level of similarity between the sounds.

below.

3.6.1 Task A: Original pitch re-synthesis with each vocoder compared to the Original sound

In the first task (TA), a comparison is made between the original sound and the resynthesized sounds by forcing the pitch of the original sound file using the four available vocoders. In other words, it involves resynthesis with dynamic pitch, following the protocols defined in the preceding section. The question about similarity provides information about the transparency of the vocoding process with each device. That is, it determines whether each technique is transparent or if, on the contrary, we can perceive an intrinsic coloration (timbre) due to vocoding, given that the melody remains the same. Additionally, we can characterize this difference to classify the vocoders from the most transparent to the least transparent for resynthesis with dynamic pitch.

3.6.2 Task B: Extreme autotuning with each vocoder compared to original sound

In the second task (TB), the reference remains the natural, unmodified sound, but the comparables are sounds resynthesized with automatically adjusted pitch (autotuned pitch). The question of similarity in this case revolves around the preservation of vocal quality after a drastic pitch modification with each vocoder. In other words, it will be investigated whether the vocoders are capable and to what extent they can preserve the timbre. Additionally, it will be examined whether there is consistency for all values within the same vocoder or, conversely, if the response is arbitrary.

3.6.3 Task C: Extreme autotuning with each vocoder compared to Extreme autotuning with ATA

The third task (TC) also relates to extreme autotuning; in this case, the reference is the sound with extreme autotuning using ATA, and the comparables are the resynthesized sounds with extreme autotuning curves using the vocoders World, Retune, and Circe. The objective of the comparison in this task is to determine if we can distinguish between the different vocoders under this melody. The hypotheses will then confirm the predominance or lack thereof of the melody over the vocal timbre generated by the vocoder.

3.6.4 Task D: Soft autotuning with each vocoder compared to Soft autotuning with ATA

In the fourth task (TD), the question of melodic predominance versus vocoder is addressed with a gentle vocal correction using ATA. An autotuned file with a smooth value of $retune - speed = 50$ is taken, and that pitch curve is imposed on the original audio file using the other vocoders. In this task, the question has the same meaning as in the third task.

3.7 Test preparation

The design of the subjective test is carried out using the MUSHRA interface, with elements previously defined for the hybrid ranking and discrimination test, including:

- Objectives: Distributed in four previously described tasks, each with specific objectives
- Stimuli: Audio files vocoded generated by the 4 vocoders (systems to tets) for task.
- Interface and Data Collection: Utilization of the MUSHRA interface and data collection in CSV format provided by MUSHRA tools.

3.8 Audio Support

The audio support for the comparison of the vocoders consists of 16 samples. These samples are edited with vocoders and autotuning cases according to Table 3.4 for tasks A,B,C and D.

3.8.1 Subject Panel

The scope of subjects involved in psychoacoustic analysis typically undergoes limitations. Within the framework of our experiment, a cohort comprising

Table 3.4: Audio Support for Comparison of Vocoders Test

Comparable		**ATA**			**World**			**Circe**			**Retune**		
Pitch / File	Original File	Original Pitch	Extreme ATA	Soft ATA	Original Pitch	Extreme ATA	Soft ATA	Original Pitch	Extreme ATA	Soft ATA	Original Pitch	Extreme ATA	Soft ATA
va part 1	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
va part 2	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vb part 1	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vb part 2	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vc part 1	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vc part 2	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vc part 3	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vd part 2	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
ve part 1	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
ve part 3	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
ve part 4	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vf part 2	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vf part 3	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vf part 4	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vg part 1	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D
vg part 2	A	A	B,C	D	A	B,C	D	A	B,C	D	A	B,C	D

21 subjects has been employed, meticulously divided between individuals possessing musical proficiency and those lacking such aptitude. As substantiated by prior studies (cf. [ITU-R-BS.1534-3, 2015]; [ITU-R-BS.1116-3, 2015]), a sample size of 20 subjects suffices for the particular evaluative paradigm undertaken in this study.

3.8.2 Test Contents

In each task, four vocoders (audio systems as delineated in [ITU-R-BS.1534-3, 2015]) are juxtaposed for comparison. According to the guidelines stipulated in reference [ITU-R-BS.1534-3, 2015], the requisite number of samples should exceed 1.5 times the count of systems under evaluation. Consequently, considering six samples as prescribed, which substantially meets this criterion, we have opted to incorporate 16 samples per task.

The assessment protocol encompasses four distinct tasks, with each task comprising 16 queries. Within each query, participants are tasked with discerning the degree of similarity between the reference stimulus and the comparable stimuli (vocoders). Consequently, each vocoder is subjected to evaluation across 16 discrete sound instances per participant, thereby resulting in a cumulative tally of 320 scores per vocoder. This methodological approach,

characterized by a diverse array of examples, serves to engender test variance while mitigating the likelihood of participant fatigue. Furthermore, the randomization of the four tasks is enacted to ensure a diversified presentation format.

3.8.3 Data Treatment

The data is gathered in CSV (Comma-Separated Values) format, encapsulated within a file that delineates identifiers corresponding to each test subject, descriptors denoting the type of comparables, subjective assessments assigned by the participants to each comparable, classifications pertaining to the type of auditory stimuli, and comprehensive timestamps detailing the duration of each stage throughout the test process. Subsequent statistical scrutiny of the amassed dataset is conducted employing the R programming language, in accordance with the statistical power analysis guidelines elucidated in [ITU-R-BS.1534-3, 2015] and [Rogers, 2017].

Incorrectly rejecting a true null hypothesis is called a Type I error [Rogers, 2017]. The proportion of decisions in which a Type I error is made is called the significance level and denoted by α. Within the MUSHRA [ITU-R-BS.1534-3, 2015] standart it equals 0.05, according to [Rogers, 2017] such value implies that "when the null hypothesis is true, then we correctly decide in favor of the null hypothesis 19 out of 20 times, and incorrectly reject the null hypothesis 1 out of 20 times."

A Type II error occurs when a false null hypothesis is not rejected [Rogers, 2017]. The proportion of decisions in which a Type II error is made is controlled at a predetermined level β (also called error rate). Conventionaly $\beta = 0.20$, such value means that when the null hypothesis is incorrect, it is retained incorrectly in one out of five tests. The statistical power of the test is defined as $1-\beta$, which is the probability of correctly rejecting the null hypothesis when it is false and the alternative hypothesis is true. For $\beta = 0.20$, the statistical power is 80%.

3.8.4 Room and sound

The experimental procedures are conducted within the confines of recording studio room 519, at the Institut Jean le Rond d'Alembert. An operational framework is established utilizing a laptop operating on macOS Big Sur (version 11.7), serving as the medium for test administration. Access to the test interface is facilitated through a dedicated website hosted in the Institut Jean le Rond d'Alembert servers. Data acquisition transpires upon the conclusion of each test segment, with each segment typically extending over a duration of approximately 20 minutes. Participants are furnished with *Sennheiser HD 205* over-ear headphones to facilitate auditory perception during the assessment process. Before initiating each test session, a examination of volume levels and equipment functionality is conducted to ensure optimal testing con-

ditions.

3.8.5 Planning

The test consists of four parts, each lasting 20 minutes, and is organized in a single session. Participants are summoned to the Jean Le Rond d'Alembert institute and are grouped into musicians and non-musicians. All tests are conducted over a one month, with planning done directly with participants. Participants were required to be well-rested before the test, and the test was compensated with a 40-euro Amazon gift card. Thomas Lucas, LAM engineer, organized planning and payment. He also collaborated in cutting some of the audio samples and verifying the HTML code.

3.9 Test procedure summary

On the day of the test, participants are summoned to the designated room, provided with the agreed-upon materials. The session begun with a presentation summarizing the test, as it follows :

- *The term vocoder technically refers to a software device designed for transparent voice encoding, transmission, and natural transformation, and which can be used in musical applications, especially for pitch auto-tuning.*
- *The purpose of this work is to establish a benchmark that facilitates the musical discussion about the vocoder, seeking to understand the perceptual impact of the vocoder and melodic modification and determine whether we can truly speak of a "vocal quality" of the vocoder. In this test, we present an audio repository that supports the comparison between different vocoders in cases of resynthesis, subtle and extreme vocal tuning.*
- *The test is divided into four segments, each lasting approximately 20 minutes. For every segment, you will be asked to assess the degree of similarity between a reference and various conditions. Following each segment, you will need to complete an information sheet, ensuring that you include the same id at the end of each section.*

Subsequently, participants were asked to read and sign a document allowing the retrieval of their response results and certain anonymous data (such as age and musical knowledge). This form was used for both: the vocoder test and the pitch correction methods test (next chapter of this thesis). The document also outlined the option to withdraw from the test at any time if they wish to do so, ans it is presented in appendix C.

Once the explanation is concluded, participants are provided with a participant id that they need to enter at the end of each test part to save their

data. Before initiating the test, participants are asked to set a comfortable volume with a sound sample from the catalog played through the website. An example question is posed to familiarize participants with the test format and ensure their understanding, as Figure 3.4. To end, participants enter some information about themselves in anonymity format (age, musical experience, etc) as shown also in appendix C.

After this, participants are allowed to proceed with the test. Once the 16 questions of the first part are completed (approximately 20 minutes), the participant sees a page where they enter their participant ID, then data is automatically saved on the server. They then move on to the second part of the test, which also lasts approximately 20 minutes. Upon completion, participants are asked how they have experienced the test so far, and if desired, they can take a longer break of up to 30 minutes. Subsequently, the participant continues with the third pard, a 5-minute break (if desired), and the fourth part. Finally, an interview is conducted to gather more details about the test.

3.9.1 Summary of tasks for vocoders comparison

The table 3.5 presents a detailed summary of the various aspects addressed in the psychoacoustic test, including the description of the references and comparables used, as well as the specific objectives of each task. Through this arrangement, the aim is to provide a comprehensive understanding of the controlled variables and parameters assessed at each stage of the study. In addition to the primary objectives of each task, the table also highlights additional conclusions that can be drawn from the collected data, further enriching the interpretation and utility of the results obtained. For the statistical analyses, the ANOVA and post-hoc Tukey HSD (honestly significant difference) protocols are used, with the null hypothesis being equality and the alternative hypothesis being the difference.

3.10 Results

The results of the subjective psychoacoustic test we conducted are based on the analysis of the CSV files generated by the WebMushra package. These data have been collected to allow the retrieval of all available data, including the tasks performed, the subjects (anonymous but identifiable by code), the audio samples under study, the comparables, and the scores given to each comparable sample. Once this data was obtained (in random), it was organized by task, corresponding to the previously defined organization (A,B,C,D). Subsequently, the means and standard deviations of the scores given to each comparable sample for each task were calculated. This calculation was performed considering all subjects and audio samples used in the study. The results obtained are presented clearly and concisely elaborating tables and graphs detailed in this section for each task. Furthermore, the possibility of conducting intra-task

Task	Reference	Melody(F0)	Comparable	Melody(F0)
A	Natural	$f_o = f_{original}$	ATA Circe World Retune	$f_{natural}$
	Comparing identity/timbre of the original with the re-synthesized voice			
B	Natural	f_o	ATA Circe World Retune	$int(f_o)$
	Comparing identity/timbre of the original with the autotuned (with several vocoders) voice. Aiming to determine whether the original vocal timbre persists after extreme autotuning.			
C	ATA	$f_{ata,extreme} = int(f_o)$	Circe World Retune	$int(f_o)$
	Comparing extreme autotuning between ATA and other vocoders. Similarity suggests that the melody dominates perceptually. Dissimilarity suggests the vocoding technique perceptually prevails over the imposed melody. *Additional: assessing similarity between World and ATA for later use as a replica for pitch correction methods subjective evaluation*			
D	ATA	$f_{ata-soft}$	Circe World Retune	$f_{ata-soft,praat}$
	Comparing soft autotuning between ATA and other vocoders. Similarity suggests that the melody dominates perceptually. Dissimilarity suggests the vocoding technique perceptually prevails over the imposed melody. *Additional: assessing similarity between World and ATA for later use as a replica for pitch correction methods subjective evaluation*			

Table 3.5: Summary of tasks for vocoders comparisson

and inter-task analyses will be highlighted for a deeper understanding of the collected data.

3.10.1 Task A: Original pitch resynthesis with each vocoder compared to Original sound

For Task A, we compared re-synthesis using various vocoders to the original sound (reference). Original pitch re-synthesis involves imposing the original pitch onto the audio. In the case of ATA, we selected the preset to make it operate transparently. Our aim with this test is to understand how each vocoder technique affects vocal timbre, given that we expect the same melody in the re-synthesized audio. The predominant difference would primarily arise from the coloration introduced by each vocoder.

The obtained data were processed in R. We conducted an analysis of variance (ANOVA), calculated the means for the groups (i.e., the comparables), and performed a post-hoc analysis Tukey's HSD test. The ANOVA was performed using the aov() function, and the results are shown in Table 3.6. Before explaining the significance of these results, we first define the number of groups (the comparables). This task has five groups: one natural sound and four vocoded sounds. Secondly, ANOVA is conducted under a null hypothesis (denoted as H0), which states that all means of the different groups are statistically equal. An ANOVA table (like Table 3.6) distinguishes between two rows: one for the groups, which refers to the variations between the comparable groups, and the other, Residuals, which relates to the variations within all observations (i.e., within the groups).

Table 3.6: ANOVA for Task A - before excluding subjects deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	4	610861	152715	576.6	$< 2e-16$
Residuals	1675	443644	265		

The first variable observed is Df, which stands for Degrees of Freedom. For the groups, it equals the number of groups minus one (the reference), so it equals 4. For residuals (within the groups), Df equals the total number of observations minus the number of groups, which in this case is 1675.

Next, we have the Sum of Squares (Sum Sq). For Groups, it indicates the variance between groups and is equal to $\sum n_i(\bar{Y}_i - \bar{Y})^2$, the sum of the variances of each group i with respect to the overall mean, where n_i is the number of observations in group i, $\bar{Y}_i$ is the mean of group i, and $\bar{Y}$ is the overall mean. Sum Sq for Residuals indicates the variance within each group and is equal to $\sum\sum(Y_{ij} - \bar{Y}_i)^2$, the sum of the variances with respect to the mean of each group, where Y_{ij} is the observation j in group i and $\bar{Y}_i$ is the mean of group i.

The Mean Square (Mean Sq) is calculated by dividing the sum of squares by the degrees of freedom, $SumSq/Df$. The F-value is calculated as the ratio

of the Groups Mean Sq divided by the Residuals Mean Sq. A small p-value, less than 0.05, indicates that the observed differences between the groups of vocoders are highly statistically significant, meaning the null hypothesis is rejected. The limitation of ANOVA is that it does not indicate which groups are different.

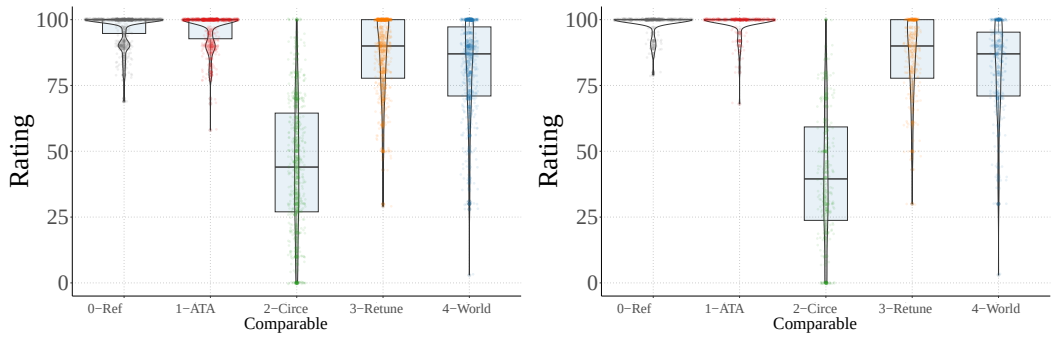


Figure 3.5: Results for Task A - before (left) and after (right) excluding subjects deemed unsuitable. Labels 1 to 5 indicate: Reference (Original Sound) and ATA, Circe, Retune and World Resynthesis

Table 3.7: Means and Tukey HSD post-hoc resume analysis for Task A: Original pitch resynthesis with each vocoder compared to the original sound.

Comparable	Mean	SD	Classification	Diff. to Ref.
0-Ref (original)	96,6	6,1	Identical	
1-ATA	96,3	6,6	Identical	-0,3 *
2-Circe	44,7	25,6	Slightly Similar	-51,9
3-Retune	85,9	15,2	Similar	-10,7
4-World	81,0	18,8	Similar	-15,7

p-value < 0.001 by ANOVA and Tukey HSD post-hoc
Except for (*) p-value $= 0.9989$

We now proceed to the analysis of Table 3.6. The between-group variance is high (for the row groups, Sum Sq equals 610861 and Mean Sq equals 152715), indicating that the differences between the groups of vocoders are significant. On the other hand, the within-group variance (for the row residuals, Mean Sq equals 265) also reflects that the average variability within the groups is high. The F-value (576.6) shows that the variance between the groups is much greater than the variance within the groups. Finally, the extremely low p-value ($< 2e - 16$) indicates that the observed differences between the groups of vocoders are highly statistically significant.

We did graphic with the means, as shown in Figure 3.5. The means correspond to the mean values obtained for each vocoder, displayed in Table 3.7. A note with the Tukey's HSD test p-values is provided at the bottom of the table. Tukey's HSD (Honestly Significant Difference) test is a post-hoc analysis used

to find means that are significantly different from each other. This analysis is also performed in R using the function TukeyHSD() and summarized in Table 3.8.

Table 3.8: Tukey HSD Analysis for Task A - full panel

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
ATA-Ref	−0.3	−3.8	3.1	0.9989
Circe-Ref	−51.9	−55.3	−48.5	<0.0001
Retune-Ref	−10.7	−14.1	−7.3	<0.0001
World-Ref	−15.7	−19.1	−12.2	<0.0001
Circe-ATA	−51.6	−55.0	−48.1	<0.0001
Retune-ATA	−10.4	−13.8	−6.9	<0.0001
World-ATA	−15.3	−18.8	−11.9	<0.0001
Retune-Circe	41.2	37.8	44.6	<0.0001
World-Circe	36.2	32.8	39.7	<0.0001
World-Retune	−5.0	−8.4	−1.6	0.0007

The table 3.8 shows the difference between group means (comparables), the lower and upper limits of that difference within a 95% confidence level, and the adjusted p-value for the pairwise comparison. If the p-value is less than the significance level (0.05), the difference between the means of the two groups (in the pair) is considered statistically significant. In our case, Tukey's HSD indicates a statistically significant difference between all pairs except the pair ATA-Ref (p-value = 0.9989), meaning there is no statistically significant difference found for this pair.

The obtained data can be graphically represented by histograms per trial and per subject, as shown in Figure 3.6. The complete graphics are shown in Appendix C.2. There are no indications of trials or subjects with results divergent from the other trials and subjects. However, considering the guidelines of MUSHRA [Schoeffler et al., 2018], two subjects have been removed from the panel for a more precise analysis. The ANOVA tables and Tukey's HSD test for the panel, excluding the outlier subjects, are included in Appendix C.2. This appendix contains all the histograms per subject (Figures C.1 and C.2) and per trial (Figures C.3 and C.4) and Tukey's HSD test (Table C.3) whose summary appears as note in Table 3.7.

A graphical representation of the results is presented in Figure 3.5, showing means, standard deviations, and data distributions before and after excluding outlier subjects (two subjects). The statistical analysis excluding the two subjects is shown in Appendix C.2: ANOVA in Table C.4 and Tukey's HSD test in Table C.5. Additionally, analyses for non-musicians (ANOVA in Table C.6, Tukey's HSD test in Table C.7) and musicians (ANOVA in Table C.8, Tukey's HSD test in Table C.9) were conducted. The corresponding graphics showing means, standard deviations, and data distributions are shown in Figure 3.7.

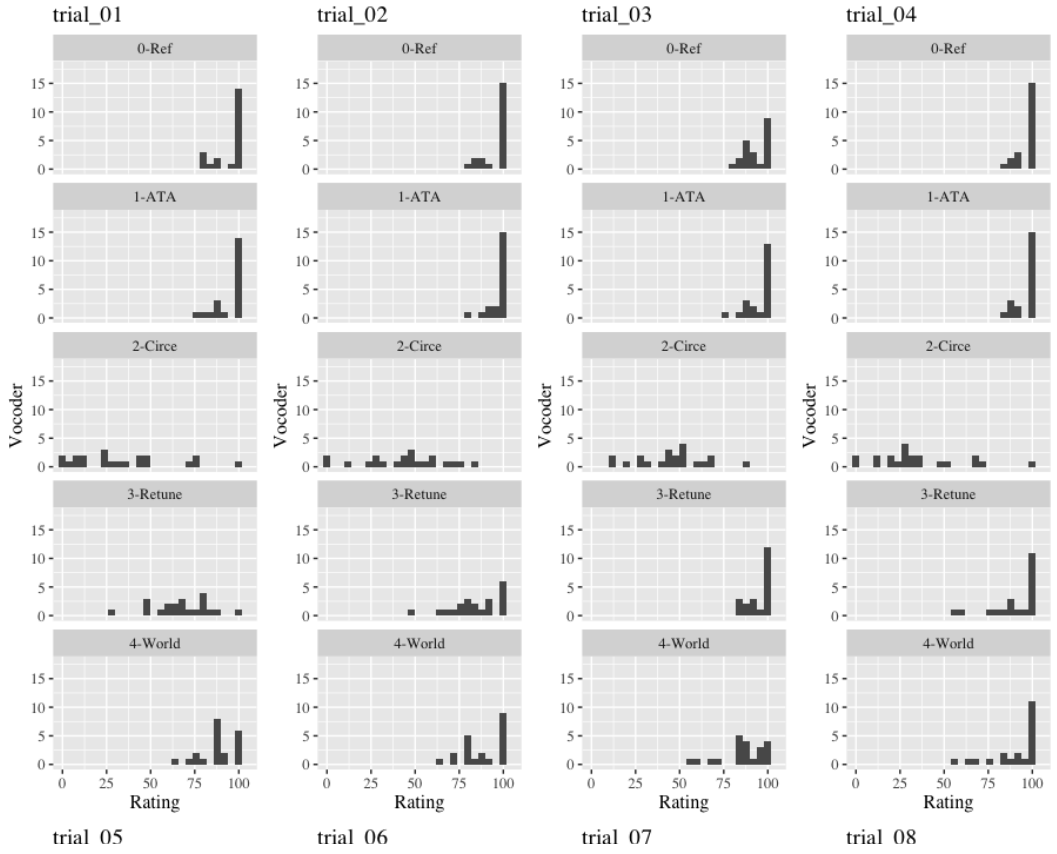


Figure 3.6: Histograms per trial for Task A - only trials 1 to 4, as illustrative example, full histograms can be found in Appendix C.2

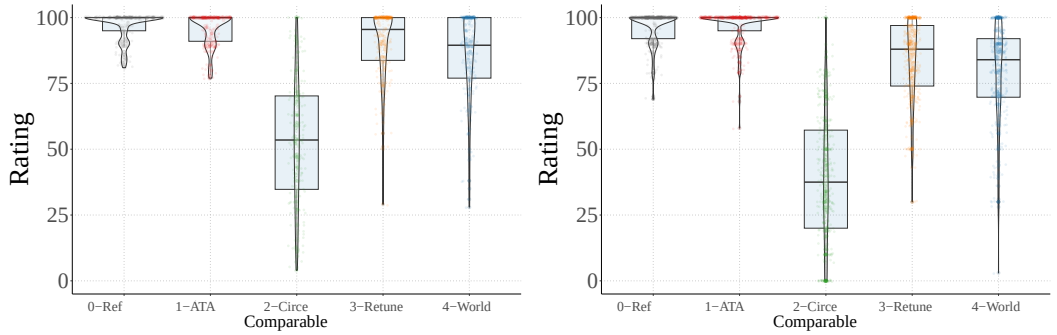


Figure 3.7: Results for Task A - Non-Musicians (left) and Musicians (right). Labels 1 to 5 indicate: Reference (Original Sound) and ATA, Circe, Retune and World Resynthesis

The primary observation is the panel's adeptness in discerning statistical discrepancies, as evidenced by ANOVA (see Table 3.6). Regarding means, the panel effectively identifies the reference, which yields the highest score. Nevertheless, ATA emerges as statistically the most transparent option for resynthesis. As per the Tukey Honest Significant Difference (HSD) analysis (see Table 3.7), ATA demonstrates statistical indistinguishability from the

original sound (reference). Noteworthy is the fact that ATA, functioning within the temporal domain, maintains spectral content integrity, thus precluding coloration introduction. This assertion is substantiated through statistical analyses conducted across musician and non-musician cohorts, as well as upon exclusion of unsuitable subjects. It is pertinent to highlight that not only are mean values retained but also the shape of the data distribution, ensuring the preservation of statistical similarities and differences across these groups.

The second observation pertains to the discernible coloration exhibited by alternative techniques such as Circe, Retune, and World, as demonstrated by Tukey's HSD analysis and distributions shapes. This phenomenon is likely attributed to modifications in spectral content inherent to these techniques. Notably, Retune and World mean values manifest similar deviations from the original sound, then we can verified it exist a statistical similarity for the pair Retune-World because p value is equal to 0.0007. The Circe vocoder, however, introduces distinctive coloration, positioning it as notably dissimilar in mean values and statistically from the original sound. Despite this discrepancy not being previously documented by its authors, it is plausible that the Circe vocoder's testing solely for constant transposition, without consideration of dynamic transposition— the mechanism employed herein to impose the original pitch—could underlie this observation.

Statistically, upon the removal of unsuitable subjects, no significant differences are observed in terms of means and distributions. Upon division of the panel between musicians (or professionals in the field) and non-musicians, it is evident that the shapes and relative positions are generally preserved, except for the retune vocoder. Musicians generally tend to rate differences lower, thereby resulting in overall lower values across all comparables. The data distribution shape remains similar for both musicians and non-musicians groups. Notably, as previously mentioned, the Retune vocoder exhibits a difference, being rated higher by the non-musician group, who may be less adept at discerning differences compared to the musician group.

The specific shape of the data in indistinguishable comparables, such as ATA and the natural sound, remains consistent. This shape is consistently observed when the subject perceives a comparable as highly similar to the reference across all tasks. The distributions of the reference and ATA show some lobes, which could be attributed to subjects rating them in a certain way, or to one or several trials inducing ratings in the ranges where these lobes occur. However, upon scrutinizing the histograms of trials and subjects in detail, no such issues are observed. It may be more closely associated with the type of scale utilized. Notably, the scale employed is partially discretized through the use of a MUSHRA standard double scale.

3.10.2 Task B: Extreme autotuning with each vocoder compared to original sound

For Task B, we have compared autotuning realized by each vocoder to the original sound (reference). Auto-tuning is achieved through re-synthesis by imposing a pitch equal to the integer part of the original pitch (in the semitone scale). Our aim with this test is to understand how each vocoder technique affects vocal timbre when autotuning is performed. The difference arises from the coloration introduced by each vocoder and autotuning. The results corresponding to the mean values obtained for each vocoder are displayed in Table 3.10, where the identified differences have also been classified. The obtained data can be graphically represented, as shown in Figure 3.8 and analyzed by ANOVA (Table 3.9) and Tukey's HSD test (Summarized in Table C.10). The statistical support can be found in Appendix C.3 and includes: histograms per subjects (Figures C.5 and C.6) and per trials (Figures C.7 and C.8).

Table 3.9: ANOVA for Task B - before excluding subjects deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	4	584859	146215	346.7	$< 2e-16$
Residuals	1465	617916	422		

Table 3.10: Tukey HSD post-hoc analysis for Task B:
Extreme autotuning with each vocoder compared to the original sound.

Comparable	Mean	SD	Classification	Diff. to Ref.	Diff. to TA
0-Ref (original)	97,0	6,1	Identical		0,3
1-ATA	55,3	23,2	Slightly Similar	-41,7	
2-Circe	38,1	23,6	Dissimilar	-58,8	-17,2
3-Retune	72,7	21,1	Similar	-24,3	-16,8
4-World	54,9	22,9	Slightly Similar	-42,0	-0,4*

p-value < 0.001 by ANOVA and Tukey HSD post-hoc
Except for (*) ATA-World pair p-value $= 0.9997$

Another graphical representation of the results can be found in Figure 3.8 showcasing means, standard deviations and distributions of data before (left) and after(right) excluding two subjects considered unsuitable according to MUSHRA [Schoeffler et al., 2018]. Statistical suppor after removing unsuitable subjects is in Appendix C.3. It includes: ANOVA (Table C.11), Tukey's HSD Test (Table C.12). Also we divided the panel in two groups: musicians and non-musicians giving the results showcased in 3.9, the statistical supports can be found in Appendix C.3: for non-musicians in tables C.13 (ANOVA) and C.14(Tukey's HSD test) and for musicians in tables C.15(ANOVA) and C.16) (Tukey's HSD test).

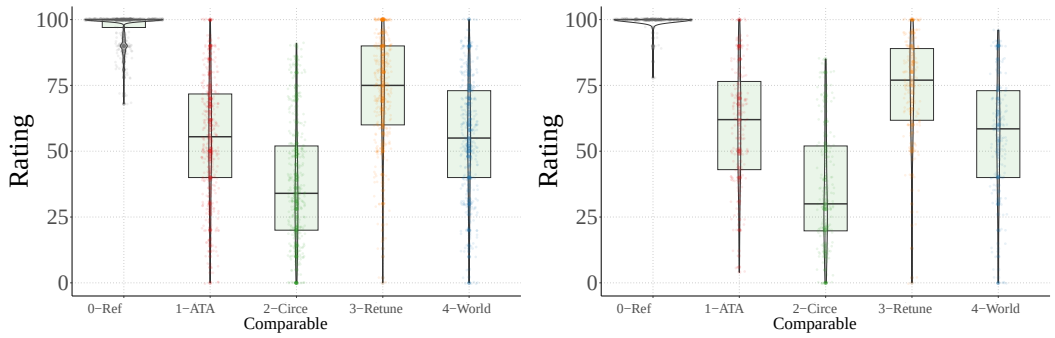


Figure 3.8: Results for Task B - before (left) and after (right) excluding subjects deemed unsuitable. Labels 1 to 5 indicate: Reference (Original Sound) and ATA, Circe, Retune and World extreme-tuning

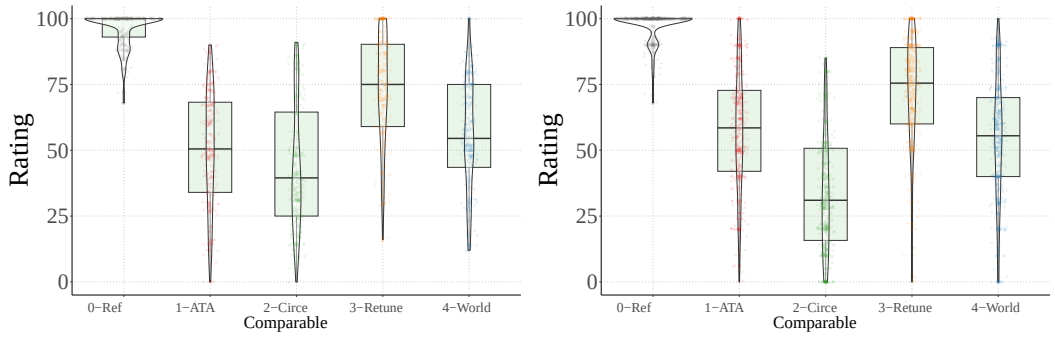


Figure 3.9: Results for Task B - Non-Musicians (left) and Musicians (right). Labels 1 to 5 indicate: Reference (Original Sound) and ATA, Circe, Retune and World extreme-autotuning

Thanks to ANOVA, we can say that there are statistically significant differences between the groups of comparables (p-value) and that the differences between groups are more important than those within groups (F-value). Regarding the mean values, the first observation is that the original sound (the reference) has a mean value similar to the mean value obtained in Task A (despite the different contexts of Tasks A and B). On the other hand, the data distribution of the reference compared to itself, as we will see, is the same in all tasks regardless of whether it is a natural or autotuned sound. This distribution shape may be due to a perceptual phenomenon of similarity and the type of MUSHRA scale used. After an analysis of the mean values and the Tukey HSD analysis, all the pairs of p-values are less than 0.001 except for the ATA-World pair; this means there is no statistically significant difference between ATA and World when both compared to the original sound. And that for all other pairs, there is a statistically significant difference.

Additionally, cuando se comparan los task A y B. tal como se muestra en la figura 3.10, it is observed that for the reference, the data dispersion is lower, which is because there is a clear difference between the natural sound and the self-tuned sound. Consequently, it is more likely that subjects will classify the

natural sound closer to 100 among the comparable sounds. Note the remaining comparables (Circe, Retune, and Word) not only exhibit statistical differences but also display more pronounced disparities in mean ratings compared to Task A, as depicted in Figure 3.10. This phenomenon concerning coloration can be interpreted as indicative of greater coloration and may stem from two primary reasons. Firstly, there is the melody imposed by the autotuning process, which adjusts all notes to whole semitones. Secondly, there is the vocoder technique and its interpretation and management of pitch data for resynthesis. The most significant degradation in terms of means is observed with Circe, transitioning from being considered slightly similar to being clearly perceived as different. Retune, however, demonstrates minor pitch alterations compared to the other vocoders, potentially due to a slower interpretation of the pitch curve.

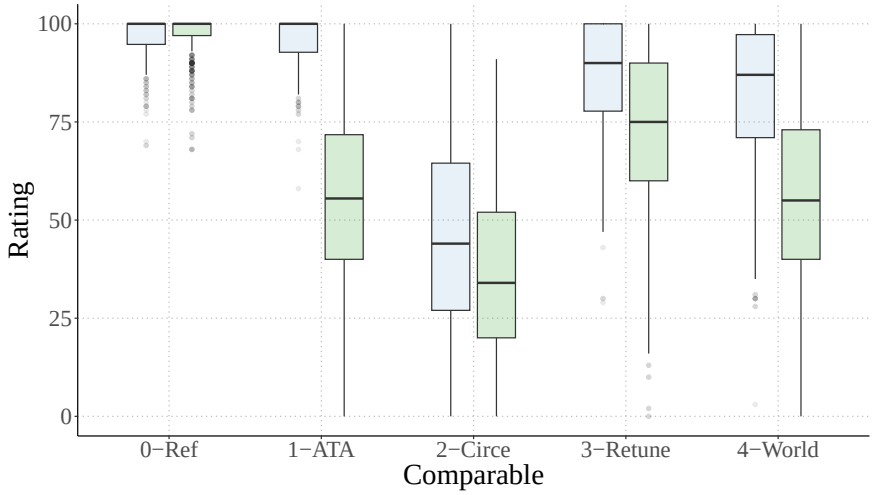


Figure 3.10: Task A and B mean values. In Blue: Task A (Original pitch resynthesis with each vocoder compared to Original sound). In green: Task B (Extreme autotuning with each vocoder compared to original sound)

In terms of means, ATA and World deviate by a similar amount from the original sound, as observed in Figures 3.5 3.8. When backed by statistical support using the post-hoc analysis of Tukey HSD in the summary table 3.10 and in annex C.10, it is found that these two systems for autotuning are statistically indistinguishable. This also provides insights into the relevance of autotuned melody in the perception of these two systems.

If we observe the degradation due to vocoding in Task A, and the degradation due to autotuning in Task B with respect to vocoding for the ATA and Word vocoders, we find that the degradation due to autotuning is greater (figure 3.10). Of course, this degradation is not the same with all vocoders as each system operates differently, but it undoubtedly shows a significant contribution perceptually speaking.

According to the post-hoc statistical analysis, all other possible combinations are significantly different. Additionally, no relevant differences are observed in the distributions after excluding unfit subjects. When the panel is

divided between musicians (or professionals in the field) and non-musicians, it can be observed that in general, the forms and relative positions are preserved. As additional annotations, it can be mentioned that musicians rate CIRCE slightly lower, possibly because they rate the difference lower than non-musicians do. Also, musicians rate ATA slightly higher, probably because they are more familiar with its sound. Finally, it is worth noting that the particular shape of the data distribution obtained for the reference is similar to that obtained in Task A; this shape may be due to the scale and a perceptual effect when the comparator is very similar to the reference stimulus, which was not previously mentioned and is systematic in the reference in all tasks.

3.10.3 Task C: Extreme autotuning with each vocoder compared to Extreme autotuning with ATA

For Task C, we compared how each vocoder performs autotuning in relation to the reference autotuning done by ATA. Autotuning is achieved by resynthesizing the original pitch, imposing a pitch equal to the integer part in a semitone scale. In ATA, this process is accomplished by configuring $retunespeed = 0$. Our objective is to understand how each vocoder technique uniquely affects the coloration of autotuning. The difference in coloration arises from both the intrinsic characteristics of each vocoder and the autotuning pitch curve.

We want to understand how people perceive extreme autotuning, i.e., if they perceptually prioritize the extreme melody (the pitch curve of extreme autotuning) or the vocoder used. Suppose no statistically significant difference is found (p-value ¿ 0.95) as the melody is the only commonality between the groups; we can consider that melody perceptually dominates extreme autotuning, which would happen. If a statistically significant difference is found (p-value ¡ 0.05), as the vocoder is the only difference between the groups, we can consider the vocoder perceptually dominates extreme autotuning. The current reference method for autotuning is ATA, which is why we use it as a reference.

The results, presented in Table 3.14, display the mean values for each vocoder, along with the identified differences. Furthermore, the data can be visually represented, as shown in Figure 3.11, and the statistical support can be found in Appendix C.4.

Table 3.11: ANOVA for Task C - before excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	351753	117251	334.5	$< 2e-16$
Residuals	1172	410807	351		

The histograms have been calculated per subjects (Figures C.9 and C.10) and per trials (Figures C.11 and C.12). The statistical analysis includes ANOVA (Table 3.11) and Tukey's HSD test (Table C.17).

Table 3.12: Tukey HSD post-hoc analysis for Task C:
Extreme autotuning with each vocoder compared to ATA extreme autotuning

Comparable	Mean	SD	Classification	Diff. to Ref.
1-Ref (ATA)	95,2	7,9	Identical	
2-Circe	49,0	25,3	Slightly Similar	-46,3
3-Retune	76,6	20,4	Similar	-18,6
4-World	85,9	16,7	Similar	-9,4

p-value < 0.001 by ANOVA and Tukey HSD post-hoc

A graphical representation of the results can be found in Figure 3.11, showcasing means, standard deviations and distributions of data. Furthermore, we've conducted these analyses under various conditions:

1. After removing unsuitable subjects (ANOVA in Table C.18, Tukey's HSD in Table C.19, graphical representation in Figure 3.11)
2. Non-musicians (ANOVA in Table C.20, Tukey's HSD in Table C.21, graphical representation in Figure 3.12)
3. Musicians (ANOVA in Table C.22, Tukey's HSD in Table C.23, graphical representation in Figure 3.12)

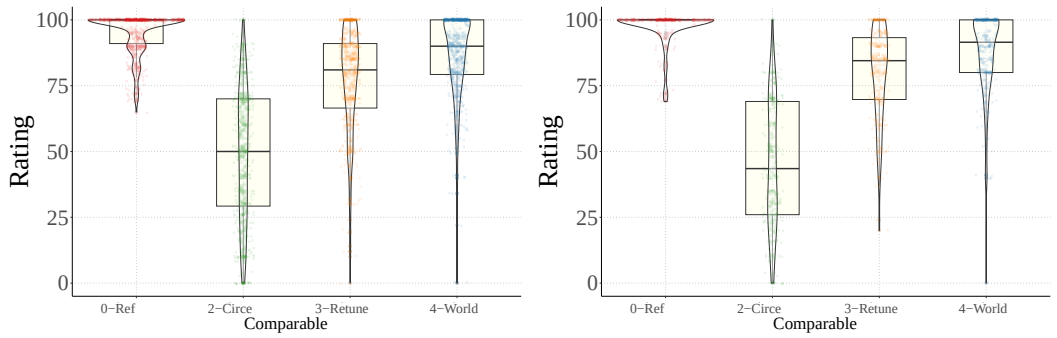


Figure 3.11: Results for Task C - before (left) and after (right) excluding subjects deemed unsuitable. Labels 1 to 4 indicate: ATA, Circe, Retune and World extreme-autotuning

The first observation indicates that in this extreme autotuning experiment, subjects statistically exhibit similarity in terms of the relative placement of the means compared to Task A (considering now that there is no original sound). This similarity can be attributed to the fact that in both Task A and Task C, the pitch curve condition of the comparables is the same within each task. Our observation is supported by the ATA scoring values, which are 95.2 ± 7.9 and 96.6 ± 6.1 for Tasks C and A, respectively. Additionally, the shape of the ATA data distribution is statistically similar to that observed in Task A, suggesting

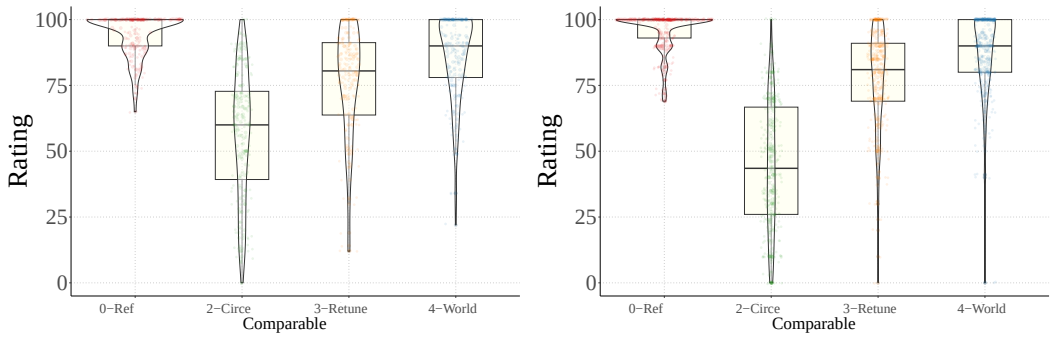


Figure 3.12: Results for Task B - Non-Musicians (left) and Musicians (right). Labels 1 to 4 indicate: ATA, Circe, Retune and World extreme-autotuning

again that it may be due to the scale and a perceptual phenomenon for very similar vocal samples, regardless of whether they are natural or vocoded.

The second observation is that statistically revealed by Tukey HSD post-hoc analysis, there is a significant difference between the vocoders (CIRCE, TEUNE, and WORLD) regarding ATA for the entire panel, but also excluding subjects and dividing the panel into groups of musicians and non-musicians. On the other hand, Task C differs from what is observed in Task A, as in terms of means, the positions of Retune and World are reversed. This suggests that different pitch scenarios (resynthesis and extreme autotuning) lead to discrepancies between vocoding processes, possibly because systems interpret pitch curves differently. If that is the case, then we can say that Retune is slower to perform a transposition and that in Task A, the proximity is due to its slowness in processing rather than the fidelity of the transposition. On the other hand, Circe has a difference in means and distribution shape similar to ATA, which is similar to that obtained in Task A.

The third observation highlights that the WORLD vocoder is the vocoder closest to ATA in terms of means. The last observation is that the shape and relative positions of the comparable ratings are consistent when dividing the panel into various groups. Once again, it is observed that lower ratings given by musicians result in lower means and an elongation of the data distribution, similar to the cases of Tasks A and B.

3.10.4 Task D: Soft autotuning with each vocoder compared to Soft autotuning with ATA

For Task D, we compare how each vocoder performs soft-autotuning compared to the reference autotuning done by ATA. Soft-autotuning is achieved by resynthesizing the original sound, imposing the soft-autotig curve recovered from ATA using Praat. In ATA, this process is carried out by configuring $retune - speed = 50$. Our goal is to understand how each vocoder technique uniquely affects the coloration of autotuning. The difference in coloration arises from both the inherent characteristics of each vocoder and the autotun-

ing process itself. The results presented in Table 3.12 show the mean values for each vocoder, along with the identified differences. Additionally, the data can be visually represented, as shown in Figure 3.13.

Table 3.13: ANOVA for Task D - before excluding subjects deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	496532	165511	503.3	$< 2e-16$
Residuals	1340	440680	329		

Table 3.14: Tukey HSD post-hoc analysis for Task D: Soft autotuning with each vocoder compared to Soft autotuning with ATA

Comparable	Mean	SD	Classification	Diff. to Ref.
1-Ref (ATA)	96,9	5,7	Identical	
2-Circe	45,0	25,5	Slightly Similar	-51,9
3-Retune*	82,3	17,9	Similar	-14,6
4-World*	82,7	17,5	Similar	-14,3

p-value < 0.001 by ANOVA and Tukey HSD post-hoc. Except:
(*) World vs Retune have a means diff. of 0.3, p-value $= 0.9937$ (full panel), and not significant for non-musicians and musicians

The histograms have been calculated per subjects (Figures C.13 and C.14) and per trials (Figures C.15 and C.16). The statistical analysis includes ANOVA (Table 3.13) and Tukey's HSD test (Table C.24).

A graphical representation of the results can be found in Figure 3.13, showcasing means, standard deviations and distributions of data. Furthermore, we've conducted these analyses under various conditions:

1. After removing unsuitable subjects (ANOVA in Table C.25, Tukey's HSD in Table C.26, graphical representation in Figure 3.13)

2. Non-musicians (ANOVA in Table C.27, Tukey's HSD in Table C.28, graphical representation in Figure 3.14)

3. Musicians (ANOVA in Table C.29, Tukey's HSD in Table C.30, graphical representation in Figure 3.14)

It is evident that Retune and World closely approximate ATA in terms of rating; however, this proximity in rating does not translate into statistical similarity for the full panel, as indicated in Appendix 3.14 and C.5. Despite their close mean values, they remain statistically distinct for the full panel. Concerning the similarity between Retune and World, it is noteworthy that they appear statistically indistinguishable when considering the entire panel or excluding certain subjects. However, upon dividing the panel between musicians

and non-musicians, no definitive conclusion can be drawn regarding their significant similarity or difference. Additionally, the relative positions of Retune and World, while similar, become inverted for the two groups. Conversely, Circe consistently exhibits a significant statistical difference and disparity in mean values across all panel groups.

There are three important considerations to note related to the task done until now:

1. In Task A, the closest re-synthesis to the original sound (apart from ATA) is achieved by Retune (-10.7 compared to the reference), followed by World (-15.7 compared to the reference).

2. In Task B (extreme autotuning), both World and ATA are equally distant from the original sound, and are indistinguishable with a confidence interval greater than 95%.

3. In Task C, extreme autotuning with World is the closest to ATA, followed by Retune.

Based on these observations, it can be concluded that overall, the World vocoder exhibits slightly higher similarity to ATA compared to Retune across the four tasks. Therefore, World may prove to be a more suitable option for evaluating pitch changes, such as in the psychoacoustic evaluation of pitch correction methods.

Finally, it is important to consider that the reference audio consistently maintains the same systematic shape across all four tasks, regardless of the pitch condition or vocoder. The distribution of the reference always presents a kind of lobes. We cannot assert that the reason is the MUSHRA scale. Nevertheless, remember that the MUSHRA scale is semi-discretized (Figure 3.4), meaning there may be cognitively preferred points within the "identical" and "similar" intervals that subjects use, or the lobes may be due to a perceptual phenomenon and different levels of subjects' discriminative ability. Although this question did not exist before, it is interesting to propose it as it is not possible to answer it with the current data.

3.11 Participants Interviews and Feedback

During and following each trial, interviews were conducted with the respective participants. These interviews were open-ended and centered around the test in its entirety. For instance, participants were queried about their overall impressions and comprehension of the questions, to which all subjects responded affirmatively. Regarding attentiveness throughout the trial, it was observed that individuals who had previously engaged in similar tests encountered no significant impediments. At the same time, first-time participants exhibited some fatigue, albeit mentioning their endeavor to perform to the best of their ability.

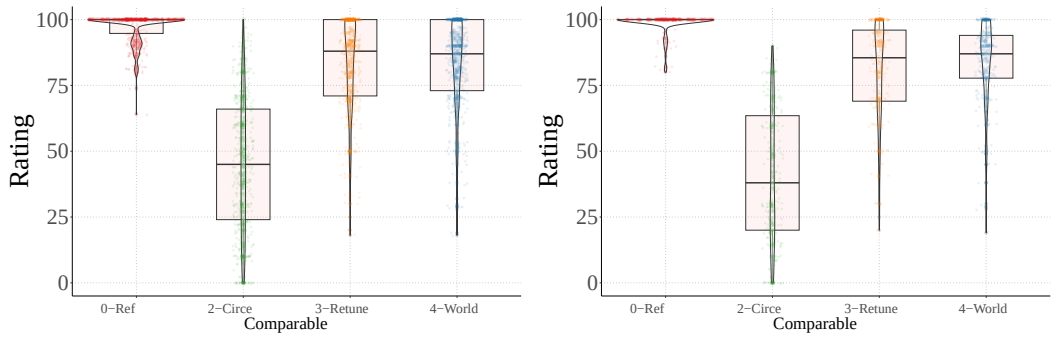


Figure 3.13: Results for Task D - before (left) and after (right) excluding subjects deemed unsuitable. Labels 1 to 4 indicate: ATA, Circe, Retune and World soft-autotuning

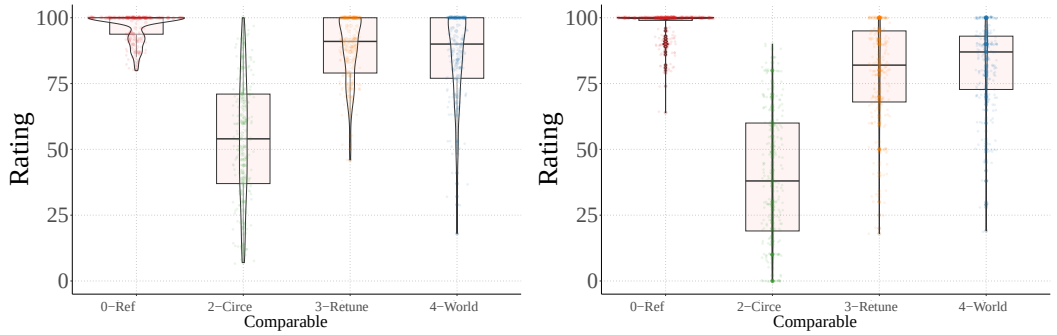


Figure 3.14: Results for Task D - Non-Musicians (left) and Musicians (right). Labels 1 to 4 indicate: ATA, Circe, Retune and World soft-autotuning

Regarding the ease or difficulty encountered in specific cases, there was a disparity in the type of samples that were perceived as easier. For some participants, it was easier to identify differences in stimuli with consonants, while for others, it was easier when the sound had fewer consonants. These observations were equally varied among both musicians and non-musicians. This may be attributed to perceptual scanning strategies, which could be of the COD (comparison over distance) type or through skimming. This confirms that the panel is diverse and that no particular type of sample is favored.

3.12 Conclusions

Completing this vocoder psycho-acoustic study allows us to corroborate some ideas we initially held as experimenters while challenging others that turned out slightly differently than expected. Below, we summarize the main points derived from compiling the results obtained from each task:

- The perception of the reference remains consistent across different contexts and comparables (Tasks A, B, C, D).

- ATA is identified as transparent in the softest possible configuration (Task A), statistically indistinguishable from the original sound.
- World and Retune exhibit similarities with the original audio in pitch resynthesis (Task A), but a statistically significant difference prevails.
- Extreme autotuning with ATA and World equally deviates from the original audio (Task B), and they are statistically indistinguishable from each other.
- Extreme autotuning with World is the closest to ATA (Task C) concerning the mean values, but a statistically significant difference exist.
- Soft autotuning with World and Retune closely resembles that of ATA (Task D), and they are statistically indistinguishable from each other.
- World exhibits a slightly superior resemblance to ATA regarding mean values compared to Retune.

Now we will discuss the implications of some of these results in more detail. Firstly, the fact that the perception of the reference remains consistent across different contexts and comparisons (Tasks A, B, C, D) indicates that the shape and mean value of the distribution are very similar, regardless of whether the reference is a natural or vocoded sound. This result suggests the existence of a vocal difference perception threshold, as the obtained value never reaches 100%. In future studies, it would be helpful to investigate this perception threshold with minimal sound variations and a larger number of vocoders to confirm that it is indeed a perceptual phenomenon. Additionally, it would be interesting to apply this approach beyond pitch, exploring other vocal characteristics. Such experiments could determine if this perceptual phenomenon varies more or less within the vocal context than the perception of pure tones or levels. Other vocal characteristics could include roughness, breathiness, etc.

ATA is identified as transparent in the softest possible configuration (Task A), statistically indistinguishable from the original sound. This result indicates that the vocoding quality of ATA is unmatched and that none of the other systems we used reach that level of quality. Additionally, tests with variable transpositions should be included to verify the sound quality of a vocoder, not just constant transpositions. This proposition is relevant for the vocoder's quality evaluation, as the systems studied have undergone quality tests with constant transposition, but variable transposition presents new challenges regarding realism and quality. Such a kind of test would also reveal that specific timbral characteristics only become evident when using variable transpositions instead of constant ones. These results suggest that any autotune preset will vary depending on the vocal transformation system used, as autotune itself involves variable transposition.

World and Retune exhibit similarities to the original audio in pitch resynthesis (Task A), but a statistically significant difference prevails. Initially,

we thought we could use one of these two vocoders as a replica of ATA to evaluate ATA's pitch correction relative to DPW, but no. This result means that the sound differences, which cannot be determined by individual listeners (the experimenter, the supervisor, the assistant), can be identified through a psychoacoustic test. Moreover, this comparison is helpful because we wanted to determine which of the two systems could eventually be used to replicate ATA, a closed system. Since ATA is indistinguishable from natural sound in Task A, we see that Retune and World cannot be used as replicas of ATA for resynthesis.

Extreme autotuning with ATA and World deviates equally from the original audio (Task B) and are statistically indistinguishable. Unlike the previous case, this shows that for the specific case of extreme autotuning, World could potentially be used as a replica of ATA (when compared to the natural sound, there is no difference). However, a specific comparative verification between ATA and World is necessary. For this, we can use Tasks C and D, but in the future, we could also go further, for example, with an ABX study comparing only ATA and World.

Extreme autotuning with World is the closest to ATA (Task C) concerning the mean values, but a statistically significant difference exists. This means that, although no difference is perceived when compared to the natural sound, World and ATA are perceived as different when compared to each other. Therefore, the idea of using World as a replica to evaluate ATA's pitch correction method in isolation is ruled out.

Soft autotuning with World and Retune closely resembles that of ATA (Task D), and they are statistically indistinguishable from each other. Similar to the previous case, this result allows us to rule out the idea of using World as a replica of ATA to evaluate ATA's pitch correction method in isolation.

World exhibits a slightly superior resemblance to ATA regarding mean values compared to Retune. This result means that if we have ATA's correction curve (tracked with Praat), we can compare ATA and DPW methods almost in isolation, at least to estimate the difference. Given the context, in which we already have a comparison of ATA against World in several cases, this would give us an idea of whether there is any additional coloration due to the pitch correction method (ATA or DPW).

With the previous clarifications, we can proceed to the Subjective Evaluation of Pitch Correction Methods, where we will compare the implemented ATA and DPW methods. We will use a copy of ATA with World, which, although not serving as a replica, can provide indications of whether the coloration due to the pitch correction method is similar in the cases already studied in tasks A, B, C, and D. This will help us determine if there is an additional impact due to pitch correction or if the coloration is primarily due to the vocoder.

Chapter 4

Subjective Evaluation of Pitch Correction Methods

In this chapter, we conduct a discriminative-based protocol comparison between ATA and DPW pitch correction methods. The psycho-acoustical evaluation test proposed here employs a similar interface and set of questions to the vocoders' comparative test. We have two methods to study: ATA and DPW. As for the vocoder evaluation, we also use a DFC test; the advantage of the DFC test over purely discriminating tests lies in its ability to analyze the statistical behavior of the panel across different scenarios. DFC not only helps ascertain whether samples are distinguishable for each task but also helps to identify differences and their consistency across various vocoders or configurations of pitch correction methods in different tasks.

Ideally, pitch tracking and vocoder variables should remain constant, with only the pitch correction algorithm varying. However, maintaining such consistency is unfeasible due to ATA's closed software nature. Thus, we propose two approaches. Firstly, a portion of the comparisons utilizes ATA's complete protocol, including its pitch tracker, pitch correction method, and "vocoder". In comparison, DPW is treated as a separate protocol, using Praat as the pitch tracker and World as the vocoder. Secondly, the remaining comparisons utilize a non-exact replica of ATA, employing World as the vocoder and imposing the tone curve from an ATA-corrected audio file, where the tone curve was extracted using Praat and Python. These comparisons allow us to draw conclusions regarding the vocoder-induced coloration and melodic effects, their prevalence across various cases, such as soft and extreme correction, and the perception of differences between the two methods.

4.1 Tasks

Each task comprises 19 questions, totaling 95 questions (trials) across all five tasks. Participants are prompted with a similarity question identical to that in the vocoder DFC test, wherein they rate the level of resemblance between the stimuli (comparables) and the reference using the provided interface. The

questions within each task are fully randomized, and the sequence of tasks is also randomized. Further elaboration on each task is provided below.

4.1.1 Task 1: extreme pitch correction ATA+ATA compared to DPW+World

In the first task, we compare, for a given source sound, audio files autotuned with ATA using extreme pitch correction (as the reference) versus audio files vocoded with World using DPW extreme pitch correction (as the comparable). This entails comparing the full protocol utilizing DPW (vocoding with World) with the ATA protocol. We hypothesize that extreme cases of both pitch correction methods are very similar. This would imply that both methods achieve the extreme retuning effect in a closer manner, thus preserving the stylistic properties of autotune. To verify this, we would check whether the extreme correction yields a similar outcome to that obtained in Task C of the vocoder's comparative test, which compared autotuning across different vocoders.

4.1.2 Task 2: extreme pitch correction ATA+World compared to DPW+World

In the second task, we compare sounds vocoded with World using both pitch correction methods (ATA and DPW) in extreme configuration. The original pitch is obtained through Praat and adjusted using the ATA and DPW algorithms. According to patent [Hildebrand, 1998], the ATA algorithm is equivalent to the integer part of the value obtained on the semitone scale (for $retune - speed = 0$). This implies that the only virtual difference between the files is the tone correction method employed (DPW and ATA as reference) to obtain the imposed pitch curve. Our hypothesis is that the extreme cases are very similar, which could be interpreted as both methods achieving the extreme correction effect in the same manner.

4.1.3 Task 3: soft pitch correction ATA+ATA compared to DPW+World

In the same manner as conducted in Task 1, in the third task, we compare the complete protocols of ATA (reference) and DPW. For this task, we employ a soft correction. The ATA protocol consists of its own pitch tracker, pitch correction algorithm, and vocoder. The DPW protocol comprises tracking with Praat, pitch correction with DPW, and vocoding with World. Through this test, we aim to ascertain whether subjects perceive differences between the methods for soft correction.

4.1.4 Task 4: soft pitch correction ATA+World compared to DPW+World

The fourth scenario involves comparing soft versions of the audio vocoded with both pitch correction methods using the same vocoder, as done in T2. Therefore, we compare the audio softly autotuned with ATA (as reference) with the audio vocoded with World for a soft DPW pitch correction. The audio corresponding to ATA is obtained by recovering the pitch from a file treated with the actual ATA protocol using Praat and then resynthesizing it with World. Task 4 is exploratory, similar to the third task, where we aim to determine whether subjects can perceive a difference or not.

4.1.5 Task 5: Source audio compared to Soft pitch corrections

The final task entails discriminating between the authentic sound (reference) and the subtle corrections applied with ATA and DPW. This will involve employing three comparisons:

1. The full ATA protocol;
2. A replication of ATA utilizing World as a vocoder, by imposing the pitch curve of a file treated with the authentic ATA protocol for subtle correction; and
3. DPW applied over the pitch curve of the original file using World as a vocoder.

Our objective here is to ascertain whether subjects establish an association based on the sound scores according to the vocoder (associating those utilizing World) or based on the melody (associating those with ATA). Moreover, the variance in assessment will enable us to determine if there is a discernible distinction between the comparables.

4.1.6 Summary of tasks for pitch correction methods comparison

The table 3.5 presents a detailed summary of the various aspects addressed in the psychoacoustic test, including the description of the references and comparables used, as well as the specific objectives of each task. Through this arrangement, the aim is to provide a comprehensive understanding of the controlled variables and parameters assessed at each stage of the study. In addition to the primary objectives of each task, the table also highlights additional conclusions that can be drawn from the collected data, further enriching the interpretation and utility of the results obtained. For the statistical analyses, the ANOVA and post-hoc Tukey HSD protocol are used, with the null hypothesis being equality and the alternative hypothesis being the difference.

Task	Reference	Melody(F0)	Comparable	Melody(F0)
1	ATA	$f_{extreme,ata}$	World	$f_{extreme,dpw}$
	Comparing pitch correction for extreme correction between ATA and world+DPW protocols			
2	World	$int(f_{nat}) \approx_{extreme,ata}$	World	$f_{extreme,dpw}$
	Comparing pitch correction for extreme correction keeping the same vocoder: world			
3	ATA	$f_{soft,ata}$	World	$f_{soft,dpw}$
	Comparing pitch correction for soft correction between ATA and world+DPW protocols			
4	World	$f_{soft,ata}$	World	$f_{soft,dpw}$
	Comparing pitch correction for soft pitch correction keeping the same vocoder: world. Praat is used to track $f_{soft,ata}$, world is used to re-synthetize the original audio with $f_{soft,ata}$.			
5	Natural	$f_{natural}$	World ATA	$f_{soft,dpw}$ $f_{soft,ata}$ $f_{soft,ata}$
	Comparing original sound with soft pitch correction for several cases. Purpose is check possible timbral difference due to pitch correction method.			

Table 4.1: Summary of tasks for pitch correction methods comparison

4.2 Test preparation

The design of the subjective test is carried out in the same path as the vocoder test: using the MUSHRA interface, with elements previously defined for the hybrid ranking and discrimination test, including:

- Objectives and Hypotheses: Distributed in the five previously described tasks, each with specific objectives and hypotheses.

- Stimuli: Audio files vocoded generated by the two pitch correction methods for 3 cases: extreme correction, soft correction, correction vs. natural sound.

- Interface and Data Collection: Utilization of the MUSHRA interface and data collection in CSV format provided by MUSHRA tools.

4.3 Audio Support

The audio support for the comparison of the pitch correction methods consists of 19 samples. These samples are edited with vocoders and pitch cases according to Table 3.4 for tasks 1,2,3,4 and 5.

Table 4.2: Audio Support for the Pitch Correction Comparison Psycho-Acoustical Test (abbreviated as p)

Method	**ATA**				**DPW**		None
Vocoder	ATA		World		**World**		None
Case	Extreme	Soft	Extreme	Soft	Extreme	Soft	Original File
Abbv. Audio File Name							
pa part1: part1-real3Maleintervals	1	3,5	2	4,5	1,2	3,4,5	5
pa part2: part2-real3Maleintervals	1	3,5	2	4,5	1,2	3,4,5	5
pb part1: part1-real19Malevoicelegatto	1	3,5	2	4,5	1,2	3,4,5	5
pb part2: part2-real19Malevoicelegatto	1	3,5	2	4,5	1,2	3,4,5	5
pc part1: part1-real23Femalelegattovirtuoso	1	3,5	2	4,5	1,2	3,4,5	5
pc part2: part2-real23Femalelegattovirtuoso	1	3,5	2	4,5	1,2	3,4,5	5
pc part3: part3-real23Femalelegattovirtuoso	1	3,5	2	4,5	1,2	3,4,5	5
pd part1: part1-realJF-mem-6-a-male2	1	3,5	2	4,5	1,2	3,4,5	5
pd part2: part2-realJF-mem-6-a-male2	1	3,5	2	4,5	1,2	3,4,5	5
pe part1: part1-realLP-mem-6-a-fem2	1	3,5	2	4,5	1,2	3,4,5	5
pe part2: part2-realLP-mem-6-a-fem2	1	3,5	2	4,5	1,2	3,4,5	5
pe part3: part3-realLP-mem-6-a-fem2	1	3,5	2	4,5	1,2	3,4,5	5
pe part4: part4-realLP-mem-6-a-fem2	1	3,5	2	4,5	1,2	3,4,5	5
pf part1: part1-realms-celinedion	1	3,5	2	4,5	1,2	3,4,5	5
pf part2: part2-realms-celinedion	1	3,5	2	4,5	1,2	3,4,5	5
pf part3: part3-realms-celinedion	1	3,5	2	4,5	1,2	3,4,5	5
pf part4: part4-realms-celinedion	1	3,5	2	4,5	1,2	3,4,5	5
pg part1: part1-realrt-yvesmontand	1	3,5	2	4,5	1,2	3,4,5	5
pg part2: part2-realrt-yvesmontand	1	3,5	2	4,5	1,2	3,4,5	5

The content of the samples can be described according to the types: staircase, vibratos and free-path, as shown in Figure 1.4.3.

4.3.1 Subject Panel

The number of subjects in the psycho-acoustic analysis is generally limited. In our experiment, we used the same panel for the vocoders comparison test: 20 subjects divided between musicians and non-musicians.

Table 4.3: Content of the Samples for the Pitch Correction Comparison Psycho-Acoustical Test

Sample	Content			Duration [ms]	Sample	Content			Duration [ms]
	Staircase	Vibratos	Freepath			Staircase	Vibratos	Freepath	
pa part 1		✓		2397	pe part 1	✓	✓	✓	2995
pa part 2		✓	✓	2397	pe part 2	✓	✓	✓	4193
pb part 1	✓	✓		2995	pe part 3	✓	✓	✓	4792
pb part 2	✓	✓		4792	pf part 4	✓	✓	✓	4193
pc part 1	✓	✓	✓	3294	pf part 1	✓	✓	✓	3594
pc part 2	✓	✓		2397	pg part 2	✓	✓	✓	3294
pc part 3	✓	✓		3594	pg part 3	✓	✓	✓	3145
pd part 1	✓		✓	3145	pg part 4	✓	✓		2396
pd part 2	✓		✓	4792	pg part 1	✓	✓		4193
					pg part 2	✓	✓	✓	7188

4.3.2 Test Contents

The test consists of five tasks, each with its respective hypotheses. Each task includes 19 questions, and in each question, subjects are asked to determine the level of similarity between the reference and the comparables (vocoders). This means that each pitch correction method is evaluated with 19 different sound examples by each subject, totaling 380 scores related to each pitch correction method case. This diversity of examples ensures test variability and prevents fatigue effects.

4.3.3 Data Treatment

Data is collected in CSV format as in the vocoder comparison test, in a file that includes identifiers for each test subject, identifiers for the type of comparable, values assigned by the subject to each comparable, the type of sound example, and the total time for each stage of the test. As in the vocoder comparison test, statistical data analysis is performed using R.

4.3.4 Room and sound

The setup closely follows the procedures used in the vocoder comparison test. To conduct the test, a dedicated recording studio room (room 519) at the Institut Jean le Rond d'Alembert has been used. The tests are administered

using a MacBook running macOS Big Sur (11.7). Participants access the test through a website hosted on the Institut Jean le Rond d'Alembert server. Test data is recorded at the end of each 20-minute test segment. *Sennheiser HD 205* over-ear headphones are provided for participants, and a pre-test check ensures optimal volume and equipment functionality.

4.3.5 Planning

The procedure mirrors that of the vocoder comparison test. The test comprises five parts, each lasting 15 minutes, conducted in a single one hour and half session. Participants, grouped into musicians and non-musicians, are summoned to the Jean Le Rond d'Alembert Institute. The testing period spans one month, with scheduling coordinated directly with participants. Rested participants are compensated with a 40-euro Amazon gift card for their participation.

4.4 Test procedure

In preparation for the test, individuals are called to the allocated room furnished with the predetermined materials. The session commences with an introductory presentation outlining the test in a similar way to the test for the vocoder comparison. The presentation contains the following information:

- *Pitch correction algorithms are widely utilized in contemporary music production. Initially designed to discreetly rectify a singer's off-key notes, some artists now employ them as a deliberate effect. To maintain the "naturalness" of the voice throughout these alterations, vocoders are needed into these algorithms. The objective of this test is to assess the impact on coloration caused by the vocoder and the pitch correction algorithms.*
- *The test is divided into five segments, each lasting approximately 20 minutes. For every segment, you will be asked to assess the degree of similarity between a reference and various conditions. Following each segment, you will need to complete an information sheet, ensuring that you include the same ID at the end of each section.*

The comparison between the vocoder and pitch correction methods was conducted jointly. Therefore, participant consent and data collection were carried out only once using the formats outlined in the preceding chapter. First, it appears a volume adjustment window, followed by an illustrative question, as done in the vocoder test (Figure 3.4). Then, participants proceed to complete the 19 questions of the task, which lasts around 15 minutes. Upon

completion, participants enter their participant ID on a page, and data is automatically saved on the server. They then progress to another task, which also lasts approximately 15 minutes. Following this, if desired, they can take a more extended break of up to 30 minutes. Subsequently, the participant continues with three other tasks, and at the end, an interview is conducted to obtain further insights into the test experience.

4.4.1 Results

4.4.2 Task 1: extreme pitch correction ATA+ATA compared to DPW+World

In this section, we present the results obtained for Task 1. Histograms have been calculated per subject (Figure C.17) and per trial (Figures C.18 and C.19), allowing for the examination of consistent results across subjects and trials. The result are presented in Figure 4.1, showcasing means and standard deviations, as well as data distributions. And four subdivisions of the panel have been studied, the results for each group are presented as follows:

1. Entire panel, before removing two unsuitable subjects (ANOVA in Table 4.4, Tukey's HSD in Table (Table C.32), graphical representation in Figure 4.1)
2. After removing two unsuitable subjects (ANOVA in Table C.33, Tukey's HSD in Table C.34, graphical representation in Figure 4.1)
3. Non-musicians (ANOVA in Table C.35, Tukey's HSD in Table C.36, graphical representation in Figure 4.2)
4. Musicians (ANOVA in Table C.37, Tukey's HSD in Table C.38, graphical representation in Figure 4.2)

Table 4.4: ANOVA for Task 1 - before excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	18761	18761	92.72	$< 2e-16$
Residuals	758	153374	202		

Our hypothesis posited that the extreme cases of both pitch correction methods would exhibit high similarity. The observed difference in mean scores in the results is less than 10% of the score obtained by the reference, suggesting a minor discrepancy. However, this outcome does not elucidate whether this difference stems from the vocoder or the correction method, as both vary for the two comparable elements of Task 1. Statistically, as evidenced in Appendix C.7, it is apparent that the distributions are disparate, and the null hypothesis

Table 4.5: Tukey HSD post-hoc analysis for Task 1:
Extreme correction, ATA+ATA compared to DPW+World

Task 1

Comparable	Mean	SD	Classification	Diff. to Ref.
DPW extreme + World	86,2	18,5	Very similar	-9,9
Ref. ATA extreme + ATA	96,2	7,9	Identical	

p-value < 0.001 by ANOVA and Tukey HSD post-hoc

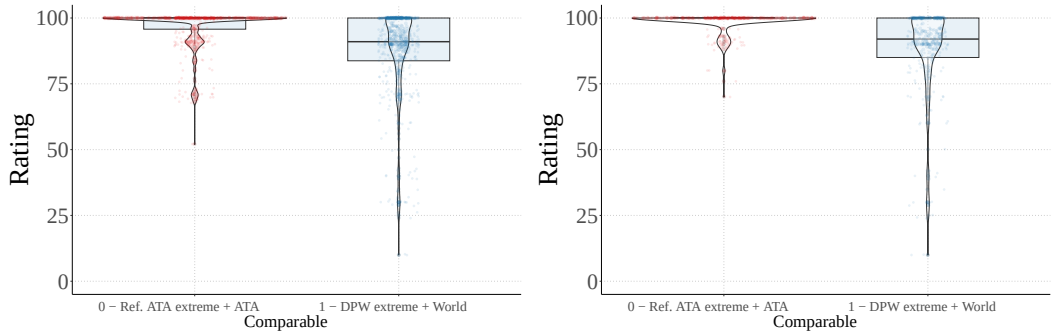


Figure 4.1: Results for Task 1 - before (left) and after (right) excluding subjects deemed unsuitable. Extreme correction for ATA+ATA (1) and DPW+World (2)

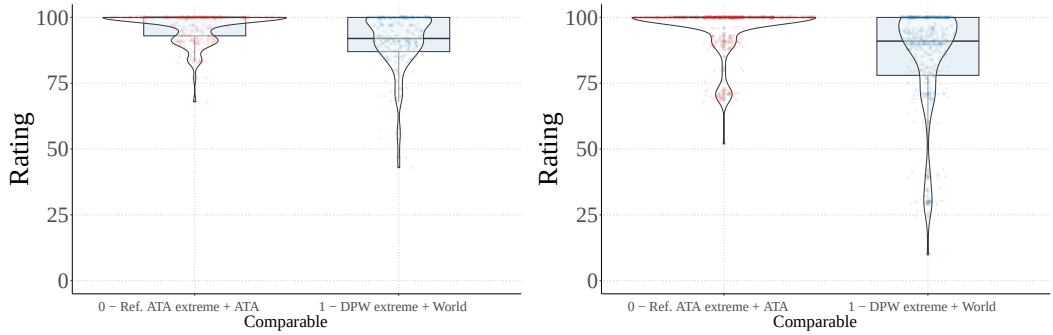


Figure 4.2: Results for Task 1 - Non-Musicians (left) and Musicians (right). Extreme correction for ATA+ATA (1) and DPW+World (2)

of equality is rejected. This indicates that despite the proximity in terms of rating, the panel subjects are capable of distinguishing between them.

Another noteworthy observation arises from comparing the results of the vocoder comparison for the re-synthesis case (Task A) with those obtained here. It becomes apparent that the mean values obtained for ATA and World vocoders demonstrate a comparable magnitude of difference, both in the cases of re-synthesis and extreme autotuning. However, it is crucial to emphasize that this difference is statistically significant. This suggests that the observed disparity may be attributed more to the vocoding process rather than the specific pitch correction method employed for this task.

Last aspect worth considering is the shape of the control stimulus lobe (the same reference), which can be reexamined. It is observed that the data distribution shapes are preserved for both non-musicians and musicians, albeit slightly more dispersed in musicians. This discrepancy in dispersion may be attributed to factors elucidated in the previous chapter tasks.

4.4.3 Task 2: extreme pitch correction ATA+World compared to DPW+World

In this section, we present the results for Task 2. The results, including means, standard deviations, and data distributions, are shown in Figure 4.3. Histograms have been calculated per subject (Figure C.20) and per trial (Figures C.21 and C.22), and four panel subdivisions are presented as follows:

1. Entire panel, before removing two unsuitable subjects (ANOVA in Table 4.6, Tukey's HSD in Table C.39, graphical representation in Figure 4.3)
2. After removing two unsuitable subjects (ANOVA in Table C.40, Tukey's HSD in Table C.41, graphical representation in Figure 4.3)
3. Non-musicians (ANOVA in Table C.42, Tukey's HSD in Table C.43, graphical representation in Figure 4.4)
4. Musicians (ANOVA in Table C.44, Tukey's HSD in Table C.45, graphical representation in Figure 4.4)

Table 4.6: ANOVA for Task 2 - before excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	12192	12192	79.91	$< 2e-16$
Residuals	758	115655	153		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.7: Tukey HSD post-hoc analysis for Task 2:
Extreme correction, ATA+World compared to DPW+World

Comparable	Mean	SD	Classification	Diff. to Ref.
DPW extreme + World	87,8	15,5	Similar	-8,0
Ref. ATA extreme + World	95,8	8,0	Identical	

p-value < 0.001 by ANOVA and Tukey HSD post-hoc
Except for asterisk (*) cases.

The resemblance between the results of Task 1 and the present study is remarkable. In terms of means, both protocols (ATA extreme + World and

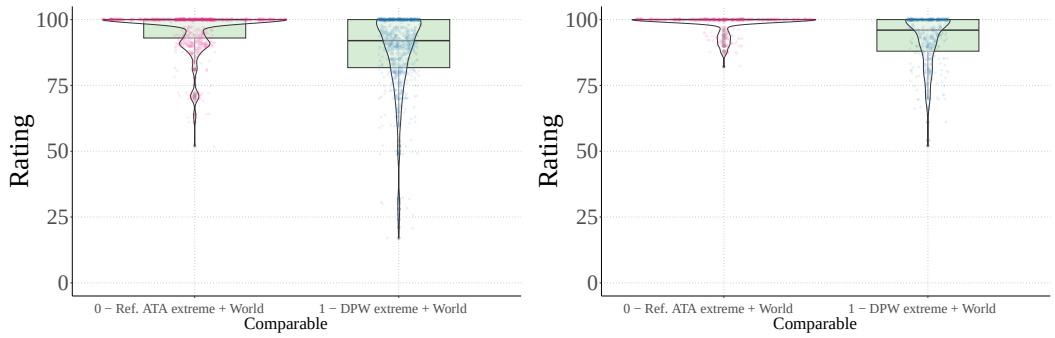


Figure 4.3: Results for Task 2 - before (left) and after (right) excluding subjects deemed unsuitable. Extreme correction for ATA+World (1) and DPW+World (2)

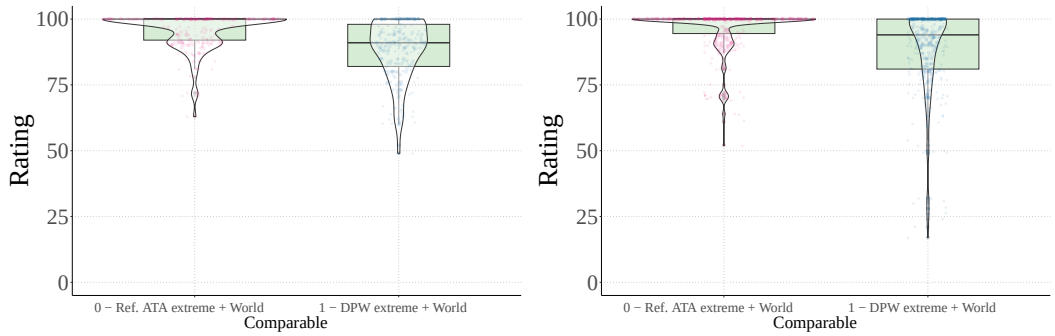


Figure 4.4: Results for Task 2 - Non-Musicians (left) and Musicians (right). Extreme correction for ATA+World (1) and DPW+World (2)

DPW extreme + World) exhibit a high degree of similarity. There is a slight remaining discrepancy (approximately 3), which is likely attributed to the pitch correction method rather than the vocoder.

Statistically and based on Appendix C.8, it can be observed that the distributions are different and the null hypothesis of equality is rejected despite the mean values being very similar. So, although there is a sound proximity in terms of scoring, the difference is discernible. It can be verified that the distribution shape for the reference stimulus is similar to the cases seen previously. The shapes of the distributions for both comparables remain consistent for both non-musicians and musicians, even when removing subjects. In other words, there is a small but consistent difference between the data. This difference is attributed to the pitch correction method, given that the vocoder remains the same.

4.4.4 Task 3: soft pitch correction ATA+ATA compared to DPW+World

In this section, we present the results for Task 3. The results, including means, standard deviations, and data distributions, are shown in Figure 4.5. His-

tograms have been calculated per subject (Figure C.23) and per trial (Figures C.24 and C.25), and four panel subdivisions are presented as follows:

1. Entire panel, before removing two unsuitable subjects (ANOVA in Table 4.8, Tukey's HSD in Table C.46, graphical representation in Figure 4.5)
2. After removing two unsuitable subjects (ANOVA in Table C.47, Tukey's HSD in Table C.48, graphical representation in Figure 4.5)
3. Non-musicians (ANOVA in Table C.49, Tukey's HSD in Table C.50, graphical representation in Figure 4.6)
4. Musicians (ANOVA in Table C.51, Tukey's HSD in Table C.52, graphical representation in Figure 4.6)

Table 4.8: ANOVA for Task 3 - before excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	22172	22172	130.6	$< 2e-16$
Residuals	758	128674	170		

Table 4.9: Tukey HSD post-hoc analysis for Task 3:
Soft correction, ATA+ATA compared to DPW+World

Comparable	Mean	SD	Classification	Diff. to Ref.
DPW soft + World	86,3	16,8	Similar	-10,8
Ref. ATA soft + ATA	97,1	7,4	Identical	

p-value < 0.001 by ANOVA and Tukey HSD post-hoc

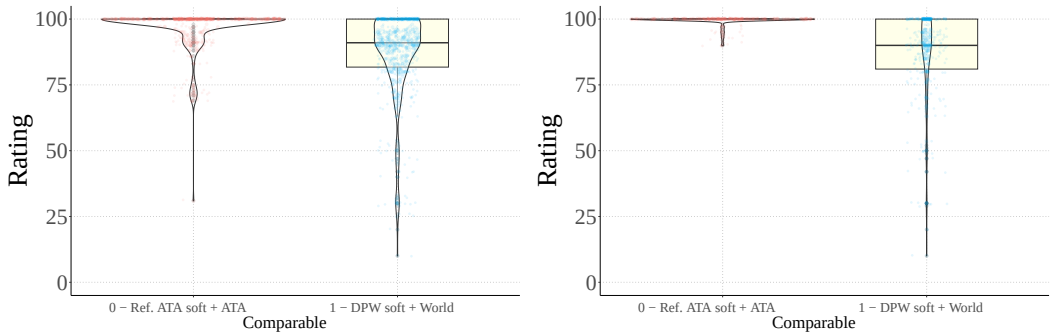


Figure 4.5: Results for Task 3 - before (left) and after (right) excluding subjects deemed unsuitable. Soft correction for ATA+ATA (1) and DPW+World (2)

Following the same methodology as in Task 1, in the third task, we compare the complete protocols of ATA (reference) and DPW+World for smooth correction. Similar to the findings in Task 1 for extreme correction and Task A for

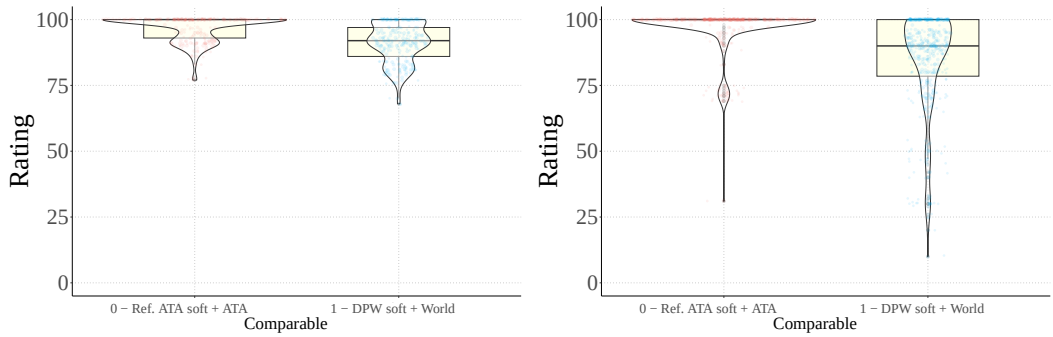


Figure 4.6: Results for Task 3 - Non-Musicians (left) and Musicians (right). Soft correction for ATA+ATA (1) and DPW+World (2)

resynthesis, a comparable difference in means is observed. Consequently, it can be inferred that the World vocoder introduces a consistent type of coloration across extreme autotuning (Task 1), soft correction (Task 3), and resynthesis (Task A). The discerned difference primarily stems from the vocoder; thus, no definitive conclusion regarding the coloration attributable to the pitch correction method can be drawn.

As per Appendix C.9, statistically, a significant difference akin to that in Task 1 is observed, likely attributed to the vocoder. Upon removing subjects from the panel, there is reduced dispersion. oreover, dividing the panel between musicians and non-musicians shows that the lobes are more similar and closer to 100 for non-musicians. In contrast, for musicians, the lobes are more dispersed, and the difference in shape is more pronounced.

4.4.5 Task 4: soft pitch correction ATA+World compared to DPW+World

In this section, we present the results for Task 4. The results, including means, standard deviations, and data distributions, are shown in Figure 4.7. Histograms have been calculated per subject (Figure C.26) and per trial (Figures C.27 and C.28), and four panel subdivisions are presented as follows:

1. Entire panel, before removing two unsuitable subjects (ANOVA in Table 4.10, Tukey's HSD in Table C.53, graphical representation in Figure 4.7)

2. After removing two unsuitable subjects (ANOVA in Table C.54, Tukey's HSD in Table C.55, graphical representation in Figure 4.7)

3. Non-musicians (ANOVA in Table C.56, Tukey's HSD in Table C.57, graphical representation in Figure 4.8)

4. Musicians (ANOVA in Table C.58, Tukey's HSD in Table C.59, graphical representation in Figure 4.8)

Table 4.10: ANOVA for Task 4 - before excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	354	354.4	4.425	0.0357
Residuals	758	60711	80.1		

Table 4.11: Tukey HSD post-hoc analysis for Task 4:
Soft correction, ATA+World compared to DPW+World

Comparable	Mean	SD	Classification	Diff. to Ref.
DPW soft + World	95,3	9,9	Similar	1,4
Ref. ATA soft + World	96,6	7,8	Identical	

By ANOVA and Tukey HSD post-hoc. p-value equal to 0.0357 (full panel), = 0.0067 (excl. subj.), 0.0067 (musicians) and 0.2730 (non-musicians)

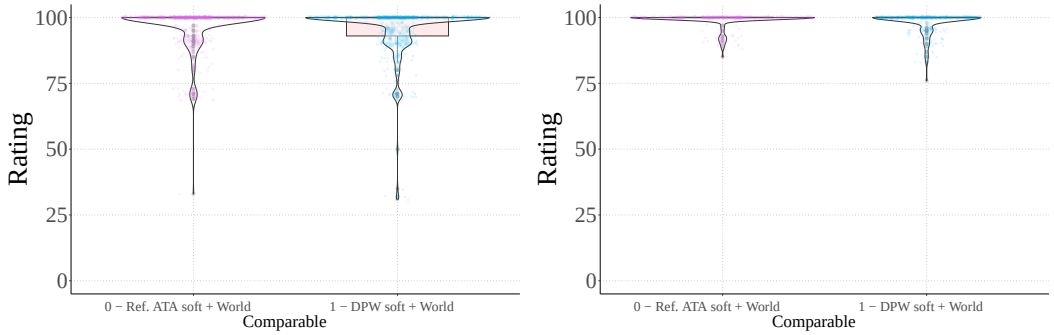


Figure 4.7: Results for Task 4 - before (left) and after (right) excluding subjects deemed unsuitable. Soft correction for ATA+World (1) and DPW+World (2)

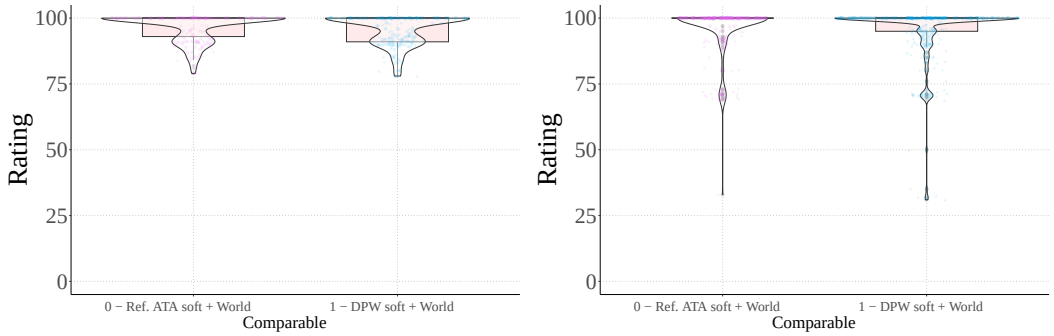


Figure 4.8: Results for Task 4 - Non-Musicians (left) and Musicians (right). Soft correction for ATA+World (1) and DPW+World (2)

In the fourth task, comparisons are made between smooth corrections using ATA and DPW with the same WORLD vocoder. Although the means indicate a high similarity between both cases, analyzing the statistical support of the entire panel confirms that there is a statistically significant difference. This leads us to conclude that smooth correction generates perceptible differ-

ences, even when they are very small, even when the data distributions are also similar. As shown in Table 4.11 and Appendix C.10, after removing inadequate subjects from the panel, the p-value changes slightly but remains within the confidence interval. The same occurs when dividing the panel between musicians and non-musicians, although in this case, the p-value falls outside the confidence interval, which prevents us from reaching a definitive conclusion. For the musician group, a p-value of 0.09 is obtained and for the non-musician group, 0.14, indicating that non-musicians are less capable of perceiving such subtleties in the difference than musicians. Nevertheless, the similarity between the data distributions for all four groups (complete panel, subjects excluded, musicians, and non-musicians) is noteworthy. Further study is required, which will be conducted with Task 5.

Note also that this did not occur in Task 2 regarding extreme pitch correction; the differences were statistically more robust in that task. We can therefore conclude that the differences in pitch correction methods are more perceptible in extreme correction than in mild correction. This is a significant conclusion, as according to [Perrotin and D'Alessandro, 2016] and the initial proposal of this thesis, the difference should have been found in soft correction, not necessarily in extreme correction. In other words, we have demonstrated the exact opposite. DPW cannot really provide a better expressive correction than ATA because, as seen in Chapter Three, the third degree of freedom is useless. Moreover, according to the results presented in this chapter, the use of other degrees of freedom in DPW is also not helpful, as the results do not show a statistically significant difference compared to ATA in the case of soft autotuning, which was the interesting aspect when this thesis was proposed.

4.4.6 Task 5: Source audio compared to Soft pitch corrections

Task 5 is a support to Task 1. For this task, several graphs and statistical calculations have also been done. The statistical analysis for the entire panel includes ANOVA (Table 4.12) and Post hoc tests using the Tukey HSD method (Table C.60). A graphical representation of the results for the full panel can be found in Figure 4.9, showcasing means and standard deviations, as well as data distributions. Additionally, histograms have been calculated per subject (Figures C.29 and C.30) and per trial (Figures C.31, C.32 and C.33), allowing for the examination of consistent results across subjects and trials. 4our panel subdivisions are presented as follows:

Table 4.12: ANOVA for Task 5 - before excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	3	56866	18955	93	<2e-16
Residuals	1516	308974	204		

Table 4.13: Tukey HSD post-hoc analysis for Task 5: Source audio compared to Soft Corrections with ATA (+ATA and +World) and DPW (+World)

Comparable	Mean	SD	Classification	Diff. to Ref.	Diff. to ATA soft + World
ATA soft + ATA	96,8	7,8	Identical	-0.3*	12.2
DPW soft + World	84,8	18,9	Similar	-12,2	0.3**
ATA soft + World	84,5	18,7	Similar	-12,5	
Ref. original	97,0	6,7	Identical		12.5

p-value < 0.001 by ANOVA and Tukey HSD post-hoc. Except for (*) p-value $= 0.9950$ and (**) p-value $= 0.9940$

1. Entire panel, before removing two unsuitable subjects (ANOVA in Table 4.12, Tukey's HSD in Table C.60, graphical representation in Figure 4.9)
2. After removing two unsuitable subjects (ANOVA in Table C.61, Tukey's HSD in Table C.62, graphical representation in Figure 4.9)
3. Non-musicians (ANOVA in Table C.63, Tukey's HSD in Table C.64, graphical representation in Figure 4.10)
4. Musicians (ANOVA in Table C.65, Tukey's HSD in Table C.66, graphical representation in Figure 4.10)

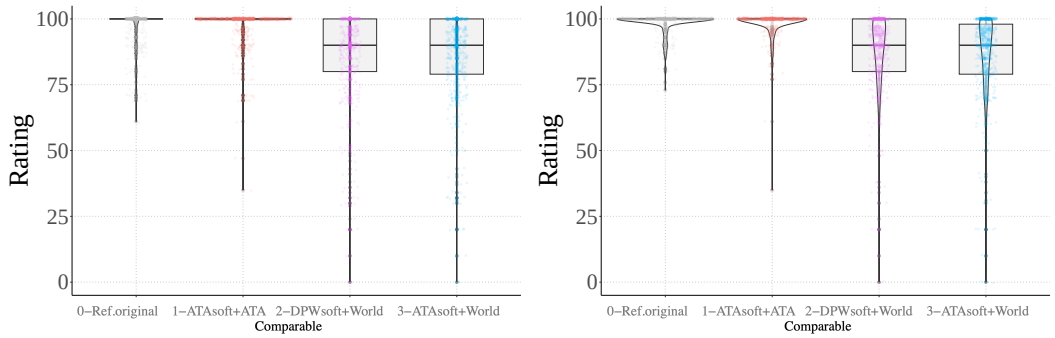


Figure 4.9: Results for Task 5 - before (left) and after (right) excluding subjects deemed unsuitable. Resynthesis (1) and Soft correction for: ATA+ATA(1), ATA+World (1) and DPW+World (2)

Initially, concerning means, the shape of data distribution, and statistical significance, it is noteworthy that the World vocoder consistently introduces a similar coloration across all comparisons conducted thus far (tasks A, B, C, D, 1, 2, 3). This aspect is reconfirmed herein, as samples utilizing the World vocoder (despite featuring different pitch curves) are scored almost identically, and the data distributions are also similar to those observed in previous tasks.

According to the data in Appendix C.11, summarized in Table 4.13, it can be concluded that when using the original sound as a reference, the ATA and

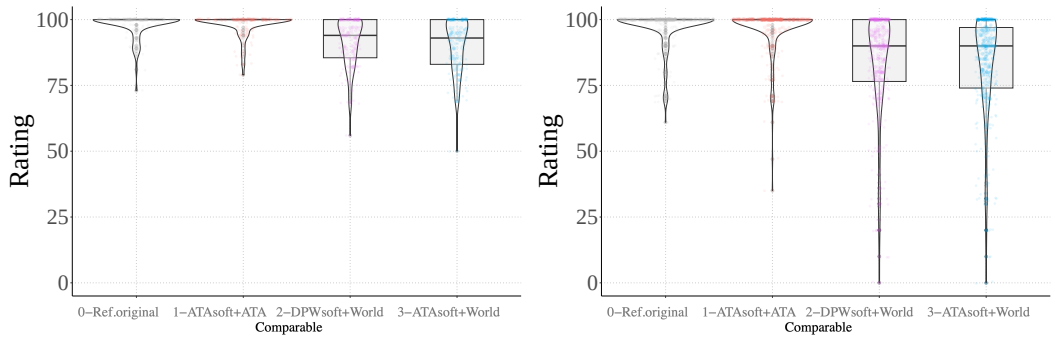


Figure 4.10: Results for Task 5 - Non-Musicians (left) and Musicians (right). Re-synthesis (1) and Soft correction for: ATA+ATA(1), ATA+World (1) and DPW+World (2)

DPW pitch correction methods using the World vocoder for mild correction are statistically indistinguishable with a confidence interval greater than 95%. Additionally, it can be observed that ATA's soft correction is indistinguishable from the original sound with a confidence interval higher than 95%. These findings are consistent across the panel, whether some subjects are removed or when the panel is divided between musicians and non-musicians.

We conclude that a predominance of coloration is attributable to the vocoder over the pitch correction method's coloration. Furthermore, it is verified that the coloration resulting from the pitch correction method for a soft case, as partially evidenced in Task 4, is very small compared to the vocoder-induced coloration.

4.5 Participants Interviews and Feedback

The test was divided into five parts; the participants were allowed to take breaks between each part. They were asked about their well-being, understanding, and any difficulties they encountered. Each stage lasted approximately 15 minutes. Participants were also asked for their overall assessment of the test, and responses varied. Some found the overall perception to be more challenging, while others stated that focusing on fewer samples made it easier. However, it was unanimously agreed that differences were subtler and, therefore, harder to detect than in the first test. According to the results obtained, participants demonstrated a very similar performance to that of the vocoder test, both in terms of reference assessment and data deviation.

4.6 Conclusions

After conducting the analyses about the psycho-acoustic of the pitch correction methods, certain overarching conclusions can be drawn. Below, we summarize the main points derived from compiling the results obtained from each task:

- The World vocoder introduces consistent coloration from different perspectives: extreme autotuning (Task 1), gentle correction (as per Task 3), and resynthesis (Task A). This premise is corroborated in Task 5, where participants group the World vocoder audio samples and rate them with very similar scores.

- According to Task 2, there is a small difference in means for extreme pitch correction between the ATA and DPW pitch correction methods; statistically, a difference can be observed.

- The observed disparity between the full ATA protocol and the DPW+World protocol (Task 3) is primarily attributed to the vocoder rather than the pitch correction algorithm.

- Contrasting results from task 2 to tasks 4 and 5, it can be seen that, although small, the coloration due to the pitch correction method exists but is more noticeable for extreme correction than for soft correction.

The World vocoder introduces consistent coloration from different perspectives. Such coloration is similar not only in extreme autotuning (Task 1) and gentle correction (as per Task 3) but also in pitch resynthesis (Task A). This premise is corroborated in Task 5, where participants group the World vocoder audio samples and rate them with similar scores. This conclusion is significant, as it means that the coloration provided by the World vocoder remains constant despite the different pitches imposed in each task, representing different scenarios. This result indicates that it is valid to discuss the colouration induced by a vocoder and that, although small, it will be statistically manifested. Furthermore, this means that performing autotuning (transparent, gentle, or extreme) with the World vocoder is not the same as doing it with ATA. The complete autotuning system (tracking+correction+vocoding) will vary according to the vocoder used. Therefore, we could have various autotuner options by changing the vocoder, much like the variety found with pedals or microphones in their respective sound contexts.

According to Task 2, there is a slight difference in the means for extreme pitch correction between the ATA and DPW pitch correction methods; statistically, a difference can be observed. This result means that although the visual differences for extreme correction are minimal, they are perceptible. The transitions in extreme correction occur at the note's attack, and since the attack is not necessarily stable in terms of pitch and we are more sensitive to the attack, we can perceive the slight sonorous differences in that region more easily. Therefore, differences in extreme correction that are almost indiscernible visually are perceptible sonorously.

The observed disparity between the full ATA protocol and the DPW+World protocol (Task 3) is primarily attributed to the vocoder rather than the pitch correction algorithm. The following facts support this proposition: (i) When using the World vocoder for both pitch correction methods (ATA and DPW)

(Task 4), the stimuli are perceived very similarly, with differences only noticeable to musicians, indicating that ATA and DPW have a minimal sonorous difference. (ii) In Task 5, it is verified that the difference is indeed due to the vocoder. Participants group assign the same mean value and distribution to the samples using the World vocoder, despite different pitch correction methods; moreover, both (World+DPW and World+ATA) are statistically indistinguishable from each other when compared to the full ATA treatment and the original sound. This result implies that the samples are perceived differently even with the same pitch correction method because the vocoder differs. Thus, pitch correction methods become secondary when a soft correction is compared with the original sound as a reference. So, the vocoder's coloration prevails over the pitch correction method's coloration. Therefore, regardless of the vocoder used, we can generate new autotune variants simply by changing the vocal transformation method.

Contrasting the results of Task 2 with Tasks 4 and 5, it can be observed that, although small, the coloration due to the pitch correction method exists but is more noticeable for extreme correction than for soft correction. We are comparing pitch correction methods: in Task 2 (extreme correction), the same vocoder is used, and the difference is small but statistically significant with a p-value ¡ 0.001 for all divisions of the panel of auditors. In Task 4 (soft correction), the difference is not statistically significant for non-musicians but is significant for other cases, although the p-values are higher than those obtained in Task 2. From this, the difference between ATA and DPW is more difficult to perceive for soft than extreme correction. Furthermore, as mentioned earlier, in Task 5, vocoders are grouped, but not the correction methods, confirming that the difference between correction methods is secondary and that the coloration due to correction methods is more important for extreme correction. Based on this result, developing pitch correction systems focusing on the attack is more interesting, as we are perceptibly more sensitive to that region. If we want to innovate in autotune, the most interesting region to work on would be the attack.

The main conclusions drawn from the two tests can be summarized as follows:

1. The coloration produced by the vocoder is dominant (Tasks A, B, C, and D) and can be affected more by extreme pitch correction (Tasks B and C) than by soft pitch correction. This conclusion implies that if we are interested in developing new autotuning systems, we can discover new sounds by first working with different vocoders and then innovating the extreme correction method; work on soft correction is secondary in terms of sonorous perception.

2. ATA is the most transparent vocoder in soft correction and pitch resynthesis, which is statistically indistinguishable from the original sound. This conclusion implies that, unlike the others, the ATA vocoder does not vary with the variable transposition over which it actually operates.

So, ATA is unbeatable in terms of quality. However, it also indicates the need to evaluate vocoders not only by constant transposition but also by variable transposition, which is not currently a standard method for assessing vocoder quality because it more pertinently reveals the coloration of each vocoder.

3. Although pitch correction (ATA or DPW) contributes to coloration and there is a statistically significant difference between these methods, the predominant influence in terms of coloration comes from the vocoder itself (Task 5). The results show that there is coloration due to the pitch correction methods. However, this is minor, to the point that listeners group the sounds of the World vocoder (despite the different correction methods) together, and ATA (soft correction) and the natural sound together. While improvements in soft correction can be made, they are more difficult to perceive and challenging to manage (in terms of developing a correction algorithm) than extreme correction improvements could be made.

4. The coloration associated with the pitch correction method is more noticeable in extreme autotuning than in soft autotuning (Tasks 1 and 2 versus 3 and 4). This conclusion is significant as it defines the most musically interesting path. Artists are often interested in creating new sound textures, hence their interest in autotune. Therefore, knowing that pitchcorrection methods and vocoders are more easily distinguishable in extreme correction than in soft correction, we can say that the most musically interesting approach would be to innovate pitch correction methods for extreme correction.

5. Using a vocoder like World facilitates the comparison of correction methods, allowing for a similar approach to the test presented here to evaluate other tone correction methods different from ATA and DPW. The fact that the World vocoder has consistent coloration under different experimental perspectives (tasks) allows us to evidence that, although we cannot use it as a replica, it does allow us to estimate to some extent if the coloration of any other varied parameter (in our study, only pitch) adds some coloration to the sound.

Chapter 5

Perspectives and Conclusions

This chapter is composed of three parts that integrate the research perspectives and the conclusions of our work. Regarding the perspectives, we have explored two topics. One of which is the sonorous description of the vocoder. We explored this topic motivated by the relevance of the vocoder as a transforming element of vocal quality, but above all because it is a primary element used to perform autotuning. The sonorous description seeks to give a guideline to propose a concept of vocoder quality analogous to vocal quality applicable to the vocoder used for tuning. At the end of the first section, there is a discussion on the perspectives of psycho-acoustical research on this topic.

The second section refers to interactive effects and their use for voice. This topic is based on developing new effects using hand gestures. Initially motivated by the desire to make a new interactive DPW, it was enriched by proposing other effect prototypes. The possibilities for future research on this topic are also discussed.

Finally, the contributions and conclusions of this thesis are addressed. This doctoral project contributes conceptually, algorithmically, and psychoacoustically to the topic of vocal audio effects (tuning, vocoders, and interaction). These contributions can be described in four subsections:

- The taxonomy of vocal effects
- The pitch correction methods
- Vocoders and their psycho-acoustical evaluation
- Psycho-acoustical Evaluation of pitch correction methods

5.1 Sonorous Description of the Vocoder

This section will examine the vocoder as a vocal effect from a transdisciplinary perspective. Our aim is to define a musically and technically applicable vocabulary (sonorous descriptors). From this foundation, we aim to demonstrate that the vocoder can be understood as an effect with its own distinct musical

identity (quality) derived from its sonorous descriptors. This approach can offer guidelines that serve as a roadmap for the vocoder's future development as a musical effect. As outlined in previous sections of Chapter 3, scientific and musical approaches to vocoder use differ. This divergence has implications for how the vocoder is perceived as an effect—specifically, what we can or cannot identify as its sonorous descriptors—and necessitates narrowing the discussion to the tuning of the vocoder's application.

The voice is an extremely complex instrument that has been modeled through the source-filter, which according to [Doval et al., 2003] "is made of a non-linear volume velocity source, which represents the glottal signal, a time-varying linear filter, associated to the vocal tract, and a radiation component, which relates the volume velocity at the lips to the radiated pressure in the far acoustic field". Such definition is compatible with the vocal generation concept and the description of the source-filter model given by [Henrich Bernardoni, 2015] (page 21 - section 1.1, page 23 sections 1.2 and 1.2.1) and the review of the model in [Ardaillon, 2017] PhD thesis.

From a musical perspective, vocal effects add a coloration to the vocal timbre, and artists seek this coloration as a musical ornament with which they can experiment and interpret in a unique way. Therefore, this aspect of coloration is of musical interest. The modification of timbre may be more evident depending on the technique used and can be analyzed through signal re-synthesis with the original pitch. In addition, the case of autotuning (integer part pitch) involves imposing given pitch curve on the vocal signal with a vocoder, it acts as an automatic gesture (within some harmonic rules) and and highlights other possibilities not visible in the original pitch re-synthesis.

5.1.1 Signal shape

Below, we will provide a qualitative description of elements to consider from a *qualitative* perspective regarding the changes evidenced by vocoder use. The most notable descriptor is **latency**, which is present in both cases of resynthesis with the original pitch and extreme autotuning pitch. After generating the output files, it can be observed that each vocoder exhibits latency, with retune having the highest latency and ATA having the lowest. Another evident changes that can occur are *signal shape alterations*. It is interesting to note, for instance, that while ATA for resynthesis or even for extreme autotuning preserves the waveform well at large scale but also zooming, inversely CIRCE, in both resynthesis and autotuning cases generates a visually different signal with several modulations along with pitch changes. Such changes are likely internally related to the logic of its algorithm. These examples can be seen in Figure 5.1. However, since many algorithms can generate waveform changes, we do not consider these changes as vocoder descriptors.

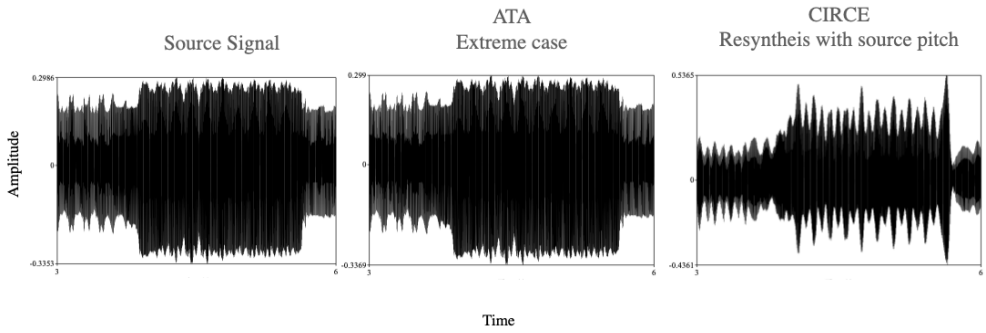


Figure 5.1: Signal shape differences for a audio segment on original file, ATA (extreme autotuning case), CIRCE (transparent case). While ATA preserve signal shape even for the case of extreme autotuning, CIRCE modify compeltely the signal for the resynthesis of original pitch case

5.1.2 Fidelity to the imposed dynamic pitch

The **fidelity to the imposed dynamic pitch** is perceptually verifiable, especially in the case of resynthesis with the original pitch. However, it has consequences for dynamic-melodic modifications when the pitch curve has been subjected to correction or modification methods such as ATA or DPW. Even the discretization present on the extreme autotuning differs between vocoders, as can be seen in Figure 5.2. Extreme autotuning can be helpful in exploring whether the loss (or preservation) of vocal quality. A parameter derived from the f0 modification is **f0-spreading**. When f0-spreading is high, the main note and all its harmonics spread, meaning that the perception of the f0 is less precise. In our sound catalog, f0-spreading is visible at several degrees for all vocoders, it is more evidently seen for extreme autotuning and it an be observed as grey vertical regions on the spectrogramme as it is shown in Figure comparing a source file with an extreme autotuning world re-synthesis.

5.1.3 Harmonic and non-harmonic changes

The spectral content can change drastically between vocoders because of the vocoder's own processing, as can be seen in Figure 5.3. The changes induced by the vocoder can affect the vocal timbre to varying degrees, prompting us to consider two descriptors for this purpose: harmonic coloration and inharmonic coloration.

Harmonic coloration considers both the amplification and/or modification of sub-harmonics and/or upper harmonics. Vocoder systems influence the output pitch, leading to changes in harmony that become evident due to such alterations. Additionally, amplified or degraded harmonics or replicas of them, may become audible, as is sometimes the case with the Retune vocoder. When

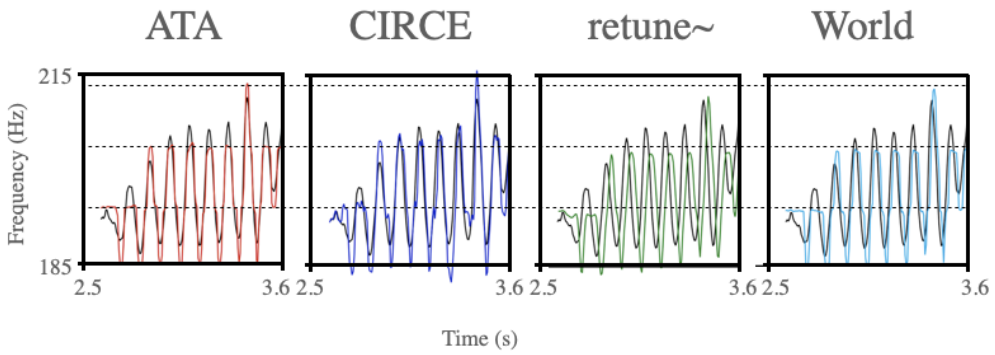


Figure 5.2: The fidelity to the imposed pitch is different for each vocoder as it is evident for the extreme autotuning case

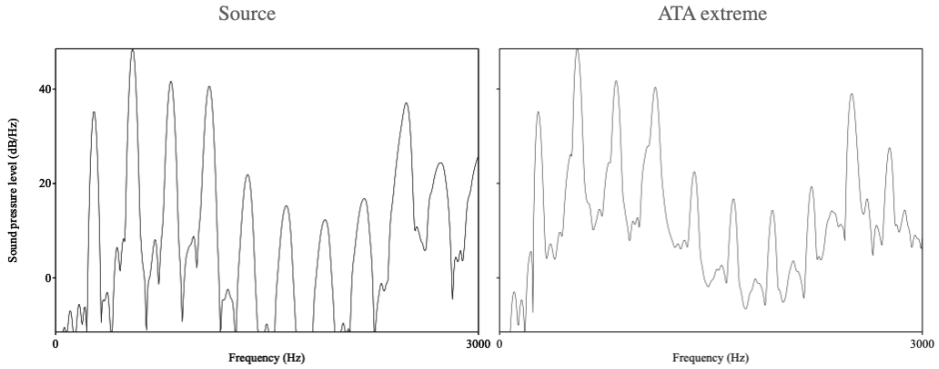


Figure 5.3: Spectral slice differences at time = 10s between the original sound and ATA re-synthesis in the case of extreme autotuning. Applying autotuning alters the spectral content; even though ATA is a high-fidelity system, changes in the spectral slice still occur because of the extreme autotuning.

vocoding and particularly when autotuning is applied, F0 and all harmonics are shifted, which modifies the spectral content and its variation over time. Additionally, autotuning can affect the intensity of the harmonics or the spectral background, as seen in the case of Circe, as visible in 5.4 for the spectral noise and the upper spectral content.

Moreover, the abrupt transition between notes generates an F0 dispersion in the neighborhood of the transition instant. As shown in the violet strips in figure 5.5, a blurry vertical gray line can be observed on the right side at the location of the transition due to autotuning f0-dispersion.

Harmonic coloration also encompass **formant modifications**, which are crucial as they relate to vowel articulation. Certain techniques are more or less transparent for formants, such as ATA and World when subjected to dynamic transpositions, as shown inf Figure 5.6. CIRCE is a vocoder that has already undergone quality verification for constant transposition values.

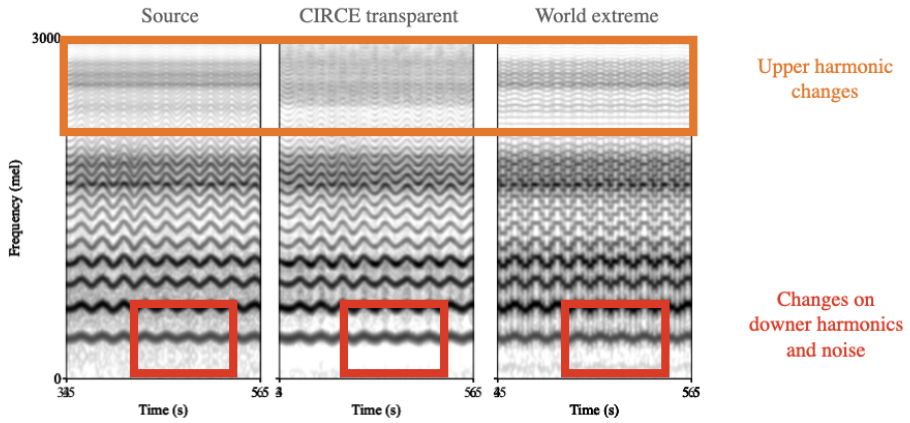


Figure 5.4: In orange, the difference in harmonics and spectral contact for upper frequencies. In red the variation of fundamental F0 in the cases: original, re-synthesis with CIRCE and autotuning extreme with ATA and the spectral noise variation.

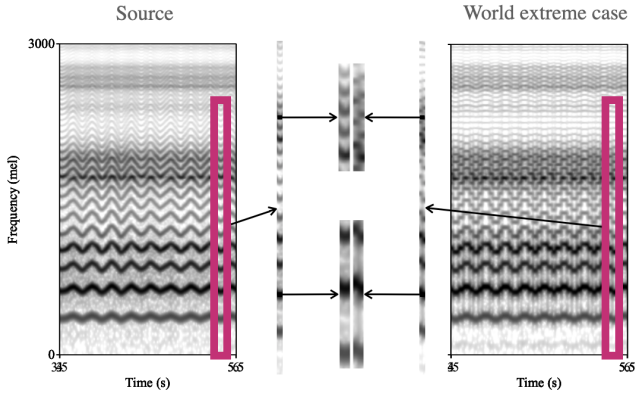


Figure 5.5: Spectrogram of the source file and extreme autotuning resynthesis compared for f0-spreading; the slices correspond to the moment of the sharp transition in extreme autotuning.

However, when subjected to dynamic transposition (extreme autotuning and original pitch), it is not transparent for formants. Thus, formant modifications provide another means of checking vocoder transparency, revealing an implicitly untreated difference between types of transposition and treatment through vocoders.

The other descriptive parameter we can utilize is **inharmonic deformation**, which involves residual noise in both the low and high-frequency regions of the spectrum, visible in 5.4. Furthermore, this descriptor significantly impacts the presence of noise around silences. A particularly visible example is the Circe vocoder, which erases all content during silent parts at the beggining and at the end of the audio samples.

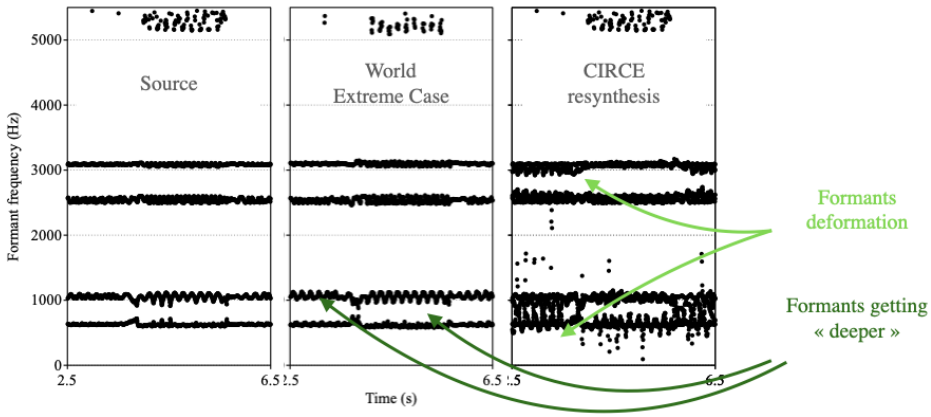


Figure 5.6: Formants modification for different techniques of vocoding and the corresponding deformation in formants, from Praat

5.1.4 Perspectives

The vocoder has evolved in two main directions: improving vocal transformation and parametric control of vocal characteristics. While the vocoder's artifacts are the more appreciable and applicable musically in modern music (for example, backing vocals with vocoded layers and autotuned voices), the human voice can be defined by its pitch range and timbre, which can be characterized by descriptors like dark, bright, soft, or noisy. The characteristics of the vocal sound, and also the vocal modes and styles, such as whispering, breathiness, and roughness [Loscos and Bonada, 2004], can be modeled and simulated using various vocoding techniques. Such research aims to improve the realism of vocal corrections with the vocoder, benefiting vocal disabilities, compression, communication, and music applications. However, the contemporary musical use of the vocoder often seeks its robotic and artificial sound as an intentional effect. The point we look to state is that musical use focuses on the scientific "defaults" and not on the perfection of the vocal transformation.

We propose four principal descriptors. The first one is related to signal shape changes as the modulations in amplitude and extreme changes seen in CIRCE, and almost nonexistent for ATA. The second is the fidelity to the f0-transitions, which we describe as a characteristic of the vocoder technique. This descriptor includes the f0-dispersion, which can produce more or less precise and realistic sound and is principally visible in autotuning. Still, it can also be present for the re-synthesis (affecting the weight of each harmonic). The third is harmonic coloration, which refers to sub-harmonic and upper-harmonic amplification. It can lead to a cleaner sound or add backing up or down voice (in the case of retune, it is sometimes audible). The fourth descriptor is inharmonic coloration, which concerns the amplification or coloration of parts where there is no vocal or instrumental sound signal, such as the more silent parts of the recording or the background noise. For example, the vocoder Circe shows a particular difference here compared to the other systems. It converts

background noise into complete silence and improves spectral noise, making it less present in the output audio file.

Beyond the significance of the sound descriptors, the work developed in this section allows us to understand that it is possible to develop a sound description of the vocoder separate from the vocal description of the voice in which the vocoder is used. Additionally, by combining this proposition with the results of the psycho-acoustic evaluation, we can explore new research paths. For example, we can define whether the proposed sound descriptors are perceptible by a panel of experts. Such a psycho-acoustic study could first identify if expert listeners perceive differences in the descriptors for different levels of sound treatment (ranging from soft to extreme correction), thereby contributing to developing new vocal effects different from those we already know. It could also help us link the technical terms used for the proposed sound descriptors with terms from the singing lexicon, providing guidelines for musical creation within a richer and more elaborated purposes.

Furthermore, this approach to the sound description of the vocoder for autotuning helps establish guidelines for the comprehensive description of the vocoder. For example, what happens when formants or breathing are modified? Do unique coloration artifacts from these processes appear? If that were eventually the case, would such coloration artifacts be useful for improving qualitative evaluations of the vocoder? Recall that this thesis has shown that using variable transposition is interesting for evaluating vocoder quality. Hence, the question arises: Would such artifacts from varying parameters other than pitch improve the qualitative evaluation of the vocoder? These questions open new research paths and could lead to a new era of vocal effects, introducing new lexical terms related to vocal effects and new ways to evaluate the quality of vocoding systems.

5.2 Exploration of Interactive Effects Using Motion Sensors

This section adds to the contemporary relevance of gestural control of interactive vocal effects. This mode of control is an intriguing component within the transformative process in modern music creation, and facilitating the diffusion of new technologies in what has been called the new digital lutherie. From an exploratory perspective, we are just beginning the examination of devices useful for capturing hand movements and presenting prototypes designed for the application of vocal effects. Even this section is at an early stage, we have developed a system that uses a sensor to track arm movements in the production of vocal effects. Through modular sound processing techniques, we have begun to capture and analyze users' body movements to create interactive vocal effects in real time. This approach allows us to tap into the creative potential of vocal effects, thus contributing to what could be called an augmented sound reality in the vocal domain.

5.2.1 A Brief Review of the State of the Art

As its name indicates, creating interactive effects through gestural control is based on the connection between movement and the modification of an audio effect. This concept closely aligns with the mapping definition given by [Rovan et al., 1997]. Our research focuses on explicit mapping [Hunt and Wanderley, 2002] and one-to-one strategies [Wanderley, 2001]. One-to-one strategies directly connect the gesture to the control variable, making them more comprehensible for the user and easier to implement, as described by Verfaille [Verfaille et al., 2006b]. The expressive power and consistency of direct relationships between gesture and sound effect are perceptibly limited [Rovan et al., 1997], but they have a tangible significance in the sound result. The advantage of explicit mapping is that it directly connects movement to sound, making it efficient and expressive. Therefore, we have decided to use explicit mapping with a one-to-one strategy. It should be noted that implicit mapping is an alternative approach that requires intermediate models to encode complex behaviour in the gestural interface. This approach has not been studied due to the organization of this thesis project.

It is also important to mention other techniques. For instance, authors like [Bowler et al., 1990] use point-based mapping, where tracking a finite set of points ensures the continuity of the mapping and, consequently, the desired gesture or sound effect. The limitation of point-based methods is that each preset must be specified manually rather than through continuous movements [Françoise, 2015]. However, the idea of creating a mental map of the gesture is important for calibrating a system based on explicit mapping and for the user to perform a geometric gesture linked to the sound, even if not all gesture variables are used (e.g., in the reverberation effect we address later), it allows for observing the parameter variations given by the sensor used and mapping it according to the established gesture, defining a reference and limit values. As stated by [Arfib et al., 2002] the transformation of related-to-gesture-perception parameters into related-to-sound-perception parameters is called the second mapping level.

Additionally, we want to mention other relevant works in this field, as the last two decades have been crucial for developing interactive instruments and effects. Technologies like Kinect or the phone have been adapted to protocols such as OSC and thus are musically usable on platforms like MAX. Also, there have been many conceptual advances. As some researchers mention [Godøy and Leman, 2009] [Delle Monache et al., 2018], understanding sound imagery helps reinforce action-sound relationships, which are fundamental for associating objects, actions, movement, and sound. Projects like MO have articulated this knowledge. Modular Musical Objects [Bevilacqua et al., 2013] [Rasamimanana et al., 2011] [Schnell et al., 2011] are based on devices with a variety of sensors such as accelerometers, gyroscopes, and piezo sensors. MO use machine learning models such as Hidden Markov Models [Bevilacqua et al., 2013], which enable the generation of sound textures (grain, shaking, sounding surface, and throwing a ball) linked to gestures. Machine learning models are

also helpful in addressing the problem of movement coarticulation [Bevilacqua et al., 2016], analogous to vocal coarticulation, meaning that the significance of a gesture depends on the context, especially in music. Coarticulation is one of the limitations of explicit mapping effects.

We want to highlight one of the key works in gestural control through movement. MI.MU gloves [Mitchell et al., 2012] [Mitchell and Heap, 2011], a project that results from the collaboration between Tom Mitchell and Imogen Heap. Initially presented as Sound Grasp [Mitchell and Heap, 2011], it is a gestural interface for live music performance using body movement and manual gestures with a glove. The gloves allowed manipulation of digital musical processes and created a direct connection between gesture and sound outcome through the use of finger flexion and abduction and Kinect.

Additionally, they required a neural network for their operation . This project was presented at TED Global 2011 and is available on YouTube under the title "Sculpting music with Mi.Mu gloves - Imogen Heap - TEDxCERN".

Finally, we emphasize that although this subsection addresses some prototypes based on explicit mapping, we are aware of its limitations. Nevertheless, we have sought to mitigate these limitations by establishing references for maximum and minimum values according to the gestures and effects used. Our objective is purely exploratory, and we have attempted to work with interactive autotuning, as it is interesting to apply it in different contexts of use, in this case, an interactive environment.

5.2.2 Gestural Control Devices

To investigate the musicality of the voice from an interactive control perspective, it is essential to recognize that, like any other sound, the voice has a spatial and morphological mental representation. This representation significantly contributes to sound perception through descriptors such as dynamics, space, time, and pitch. Therefore, a first step in exploring gestural control involves directing toward modifying these descriptors. Our objective is to establish a relationship between the performed gestures and the events of sound descriptor modification. This task naturally requires the utilization of effective real-time motion-tracking methods.

In our research group, we have conducted various studies focusing on gestural control within the realm of New Interfaces for Musical Expression (NIME). For instance, we have explored the use of handwriting gestures in projects such as Cantor Digitalis [Feugère et al., 2017], as well as the employment of the theremin and manual triggering in T-Voks [Xiao et al., 2019], and rhythm control through movement in Mono-Replay [Lucas et al., 2021].

Our approach has centered on the interactive control of modular digital vocal effects using the singer's own movement. This approach has been proposed due to its potential practicality within musical performance, allowing the lead singer or chorister to modify their voice through their own gestures. We have utilized motion tracking systems to achieve this interactive control, precisely

two Hot Hand devices and the BITalino R-IoT module. Next, we will proceed to explain the functioning of these devices.

5.2.2.1 Hot Hand

The Hot Hand USB is a device manufactured by Source Audio. It is a small wireless motion and tilt sensing ring in two axes, as shown in Figure 5.7. This device connects to a USB receiver via Bluetooth. It does not require drivers and appears as a MIDI device on the computer. The data obtained by the ring can be tracked in real-time from a DAW or from a dedicated monitoring application provided by the manufacturer. Additionally, this application allows filtering the signal to obtain a smoother output. It is a device designed specifically for dynamic control of musical effects. This device is highly useful due to its ease of connection and tracking and it operates within a range of 5 meters. However, its limitations include longer distances, and the lack of information about the orientation of the person wearing it, as shown in Figure 5.8.

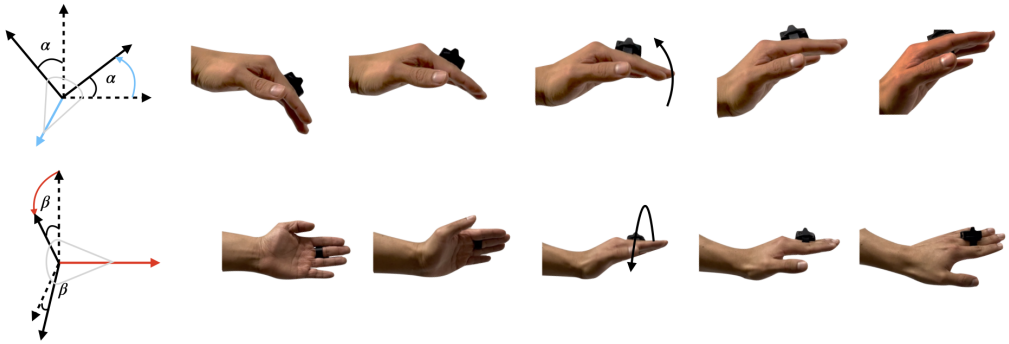


Figure 5.7: Elevation and rotation tracking with HotHand and Bitalino Ri-ot devices

Figure 5.8: Orientation tracking with Bitalino Ri-ot device but not with Hot-Hand device

5.2.2.2 Bitalino

The BITalino R-IoT module, developed by IRCAM, is an Inertial Measurement Unit (IMU) designed to capture motion wirelessly with low latency and high data speed, connected to a computer via WiFi. This unit allow to capture of body motion, including inclination, rotation, and orientation. It has fewer

limitations than the Hot Hand, as it operates within the WiFi router range and allows for elevation, rotation, and orientation capture in three axes, as show in figures 5.7 and 5.8. However, its drawback lies in being an open unit. Thus, we have made wristbands to carry the sensor. The BITalino R-IoT module incorporates a 9-axis sensor (LSM9DS1), which includes a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer. This allows for obtaining absolute orientation in space, providing precise motion tracking capabilities. By leveraging this technology, users can interact with their voice, as we will see later, by mapping the movements of their hands to control digital vocal effects.

5.2.2.3 Tracking

Data tracking is performed through two different protocols:

- The HotHand device connects via Bluetooth to a USB port, and then, with the dedicated application *Source Audio Hot Hand*, data can be received, filtered, and sent to MAX, as shown in Figure 5.9.
- To track the BITalino R-IoT device, we used the MAX BITalino packages and the configuration provided by IRCAM. The device is configured to connect via a router through Wi-Fi. The router is connected to the computer either through LAN or Wi-Fi. On the computer, data is received directly into the MAX object packet from IRCAM dedicated to connecting with the BITalino R-IoT device, as shown in 5.10.

With these two protocols, we can effectively capture hand movements and orientation, which is essential for interactive control of vocal effects.

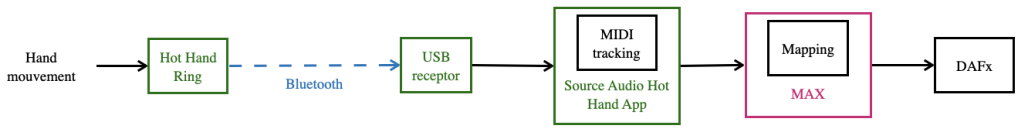


Figure 5.9: Tracking data from Hot Hand device

5.2.2.4 Mapping

The hand's elevation, rotation, and orientation can be obtained through the information from the sensors acquired in MAX. Although the sensors have a full range for each variable, in reality, the user will have a limited range for each of them. While some full-hand movements are possible, they may not be well-suited for optimal live execution. Therefore, it is necessary to consider the

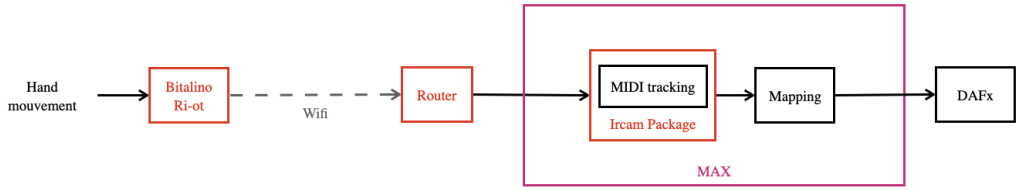


Figure 5.10: Tracking data from Bitalino Ri-ot device

constraints that the user would eventually encounter. As depicted in Figures 5.11, 5.12, and 5.13, elevation, rotation, and orientation, respectively, have their own limitations. Thus, a part of the mapping process involves identifying these constraints and adjusting the operating range according to the individual using the device. This constraint within the data range assists in filtering and controlling the captured data, thereby avoiding erroneous interpretations of gestures.

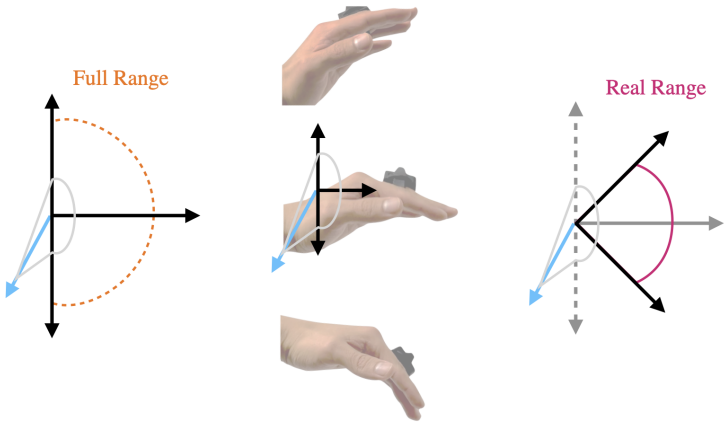


Figure 5.11: Elevation limitation

5.2.3 Pitch and Spatial Perception Exploration

We will now focus on exploring vocal modification through two sound descriptors: pitch and spatialization. Our objective is to propose an immersive performance and listening experience where physical movement directly influences the perception of the source's pitch and the surrounding sound space.

Pitch exploration has been approached through two methods. Firstly, harmonization involves creating multiple vocal layers composed of copies of the original signal transposed in real-time. The number of semitones transposed is controlled according to orientation, and the transition time of transposition is

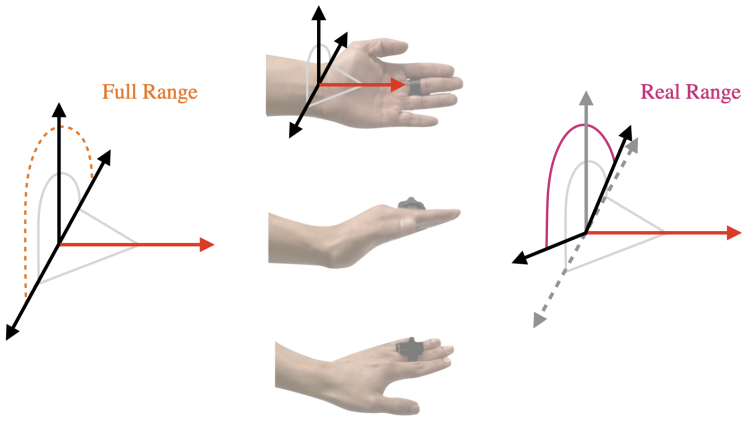


Figure 5.12: Rotation Limitation

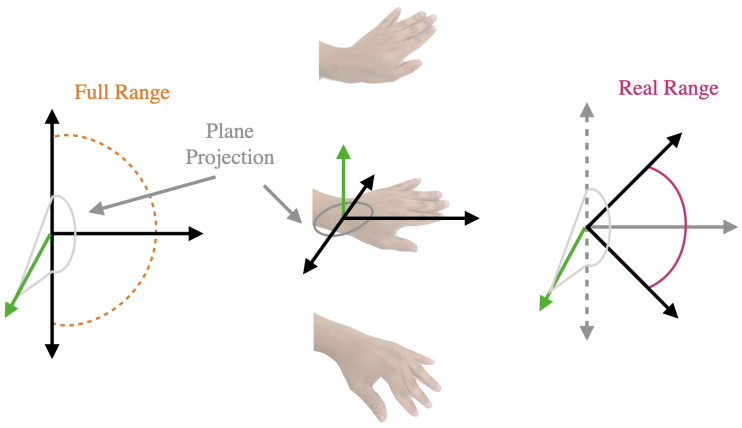


Figure 5.13: Orientation limitation

adjusted based on inclination. The transition time of transposition is the time to transpose given number of semitones, making a line from 0 to the number of semitones. This effect, known as vocal layer harmonization, influences pitch, harmony, and vocal presence. The second technique addressed is interactive retuning. While the term autotuning has been extensively discussed in this manuscript, we differentiate its employment here by utilizing a real-time version of the DPW algorithm with the audio signal and the Retune vocoder. Interactive control is achieved by relating movement to the transition time parameter.

Spatial sound exploration has been conducted in two ways. Firstly, by altering the position of the source through panning, and secondly, by modifying

the environment where the source is situated through reverberation. In the case of panning, hand orientation is directly related to panning. Regarding reverberation, we dynamically adjust the virtual room size and cutoff frequency according to the user's movement.

5.2.4 Interactive tuning

In this manuscript, we have thoroughly explored the autotune effect, a widely used tool for melodic modification, particularly in vocals, systematically applied in music production. This effect can range from extreme settings that result in noticeable pitch changes to more subtle and transparent configurations. Among the available tools for autotuning, Antares Autotune stands out as a widely disseminated and utilized system. However, we have developed our own tuner using the objects available in MAX.

For our interactive proposal, we have developed a patch in MAX utilizing fzero as a pitch tracker, DPW as a pitch correction method, and retune as a pitch modulator, as depicted in Figure 5.14. Mapping is achieved by estimating the rate of elevation changes, with values assigned to the transition time parameter; such values are restricted by hand movement. The estimation of elevation velocity is performed with a 100 ms window and is given by $\frac{x_f - x_i}{0.1}$, where x_f is the real-time elevation value and x_i is the elevation value with a 100 ms delay. Auto-tuning is only performed when the hand is in constant motion cause it depends on the elevation velocity. If the hand does not move, then autotuning is not performed.

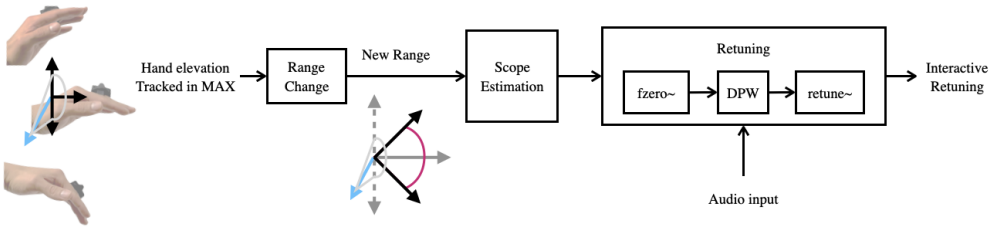


Figure 5.14: Interactive Tuning Schema. The sensor detects hand movement and receives it in Max. However, the actual range of hand movement is smaller than the range of values handled by the sensor. Therefore, the movement range limits must be restricted (using zmap function in Max). Next, the velocity is estimated over a 100 ms window and smoothed with a 100 ms filter. This signal is used to control the transition time of DPW.

5.2.5 Interactive Pitch-shifted Vocal Layer

In vocal compositions, whether in the realm of classical lyrical music or contemporary music, the use of vocal layers is fundamental. In modern music, employing layers of voice is of significant importance, often involving resynthesized, transposed, filtered, or vocoded voices. Harmonization with vocal layer

can be achieved with pitch shifts that directly produce the harmonization. The idea is to control the pitch shift through gestures, determining the number of semitones upward or downward desired in the vocal layer. This modification is essential and closely linked to musical concordance, both in performance and contextual terms. Consequently, harmony and dissonance can manifest and acquire musical significance depending on the context. The decision on this matter lies with the composer. Our goal is to provide possibilities through gestural control.

We have developed a mapping for the rotation of the user's hand, allowing rotation range limits to be incorporated. This mapping assigns hand rotation to real values ranging from 0 to 12 semitones. The calibration process for the device is as follows: First, the user agrees on a comfortable hand position, which is taken as the reference. Then, the user moves their hand to define the upper and lower limits. To handle asymmetry, the range is divided: the portion below zero is processed with one gain value, and the portion above zero with another gain value. This process is not continuous, but it is simply a prototype.

System activation occurs through elevation when it falls within a specific range, delimited according to the user. A faster or slower transition between the original audio and the transposition occurs depending on the inclination value. Transposition is carried out using DPW and a Warper based on Retune. In summary, the operation is straightforward: hand rotation controls the semitones to transpose, while elevation controls activation and transition speed. A flowchart of the implementation is shown in the Figure 5.15.

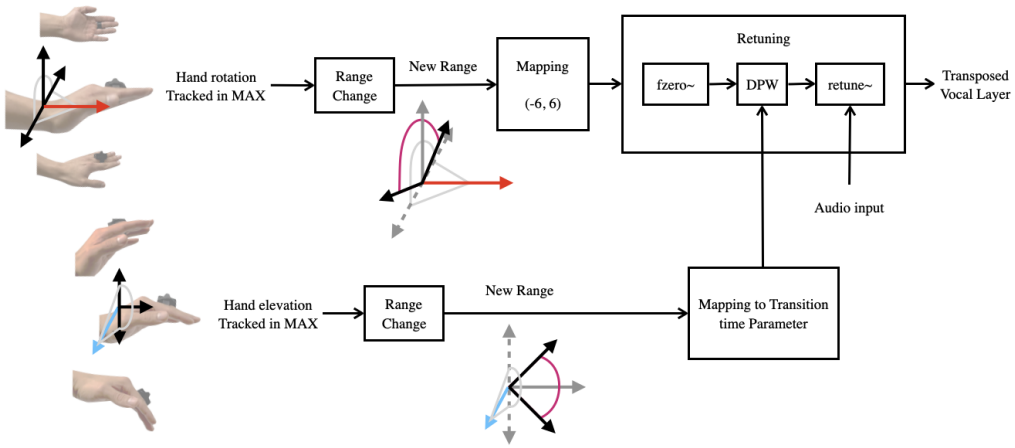


Figure 5.15: Pitch-shifted vocal layer by hand control. The movement range limits are calibrated for the individual user; both hand elevation and rotation can be used. Asymmetry is managed by applying one gain below zero and another gain above zero.

5.2.6 Interactive Reverb

Evolution has endowed our ears with the ability to discern localization, environment, and texture through reverberation. This phenomenon, characterized by sound reflected within a space, is fundamental in acoustics and has been implemented since the earliest days of music. Reverberation varies depending on materials and dimensions, and its spectral response can change due to the geometry and acoustic enhancements of the design. The study of reverberation addresses various aspects, such as sound intelligibility, segregation of sound sources, spatial localization, and auditory distance perception. In sound engineering, audio effects are utilized to alter the spatial perception of sound. Reverberation is the primary audio effect and is crucial for both performers and listeners, influencing musical expression and the perception of sonic space.

The implementation of reverberation through hand gestures is based on the observation that people naturally use hand gestures while speaking and singing [Fulford and Gingsborg, 2013]. These gestures can convey the breadth of space and location in time and space. It has been found that musicians, including the visually impaired, use gestures to communicate during performances. The hand position is directly related to the spatial position of the voice, so broad gestures indicate an expansive performance, while gestures closer to the body indicate a more intimate performance. To control the reverberation, we have relied on a simple gesture: the proximity to the hand and the projection of hand movement, as shown in Figure 5.19. These two positions are considered the minimum and maximum limit positions of reverberation. Mapping is done with rotation within a range of approximately 45 degrees, adjustable according to the individual using the effect. A flowchart of the implementation is shown in the figure 5.16.

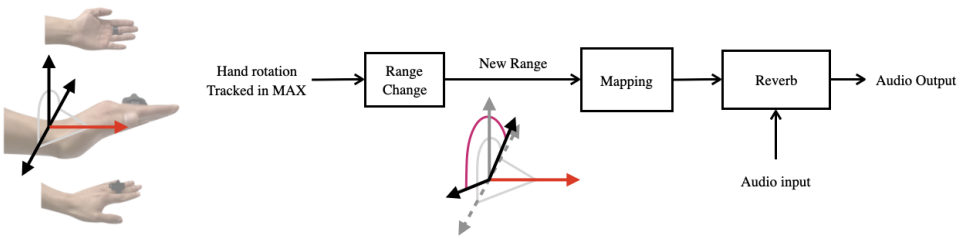


Figure 5.16: Interactive reverb schema

5.2.7 Interactive Panning

Interactive panning has been implemented based on hand rotation for the HotHand sensor (due to the inability to obtain orientation) and orientation for the BITalino R-IoT sensor. In the case of rotation, the user's range of use and the center, when the user desires equal balance in the left and right speakers, are defined. The mapped data is then sent to a MAX panning module. For the BiTalino device, the user is asked to indicate the positions of the speakers

and the center, which delineate the range of use. Orientation is mapped in this case, and the delimited data is also sent to a MAX panning module. A flowchart of the implementation is shown in the Figure 5.17.

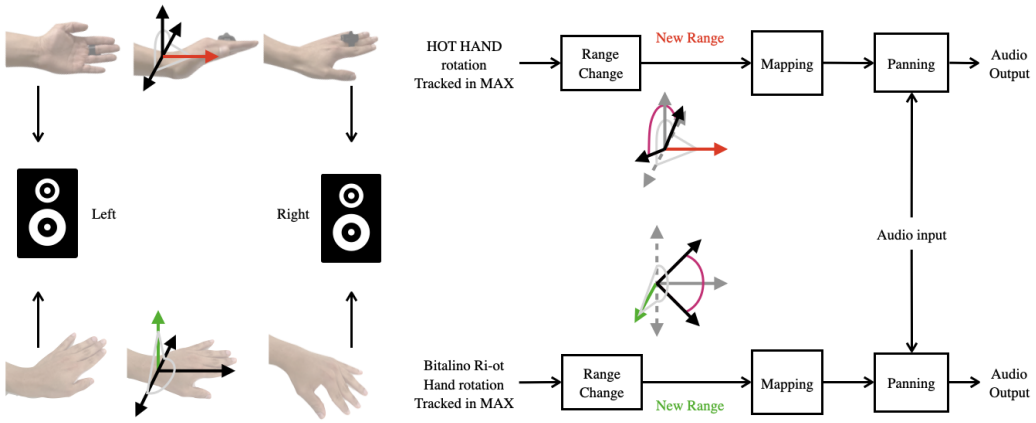


Figure 5.17: Interactive panning schema

5.2.8 Prototypes

The mentioned effects have been explored experimentally using the HotHand device. In the left side of Figure 5.18, we can see blurred image to identify the movement when doing interactive tuning. In the right side of 5.18 we can see the initial and final positions for vocal layer harmonization.

Figure 5.18: Left: Fast moving for controlling tuning by elevation tracking with Hot Hand. Right: Hand rotation to control harmonization with Hot Hand

Interactive reverb was achieved using the initial and final positions shown in Figure 5.19.

Panning has been implemented using both the HotHand device and BiTalino Ri-ot. The positions utilized for rotation mapping with the HotHand device

Figure 5.19: Rotation controlling reverb with HotHand

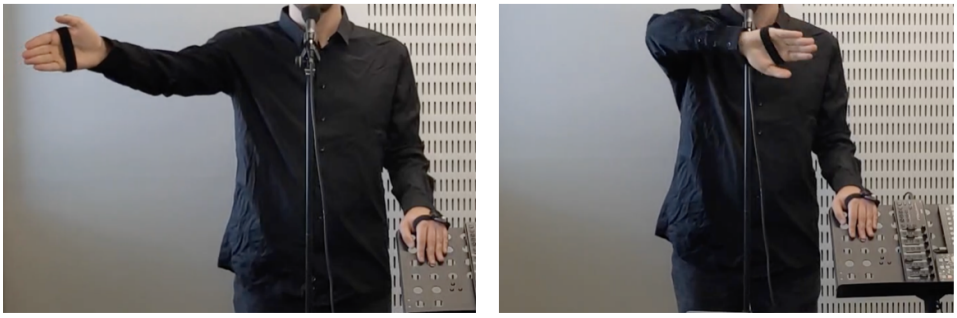


Figure 5.20: Orientation using Bitalino Ri-ot device, descriptive example

can be seen in Figure 5.18, while the positions for orientation can be observed in Figure 5.20.

Interactive tuning was worked with elevation by fast movement, so we do not show a picture of it here, but we provide a video support. The video support includes the following examples:

- Interactive harmonization by vocal layers
- Interactive Autotune
- Interactive Reverb
- Interactive Panning

Additionally, there is a supporting video of the article presented at the International Symposium on Computer Music Multidisciplinary Research.

5.2.9 Perspectives

The primary focus of this section has been exploratory, aiming to design prototypes amenable to musical exploration. Our contribution was incorporated into the demonstration presented at the CMMR2023 conference, the 16th edition of the International Symposium on Computer Music Multidisciplinary Research. In this demonstration, we utilized the musical proposal of G. Locqueville and

T.Lucas interpreting the Beatles, in collaboration with other approaches related to gesture use from our research group.

It is noteworthy that prior research has evidenced significant variability in listeners' preferences regarding sound and gesture characteristics, influenced by their criteria, such as preference for speakers' sound color. The inherently subjective nature of timbre appreciation reflects the diversity of musical genres, vocal types, and arrangements that may appeal to some individuals but not to others. This variability also extends to the realm of digital audio effects and gesture usage.

The creative process involved in the experience presented within the framework of the International Conference on Multidisciplinary Research in Computer Music (CMMR) revealed a significant disparity in the qualitative appreciation of comfort and applicability of manual gestures among device users. It is emphasized that such characteristics vary depending on the individual using the prototype and their preferences and specific usage goals. Therefore, they are not necessarily negative. A thorough and dedicated study is still needed to determine which gesture is indeed most convenient for each effect and how to optimize its musical use.

In our experience, there was an initial vision of the effects. However, a technical-creative process became necessary when confronted with the performers due to the discrepancy between the initial proposal and the execution preferences. Only through constant and fluid discussion was it possible to agree on the best way to map the gestures and effects. During the gesture selection process, it was observed that these varied slightly depending on the person using them. For example, the hand's rotation or the arm's movement from side to side differs from person to person. These variations are not necessarily due to the person's size but rather to how each individual comfortably executes the gesture. Therefore, discussing the comfort and feasibility of the gestures employed became important.

Although this is not about formal subjective evaluations but a creative process, this process itself shows that there are important variables to consider in developing interactive effects. These variables can include the desired gesture according to the performer's sound imagery of the effect they want to use and the intentionality (for example, performing a pitch shifting by speed or orientation). The variation of these gestures from person to person can be due to physiological factors, such as the individual's size, but also to the manner, personality, and precision with which the gesture is executed. Additionally, it is crucial to consider comfort, i.e., how the gesture is articulated, when, and for how long. Finally, reproducibility must be considered: if the gesture is comfortable, well-calibrated, and musically meaningful, it should be reproducible by the same person at different times without recalibration.

From our creative experience in this work, we have identified that these issues could be studied and addressed in depth. Without undermining the work done, we consider explicit mapping an interesting creative solution but is not consistent enough (as other authors mention) to be practically applied

in the musical field without a computer musician. In the future, it would be beneficial to experiment with real-time, low-latency gesture recognition techniques combined with explicit mapping.

5.3 Conclusions and Contributions

Through this research work, we have contributed to the modern use of autotuning as a vocal effect. We have comprehensively addressed the pitch correction topic and studied the vocoder and autotuning basics while managing the future research lines that may emerge from our findings. We conducted a varied theoretical review that includes vocal effects preceding the vocoder, not necessarily from a mathematical perspective but from a musical standpoint, given their use in vocal use. This approach has allowed us to position the vocoder within the taxonomy of effects and propose a perceptual classification.

Additionally, we have established a use framework for the pitch modification, clarifying concepts that, although frequently used in vocoders and autotuning systems, are not well-defined or theoretically established. This lexicon enabled us to approach the subject appropriately when conducting a psychoacoustic evaluation. Moreover, this lexicon holds potential for future, more in-depth research and the design of new autotuning systems that are more musically engaging.

The psychoacoustic evaluation of both the coloration produced by the vocoder and that caused by pitch correction methods allows us to define better how innovation in such algorithms is possible and understand the differences and similarities between ATA and DPW systems. Finally, the work related to the sonorous description of the vocoder and the development of interactive effect prototypes opens the door to expanding the field of autotuning study from a multidisciplinary perspective, integrating disciplines such as signal processing, psychoacoustics, and musicology.

5.3.1 Taxonomy of Vocal Effects

The taxonomic study aims to place the vocoder within the realm of sound effects, with a particular focus on a taxonomy of vocal effects. This classification allows us to address the subtleties of vocal use while avoiding tedious details that are more relevant to the general taxonomy of effects proposed by Verfaille. Not all effects are applicable to the voice, and not all effects carry the same level of importance in vocal applications. Hence, developing a specific taxonomy for vocal effects is convenient and practical.

The taxonomic study also takes into account the perceptual nature of audio effects. The effects applied to vocal tracks are categorized into pitch, space, time, and timbre (vocal quality). The concept of timbre is more complex in this vocal taxonomy than in the general taxonomy of effects, as it encompasses the preservation, modification, or destruction of vocal quality.

Preservation refers to making subtle modifications to vocal characteristics without losing the original vocal quality of the singer. Modification involves completely altering the vocal quality, making the singer sound like a different person or entity. Finally, the destruction of vocal quality refers to effects that partially or entirely disrupt the original vocal quality.

The primary types of effects have been exemplified through their usage in modern popular music, their primary using domain. In the future, a more in-depth study of the techniques used for each type of effect would be of great interest from a musicological perspective. Additionally, we have introduced the concept of modularity in audio effects, where effects chains are generated, resulting in the perceptual dominance of one particular effect.

Autotuning is an effect that encompasses both pitch and timbre. This effect introduces a coloration due to the imposition of a specific pitch contour. Through psychoacoustic study, we have explored which factor carries more weight in this coloration: the vocoder or the method of pitch contour correction.

The deliverable of this section is a comprehensive perceptual taxonomy that takes into account:

- The emphasis on effects used in vocal tracks.
- A perceptual standpoint related to the vocal quality.
- A sound support based on popular musical recordings.

The classification comprises five main categories of vocal effects: dynamic, temporal, spatial, melodic, and timbral. Furthermore, a special division has been established for timbral effects, subdivided into those that preserve, distort, or transform vocal identity. Additionally, it has been emphasized:

- The modularity when applying effects and the importance of their practical utility over the underlying algorithm.
- The relevance of the vocoder and melody in tuning, raising questions about their impact on vocal identity and importance on musical application with a musical support reference.

The taxonomic classification we have developed has allowed us not only to achieve immediate results but also to consolidate the application of interactive digital effects in a more structured manner compared to previous projects. New prototypes of interactive effects were designed directly in Ableton using Max mapping, and a demo presentation was successfully showcased at the 16th International Symposium on Computer Music Multidisciplinary Research in Japan.

It is plausible to advance much further based on the established taxonomy. The first step would be to consolidate the classification of effects from a psycho-acoustical perspective. Through a psycho-acoustic evaluation, modular

effects could be analyzed through several processes. The first process would involve classifying modular effects using samples created by researchers. Such classification would include the five categories outlined in our taxonomy (dynamics, pitch, space, time, timbre/vocal quality). As a result, we would have an initial indication of whether listeners can identify the various effects in the audio samples.

Additionally, samples of popular music with musicological support could be used to determine which effect of the vocal chain effects listeners perceive first, allowing for a comparison between the intended musical effect and what listeners actually hear. Finally, it could be interesting to configure the vocal effect chains with variable presets to study if there is any perceptual threshold where one effect predominates over others. A study of this magnitude would require considerable time and a specialized panel to conduct, and it could be a medium-term research proposal.

Another research perspective could be a musicological study that examines the different musical objectives of the vocal effects from a stylistic point of view according to musical genres. This study should consider the purpose of the musical message that the producer intends to convey, taking into account the musical genre and the listener's perception. Such a study would be valuable as a snapshot of current music and vocal production.

Also a deliverable related to the interactive effects has been developed. Hand movements related to elevation, rotation, and orientation have been utilized, and prototypes of the following effects have been created:

- Interactive tuning
- Interactive harmonization by vocal layer
- Interactive reverb
- Interactive panning

Regarding interactive effects, there are promising prospects in the area of gesture control. It could be interesting to explore further by employing gesture recognition techniques and combining sensors with camera-based motion recognition methods. Psychoacoustically, we could study how to improve the use of gestures and how they relate to the sound image and musical practice.

5.3.2 The Pitch Correction Methods

Through the study of Dynamic Pitch Correction, we have clarified and consolidated a framework for studying not only the DPW method but also the tuning or pitch correction concept in its entirety. Despite the variety of corrective methods available and the widespread use of effects such as ATA and Melodyne, there is no comprehensive theoretical framework for the study of tuning. We have worked to establish these foundations, including:

- Types of pitch transposition (constant and variable).
- Purposes of pitch correction (tracking, tuning, warping).
- Uses of pitch transposition and autotuning in musical recordings.
- Types of correction (extreme, transparent, expressive).

These elements have been fundamental in determining what differentiates DPW from ATA, and they can be extended to any pitch correction method. Not only that, but we can now define what is sought in a pitch correction method, and we have made corrections to previous works where there were inaccuracies regarding ATA's capabilities. Surprisingly, ATA does not differ significantly from DPW in terms of the algorithmic outcomes it generates. Additionally, DPW has several limitations due to how its parameters have been designed. Today, we can assert that the primary difference lies in the trade-off between the staircase effect and the freeform section and that both methods successfully recover the expressive component of the signal (DPW does it more symmetrically).

The study of pitch correction that has been conducted provides us with a theoretical basis for algorithmically analyzing pitch correction methods and proves useful for conducting the psychoacoustic evaluation of these methods. The definitions of baseline cases and presets serve as a foundation for designing the psychoacoustic tests in Chapters 4 and 5, aimed at exploring the coloration in tuning caused by the vocoder and the pitch correction method.

The perspective on this topic is well-defined: it is necessary to develop new pitch correction methods that truly improve upon existing methods and do not have such pronounced limitations regarding parameter variation. We recommended using platforms other than Max that allow for more effective signal processing and are more compatible with development processes. Therefore, New pitch correction methods must address, first, the trade-off between staircase and free-path effects, and second, they must manage the vibrato component as effectively as ATA or DPW.

5.3.3 Vocoders and their psycho-acoustical evaluation

The psychoacoustic study is an immediate application of the theoretical framework for pitch correction methods that we have developed and, indirectly, of the vocal effects taxonomy of vocal effects. The taxonomy allows us to place the vocoder in a privileged position as a vocal effect capable of varying timbre and thus altering vocal quality. Our study of the vocoder has allowed us to identify the intersection between the scientific study of the vocoder and its musical application. These two approaches differ significantly: while scientists strive to continuously improve vocal transformation and to search for a complete parameterization, musicians focus on modifying vocal quality. Additionally, we have emphasized the importance of studying the vocoder as a key

component of pitch correction, as imposing a specific pitch curve (or a given transposition) requires a vocal transformation algorithm, such as the vocoder.

Musically, both the vocoder and autotune saw a significant rise in usage during the same period and are now essential elements in the production of Anglo-American popular music and international genres like hyperpop. The vocoder is used to enhance vocal presence, create vocal layers to improve texture, modify formants, alter vocal quality, and perform tuning.

We have conducted a subjective psychoacoustic evaluation to compare the sound of four systems: ATA, World, Circe, and Retune. Using these systems, we created a sound catalog and prepared short samples for a psychoacoustic evaluation. The cases examined included original pitch, extreme correction, and soft correction. Four comparisons were studied: comparisons between original pitch cases, comparisons with extreme autotuned cases using the original sound as the reference, extreme autotuned cases using extreme autotune with ATA as the reference, and soft correction cases. The main conclusions drawn can be summarized as follows:

- The perception of the reference remains consistent no matter the case (resynthesis or extreme autotuning).
- ATA is identified as transparent in the softest possible configuration (Task A) for "re-synthesis".
- Extreme autotuning with ATA and World equally deviates from the original audio (Task B).
- Extreme autotuning with World is the closest to ATA (Task C), but there exists a coloration difference, which is evidenced by statistics results.
- Soft autotuning with World and Retune closely resembles that of ATA (Task D) in mean values, but statistically, there is a difference in relation to ATA. For the full panel, World and Retune are equal between them; nevertheless, when dividing the panel into musicians and non-musicians, it is not possible to conclude about the similarity between Retune and World.
- World shows a slightly superior response to Retune in terms of its similarity to ATA, considering Tasks A, B, C, and D.

The main result is that the vocoder plays a vital role in the coloration that occurs during pitch correction, regardless of the type of pitch correction applied. Even with the most transparent vocoder, some coloration is still present. The study also shows that musicians are more sensitive to detecting differences. The immediate consequence of this research is that it serves as a reference for the psychoacoustic study of pitch correction, specifically to determine whether pitch correction methods are more perceptibly distinguishable from each other than vocoders. Such a study, in return, would help to identify whether the

primary contributor to the coloration in autotune is the vocoder or the pitch correction method itself.

An indirect perspective of our study on the vocoder has been the development of a sonorous description of the vocoder for autotuning, which we have undertaken in this chapter. This approach allows us to establish a framework for defining a vocoder quality analogous to vocal quality, with specific sonorous descriptors for autotuning:

- Signal shape.
- Fidelity to the imposed dynamic pitch and f0-spreading.
- Harmonic coloration (in the harmonic content of the signal).
- Non-harmonic changes (transients and noise).

As mentioned earlier in this chapter, a medium-term research perspective for the vocoder could focus on the psychoacoustic study of the defined sonorous descriptors to assess whether they are perceptually effective and technically feasible and how they could be measured. Additionally, we could explore processes other than tuning by adopting a scientific structure similar to the one used here. For instance, if a vocoder is used to add roughness to the voice, and a study on roughness in the vocoder would be desirable, we would begin by identifying perceptual sound descriptors of vocal roughness. Subsequently, we would evaluate different vocoders and levels of any auditory vocal descriptor to compare vocoders, following a psychoacoustic comparative protocol similar to the one we have developed

5.3.4 Psycho-acoustical Evaluation of pitch correction methods

Furthermore, we have conducted a subjective psychoacoustic evaluation to compare the pitch correction methods ATA and DPW under various conditions. The objective was to determine if the methods are perceived differently and to investigate the prominence of the vocoder or the imposed melody in terms of vocal identity and using full ATA protocol and an ATA replica (using the World vocoder). The main conclusions drawn can be summarized as follows:

- The World vocoder introduces a consistent coloration in all cases.
- According to Task 2, there is an insignificant difference in means for extreme pitch correction between the ATA and DPW pitch correction methods. However, they are statistically different.
- The observed disparity between the full-ATA protocol and the DPW+World protocol (Task 3), according to statistical results, can be primarily attributed to the vocoder rather than the pitch correction algorithm. With

similar results as Task 1, the shape of the distributions is resembling while the only similarity between both tasks are the vocoders of the reference and the comparable stimuli.

- When employing the world vocoder for both ATA and DPW correction methods (Task 4), they are perceived as virtually indistinguishable by means, but the difference is less evident statistically because we get higher p-values (still in the interval of confidence for the full panel, but not for divided in musicians and non-musicians panels)

- Contrasting results from task 2 to tasks 4 and 5, it can be seen that, although small, the coloration due to the pitch correction method is more noticeable for extreme correction than for soft correction. In task 5, in particular, it is not anymore possible to differentiate pitch correction methods when using World, when the reference is part of the comparable stimulis.

The main conclusions drawn from the two tests can be summarized as follows:

- The coloring produced by the vocoder dominates (Tasks A, B, C, and D) but may be influenced by the extreme pitch case (Tasks B and C) as each vocoder can interpret the pitch curve slightly differently.

- Each vocoder has a consistent coloration over tasks A, B, C and D. ATA stands out as the most transparent vocoder in soft correction and resynthesis.

- Although pitch correction methods (ATA or DPW) contribute to coloring, the predominant influence comes from the vocoder itself.

- The coloring associated with the pitch correction method is more noticeable in the case of extreme autotuning than in soft autotuning

- There is a scale effect and likely a perceptual effect when the comparable stimulus is identical to the reference, which systematically repeats in all the results.

The study has two main conclusions. The first conclusion is that the vocoder coloration is more relevant than the pitch correction method coloration. The weight of the vocoder's coloration happens because the vocoder interprets and forces a given pitch curve in the audio sample. If new fast enough real-time pitch trackers and vocoders appear, we can have many autotuning textures depending on the vocoder employed. The second conclusion is that pitch correction methods can be differentiated more easily by extreme correction than by soft correction, and that is very important because it gives us clues for creating new vocal effects. We perceptually privilege the transient,

so everything that happens with extreme pitch transitions will be perceptually stimulating and, therefore, musically useful.

On the other hand, the soft correction does not allow us to differentiate DPW from ATA easily. So, the improvement proposed by the authors of DPW is not really perceptible. It would be necessary to make another type of evaluation that considers a tuning reference to see if there is a difference concerning intonation rather than coloration.

As a final point, we emphasize that the structure and protocol of the subjective psychoacoustic evaluation method for comparing isolated pitch correction methods with the World Vocoder can be a basis for comparing other methods and tuning conditions in the future.

Appendix

Appendix A

Signal Processing and Vocoding

A signal can be decomposed and represented in terms of its partials. These partials outline a spectral content which, in turn, can be visualized as an ordered series of frequency bands, each with different amplitudes that reflect the spectral weight within that specific band. In their early days, filters, whether infinite or finite, emerged with the purpose of modifying the spectral content of electrical signals. It was this transformation power that eventually led to their adaptation in the musical field, being initially implemented in electric guitars and later in synthesizers.

In filter design, it is crucial to analyze the signal flow. This analysis not only provides insights about the stability of a filter but also allows determining its response in the frequency domain. From this, the quality and precision with which a filter operates is derived. Parametric filters, on the other hand, are composed of key elements such as gain, cutoff frequency, bandwidth, and the Q factor, among others. This configuration provides a more detailed control of the filter's response. Additionally, by allowing the temporal variation of these parameters, what we know as time-varying filters emerge. A clear example of this category, widely popular in genres like rock and funk, is the wah-wah filter.

The digital age has also revolutionized the world of signals, promoting the emergence of digital filters. These began as simulations of acoustic and electric effects, but over time, they evolved into a fusion of traditional techniques with innovations inherent to digital electronics.

To work with digital audio effects, we must understand how to generate a signal acoustically (physically), the algorithms, and their musical applications. A musical signal must be captured (through an Analog to Digital Converter ADC); transformed (digital filters or effects), and reconstructed (through a Digital to Analog Converter DAC). This transformation is done by accomplishing Nyquist theorem. A signal can be processed sample by sample (discrete time domain) or in block processing (through a digital FFT). The FFT window or buffer is continuously wide and fed. The equation of the FFT is given by A.1:

$$X(k) = DFT\left[x(n)\right] = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} \quad k = 0, 1, ..., N-1 \tag{A.1}$$

The coefficients of each one of the FFT can be expressed as complex or as a magnitude a phase as following:

$$\mid X(k) \mid = \sqrt{(X_R^2(k) + X_i^2)^{1/2}} \quad k = 0, 1, ..., N-1 \tag{A.2}$$

$$\phi = arctan\frac{X_i(k)}{X_R(k)} \quad k = 0, 1, ..., N-1 \tag{A.3}$$

And the anti-transform is given by:

$$X(k) = IDFT\left[X(k)\right] = \sum_{n=0}^{N-1} X(k)e^{j2\pi nk/N} \quad k = 0, 1, ..., N-1 \tag{A.4}$$

In MAX MSP we usually work with a FFT of 1024 points. If we are interested in analyzing a reduced number of samples, for example 64, we can add zeros to compute the FFT [Zölzer, 2011]. For a linear time-invariant digital system, the relations intrinsic to that system are based on impulses, convolutions, and algorithm of signal flow. Convolution is a mathematical process in the time domain that, when analyzed in the Z-transform domain, corresponds to multiplication. This operation allows us to simulate the interaction between a sound and the frequency response of a room, incorporating a dimension of spatiality based on the modeling of that room. This technique is essential for creating soundscapes and environments in contemporary music. On the other hand, the Z-transform is fundamental for filter design and signal transformation. The transfer function of a digital system can be expressed through the Z-transform, and it allows to work with a discrete time given by the digital systems. The Z transform applied to a signal $x(n)$ is given by:

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \quad with \quad \omega = 2\pi f/f_s \tag{A.5}$$

Applied to the impulse response, and considering a linear system, the output is related to the input trough the Z-transform:

$$Y(z) = H(z) \cdot X(z) \tag{A.6}$$

The sonic impact achieved through the use of filters solidified their position as essential tools in modifying a wide range of musical instruments. In the case of the voice, they have been used in creating spatial effects, noise reduction, and vocal distortions. This underscores the fact that filters are the cornerstone in sound modification.

A.1 Phase Vocoder

The fundamental idea of the phase vocoder [Portnoff, 1976] is to be able to reconstruct a signal from the short-time Fourier transform (STFT). To do this, several analogies are made of the problem of deconstructing the vocal signal and its subsequent reconstruction, which we will see later in this section. Let $x(n)$ be the input signal; its Short-Time Fourier Transform (STFT) is written:

$$X_r(n) = \sum_{r=-\infty}^{\infty} x(r)h(n-r)W_N^{-rk} = |X_r(n)| \cdot e^{\varphi(n,k)} \tag{A.7}$$

For $k = 0, 1, \ldots, N-1$, where $W_N = e^{\frac{-2\pi i}{N}}$, $X_k(n)$ is derived at each time sample n by weighting $x(r)$ with the window function $h(n-r)$ and subsequently computing the Fourier transform of the resultant sequence. It is pertinent to note that the form of $h(n)$ can be constrained in such a manner that we can recover $x(n)$ from $X_k(n)$ as follows:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X_k(n) W_N^{nk} \tag{A.8}$$

The idea is that the window can be thought of as the sum of N bandpass filters $\{h_k(n)\}$ for each frequency band k, as:

$$h_k(n) = \frac{1}{N} h(n) W_N^{nk} \tag{A.9}$$

$$H_k(n) = H(e^{\Omega - \Omega_k}) \tag{A.10}$$

Where $\Omega_k = \frac{2\pi}{N}k$. For $k = 0, 1, \ldots, N-1$ bands. Following the path, the signal can be reconstructed by adding together the N responses produced by these filters, as shown in Figure A.1. the individual response of each $h_k(n)$ filter is given by the convolution:

$$y_k(n) = \sum_{r=-\infty}^{\infty} x(r)h_k(n-r) = \frac{1}{N} W_N^{nk} X_k(n) \tag{A.11}$$

The output of the filter-bank would be the sum of the individual response of each k-filter as follows:

$$y(n) = \sum_{k=0}^{N-1} y_k(n) = \frac{1}{N} \sum_{k=0}^{N-1} W_N^{nk} X_k(n) \tag{A.12}$$

This procedure of each filter can be summarize as it is shown in figure A.2.

A.1.1 Short-Time Analysis

La expresion A.11 debe poder ser calculable, para ello cambiaremos la forma de la expresion usando algunos cambios de variable, primero haaremos $s = r - n$ lo cual permite obtener:

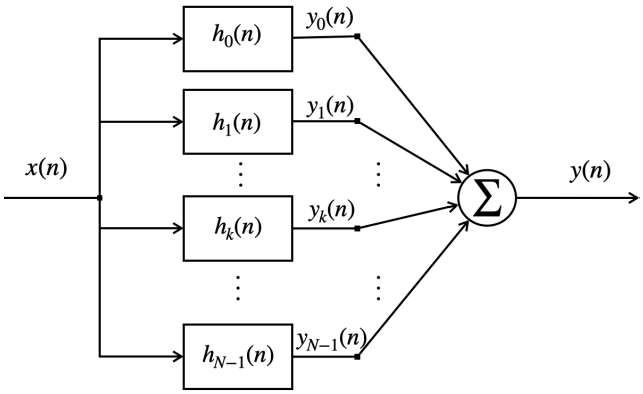


Figure A.1: Filter-Bank analogy for the STFT, adapted from [Portnoff, 1976]

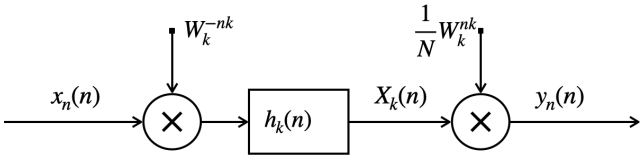


Figure A.2: K-filter analogy, adapted from [Portnoff, 1976]

$$X_k(n) = W_N^{-nk} \sum_{s=-\infty}^{\infty} x(n+s)h(-s)W_N^{-sk} \tag{A.13}$$

If the domain is divided into segments of size N, then there are two indices. One index $m = 0, 1, \ldots, N-1$ traverses each segment internally, while the other index $l = -\infty, \ldots, -1, 0, 1, \ldots, \infty$ advances to the next segment. Using the variable substitution $s = lN + m$, we can write:

$$X_k(n) = W_N^{-nk} \sum_{m=0}^{N-1} \tilde{x}_m(n)W_N^{-mk} \tag{A.14}$$

Where, for a fixed n, the sequence $\tilde{x_m}(n)$ is given by:

$$\tilde{x}_m(n) = \sum_{s=-\infty}^{\infty} x(n+lN+m)h(-lN-m) \tag{A.15}$$

The DFT of $\tilde{x}_m(n)$ would be computed using the following expression:

$$\tilde{X}_k(n) = \sum_{m=0}^{N-1} \tilde{x}_m(n)W_N^{-mk} = \tag{A.16}$$

Thus, the expression A.14 can be written as:

$$X_k(n) = W_N^{-nk}\tilde{X}_k(n) \tag{A.17}$$

If the traversal of r in equation A.11 is redistributed with the variable change $r = lN + m$ as was done previously with the variable s, we obtain:

$$X_r(n) = \sum_{m=0}^{N-1} \sum_{l=\infty}^{\infty} x(n+(l-l')N+(m-n))h(-(l-l')N-(m-n))W_N^{-mk} \quad \text{(A.18)}$$

Returning to equation A.15, if we compute $\tilde{x}_{((m-n))_N}(n)$, where $((i))_N$ symbolizes the last residue of i/N, that is, $((m-n))_N = m-n-l'N$, we obtain:

$$\tilde{x}_{((m-n))_N}(n) = \sum_{l=\infty}^{\infty} x(n+(l-l')N+(m-n))h(-(l-l')N-(m-n)) \quad \text{(A.19)}$$

For a circular shift of n samples in m, which is the term inside equation A.18, such that:

$$X_r(n) = \sum_{m=0}^{N-1} \tilde{x}_{((m-n))_N}(m)W_N^{-mk} \quad \text{(A.20)}$$

Which can be written as:

$$X_r(n) = \sum_{m=0}^{N-1} x(m)W_N^{-mk} \quad \text{(A.21)}$$

Where $x_m(n) = \tilde{x}_{((m-n))_N}(m)$.

A.1.2 Analysis/Synthesis Framework

If the STFT is sampled every R_a time samples using window h_a, also referred to as being decimated by factor R_a, and if s is the temporal index of this decimation, the resulting expression is:

$$X_k(sR_a) = \sum_{m=-\infty}^{\infty} x(m)h_a(sR_a - m)W_N^{mk} \quad \text{(A.22)}$$

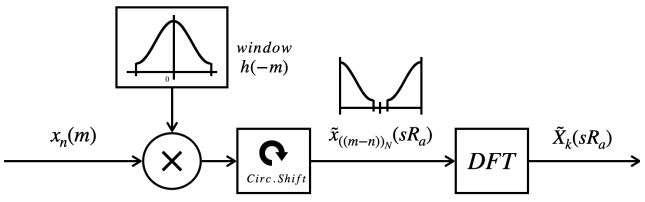


Figure A.3: Generation of $\tilde{X}_k(sR_a)$ and its equivalent, adapted from [Portnoff, 1976]

By reusing equations A.7 through A.21, a new system of equations can be generated, which can be understood with the scheme in figure A.3.

$$X_k(sR_a) = W_N^{sR_a k} \tilde{X}_k(sR_a) \tag{A.23}$$

$$\tilde{X}_k(sR_a) = \sum_{m=0}^{M-1} \tilde{x}_m(sR_a) W_N^{-mk} \tag{A.24}$$

$$\tilde{X}_k(sR_a) = \sum_{m=0}^{M-1} x_m(sR_a) W_N^{-mk} \tag{A.25}$$

$$x_m(sR_a) = \tilde{x}_{((m-sR_a))_N}(sR_a) \tag{A.26}$$

If the window used is a symmetric zero-phase FIR and is chosen such that its origin aligns with the center of the block of size M, then the rotation would be equivalent to $M/2$. Thus, the term W_N^{-mK} becomes equal to $e^{j2\pi(M/2)M} = (-1)^k$. By reusing equations A.7 through A.21, a new system of equations can be generated, which can be understood with the scheme in figure A.4.

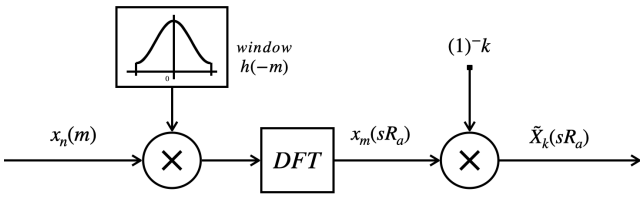


Figure A.4: Generation of $\tilde{X}_k(sR_a)$ in equivalent schema, adapted from [Portnoff, 1976]

Let us consider, then, that a similar scheme can be generated for the synthesis or reconstruction of a signal $y(n)$ from a set $\tilde{Y}_k(sR')$ that has the structure of $\tilde{X}_k(sR_a)$. According to [Portnoff, 1980], under these conditions, we can write:

$$y(n) = \sum_{s=-\infty}^{\infty} f(n - sR') \frac{1}{M} \sum_{k=0}^{M-1} Y_k(sR') W_N^{nk} \tag{A.27}$$

Where $f(n)$ is a window that has the same characteristics as $h(n)$. Note that the second summation has the form of the inverse transform of $Y_k(sR')$:

$$y_n(sR') = \frac{1}{N} \sum_{k=0}^{M-1} Y_k(sR') \tag{A.28}$$

Then:

$$y(n) = \sum_{s=-\infty}^{\infty} f(n - sR') y_n(sR') \tag{A.29}$$

Similarly to the analysis stage, $\tilde{Y}_k(sR')$ can be defined such that:

$$\tilde{Y}_k(sR') = W_M^s R' k \tilde{Y}_k(sR') \tag{A.30}$$

And analogously to the circular shifting of the analysis, it is possible to derive the following set of equations:

$$\tilde{y}_n(sR') = y_{((n+sR'))_M}(sR') \tag{A.31}$$

$$y_n(sR') = \tilde{y}_{((n-sR'))_M}(sR') \tag{A.32}$$

$$y(n) = \sum_{s=-\infty}^{\infty} f(n - sR')\tilde{y}_{((n-sR'))_M}(sR') \tag{A.33}$$

Based on this system of equations, it is possible to construct a complete scheme to obtain the spectral content of the STFT of an audio segment and reconstruct it as shown in figure A.5.

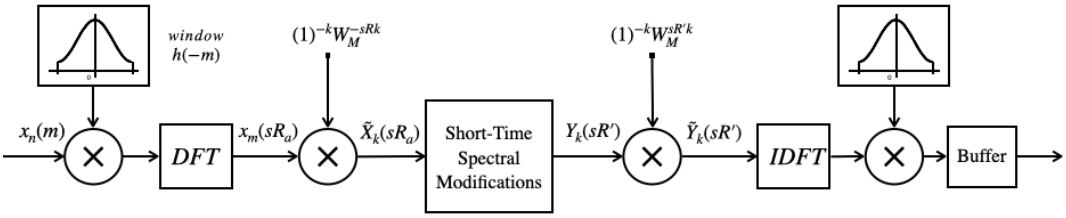


Figure A.5: Complete schema of analysis/synthesis, adapted from [Portnoff, 1976]

Each windowed segment of audio represents an audio grain; the grains obtained after resynthesis are overlapped and summed as we do for some time strechting techniques as PSOLA, in section 1.5.1.2.

A.2 Frequency-Time Implementation Fundamentals

The FFT applied over a windowed audio signal segment, allows calculating magnitudes and phases content of that segments. When we use a sliding-in-time window in time, it can be employed to modify the signal in interesting ways modifying its spectral content. The implementation of the FFT generates a so-called time-frequency representation according to the following scheme. At a given point in the time domain, a sliding window corresponds to it, the FFT is calculated for that window. The resulting matrix with the FFT calculation generates magnitude values for each frequency index k. The successive generation, for each sliding window, of such sequences of magnitudes

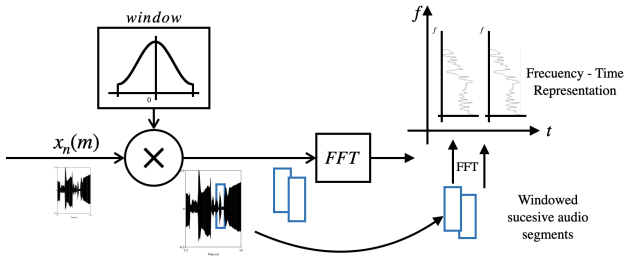


Figure A.6: Schema of frequency-time representation, adapted from [Zölzer, 2011]

distributed across the k indices forms the time-frequency representation, as shown in fig. A.6.

Sound reconstruction is possible if some conditions are accomplished, for example: the sum of overlapping windows equals unity and use of the algorithm should allow the transform of an impulse to have zero phase. Since the FFT starts from the left side, this is achieved with circular shifting that swaps the first and second parts of the buffer, as shown earlier. In the time domain, this is equivalent to $(-1)^k$. This is achieved by applying circular shifting, and in this way, the FFT becomes equivalent to a bank of filters with zero-phase filters as we explained precedently. Each horizontal line can be regarded as a filtered version of the signal for the corresponding coefficient in each frequency band. If the hop size is greater than one, interpolation must be performed between the magnitude and phase values under certain instantaneous frequency conditions, which we will mention shortly.

The other way to perform the reconstruction is through the use of small grains in the time-frequency representation called gaborets. Gaborets have an exponential windowed shape defined by $g_{\Omega_k}(n) = e^{-\Omega_k n} g_\alpha(n)$, which is based on the Gabor transform, a Fourier transform that uses a Gaussian window $g_\alpha(n) = \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{n^2}{2\alpha}}$, with $\alpha < 0$. The reconstruction is done from the following equation:

$$y(n) = \sum_{s=-\infty}^{\infty} \sum_{k=0}^{N-1} Y(sR_s, k) f(n - sR_s) W_N^{-nk} \tag{A.34}$$

which is equivalent to performing windowing plus an FFT/IFFT and another windowing. The interconnection between different frames in the time-frequency representation requires careful reconstruction, as any change or error will have consequences on the sound reconstruction. This is where phase changes become important; reconstruction must preserve the same instantaneous frequency, which is nothing more than the phase change over time for each frequency band, as mentioned in the filter bank approximation in the previous paragraph. To reconstruct the phase properly, the phase difference between successive frames must be preserved (or, in other words, there must be an equivalence of instantaneous frequency):

$$\Delta\varphi((s+1)R_a) = \Omega_k R_a + princarg\left[\widetilde{\varphi}((s+1)R_a) - \widetilde{\varphi}(sR_a) - \Omega_k R_a\right] \quad \text{(A.35)}$$

Where $\widetilde{\varphi}$ is the phase of $X_k(sR_a)$.

Appendix B

DPW and ATA Audio Support for Chapter 2

In this section, we provide a summary of the auditory support for Chapter 2. The original audio file was created using Cantor Digitalis software, based on the article by [Perrotin and D'Alessandro, 2016]. We performed an autotune process on the audio file using the ATA device, with various configurations as shown in Table B.1. Subsequently, we retrieved the obtained pitch curve and then resynthesized the audio using the World vocoder. For DPW samples, we directly retrieved the pitch curve from the stylus, modified it using the DPW algorithm in MATLAB, and then conducted a resynthesis in World. The summary of file names and their correspondence with Chapter 3 figures is shown in Table B.1.

Table B.1: Description of Audio Support for Chapter 2

Figure	Method	TT	FT	CT	File Name
3.15	ATA	00			FIG3.15-ATA-TT00.wav
3.15	ATA	15			FIG3.15-ATA-TT15.wav
3.15	ATA	50			FIG3.15-ATA-TT50.wav
3.15	ATA	100			FIG3.15-ATA-TT100.wav
3.15	ATA	200			FIG3.15-ATA-TT200.wav
3.16	ATA	00	00		FIG3.16-ATA-TT00-FT00.wav
3.16	ATA	00	40		FIG3.16-ATA-TT00-FT40.wav
3.17	ATA	15	00		FIG3.17-ATA-TT15-FT00.wav
3.17	ATA	15	30		FIG3.17-ATA-TT15-FT30.wav
3.18	ATA	50	30		FIG3.18-ATA-TT50-FT30.wav
3.18	ATA	50	60		FIG3.18-ATA-TT50-FT60.wav
3.19	ATA	00	40		FIG3.19-ATA-TT00-FT40.wav
3.19	ATA	50	40		FIG3.19-ATA-TT50-FT40.wav
3.19	ATA	100	40		FIG3.19-ATA-TT100-FT40.wav
3.19	ATA	200	40		FIG3.19-ATA-TT200-FT40.wav
3.20	DPW	100		100	FIG3.20-DPW-TT100-CT100.wav
3.20	DPW	200		100	FIG3.20-DPW-TT200-CT100.wav
3.20	DPW	400		100	FIG3.20-DPW-TT400-CT100.wav
3.21	DPW	025		250	FIG3.21-DPW-TT025-CT250.wav
3.21	DPW	200		250	FIG3.21-DPW-TT200-CT250.wav
3.22	DPW	050		100	FIG3.22-DPW-TT050-CT100.wav
3.22	DPW	050		150	FIG3.22-DPW-TT050-CT150.wav
3.22	DPW	050		250	FIG3.22-DPW-TT050-CT250.wav
3.23	ATA	50	40		FIG3.23-ATA-TT50-FT40.wav
3.23	DPW	050		250	FIG3.23-DPW-TT050-CT250.wav

Appendix C

Pyscho-Acoustical Test Support

Now will provide a comprehensive summary of the psychoacoustic test additional supporting elements. The following document was digitally signed by each participant.

- *I certify the acceptance of my participation in the research study:*
  - *Vocoder Transparency Test*
  - *Pitch Correction Methods Comparison Test*
- *Under the responsibility of Christophe d'Alessandro. This experiment will be conducted by Daniel Molina Villota and Thomas Lucas. The experiment will take place at Institut Jean Le Rond d'Alembert in Paris.*
- *We inform you that:*
  - *Your participation is completely voluntary, and you are completely free to stop participating at any time without notice.*
  - *By participating in this experience, you agree that your performance will be recorded, and you acknowledge having been informed of the nature of the recording. This data will be used without being linked to your identity, anonymously as part of a statistical analysis. No personal data will be disclosed or published as part of the research.*
  - *You will be compensated with 2*30€ as an Amazon gift card for your participation, except in the event of you not completing the experiment.*
  - *By giving consent to this form, you certify that you have read and understood the above information, and you agree to volunteer in the study.*

Also, some information was asked to the participants as follows:

- **Please complete the form.**
- *Test-Id*
- *Age*
- *Date*
- *Hearing problems*
- *Do you have experience in critical listening? How much? (None, Beginner, Intermediate, Expert)*
- *How many listening tests have you already participated in?*
- *Do you have a musical practice? If yes, which one?*
- *What is your musical practice level? (Beginner, Intermediate, Expert)*

The psychoacoustic evaluation was executed utilizing a webMUSHRA-based website. The respective code was authored by Daniel Molina and Thomas Lucas and is accessible via the Lutherie-Acoustique-Musique (LAM) website, categorized into three sections for consent, vocoder evaluation, and pitch correction methods evaluation.

- **Consent:**
  http://vocodertest.lam.jussieu.fr/?config=test-consent-v2.yaml
- **Test Vocoder (i-ii-iii-iv):**
  http://vocodertest.lam.jussieu.fr/?config=test-partA-v2.yaml
  http://vocodertest.lam.jussieu.fr/?config=test-partB-v2.yaml
  http://vocodertest.lam.jussieu.fr/?config=test-partC-v2.yaml
  http://vocodertest.lam.jussieu.fr/?config=test-partD-v2.yaml
- **Test Pitch Correction Methods (12345):**
  http://vocodertest.lam.jussieu.fr/?config=test-suite-1-v2.yaml
  http://vocodertest.lam.jussieu.fr/?config=test-suite-2-v2.yaml
  http://vocodertest.lam.jussieu.fr/?config=test-suite-3-v2.yaml
  http://vocodertest.lam.jussieu.fr/?config=test-suite-4-v2.yaml
  http://vocodertest.lam.jussieu.fr/?config=test-suite-5-v2.yaml

Below is the summary of subjects who participated in the test. Each subject was assigned a participant code. All subjects participated in both the test for vocoder comparison and the test for correction method comparison. Mixed versions of the tasks were developed to randomize the tests content (roman numerals), in the specified order, check table C.1:

Table C.1: Subjects and tests order

Subj.	Vocoder	ATA vs DPW	**Subj.**	Vocoder	ATA vs DPW
A01	ii-iii-iv-i	51423	L12	iv-i-ii-iii	31542
B02	iv-iii-i-ii	51432	M13	iv-iii-ii-i	31425
C03	iv-i-ii-iii	51243	N14	iv-iii-i-ii	31542
D04	i-iv-ii-iii	52314	O15	iii-i-ii-iv	25431
E05	ii-i-iii-iv	15432	P16	i-iii-iv-ii	15432
F06	iii-ii-iv-i	14325	R18	ii-i-iii-iv	-
G07	iii-i-iv-ii	12345	S19	i-ii-iv-iii	15432
H08	iii-i-ii-iv	24153	T20	iv-ii-iii-i	42531
I09	i-ii-iv-iii	24315	U21	iii-i-ii-iv	24531
J10	ii-i-iii-iv	25431	W23	iii-i-iv-ii	43251
K11	ii-iii-i-iv	24315			

C.1 Classification Performance for the Vocoder Comparison - Tasks A, B, C and D

A performance calculation has been conducted according to the recommendations of the MUSHRA test, wherein the percentage of trials where the reference is rated below 90 is calculated. According to MUSHRA recommendations, subjects with a percentage higher than 15% in all tasks are considered unsuitable and are removed from the data. Nevertheless, such subjects do not negatively impact our results, as their contributions exhibit data distributions (histograms) consistent with ratings provided by good subjects. In these distributions, the most similar samples display similar mean scores and distribution shapes, and the different ones also maintain distribution shapes and mean values. The results of performance can be observed in the table C.2.

C.2 Statistical Support for Task A

This section includes:

1. Histograms per subjects (Figures C.1 and C.2)
2. Histograms per trials (Figures C.3 and C.4).

The statistical analysis includes the complete calculations for the Task A analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.3).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.4

Table C.2: Classification Performance (Tasks A, B, C, D)

Subj.	% Incorrect by Task				Subj.	% Incorrect by Task			
	A	B	C	D		A	B	C	D
A01	25.0	0.0	42.9	12.5	L12	0.0	0.0	0.0	0.0
B02	43.8	50.0	78.6	56.2	M13	6.3	0.0	0.0	0.0
C03	6.3	7.1	0.0	0.0	N14	25.0	28.6	28.6	18.8
D04	6.3	0.0	0.0	6.2	O15	43.8	21.4	35.7	31.2
E05	0.0	7.1	7.1	6.2	P16	6.3	7.1	21.4	6.2
F06	12.5	0.0	0.0	0.0	R18	12.5	21.4	21.4	12.5
G07	37.5	28.6	42.9	50.0	S19	0.0	14.3	0.0	0.0
H08	18.8	50.0	64.3	18.8	T20	6.3	0.0	7.1	0.0
I09	0.0	0.0	14.3	0.0	U21	0.0	0.0	0.0	0.0
J10	0.0	0.0	0.0	0.0	W23	12.5	0.0	14.3	18.8
K11	0.0	7.1	7.1	12.5					

- Post-hoc Tukey HSD multi-comparison summarized in Table C.5

3. Non-musicians:

   - ANOVA results are summarized in Table C.6
   - Post-hoc Tukey HSD multi-comparison summarized in Table C.7

4. Musicians:

   - ANOVA results are summarized in Table C.8
   - Post-hoc Tukey HSD multi-comparison summarized in Table C.9

C.2.1 Histograms per subjects for Task A

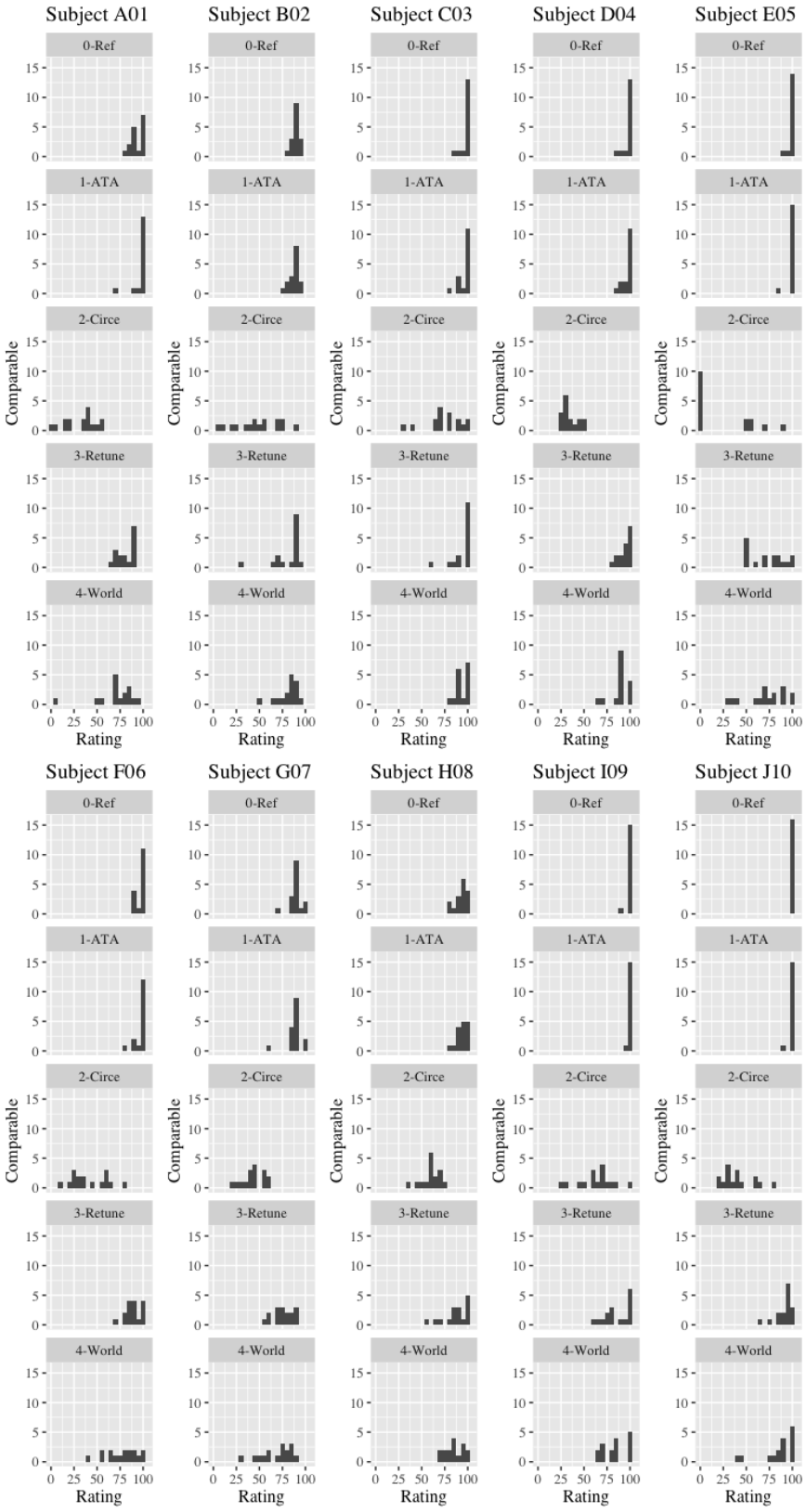


Figure C.1: Histograms per subject for Task A - part 1

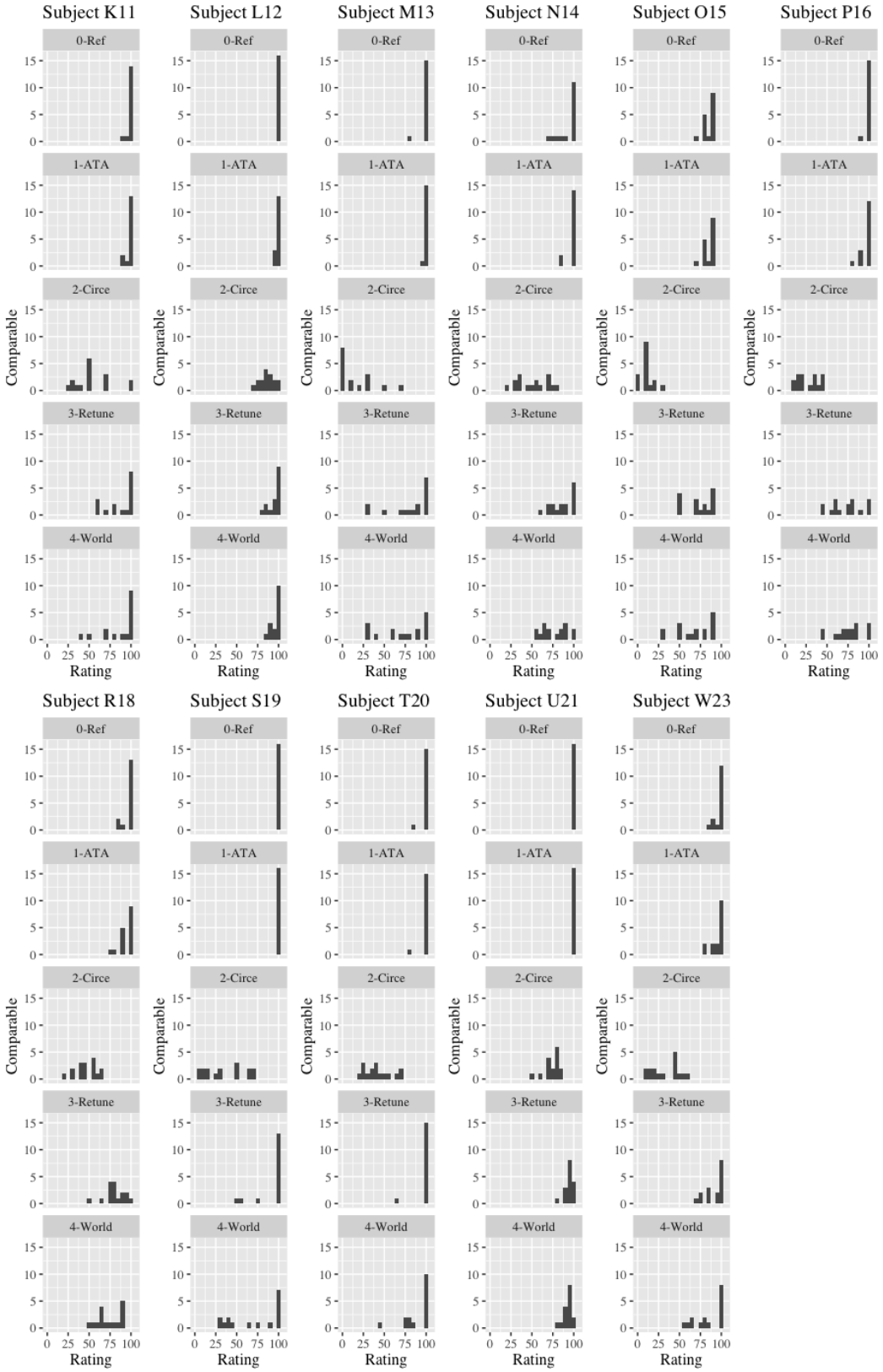


Figure C.2: Histograms per subject for Task A - part 2

C.2.2 Histograms per trials for Task A

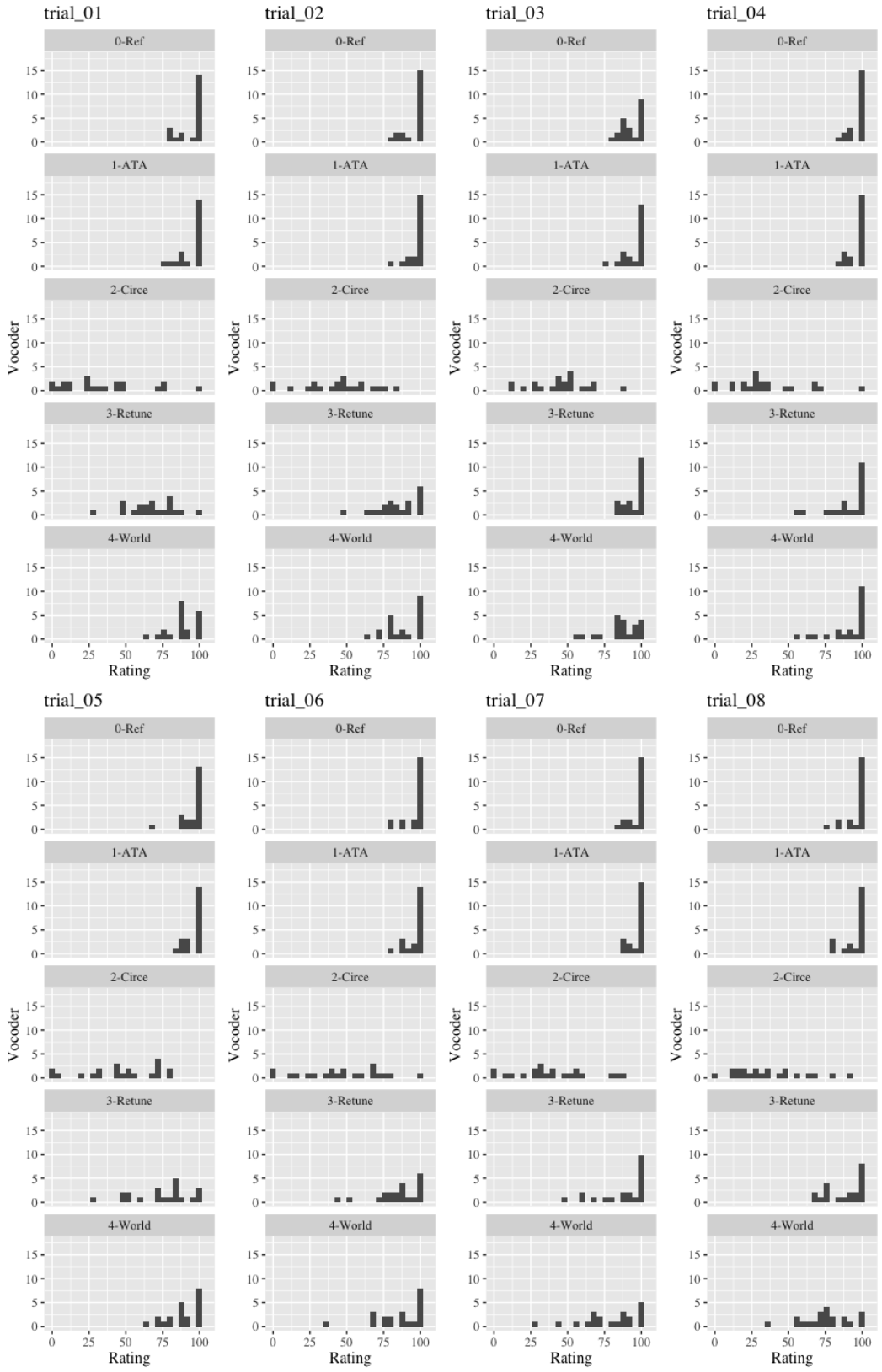


Figure C.3: Histograms per trial for Task A - part 1

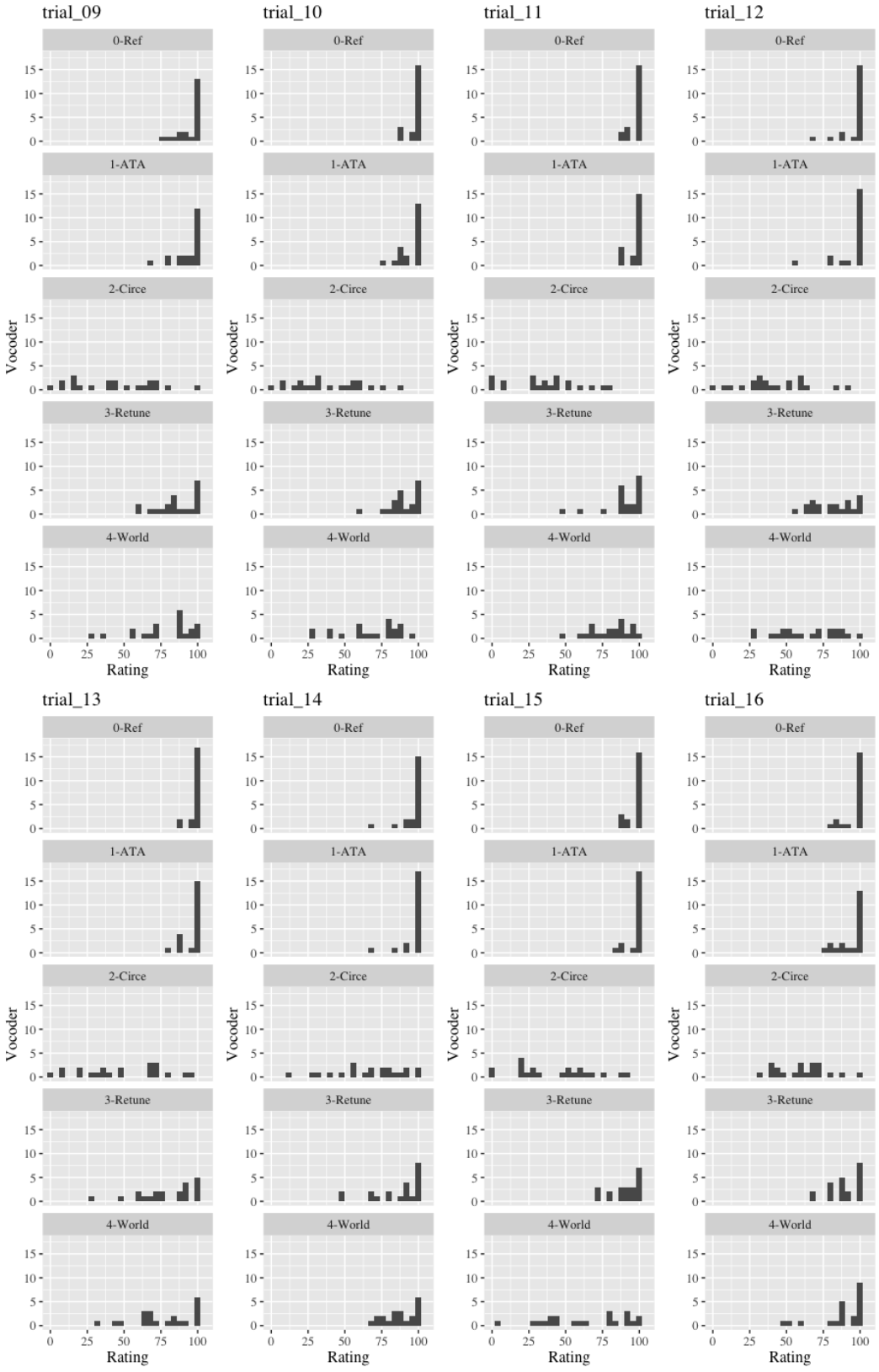


Figure C.4: Histograms per trial for Task A - part 2

C.2.3 ANOVA and Tukey HSD for Task A

Table C.3: Tukey HSD for Task A - before excluding subj. deemed unsuitable

Comparison	Difference	Lower	Upper	p-value
ATA-Ref	−0.3	−3.8	3.1	0.9989
Circe-Ref	−51.9	−55.3	−48.5	<0.0001
Retune-Ref	−10.7	−14.1	−7.3	<0.0001
World-Ref	−15.7	−19.1	−12.2	<0.0001
Circe-ATA	−51.6	−55.0	−48.1	<0.0001
Retune-ATA	−10.4	−13.8	−6.9	<0.0001
World-ATA	−15.3	−18.8	−11.9	<0.0001
Retune-Circe	41.2	37.8	44.6	<0.0001
World-Circe	36.2	32.8	39.7	<0.0001
World-Retune	−5.0	−8.4	−1.6	0.0007

Table C.4: ANOVA for Task A - after excluding subj. deemed unsuitable

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Vocoder	4	488158	122040	458.1	$< 2e-16$
Residuals	1275	339662	266		

Table C.5: Tukey HSD for Task A - after excluding subj. deemed unsuitable

Comparison	Difference	Lower	Upper	p-value
ATA-Ref	−0.5	−4.4	3.5	0.9975
Circe-Ref	−53.2	−57.1	−49.3	<0.0001
Retune-Ref	−10.9	−14.8	−7.0	<0.0001
World-Ref	−16.1	−20.0	−12.1	<0.0001
Circe-ATA	−52.7	−56.7	−48.8	<0.0001
Retune-ATA	−10.4	−14.4	−6.5	<0.0001
World-ATA	−15.6	−19.6	−11.7	<0.0001
Retune-Circe	42.3	38.4	46.2	<0.0001
World-Circe	37.1	33.2	41.1	<0.0001
World-Retune	−5.2	−9.1	−1.2	0.0031

Table C.6: ANOVA for Task A (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	4	164544	41136	177.7	$< 2e-16$
Residuals	635	147008	232		

Table C.7: Tukey HSD for Task A (Non-Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
ATA-Ref	−1.2	−6.4	4.0	0.9718
Circe-Ref	−43.7	−48.9	−38.5	<0.0001
Retune-Ref	−7.0	−12.2	−1.8	0.0025
World-Ref	−12.1	−17.3	−6.9	<0.0001
Circe-ATA	−42.5	−47.7	−37.3	<0.0001
Retune-ATA	−5.8	−11.0	−0.6	0.0205
World-ATA	−10.9	−16.1	−5.7	<0.0001
Retune-Circe	36.7	31.5	41.9	<0.0001
World-Circe	31.6	26.4	36.8	<0.0001
World-Retune	−5.1	−10.3	0.1	0.0571

Table C.8: ANOVA for Task A (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	4	457091	114273	429.7	$< 2e-16$
Residuals	1035	275217	266		

Table C.9: Tukey HSD for Task A (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
ATA vs. Ref	0.2	−4.2	4.6	0.9990
Circe vs. Ref	−57.0	−61.3	−52.6	<0.0001
Retune vs. Ref	−13.0	−17.3	−8.6	<0.0001
World vs. Ref	−17.9	−22.2	−13.5	<0.0001
Circe vs. ATA	−57.2	−61.5	−52.8	<0.0001
Retune vs. ATA	−13.2	−17.5	−8.8	<0.0001
World vs. ATA	−18.1	−22.4	−13.7	<0.0001
Retune vs. Circe	44.0	39.6	48.4	<0.0001
World vs. Circe	39.1	34.7	43.5	<0.0001
World vs. Retune	−4.9	−9.3	−0.5	0.0190

C.3 Statistical Support for Task B

This section includes:

1. Histograms per subjects (Figures C.5 and C.6)
2. Histograms per trials (Figures C.7 and C.8).

The statistical analysis includes the complete calculations for the Task B analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.10).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.11
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.12
3. Non-musicians:
    - ANOVA results are summarized in Table C.13
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.14
4. Musicians:
    - ANOVA results are summarized in Table C.15
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.16

C.3.1 Histograms per subjects for Task B

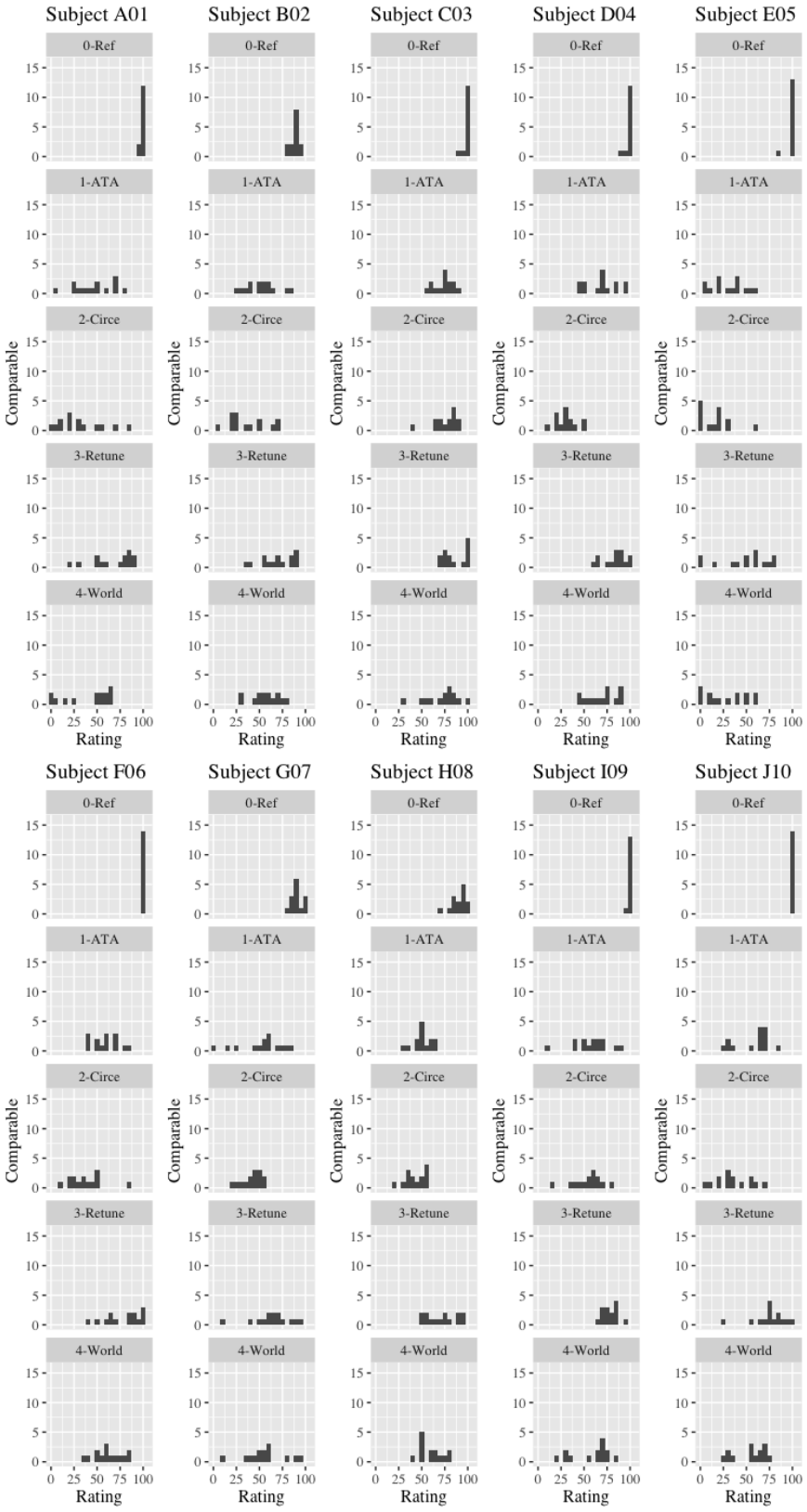


Figure C.5: Histograms per subject for Task B - part 1

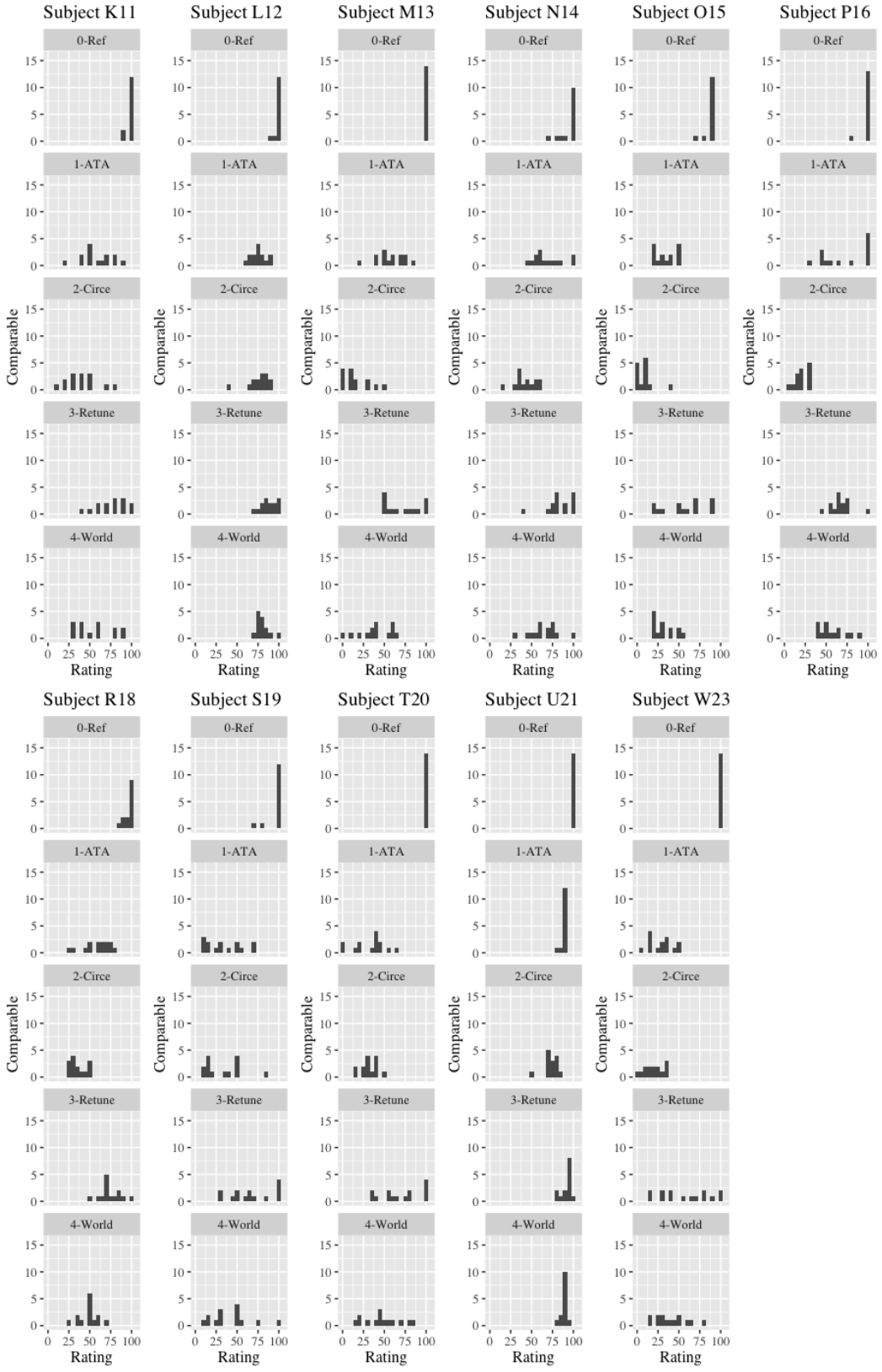


Figure C.6: Histograms per subject for Task B - part 2

C.3.2 Histograms per trials for Task B

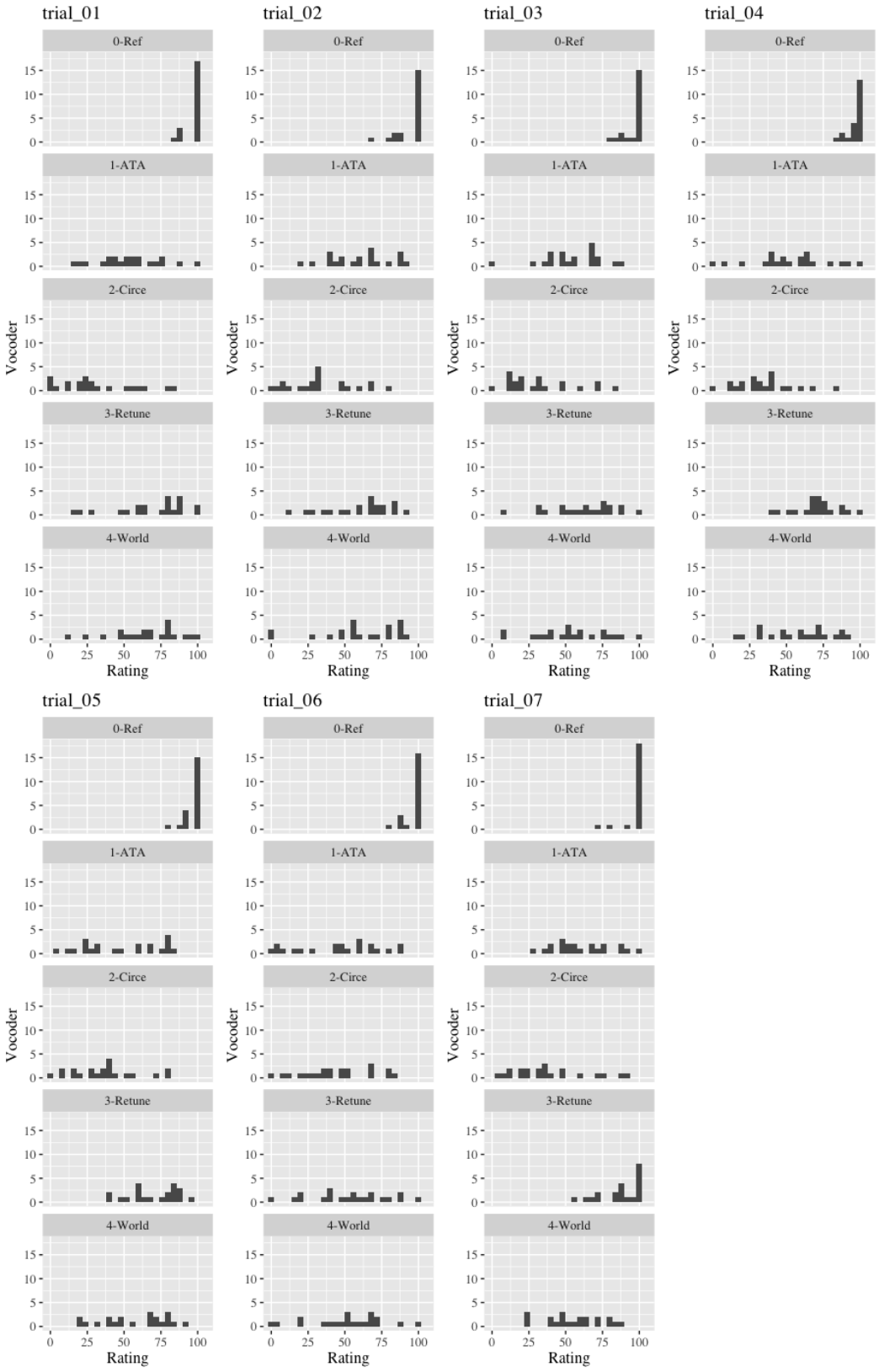


Figure C.7: Histograms per trial for Task B - part 1

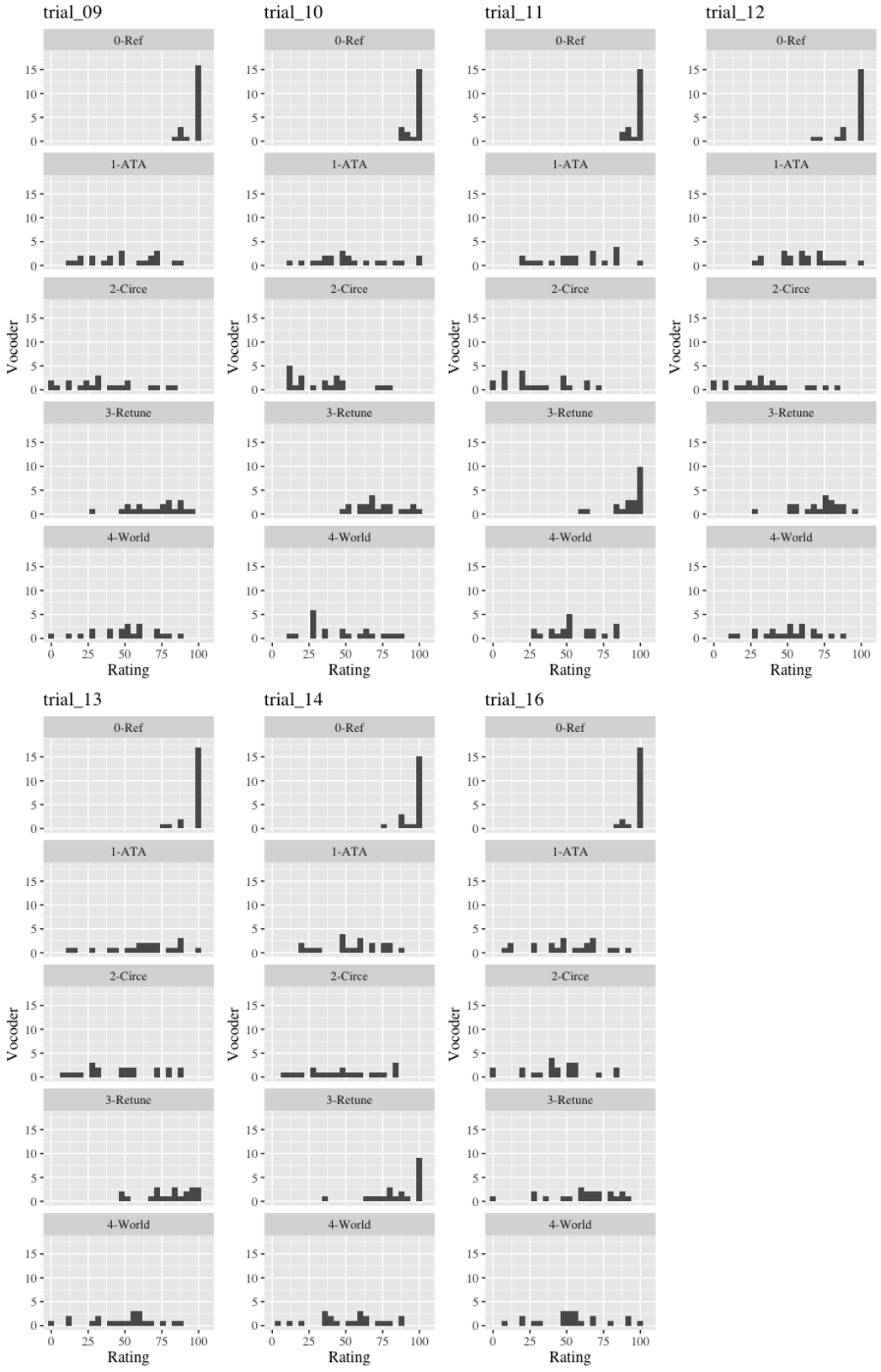


Figure C.8: Histograms per trial for Task B - part 2

C.3.3 ANOVA and Tukey HSD for Task B

Table C.10: Tukey HSD for Task B - before excluding subj. deemed unsuitable

Comparison	Difference	Lower	Upper	p-value
ATA-Ref	−41.7	−46.3	−37.1	<0.0001
Circe-Ref	−58.8	−63.5	−54.2	<0.0001
Retune-Ref	−24.3	−28.9	−19.7	<0.0001
World-Ref	−42.0	−46.7	−37.4	<0.0001
Circe-ATA	−17.1	−21.8	−12.5	<0.0001
Retune-ATA	17.4	12.8	22.0	<0.0001
World-ATA	−0.3	−5.0	4.3	0.9997
Retune-Circe	34.6	29.9	39.2	<0.0001
World-Circe	16.8	12.2	21.4	<0.0001
World-Retune	−17.7	−22.4	−13.1	<0.0001

Table C.11: ANOVA for Task B - after excluding subj. deemed unsuitable

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Vocoder	4	464517	116129	260.3	$< 2e-16$
Residuals	1115	497469	446		

Table C.12: Tukey HSD for Task B - after excluding subj. deemed unsuitable

Comparison	Difference	Lower	Upper	p-value
ATA-Ref	−42.7	−48.1	−37.2	<0.0001
Circe-Ref	−59.9	−65.3	−54.4	<0.0001
Retune-Ref	−25.2	−30.6	−19.7	<0.0001
World-Ref	−43.8	−49.3	−38.4	<0.0001
Circe-ATA	−17.2	−22.6	−11.7	<0.0001
Retune-ATA	17.5	12.1	23.0	<0.0001
World-ATA	−1.2	−6.6	4.3	0.9784
Retune-Circe	34.7	29.3	40.2	<0.0001
World-Circe	16.0	10.6	21.5	<0.0001
World-Retune	−18.7	−24.1	−13.2	<0.0001

Table C.13: ANOVA for Task B (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	4	195534	48884	119.3	$< 2e-16$
Residuals	555	227478	410		

Table C.14: Tukey HSD for Task B (Non-Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
ATA-Ref	−45.2	−52.6	−37.8	<0.0001
Circe-Ref	−51.9	−59.3	−44.5	<0.0001
Retune-Ref	−22.8	−30.2	−15.4	<0.0001
World-Ref	−40.1	−47.5	−32.7	<0.0001
Circe-ATA	−6.7	−14.1	0.7	0.0993
Retune-ATA	22.4	15.0	29.8	<0.0001
World-ATA	5.2	−2.2	12.6	0.3127
Retune-Circe	29.1	21.7	36.5	<0.0001
World-Circe	11.8	4.4	19.3	0.0001
World-Retune	−17.3	−24.7	−9.9	<0.0001

Table C.15: ANOVA for Task B (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	4	399680	99920	238.1	$< 2e-16$
Residuals	905	379865	420		

Table C.16: Tukey HSD for Task B (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
ATA vs. Ref	−39.5	−45.4	−33.7	<0.0001
Circe vs. Ref	−63.1	−69.0	−57.2	<0.0001
Retune vs. Ref	−25.2	−31.1	−19.3	<0.0001
World vs. Ref	−43.2	−49.1	−37.4	<0.0001
Circe vs. ATA	−23.6	−29.5	−17.7	<0.0001
Retune vs. ATA	14.3	8.4	20.2	<0.0001
World vs. ATA	−3.7	−9.6	2.2	0.4180
Retune vs. Circe	37.9	32.0	43.8	<0.0001
World vs. Circe	19.9	14.0	25.7	<0.0001
World vs. Retune	−18.0	−23.9	−12.1	<0.0001

C.4 Statistical Support for Task C

This section includes:

1. Histograms per subjects (Figures C.9 and C.10)
2. Histograms per trials (Figures C.11 and C.12).

The statistical analysis includes the complete calculations for the Task C analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.17).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.18
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.19
3. Non-musicians:
    - ANOVA results are summarized in Table C.20
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.21
4. Musicians:
    - ANOVA results are summarized in Table C.22
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.23

C.4.1 Histograms per subjects for Task C

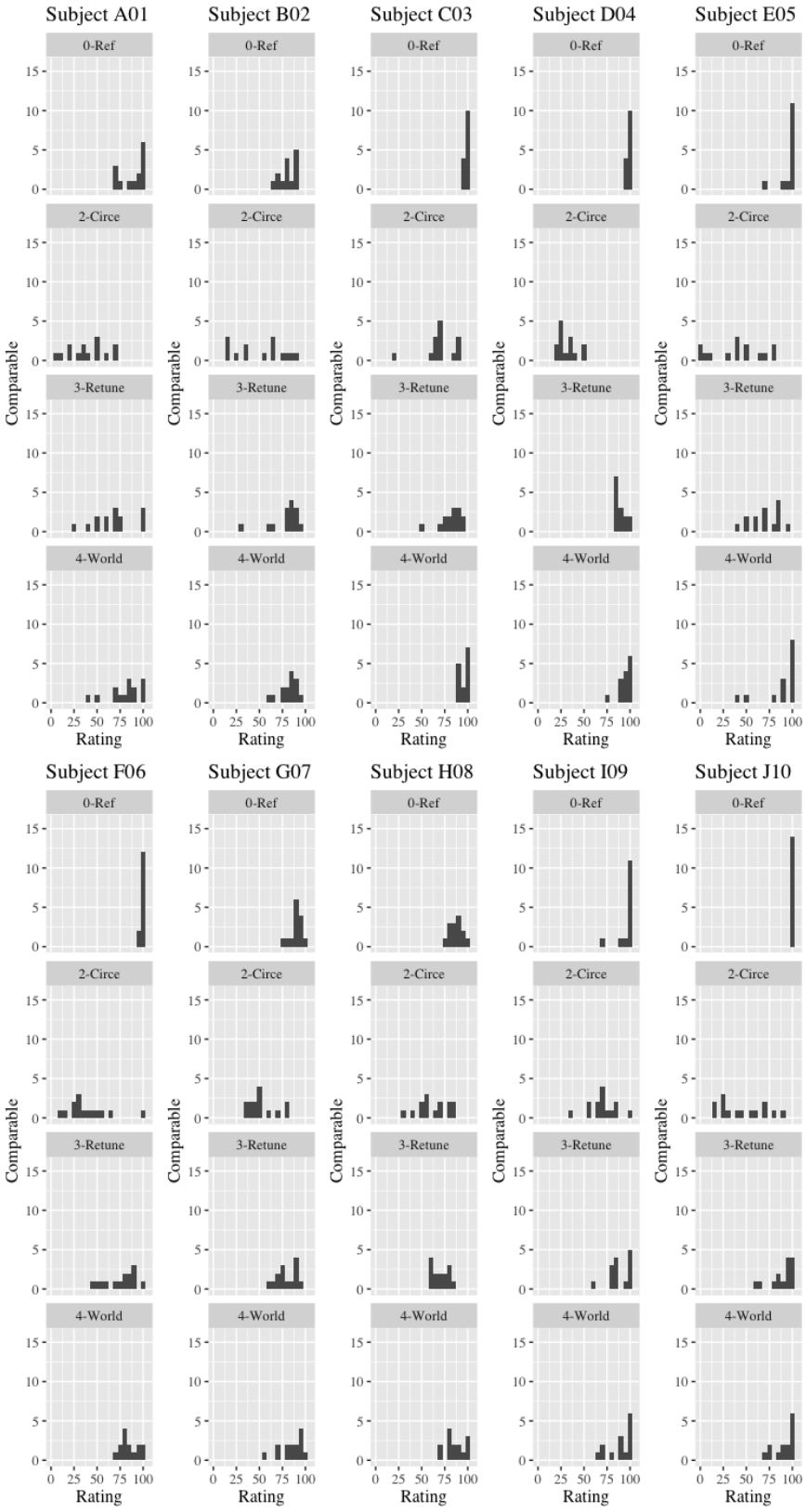


Figure C.9: Histograms per subject for Task C - part 1

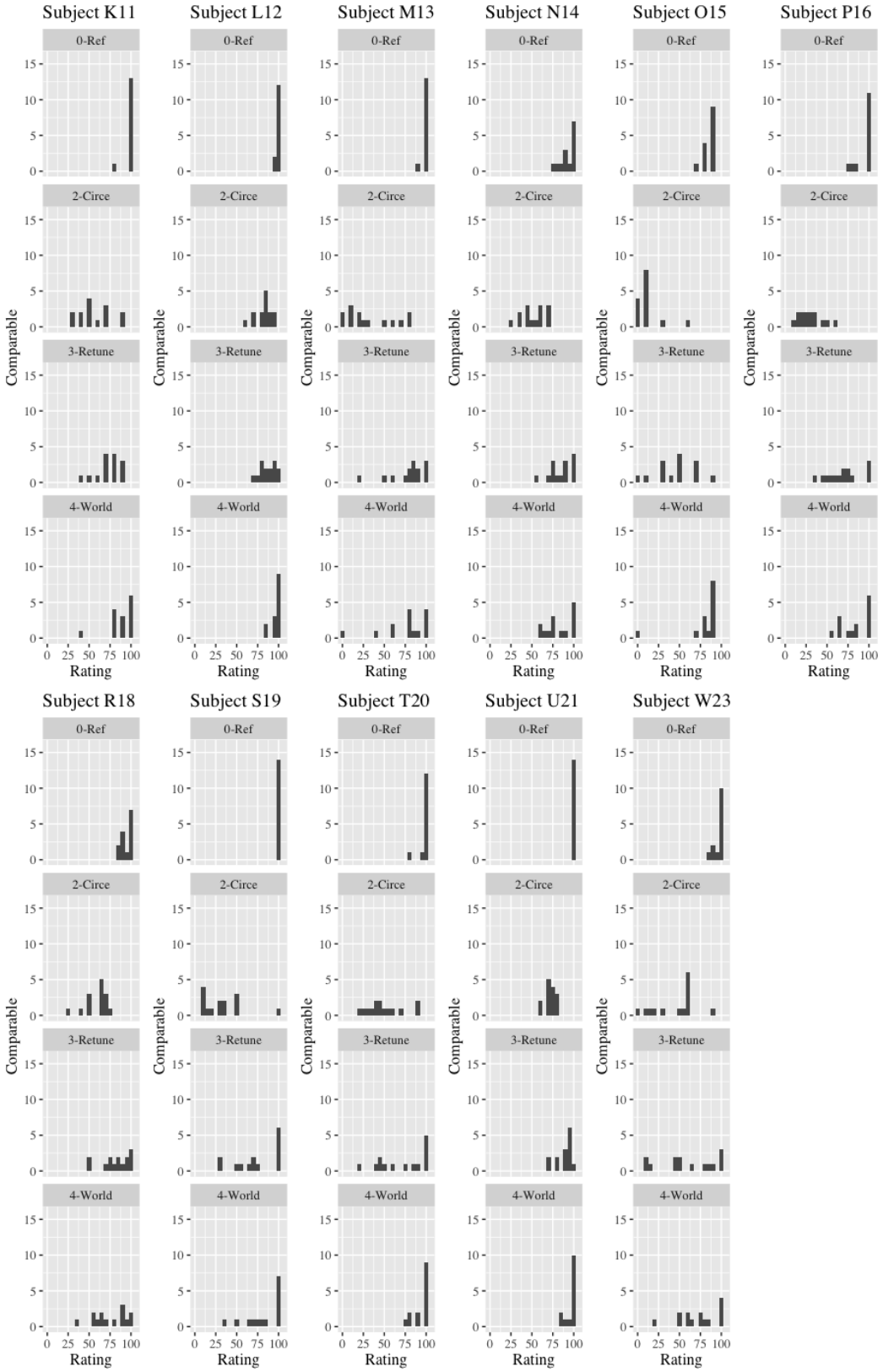


Figure C.10: Histograms per subject for Task C - part 2

C.4.2 Histograms per trials for Task C

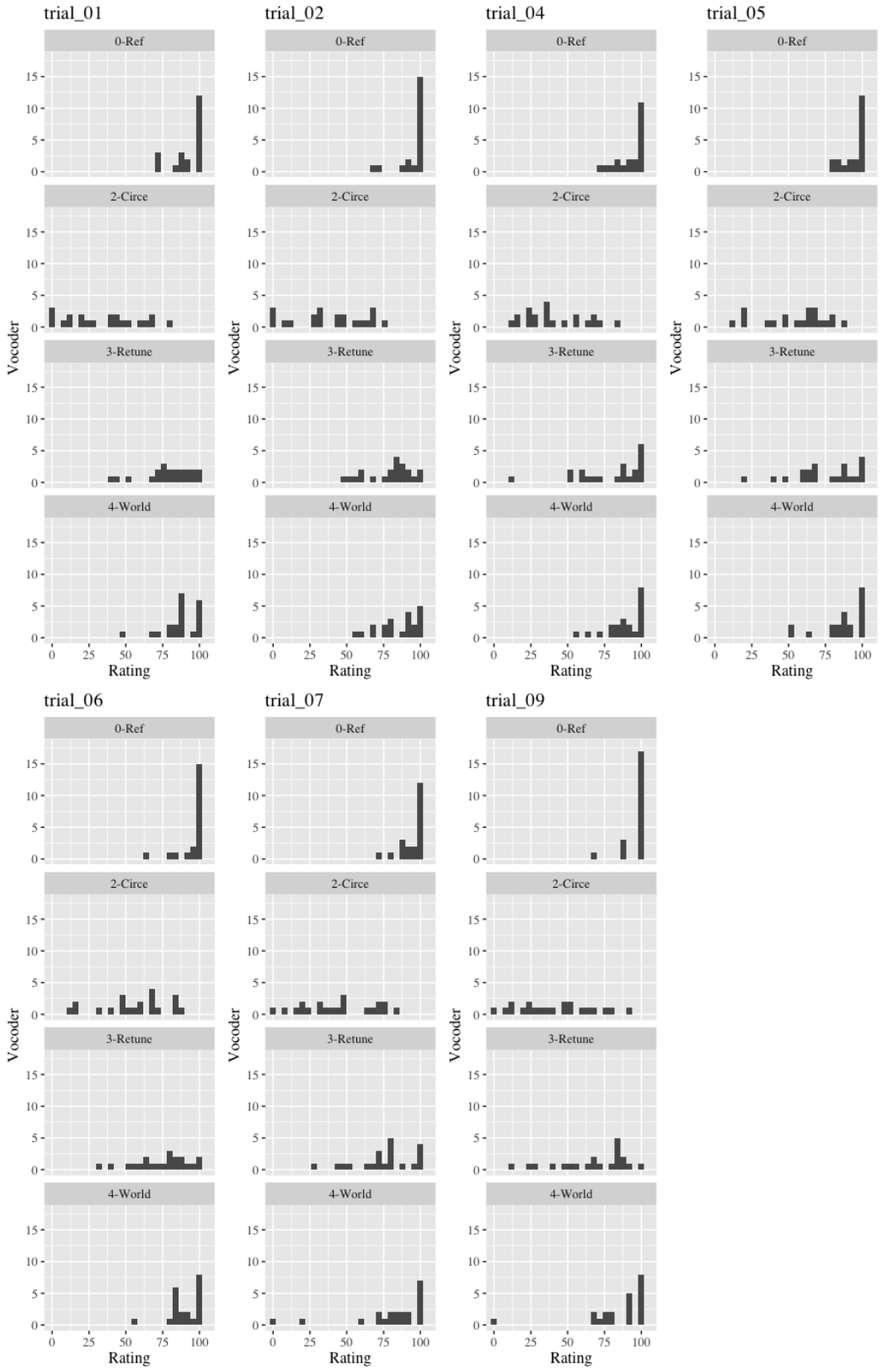


Figure C.11: Histograms per trial for Task C - part 1

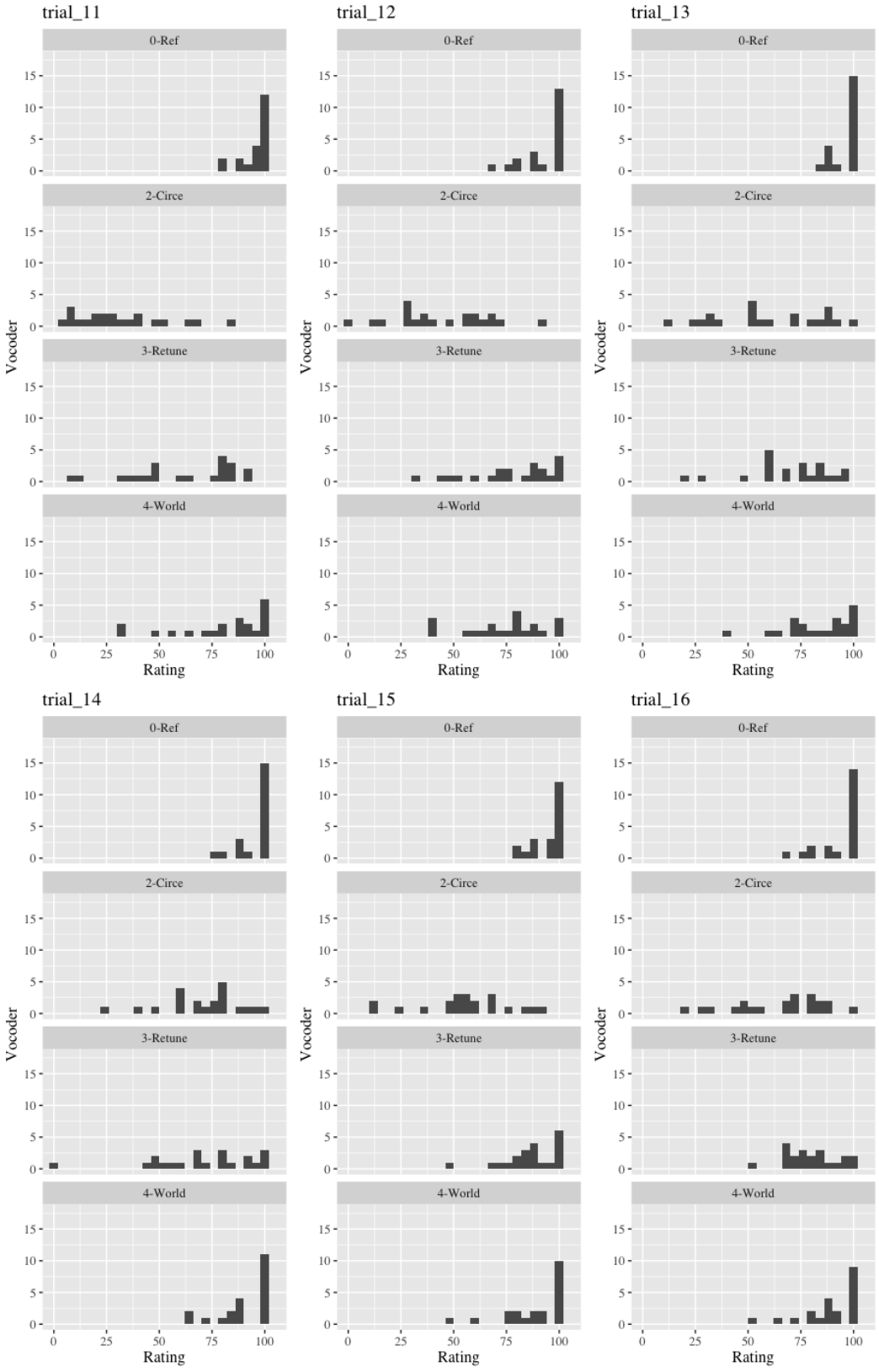


Figure C.12: Histograms per trial for Task C - part 2

Table C.17: Tukey HSD or Task C - before excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe-Ref	-46.262	-50.235	-42.289	< 0.0001
Retune-Ref	-18.619	-22.592	-14.646	< 0.0001
World-Ref	-9.364	-13.337	-5.391	< 0.0001
Retune-Circe	27.643	23.670	31.616	< 0.0001
World-Circe	36.898	32.925	40.871	< 0.0001
World-Retune	9.255	5.282	13.228	< 0.0001

Table C.18: ANOVA for Task C - after excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	280392	93464	269.7	$< 2e-16$
Residuals	892	309130	347		

Table C.19: Tukey HSD for Task C - after excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe-Ref	−47.7	−52.3	−43.2	<0.0001
Retune-Ref	−19.5	−24.0	−15.0	<0.0001
World-Ref	−10.9	−15.4	−6.3	<0.0001
Retune-Circe	28.3	23.7	32.8	<0.0001
World-Circe	36.9	32.3	41.4	<0.0001
World-Retune	8.6	4.1	13.1	<0.0001

Table C.20: ANOVA for Task C (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	88662	29554	83.25	$< 2e-16$
Residuals	444	157612	355		

Table C.21: Tukey HSD for Task C (Non-Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe-Ref	−37.9	−44.4	−31.5	<0.0001
Retune-Ref	−18.5	−25.0	−12.0	<0.0001
World-Ref	−9.1	−15.5	−2.6	0.0020
Retune-Circe	19.4	12.9	25.9	<0.0001
World-Circe	28.9	22.4	35.4	<0.0001
World-Retune	9.5	3.0	15.9	0.0011

Table C.22: ANOVA for Task C (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	272170	90723	271	$< 2e-16$
Residuals	724	242418	335		

Table C.23: Tukey HSD for Task C (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe vs. Ref	−51.4	−56.3	−46.4	<0.0001
Retune vs. Ref	−18.7	−23.6	−13.7	<0.0001
World vs. Ref	−9.6	−14.5	−4.6	<0.0001
Retune vs. Circe	32.7	27.8	37.6	<0.0001
World vs. Circe	41.8	36.9	46.8	<0.0001
World vs. Retune	9.1	4.2	14.1	<0.0001

C.5 Statistical Support for Task D

This section includes:

1. Histograms per subjects (Figures C.13 and C.14)
2. Histograms per trials (Figures C.15 and C.16).

The statistical analysis includes the complete calculations for the Task D analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.24).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.25
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.26
3. Non-musicians:
    - ANOVA results are summarized in Table C.27
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.28
4. Musicians:
    - ANOVA results are summarized in Table C.29
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.30

In this modified version, all occurrences of ”C” have been replaced with ”D” in the labels and references.

C.5.1 Histograms per subjects for Task D

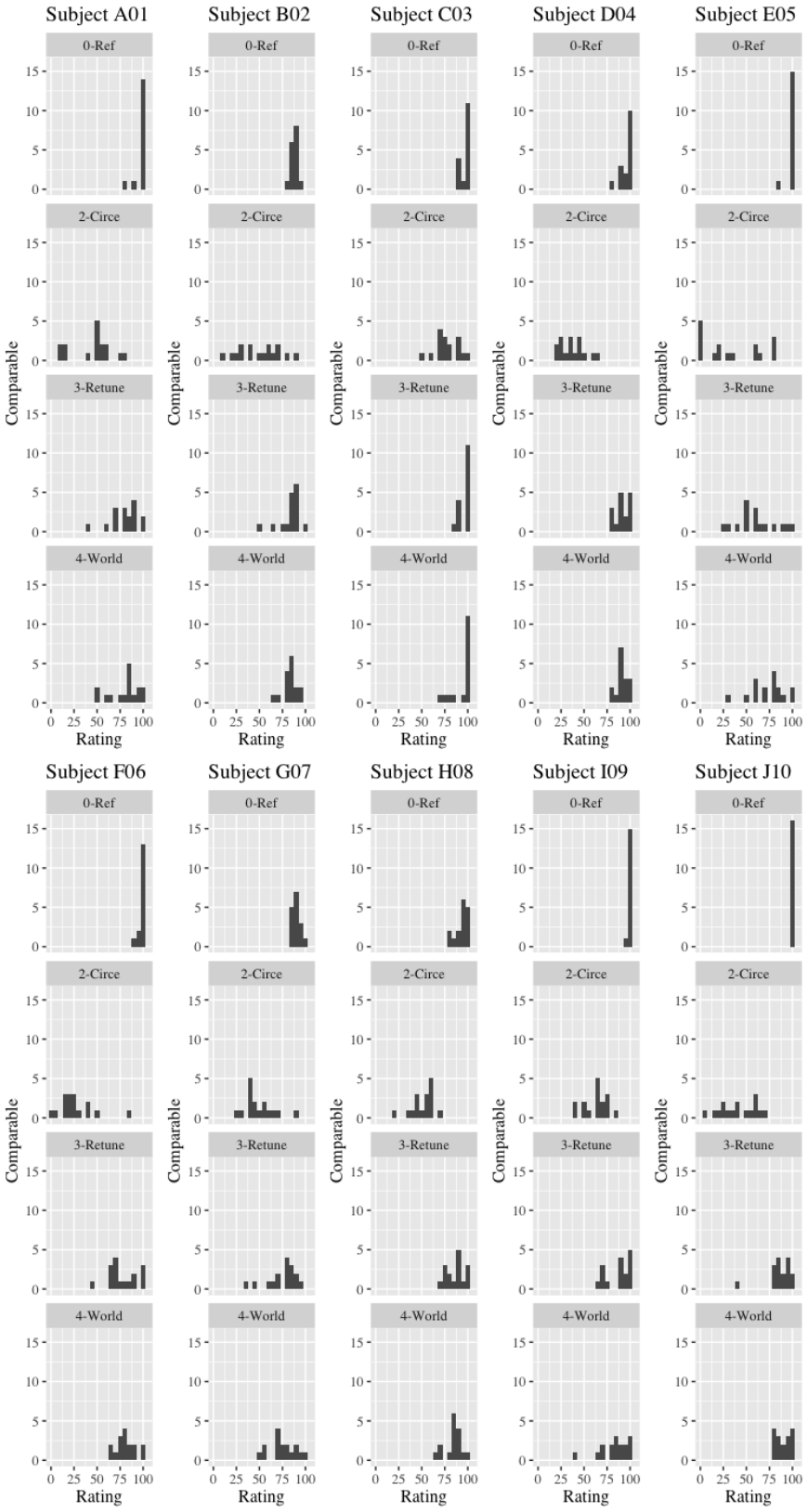


Figure C.13: Histograms per subject for Task D - part 1

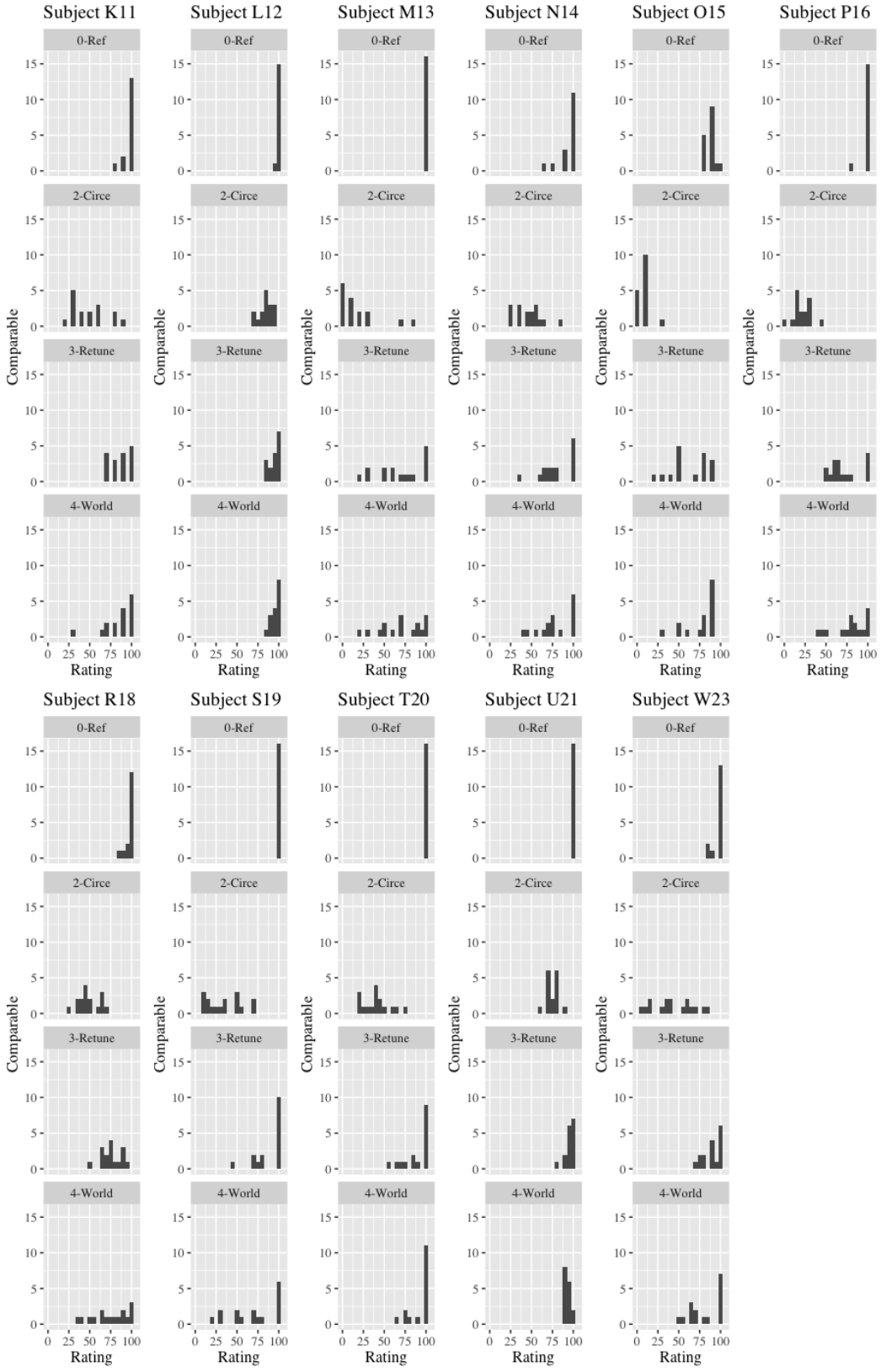


Figure C.14: Histograms per subject for Task D - part 2

C.5.2 Histograms per trials for Task D

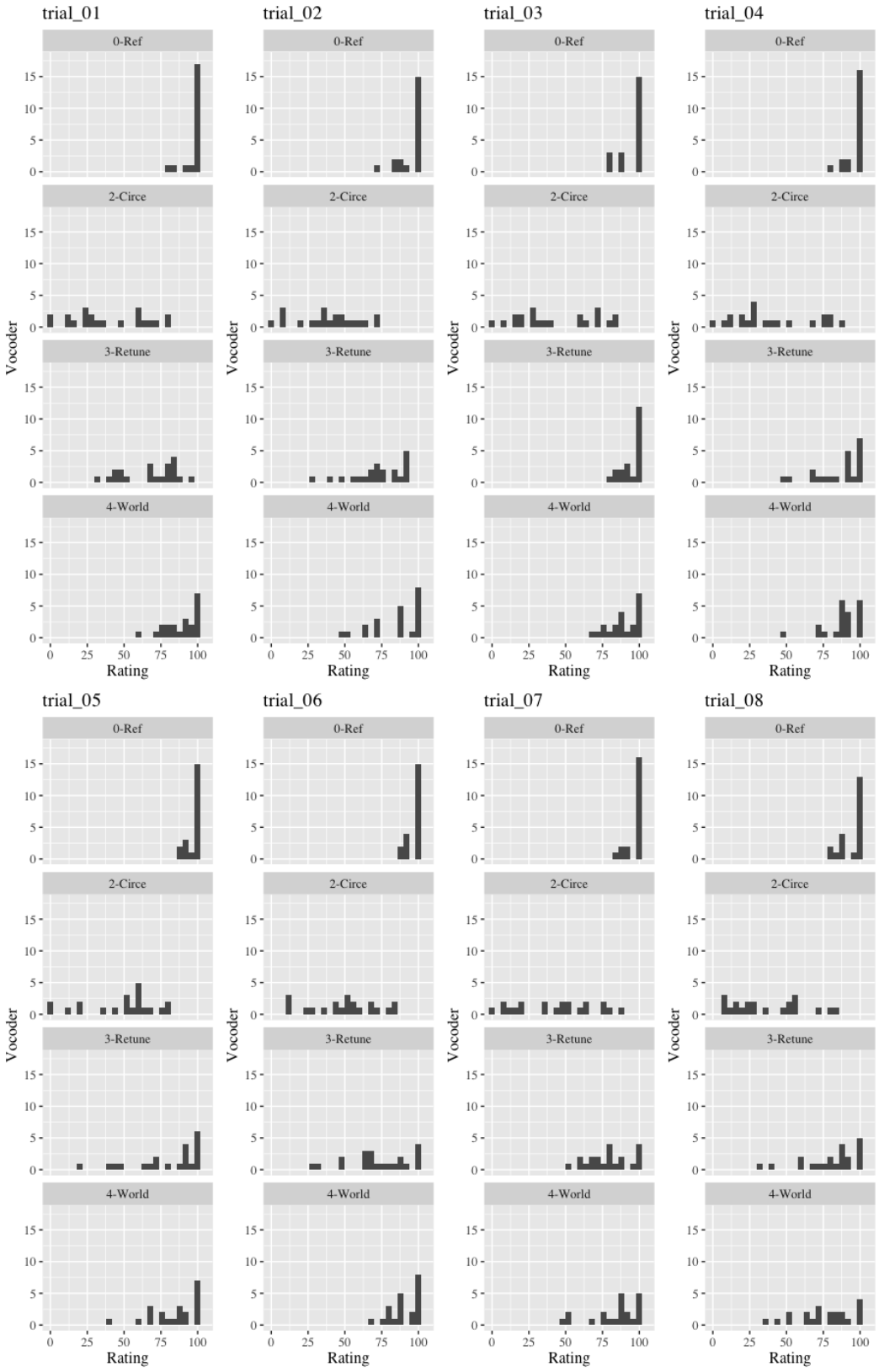


Figure C.15: Histograms per trial for Task D - part 1

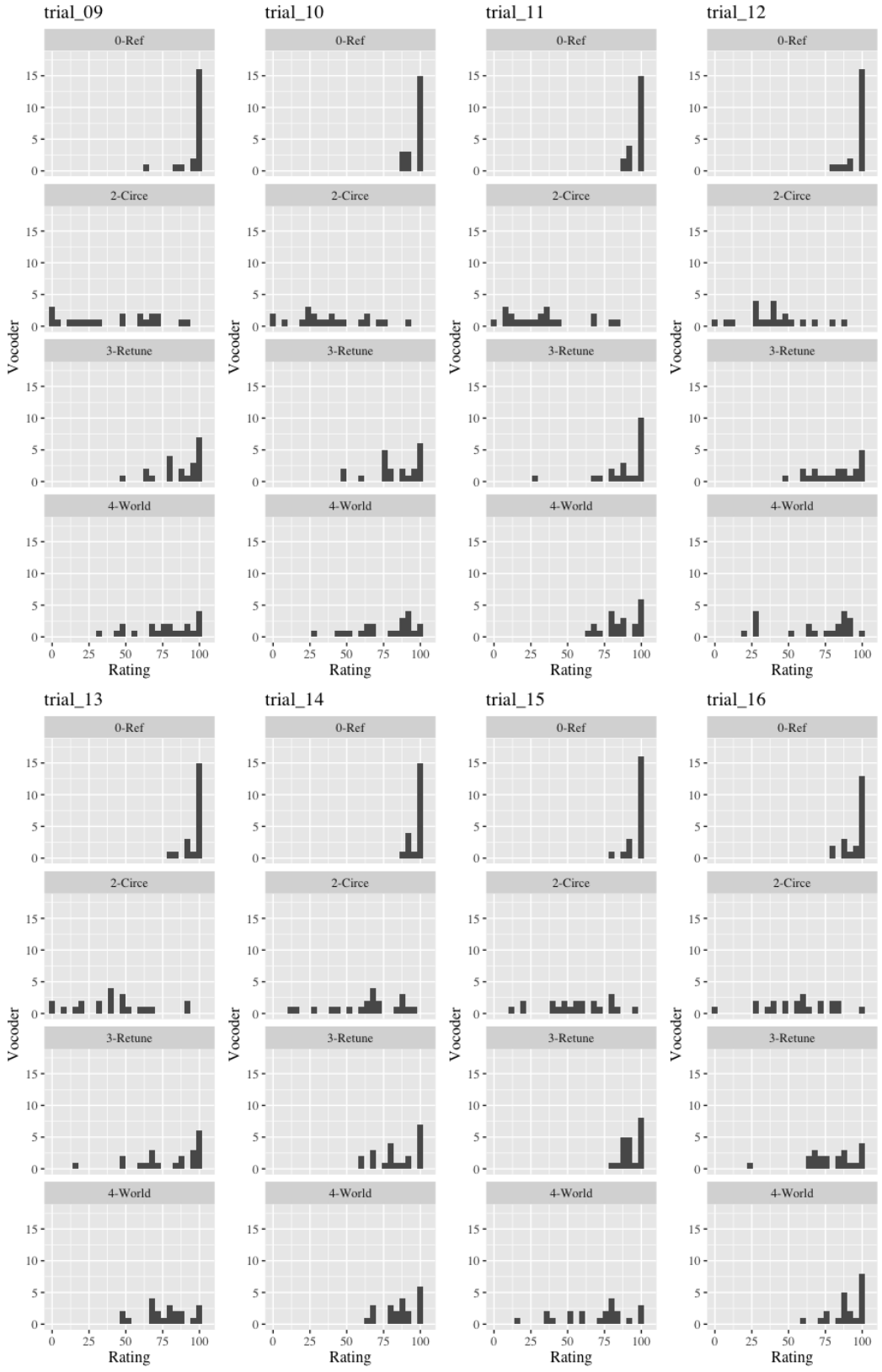


Figure C.16: Histograms per trial for Task D - part 2

C.5.3 ANOVA and Tukey HSD for Task D

Table C.24: Tukey HSD for Task D - before excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe-Ref	-51.878	-55.477	-48.279	< 0.0001
Retune-Ref	-14.631	-18.230	-11.032	< 0.0001
World-Ref	-14.265	-17.864	-10.666	< 0.0001
Retune-Circe	37.247	33.648	40.846	< 0.0001
World-Circe	37.613	34.014	41.212	< 0.0001
World-Retune	0.366	-3.233	3.965	0.9937

Table C.25: ANOVA for Task D - after excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	384535	128178	381.9	$< 2e-16$
Residuals	1020	342362	336		

Table C.26: Tukey HSD for Task D - after excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe-Ref	−52.5	−56.6	−48.3	<0.0001
Retune-Ref	−14.9	−19.1	−10.7	<0.0001
World-Ref	−15.2	−19.4	−11.0	<0.0001
Retune-Circe	37.6	33.4	41.7	<0.0001
World-Circe	37.3	33.1	41.4	<0.0001
World-Retune	−0.3	−4.5	3.9	0.9976

Table C.27: ANOVA for Task D (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	133984	44661	166.1	$< 2e-16$
Residuals	508	136589	269		

Table C.28: Tukey HSD for Task D (Non-Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe-Ref	−42.9	−48.2	−37.6	<0.0001
Retune-Ref	−8.7	−14.0	−3.4	0.0001
World-Ref	−12.0	−17.3	−6.7	<0.0001
Retune-Circe	34.2	28.9	39.4	<0.0001
World-Circe	30.8	25.6	36.1	<0.0001
World-Retune	−3.3	−8.6	2.0	0.3704

Table C.29: ANOVA for Task D (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Vocoder	3	372379	124126	366.7	$< 2e-16$
Residuals	828	280238	338		

Table C.30: Tukey HSD for Task D (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
Circe vs. Ref	−57.4	−62.1	−52.8	<0.0001
Retune vs. Ref	−18.3	−22.9	−13.6	<0.0001
World vs. Ref	−15.6	−20.3	−11.0	<0.0001
Retune vs. Circe	39.1	34.5	43.8	<0.0001
World vs. Circe	41.8	37.1	46.4	<0.0001
World vs. Retune	2.6	−2.0	7.3	0.4640

C.6 Classification Performance for Pitch Corrections Methods - Tasks 1, 2, 3, 4, 5

A performance calculation has been conducted according to the recommendations of the MUSHRA test, wherein the percentage of trials where the reference is rated below 90 is calculated. According to MUSHRA recommendations, subjects with a percentage higher than 15% in all tasks are considered unsuitable and are removed from the data. Nevertheless, such subjects do not negatively impact our results, as their contributions exhibit data distributions (histograms) consistent with ratings provided by good subjects. In these distributions, the most similar samples display similar mean scores and distribution shapes, and the different ones also maintain distribution shapes and mean values. The results of performance can be observed in the table C.31.

Table C.31: Classification Performance (Tasks 1, 2, 3, 4, 5)

Subject	% Incorrect Ref. Classification				
	Task 1	Task 2	Task 3	Task 4	Task 5
A01	5.3	0.0	0.0	15.8	10.5
B02	21.1	5.3	15.8	15.8	57.9
C03	0.0	0.0	0.0	0.0	0.0
D04	0.0	0.0	0.0	0.0	0.0
E05	0.0	0.0	0.0	0.0	0.0
F06	0.0	0.0	0.0	0.0	0.0
G07	10.5	15.8	0.0	0.0	10.5
H08	5.3	15.8	0.0	5.3	21.1
I09	10.5	10.5	0.0	0.0	10.5
J10	0.0	0.0	0.0	0.0	0.0
K11	26.3	47.4	5.3	15.8	0.0
L12	0.0	5.3	0.0	10.5	0.0
M13	10.5	0.0	0.0	0.0	0.0
N14	0.0	0.0	0.0	0.0	0.0
O15	100.0	84.2	100.0	100.0	100.0
P16	5.3	0.0	0.0	0.0	5.3
R18	NaN	NaN	NaN	NaN	NaN
S19	0.0	5.3	0.0	21.1	5.3
T20	10.5	21.1	10.5	5.3	15.8
U21	0.0	10.5	0.0	5.3	0.0
W23	31.6	31.6	26.3	15.8	0.0

C.7 Statistical Support for Task 1

This section includes:

1. Histograms per subjects (Figure C.17)
2. Histograms per trials (Figures C.18 and C.19).

The statistical analysis includes the complete calculations for the Task 1 analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.32).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.33
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.34
3. Non-musicians:
    - ANOVA results are summarized in Table C.35
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.36
4. Musicians:
    - ANOVA results are summarized in Table C.37
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.38

C.7.1 Histograms per subjects for Task 1

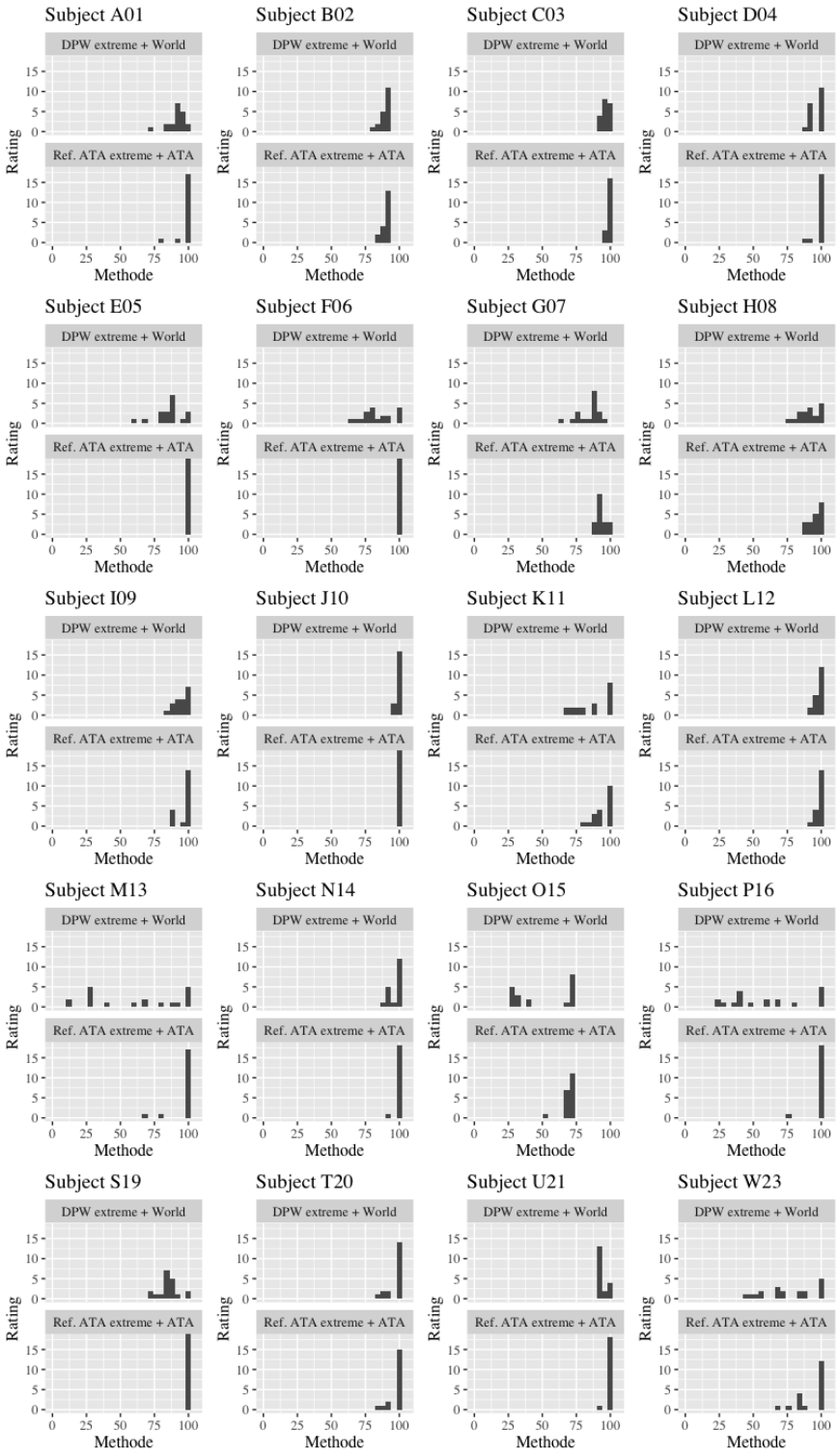


Figure C.17: Histograms per subject for Task 1

C.7.2 Histograms per trials for Task 1

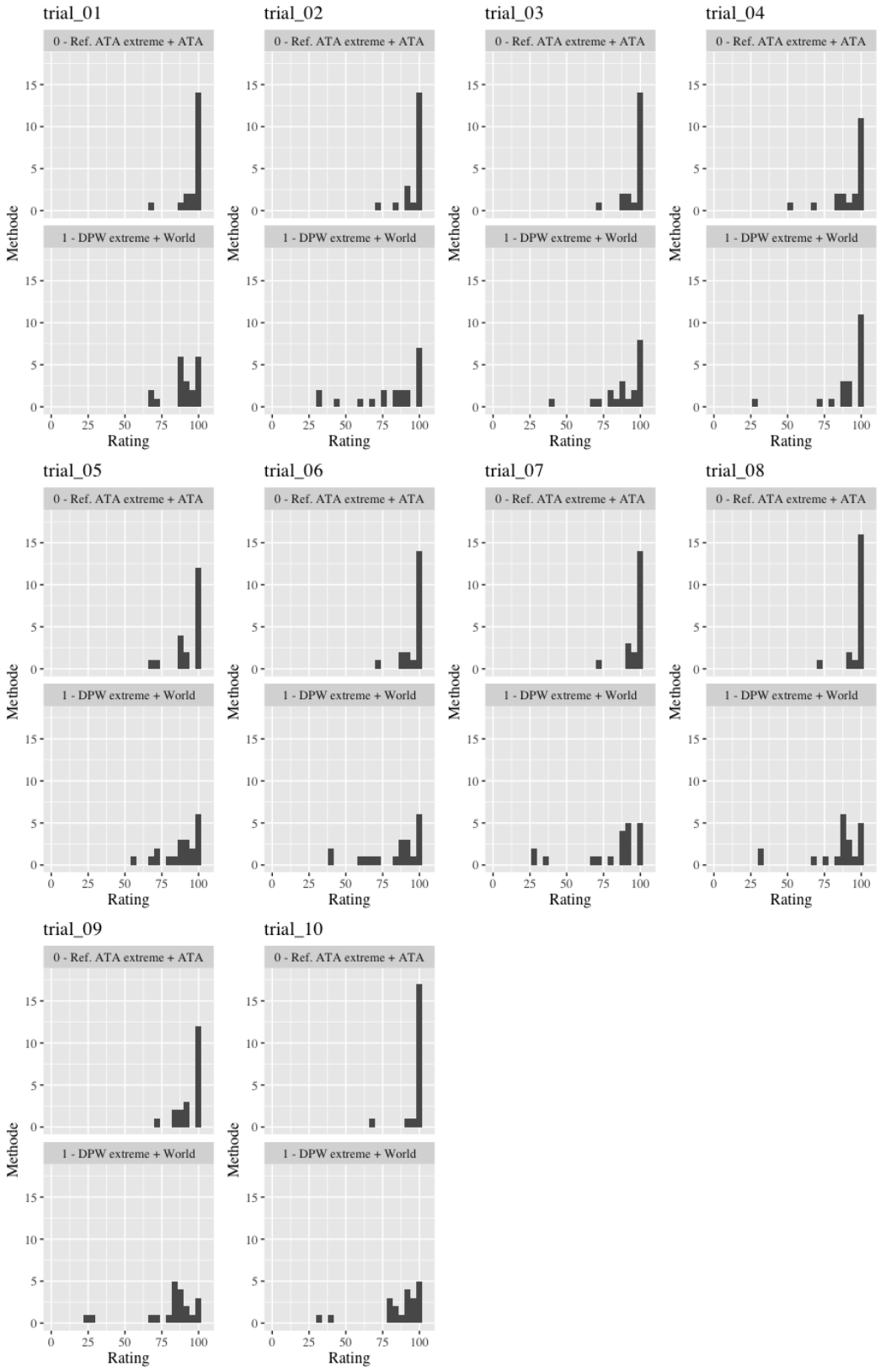


Figure C.18: Histograms per trial for Task 1 - part 1

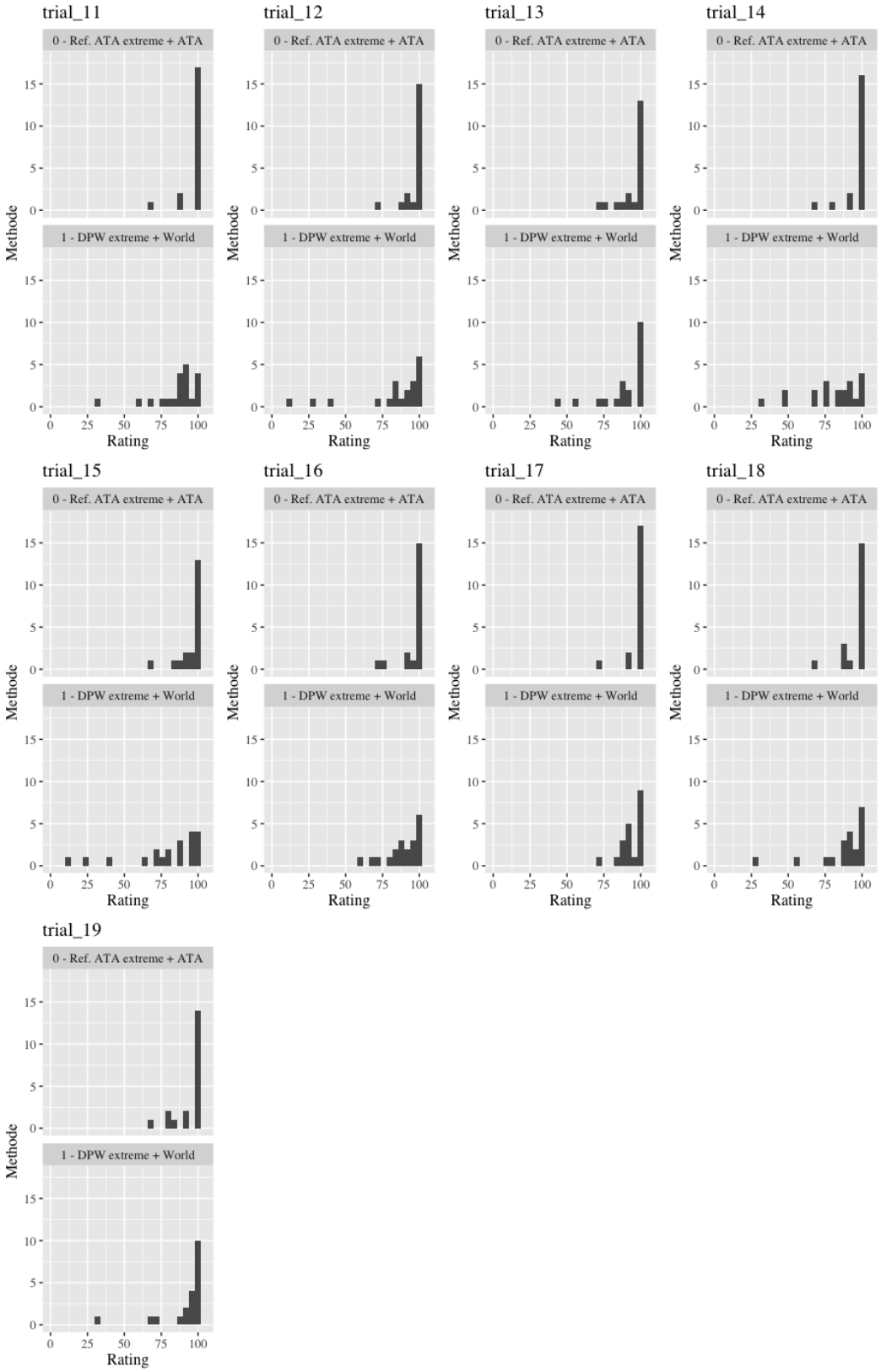


Figure C.19: Histograms per trial for Task 1 - part 2

C.7.3 ANOVA and Tukey HSD for Task 1

Table C.32: Tukey HSD for Task 1 - before excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + ATA	9.9	7.9	12.0	<0.0001

Table C.33: ANOVA for Task 1 - after excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Methode	1	14265	14265	96.96	$< 2e-16$
Residuals	606	89155	147		

Table C.34: Tukey HSD for Task 1 - after excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + ATA	−9.7	−11.6	−7.8	<0.0001

Table C.35: ANOVA for Task 1 (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	2166	2165.7	27.51	$3.21e-07$
Residuals	264	20783	78.7		

Table C.36: Tukey HSD for Task 1 (Non-Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + ATA	−5.7	−7.8	−3.6	<0.0001

Table C.37: ANOVA for Task 1 (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	18426	18426	70.62	$4.69e-16$
Residuals	492	128368	261		

Table C.38: Tukey HSD for Task 1 (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + ATA	−12.2	−15.1	−9.4	<0.0001

C.8 Statistical Support for Task 2

This section includes:

1. Histograms per subjects (Figure C.20)
2. Histograms per trials (Figures C.21 and C.22).

The statistical analysis includes the complete calculations for the Task 2 analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.39).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.40
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.41
3. Non-musicians:
    - ANOVA results are summarized in Table C.42
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.43
4. Musicians:
    - ANOVA results are summarized in Table C.44
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.45

C.8.1 Histograms per subjects for Task 2

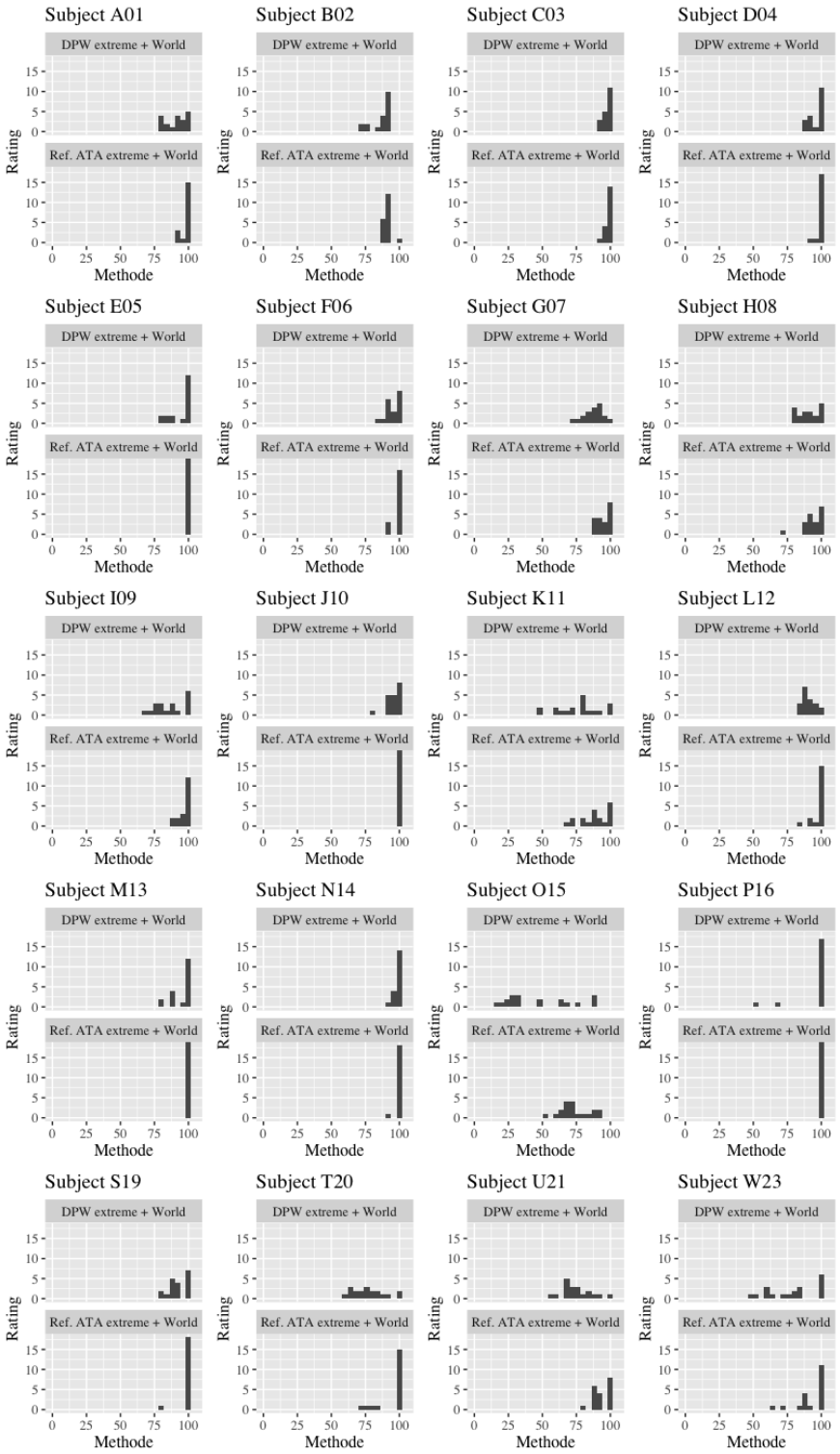


Figure C.20: Histograms per subject for Task 2

C.8.2 Histograms per trials for Task 2

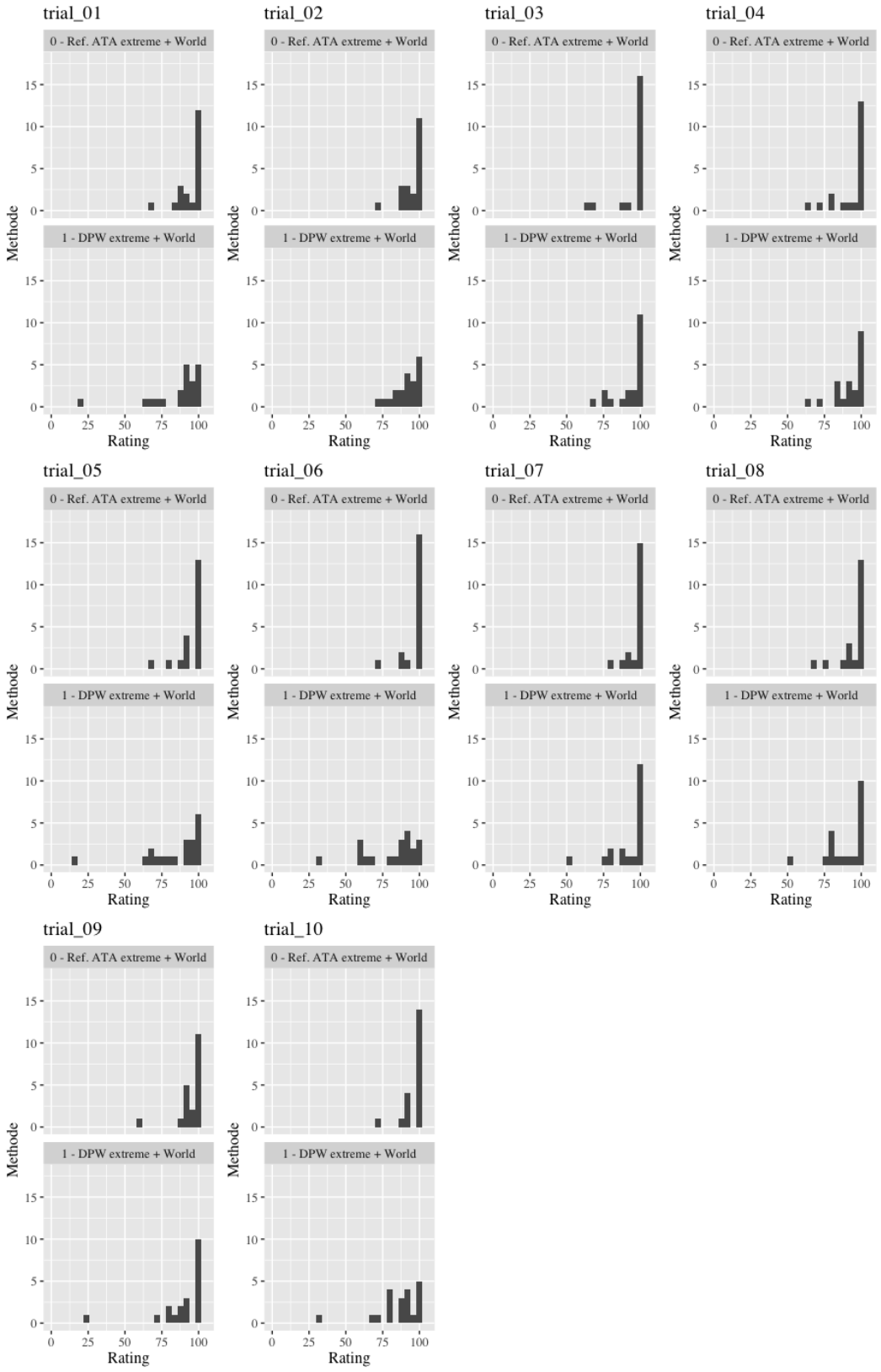


Figure C.21: Histograms per trial for Task 2 - part 1

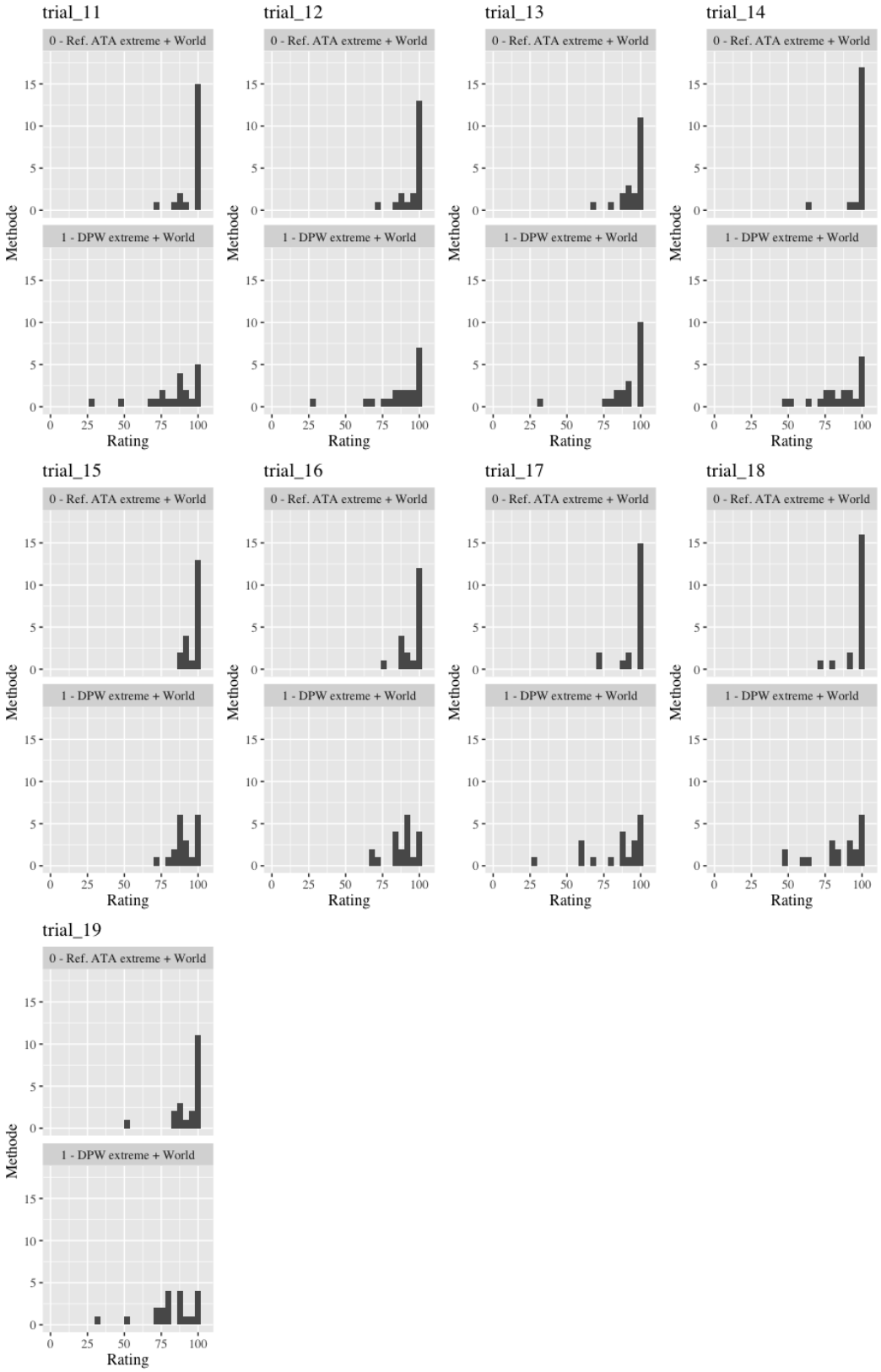


Figure C.22: Histograms per trial for Task 2 - part 2

C.8.3 ANOVA and Tukey HSD for Task 2

Table C.39: Tukey HSD for Task 2 - before excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + World	8.0	6.3	9.8	<0.0001

Table C.40: ANOVA for Task 2 - after excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Methode	1	6429	6429	106.3	$< 2e-16$
Residuals	606	36649	60		

Table C.41: Tukey HSD for Task 2- after excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + World	−6.5	−7.7	−5.3	<0.0001

Table C.42: ANOVA for Task 2 (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	3973	3973	43.8	$2.02e-10$
Residuals	264	23946	91		

Table C.43: Tukey HSD for Task 2 (Non-Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + World	−7.7	−10.0	−5.4	<0.0001

Table C.44: ANOVA for Task 2 (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	8227	8227	44.14	$8.11e-11$
Residuals	492	91700	186		

Table C.45: Tukey HSD for Task 2 (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW extreme + World vs. Ref. ATA extreme + World	−8.2	−10.6	−5.7	<0.0001

C.9 Statistical Support for Task 3

This section includes:

1. Histograms per subjects (Figure C.23)
2. Histograms per trials (Figures C.24 and C.25).

The statistical analysis includes the complete calculations for the Task 3 analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.46).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.47
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.48
3. Non-musicians:
    - ANOVA results are summarized in Table C.49
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.50
4. Musicians:
    - ANOVA results are summarized in Table C.51
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.52

C.9.1 Histograms per subjects for Task 3

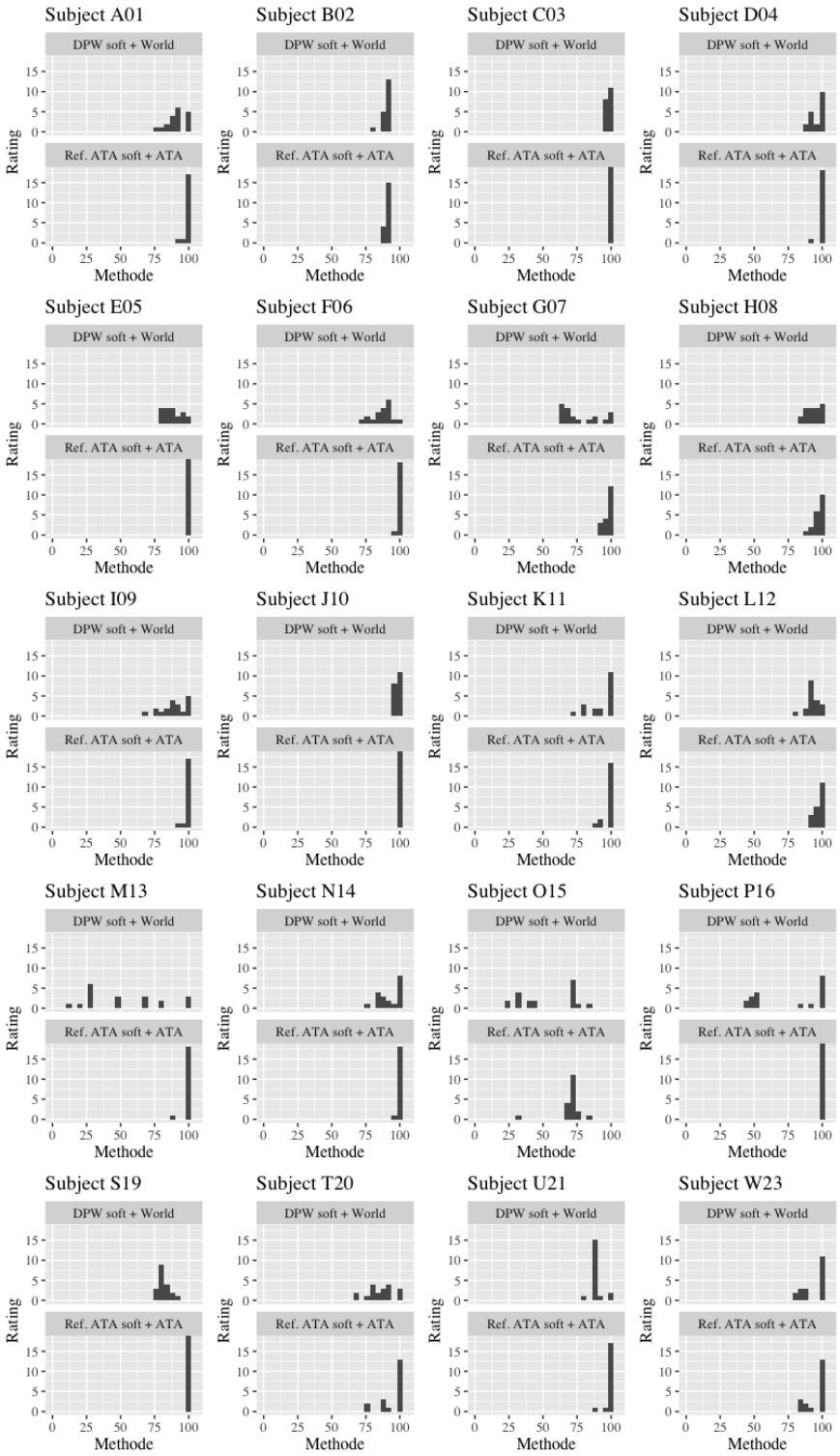


Figure C.23: Histograms per subject for Task 3

C.9.2 Histograms per trials for Task 3

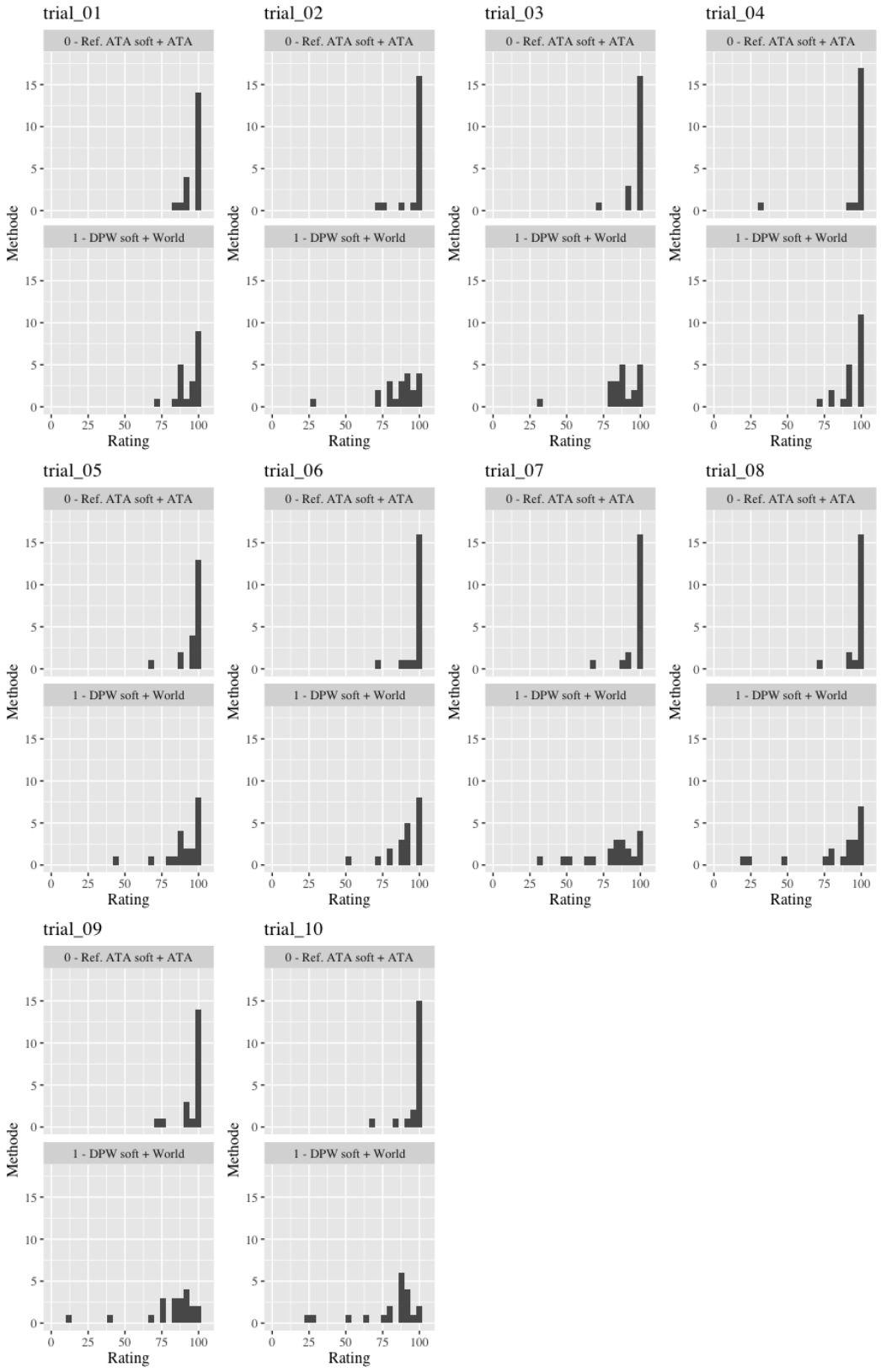


Figure C.24: Histograms per trial for Task 3 - part 1

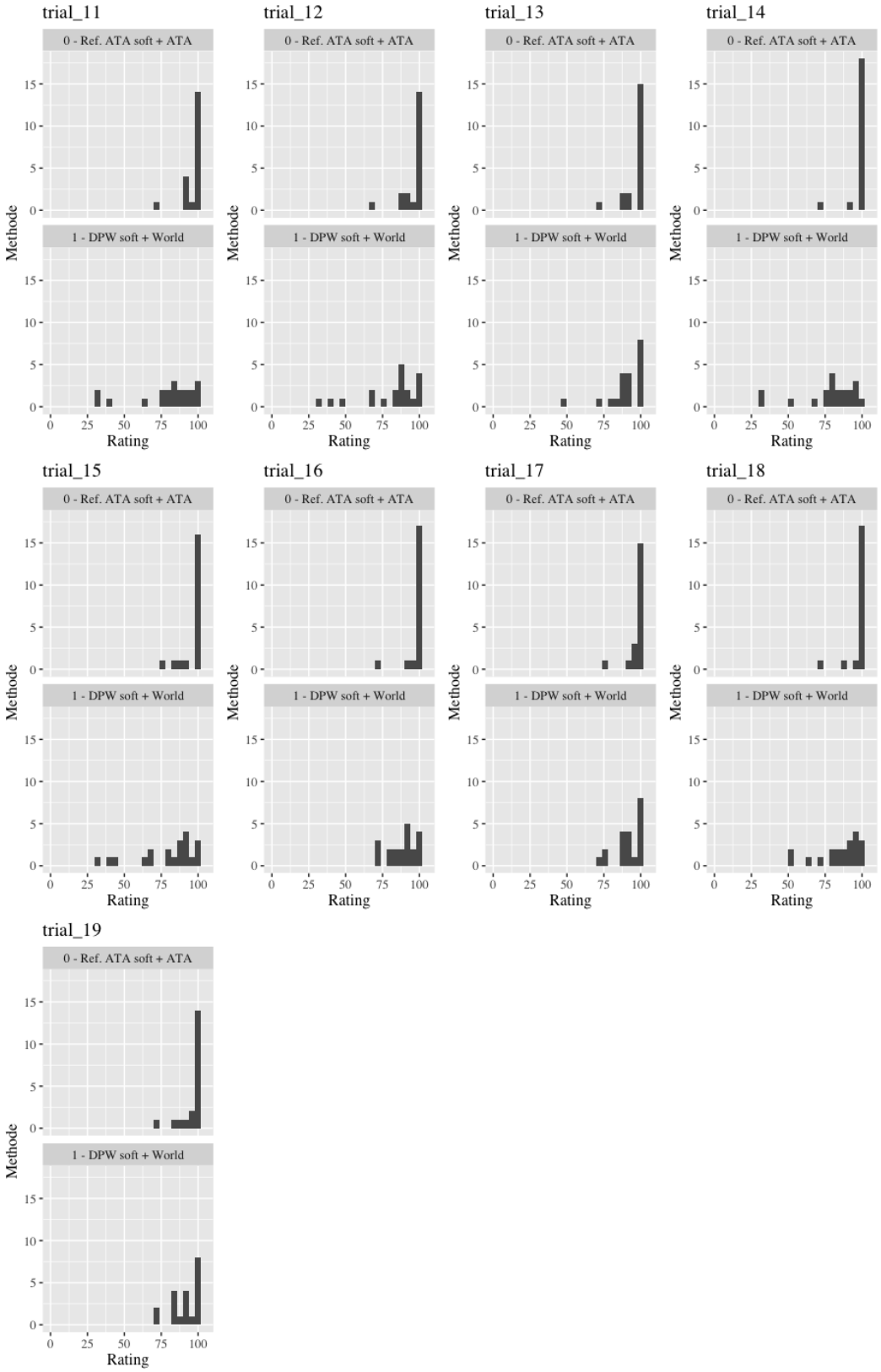


Figure C.25: Histograms per trial for Task 3 - part 2

C.9.3 ANOVA and Tukey HSD for Task 3

Table C.46: Tukey HSD for Task 3 - before excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + ATA	10.8	8.9	12.7	<0.0001

Table C.47: ANOVA for Task 3 - after excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Methode	1	21732	21732	172.3	$< 2e-16$
Residuals	606	76419	126		

Table C.48: Tukey HSD for Task 3 - after excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + ATA	−12.0	−13.7	−10.2	<0.0001

Table C.49: ANOVA for Task 3 (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	2154	2154	53.88	$2.64e-12$
Residuals	264	10557	40		

Table C.50: Tukey HSD for Task 3 (Non-Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + ATA	−5.7	−7.2	−4.2	<0.0001

Table C.51: ANOVA for Task 3 (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	22690	22690	98.6	$< 2e-16$
Residuals	492	113226	230		

Table C.52: Tukey HSD for Task 3 (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + ATA	−13.6	−16.2	−10.9	<0.0001

C.10 Statistical Support for Task 4

This section includes:

1. Histograms per subjects (Figure C.26)
2. Histograms per trials (Figures C.27 and C.28).

The statistical analysis includes the complete calculations for the Task 4 analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.53).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.54
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.55
3. Non-musicians:
    - ANOVA results are summarized in Table C.56
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.57
4. Musicians:
    - ANOVA results are summarized in Table C.58
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.59

C.10.1 Histograms per subjects for Task 4

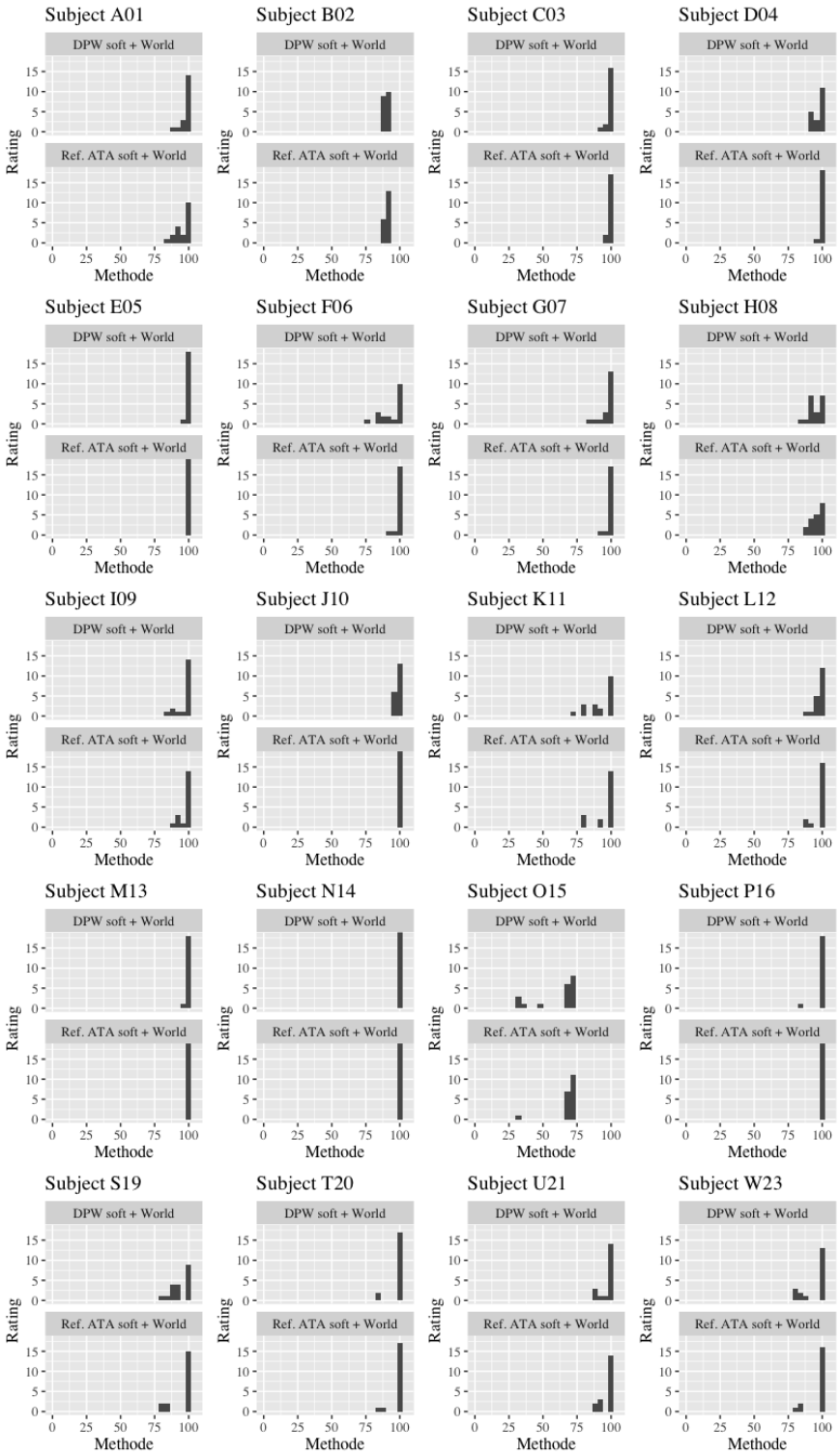


Figure C.26: Histograms per subject for Task 4

C.10.2 Histograms per trials for Task 4

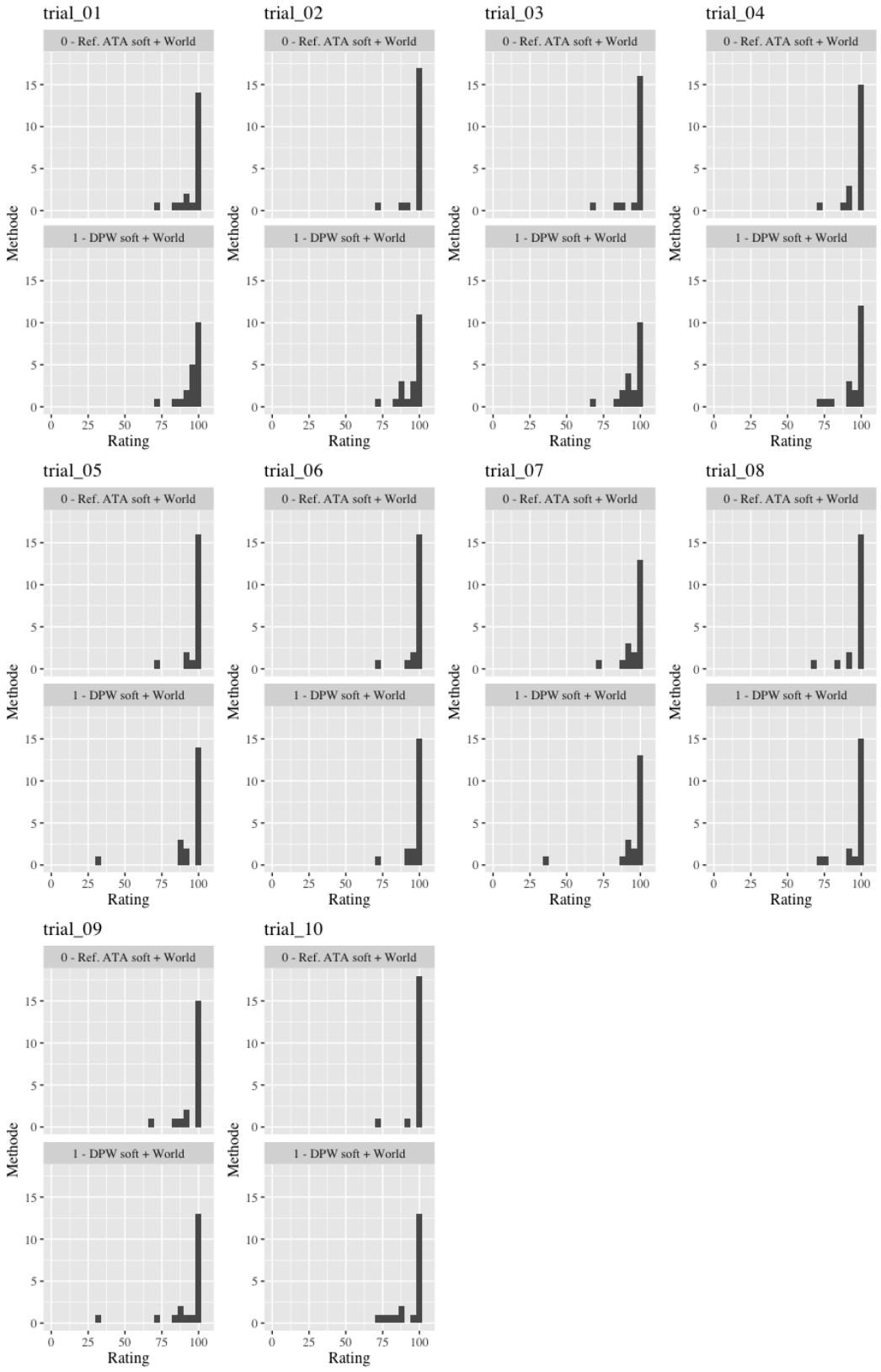


Figure C.27: Histograms per trial for Task 4 - part 1

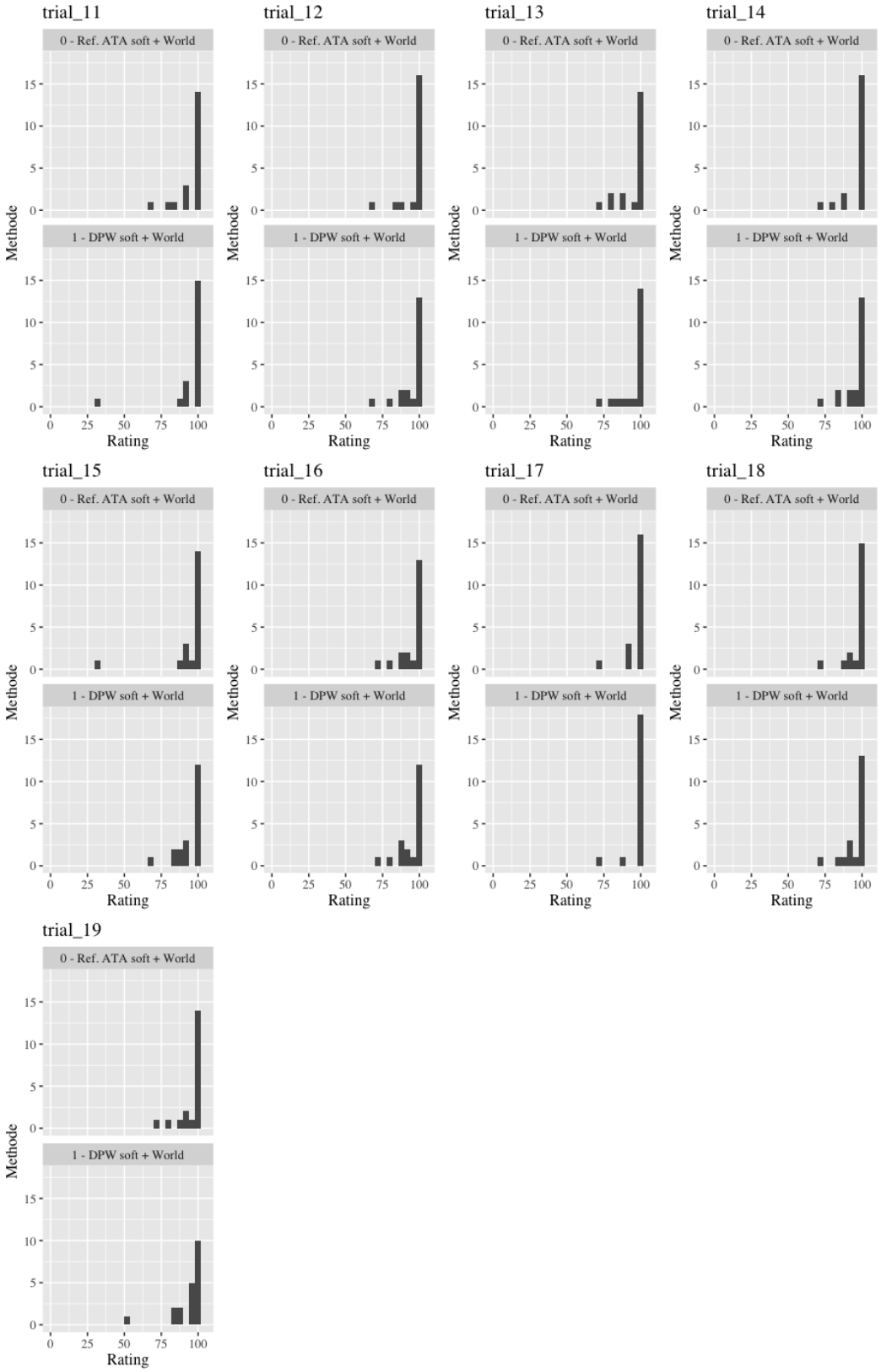


Figure C.28: Histograms per trial for Task 4 - part 2

C.10.3 ANOVA and Tukey HSD for Task 4

Table C.53: Tukey HSD for Task 4 - before excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + World	1.4	0.1	2.6	0.0357

Table C.54: ANOVA for Task 4 - after excluding subj. deemed unsuitable

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Methode	1	114	113.76	7.393	0.00673
Residuals	606	9325	15.39		

Table C.55: Tukey HSD for Task 4 - after excluding subj. deemed unsuitable

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + World	-0.9	-1.5	-0.2	0.0067

Table C.56: ANOVA for Task 4 (No Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	65	64.52	2.186	0.14
Residuals	264	7790	29.51		

Table C.57: Tukey HSD for Task 4 (No Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + World	−1.0	−2.3	0.3	0.1404

Table C.58: ANOVA for Task 4 (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	1	305	304.7	2.837	0.0928
Residuals	492	52858	107.4		

Table C.59: Tukey HSD for Task 4 (Musicians)

Comparison	**Difference**	**Lower**	**Upper**	**p-value**
DPW soft + World vs. Ref. ATA soft + World	−1.6	−3.4	0.3	0.0930

C.11 Statistical Support for Task 5

This section includes:

1. Histograms per subjects (Figures C.29 and C.30)
2. Histograms per trials (Figures C.31 and C.32).

The statistical analysis includes the complete calculations for the Task 5 analysis, encompassing the following cases:

1. Full panel: Post-hoc Tukey HSD multi-comparison (Table C.60).
2. Removing unsuitable subjects:
    - ANOVA results are summarized in Table C.61
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.62
3. Non-musicians:
    - ANOVA results are summarized in Table C.63
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.64
4. Musicians:
    - ANOVA results are summarized in Table C.65
    - Post-hoc Tukey HSD multi-comparison summarized in Table C.66

C.11.1 Histograms per subjects for Task 5

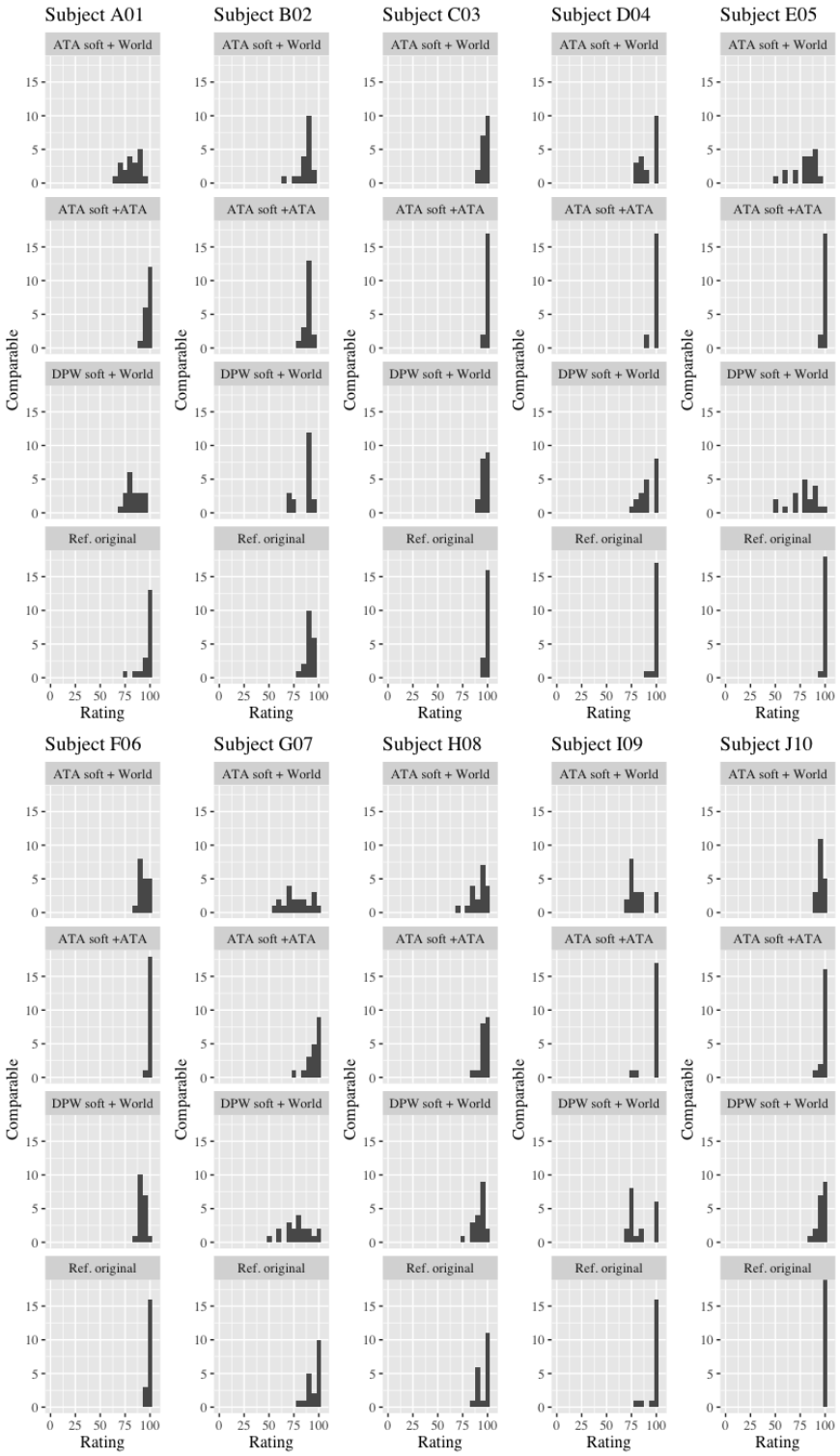


Figure C.29: Histograms per subject for Task 5 - part 1

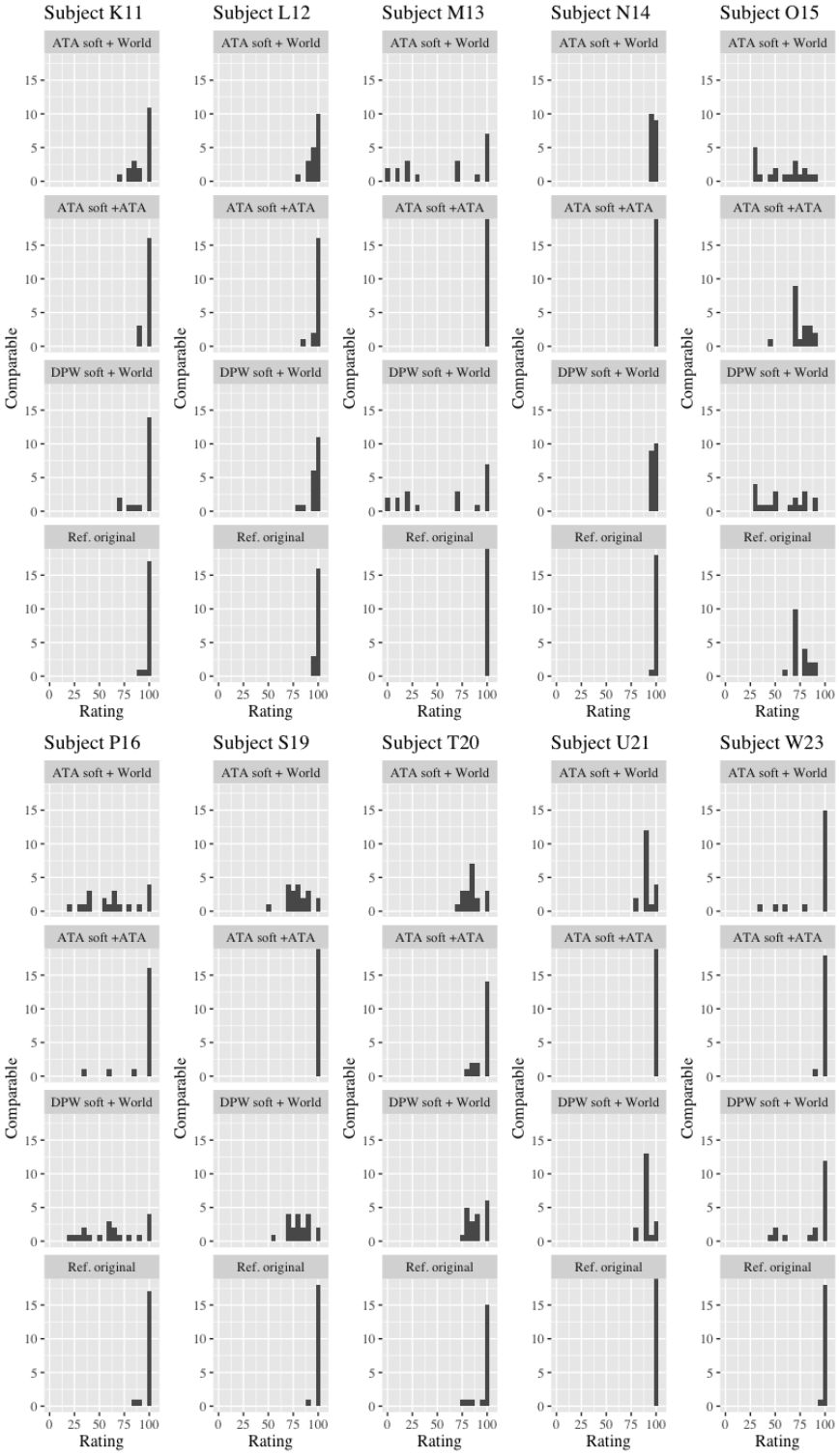


Figure C.30: Histograms per subject for Task 5 - part 2

C.11.2 Histograms per trials for Task 5

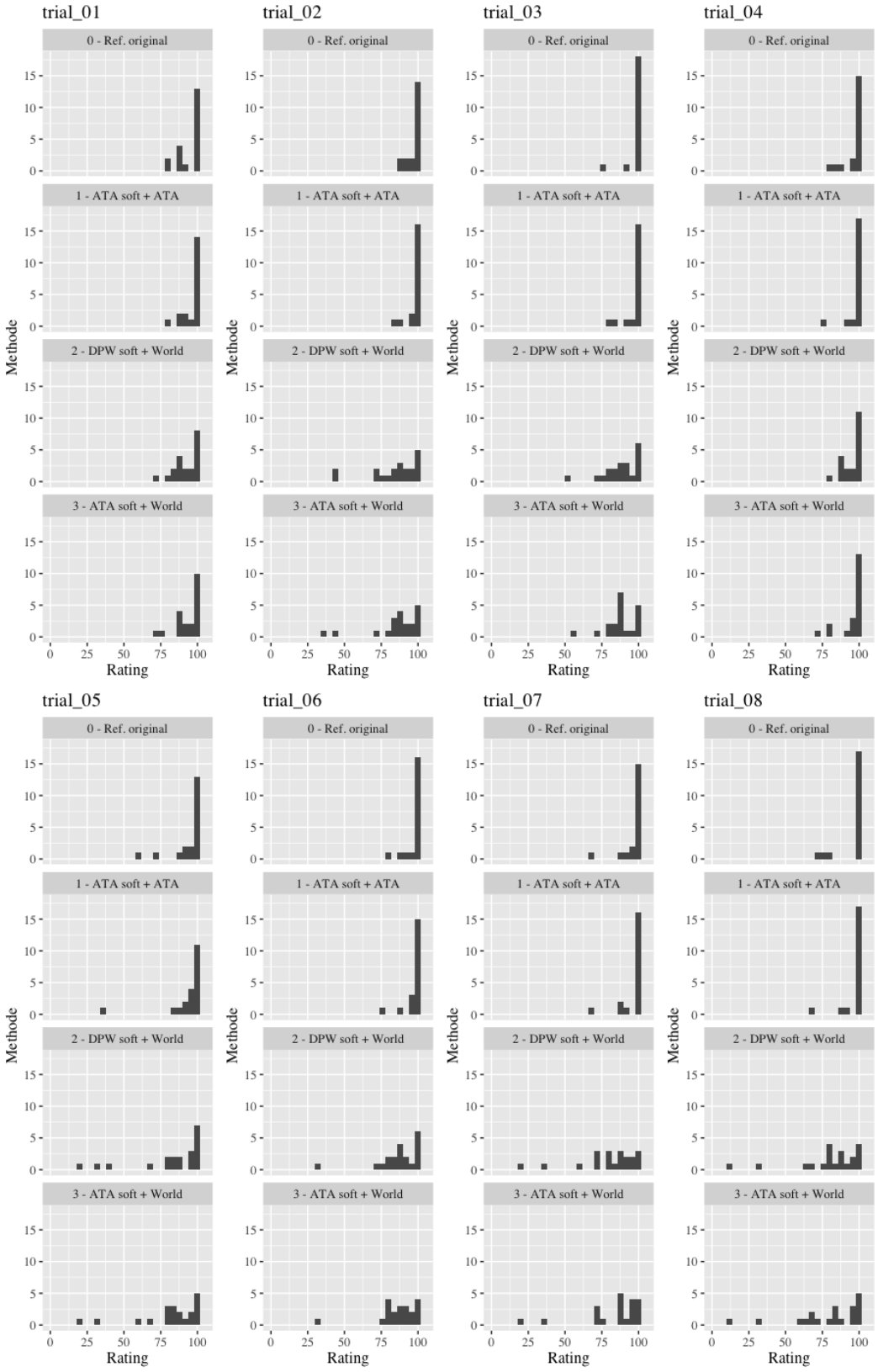


Figure C.31: Histograms per trial for Task 5 - part 1

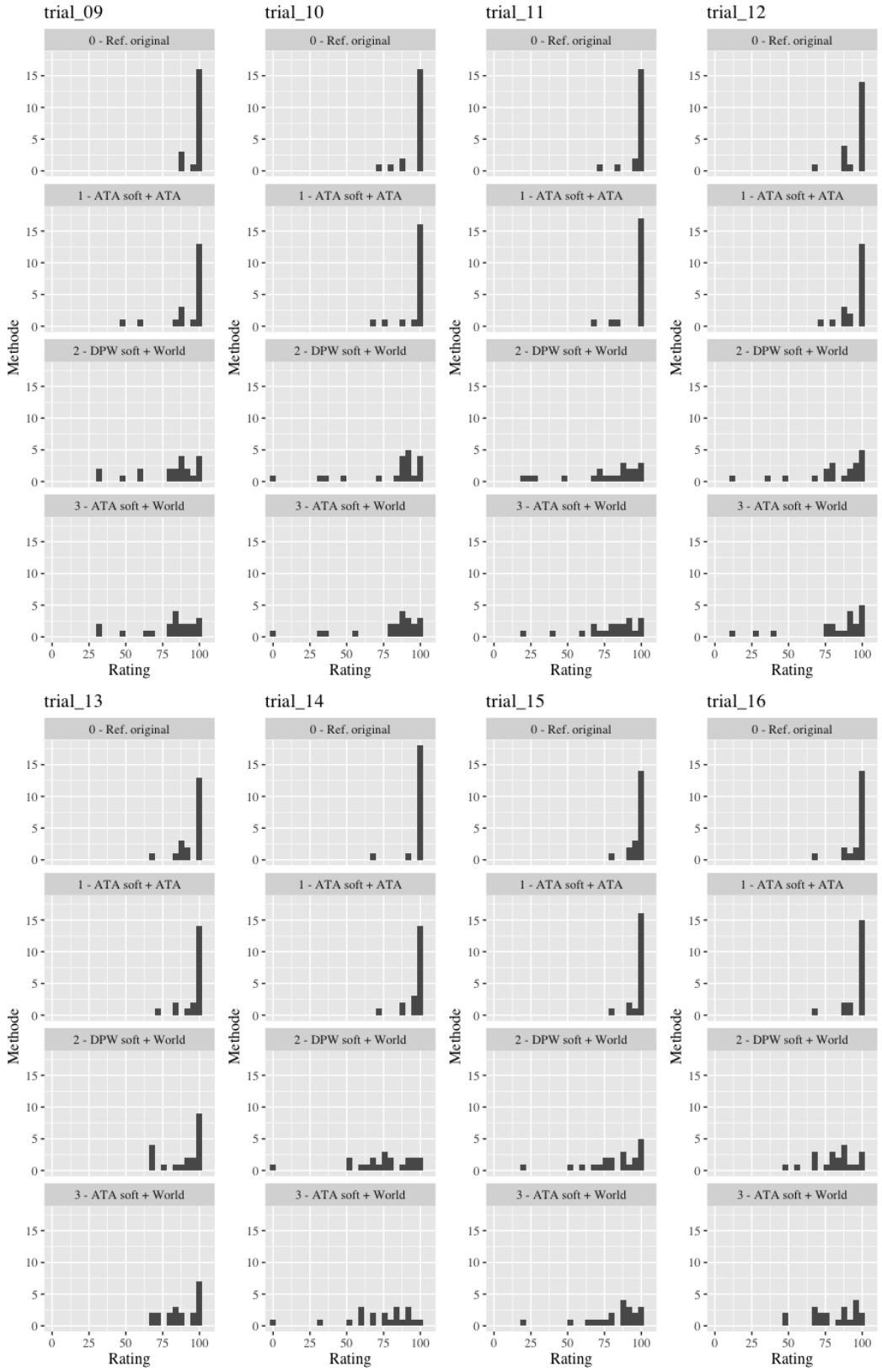


Figure C.32: Histograms per trial for Task 5 - part 2

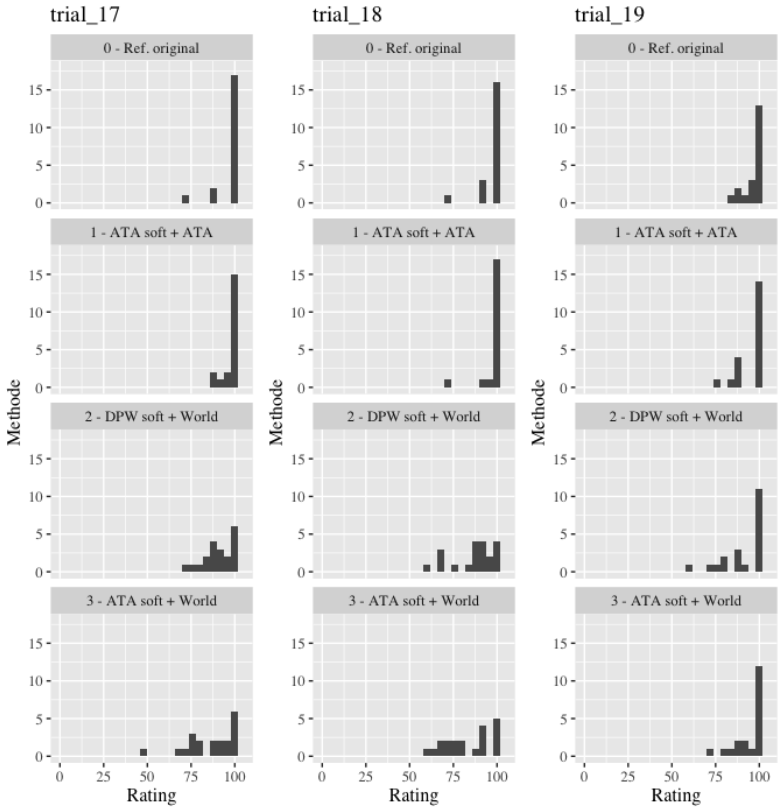


Figure C.33: Histograms per trial for Task 5 - part 3

C.11.3 ANOVA and Tukey HSD for Task 5

Table C.60: Tukey HSD for Task 5 - before excluding subj. deemed unsuitable

Comparison	Diff.	Lower	Upper	p-value
ATA soft + ATA vs Ref. original	−0.3	−2.9	2.4	0.9950
DPW soft + World vs Ref. original	−12.2	−14.9	−9.6	<0.0001
ATA soft + World vs Ref. original	−12.5	−15.2	−9.8	<0.0001
DPW soft + World vs ATA soft + ATA	−12.0	−14.6	−9.3	<0.0001
ATA soft + World vs ATA soft + ATA	−12.2	−14.9	−9.6	<0.0001
ATA soft + World vs DPW soft + World	−0.3	−2.9	2.4	0.9940

Table C.61: ANOVA for Task 5 - after excluding subjects deemed unsuitable

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Method	3	51661	17220	97.22	<2e-16
Residuals	1212	214674	177		

Table C.62: Tukey HSD for Task 5 - after excluding subj. deemed unsuitable

Comparison	Diff.	Lower	Upper	p-value
ATA soft + ATA vs Ref. original	−0.2	−3.0	2.6	0.9980
DPW soft + World vs Ref. original	−12.9	−15.7	−10.1	<0.0001
ATA soft + World vs Ref. original	−13.4	−16.1	−10.6	<0.0001
DPW soft + World vs ATA soft + ATA	−12.7	−15.5	−9.9	<0.0001
ATA soft + World vs ATA soft + ATA	−13.2	−15.9	−10.4	<0.0001
ATA soft + World vs DPW soft + World	−0.5	−3.2	2.3	0.9740

Table C.63: ANOVA for Task 5 (Non-Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	3	7116	2372.0	29.6	$< 2e-16$
Residuals	528	42308	80.1		

Table C.64: Tukey HSD for Task 5 (Non-Musicians)

Comparison	**Diff.**	**Lower**	**Upper**	**p-value**
ATA soft + ATA vs. Ref. original	0.1	−2.8	2.9	0.9990
DPW soft + World vs. Ref. original	−7.1	−10.0	−4.3	<0.0001
ATA soft + World vs. Ref. original	−7.4	−10.2	−4.6	<0.0001
DPW soft + World vs. ATA soft + ATA	−7.2	−10.0	−4.4	<0.0001
ATA soft + World vs. ATA soft + ATA	−7.5	−10.3	−4.7	<0.0001
ATA soft + World vs. DPW soft + World	−0.3	−3.1	2.6	0.9950

Table C.65: ANOVA for Task 5 (Musicians)

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Method	3	54710	18237	70.11	$< 2e-16$ ***
Residuals	984	255959	260		

Table C.66: Tukey HSD for Task 5 (Musicians)

Comparison	**Diff.**	**Lower**	**Upper**	**p-value**
ATA soft + ATA vs. Ref. original	−0.4	−4.2	3.3	0.9910
DPW soft + World vs. Ref. original	−15.0	−18.7	−11.2	<0.0001
ATA soft + World vs. Ref. original	−15.2	−19.0	−11.5	<0.0001
DPW soft + World vs. ATA soft + ATA	−14.5	−18.3	−10.8	<0.0001
ATA soft + World vs. ATA soft + ATA	−14.8	−18.5	−11.1	<0.0001
ATA soft + World vs. DPW soft + World	−0.3	−4.0	3.5	0.9980

Appendix D

Paper: Dynamic pitch warping for expressive vocal retuning

Daniel Hernan Molina Villota, Christophe d'Alessandro, Olivier Perrotin. Dynamic pitch warping for expressive vocal retuning. 26th International Conference on Digital Audio Effects (DAFx23), Sep 2023, Copenhagen, Denmark. pp.118-125. hal-04256554f.

A sound support is contained in a folder, file names correspond to number of figures and configurations.

DYNAMIC PITCH WARPING FOR EXPRESSIVE VOCAL RETUNING

Daniel Hernan Molina Villota
Institut Jean Le Rond d'Alembert
Equipe Lutheries-Acoustique-Musique
Sorbonne Université
Centre National de la Recherche Scientifique
Paris, France
daniel.molina_villota
@sorbonne-universite.fr

Christophe d'Alessandro
Institut Jean Le Rond d'Alembert
Equipe Lutheries-Acoustique-Musique
Sorbonne Université
Centre National de la Recherche Scientifique
Paris, France
christophe.dalessandro
@sorbonne-universite.fr

Olivier Perrotin
Université Grenoble Alpes
CNRS, Grenoble INP
GIPSA-lab
Grenoble, France
olivier.perrotin
@grenoble-inp.fr

ABSTRACT

This work introduces the use of the Dynamic Pitch Warping (DPW) method for automatic pitch correction of singing voice audio signals. DPW is designed to dynamically tune any pitch trajectory to a predefined scale while preserving its expressive ornamentation. DPW has three degrees of freedom to modify the fundamental frequency (f_0) signal: detection interval, critical time, and transition time. Together, these parameters allow us to define a pitch velocity condition that triggers an adaptive correction of the pitch trajectory (pitch warping). We compared our approach to Antares Autotune (the most commonly used software brand, abbreviated as ATA in this article). The pitch correction in ATA has two degrees of freedom: a triggering threshold (flextune) and the transition time (retune speed). The pitch trajectories that we compare were extracted from autotuned-in-ATA audio signals, and the DPW algorithm implemented over the f_0 of the input audio tracks. We studied specifically pitch correction for three typical situations of f_0 curves: staircase, vibrato, free-path. We measured the proximity of the corrected pitch trajectories to the original ones for each case obtaining that the DPW pitch correction method is better to preserve vibrato while keeping the f_0 free path. In contrast, ATA is more effective in generating staircase curves, but fails for not-small vibratos and free-path curves. We have also implemented an off-line automatic picth tuner using DPW.

1. INTRODUCTION

Pitch correction (or automatic pitch tuning) is nowadays one of the most commonly used digital audio effects for vocal music. Initially known as the "Cher" effect, the audible distortion produced by sharp pitch transition in retuned singing became appreciated on its own in popular electronic music. The sharp transition is a case of use where all minor expressive singing variations are flattened. Noticeable gliding appears often in the transitions between notes. The success of Autotune in the music industry has sparked much discussion and debate. Some argue that it is a tool that helps artists achieve a perfect pitch singing, while others criticise its use as it can lead to a loss of natural expression and emotion in the music. Despite this, Autotune has become a staple in modern music production and is used in various genres such as pop, hip-hop, and electronic music [1]. Although it is a common practice to use DAFx effects which involve perceptual features such as [2] melody (pitch), source (timbre, [3]), or space [4],pitch correction is one of the most commonly used. I became a stylistic signature for many popular music genre.

Antares Autotune (ATA)[1] is a digital audio effect developed by H. Hildebrand in 1997 [5] and its enduring popularity has spanned over 25 years. ATA uses an autocorrelation method that was initially developed for seismic imaging, with the help of short-time Fourier transform. Although the initial purpose of ATA was not to enrich the voice with a new vocoder-like audio effect but to correct out-of-tune melodies, the unique electronic texture produced has been embraced in popular music and has even become a hallmark of specific musical styles, often employed systematically. ATA offers two use cases: one the one hand pitch correction is used for better rendering of out of tune singing and on the other hand the distortion effect occurring extreme correction situations is appreciated on its own. The need for melodic correction also appeared in digital music instruments (DMI) [6, 7, 8, 9]. These DMIs use interfaces with particular features that involve learnability, explorability, and controllability [10]. A new pitch tuning correction, Dynamic Pitch Warping (DPW) [11], has been developed for performative vocal synthesis in Cantor Digitalis [8] where the fundamental frequency (pitch) is controlled in real-time with the help of a stylus on a graphic tablet. Pitch correction helps for singing accurate notes. However, it is very important to preserve small expressive ornaments like vibrato [12] without flattening the notes to preserve naturalness.

The purpose of this paper is to study the DPW pitch correction method. This method was designed to preserve expressive variations like vibrato while adjusting the main shape of the f_0 curve to a predefined scale. We identify three cases of particular interest: abrupt pitch transitions (staircase notes), notes with vibrato and free path curves that should not be corrected. The results of this paper allow us to open perspectives for developing dynamic and singer-controlled vocal digital audio effects that are able to preserve expressive ornaments in real-time. Section 2 presents a review of the pitch correction method studied (ATA and DPW). Section 3 compares DPW and ATA on typical pitch patterns. Section 4 presents the off-line implementation of DPW for audio signals.

2. PITCH CORRECTION SYSTEMS

An audio pitch correction system contains three parts: a pitch detection algorithm (PDA), a pitch correction algorithm, and finally

[1] https://www.antarestech.com/ last checked: 6 April 2023

a pitch warping modification (vocoder). The present paper aims to apply DPW as a pitch correction algorithm for vocal speech intonation. DPW offers three control parameters when other correction methods have one or two parameters. DPW uses an adaptive function, the term "adaptive" is related to the adaptive digital audio effects (aDAFx) that are recent solutions designed to respond to changes in the input signal and adjust specific audio parameters accordingly to it, thanks to specific denominated adaptive functions. These kind of effects are more dynamic and responsive that the traditional DAFx, some examples of aDAFx being the compressor, the expander and the limiter (auto-adaptive on loudness).

Along this line, several DMIs have introduced the use of pitch correction methods to improve the expressivity of musical user interfaces. That is the case for devices such as the Continuum Fingerboard [6, 7][2], the Seaboard[13][3], Garageband[4], TouchKeys[14], and Cantor Digitalis [8][5]. The latter is particularly interesting since it uses a Dynamic Pitch Warping method to correct the continuous position of the pitch controller relative to a pitch scale. The corresponding adaptive warping function proposed by Perrotin and d'Alessandro [11] attracts real pitch values towards integer values, using a MIDI scale. The integer values are tuned notes. DPW is based on a pitch velocity condition expressed as the pitch stability within a pitch interval during a critical time threshold before triggering the automatic correction. We will review the warping methods applied in ATA and DPW in the following two subsections.

2.1. Autotune Antares

Autotune was developed by H. Hildebrand using techniques originally developed for mapping the Earth's subsurface and is considered a time-domain vocoder that modifies the signal both on the frequency and time domain using a short-time Fourier transform with a window function to frame the inner transform. Autotune is a full pitch correction system including the three steps described above: pitch detection, pitch correction and pitch modification. We present in this section the pitch correction method. For this purpose, the sung notes are shifted to the closest note in a predefined scale, and the transition is carried out over a duration equal to a transition time (named "retune speed" on ATA). Autotune also includes the flextune parameter, which acts as a threshold for the correction and represents the size of the neighborhood of a note in which a pitch correction can be triggered.

Due to lack of detail in the patent [5], the ATA algorithm can only be reproduced for an extreme correction case, meaning a value 0 on the Decay parameter in the patent of ATA (internal parameter of the code, and related to the retune speed parameter). This case corresponds to force the input trajectory to match integer MIDI values, i.e., the target notes. For the non-zero Decay cases we cannot reproduce the algorithm as the patent doesn't describe exactly the configuration of the smoothing step. To treat cases with non-zero transition time we will apply the ATA VST on audio signals and then extract the retuned f_0 to study correction actually carried on.

[2] `https://www.hakenaudio.com/continuum-fingerboard` last checked: 25 may 2023

[3] `https://www.roli.com` last checked: 25 may 2023

[4] `https://www.apple.com/mac/garageband/` last checked: 25 may 2023

[5] `http://www.lam.jussieu.fr/cantordigitalis/` last checked: 25 may 2023

2.2. Dynamic Pitch Warping

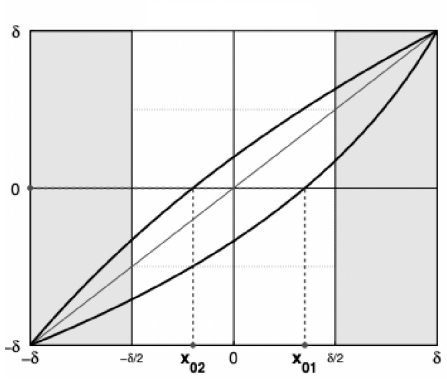


Figure 1: *The arc of curvature for the dynamic pitch correction method, took from [11]*

DPW is a real-time pitch correction method developed by Perrotin and d'Alessandro for Cantor Digitalis. Although it was originally designed to correct a driven by stylus pitch on a graphic tablet, we aim to use DPW for vocal correction. DPW relies on pitch velocity (speed) to trigger an adaptive correction that modifies the input f_0 curve gradually, enabling the output f_0 to converge to the nearest semitone on the MIDI scale. When pitch velocity falls below a threshold, DPW smoothly shifts subsequent f_0 values to converge to a tuned semitone, while preserving some expressive motion of the original f_0 value. The adaptive function remains static when the pitch velocity condition is not met, allowing intended notes to be corrected while retaining expressiveness and preserving all dynamics for non-corrected notes. To review the method, we first analyze the isolated adaptive function, as seen in Figure 1 that maps the input f_0 (x axis) to the output f_0 (y axis). On both axes, zero represents the closest target (ideal) pitch, and $-\delta$ and $+\delta$ correspond to the previous and next notes on the discrete target pitch scale, respectively. While it works on any arbitrary scale, $\delta = 1$ when working with semitones. For input pitch x_{01}, the closest target note is zero. Therefore, at the time the correction is triggered, the corresponding adaptive function that is initially diagonal will smoothly shift towards the lowest arc-shaped curve, to eventually map the input f_0 to the pitch target (zero) as output f_0. The adaptive function then becomes static until it is newly triggered. To avoid introducing a constant shift on the full pitch range, the adaptive function is arc-shaped so that if the input moves from the x_{01} value to the neighbour notes on the pitch scale ($-\delta$ or $+\delta$), the output f_0 will continuously reach $-\delta$ or $+\delta$. If those boundaries are reached, the adaptive function goes back to a linear mapping between input and output, until it is triggered again for a new input.

The adaptive function is derived from the analytic definition of an arc. To ease formulation, the inverse function is first defined:

$$x(y) = Ae^{\gamma(y+B)} + C \tag{1}$$

where the parameters A, B, C and γ can be calculated from the boundary conditions, i.e., the arc must satisfy $x(\pm\delta) = \pm\delta$. If we use this condition, we can write A and C in terms of γ, δ, and B

as follows:

$$C = -\delta\left(1 + \frac{2}{e^{2\gamma\delta} - 1}\right), A = 2\delta\frac{e^{\gamma(\delta - B)}}{e^{2\gamma\delta} - 1} \quad (2)$$

Replacing these values in the original equation 1, we find that the dependency on B disappears. Furthermore, the function is not defined for $\gamma = 0$, but it corresponds to an absence of correction, i.e., the mapping is linear. So the function of the arc curvature can be written as:

$$x(y) = \begin{cases} \delta\left[2\frac{e^{\gamma(\delta+y)}-1}{e^{2\gamma\delta}-1} - 1\right] & if \quad \gamma \neq 0 \\ y & if \quad \gamma = 0 \end{cases} \quad (3)$$

The adaptive warping function is defined as the inverse of 3:

$$y(x) = \begin{cases} 1/\gamma\left[\log\left[(e^{2\gamma\delta} - 1)(\frac{x}{\delta} + 1)\frac{1}{2} + 1\right]\right] - \delta & if \quad \gamma \neq 0 \\ x & if \quad \gamma = 0 \end{cases} \quad (4)$$

Where γ is the factor of correction, y is the output pitch after the correction, and x is the input pitch. When the correction is triggered (at that moment $x = x_o$) , the value of $\gamma = \gamma_0$ can be calculated from the input value x_0 to ensure that $y(x_0) = 0$ following the equation:

$$\gamma_0 = \frac{1}{\delta}\log\left(\frac{\delta - x_0}{\delta + x_0}\right) \quad (5)$$

The DPW has two stages that can be seen on Figure 2. One is the triggering part and the other is the warping stage. For the correction to be triggered, the pitch trajectory has to be stable enough, i.e., it has to stay within an interval of detection (ID) during a critical time (T_c) [11]. If these conditions are met, we can calculate the curvature γ_0 given the input pitch at triggering time (x_0 in the definition, f_0 for us). To ensure a smooth transition, γ is linearly interpolated from 0 (linear mapping) to γ_0. This transition spans a time interval denominated transition time (T_t). When the transition is completed, the input pitch has converged to the closest integer notes on the midi scale. This transition is carried out similarly to the static case of ATA, not over the frequency but over the γ value, then f_0 (input) is warped with the adaptive function.

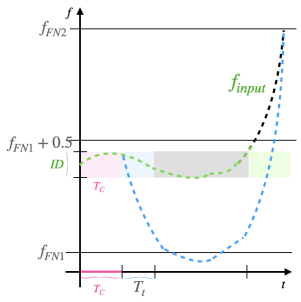


Figure 2: *Illustration of the dynamics of DPW. The input f_0 (green curve) is stable in a detection interval ID during the critical time T_c (pink region). The correction is triggered during the transition time T_t (blue region). The input f_0 can vary continuously during the transition time, until it reaches the next semitone on the pitch scale (integer, black).*

3. CASE STUDIES OF PITCH CORRECTION

In this section, we compare both ATA and DPW methods. Firstly, we want to show the difference between the methods through a simple case. We take as example a constant flat note (C♯) with a pitch shift of 0.15 semitone (ST), and we use both methods to correct it. In Figure 3, we see a DPW correction (blue) triggered with the following parameters: $ID = 0.1$ ST, $T_c = 0.5$ s, and $T_t = 0.5$ s. The ATA correction (red) has a retune speed equal to T_t. We have chosen a non-zero value for T_c to show the inclusion of the new parameter. The critical time is the main difference between both methods. While it introduces a triggering delay in DPW, we find similar results for both corrections once after that trigger.

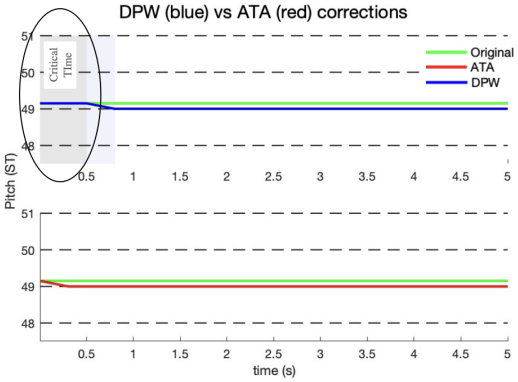


Figure 3: *DPW correction (blue curve) and ATA correction (red curve) of a constant input pitch (green curve).*

3.1. Extreme correction with zero transition time parameter

We denominate extreme correction to a full discretization of the input pitch trajectory. To check the extreme correction, we chose two typical examples: the first one is a glissando, and the second is a melody taken from [15]. After trying some configurations, we have found a combination of parameters that provides similar results with both methods. For DPW, we have chosen the parameters $T_c = 0$ s, $ID = 0.01$ ST, and $T = 0.001$ s (the minimal value). For ATA we choose just the zero retune speed the minimal value), that as described in the patent generates discrete notes (integers on ST scale). We can see the results in Figure 4 and 5. The f_o-signal treated with DPW is in blue, and the one treated with ATA is in red and the original is in green.

3.2. Expressive Correction with ATA

One of the most important artifacts of vocal expression is vibrato. Expressive Correction is the term we use here to refer to a fast transition within pitch correction that correspond to oscillatory ornaments, particularly vibrato. As we will see, a vibrato with a small amplitude can be shifted around the target pitch with DPW, while it is not well centered under the ATA correction. The expressive correction requires a non-zero transition time parameter. We don't have access to the full implementation of the transition time parameter in ATA (also referred to as the Decay parameter in the

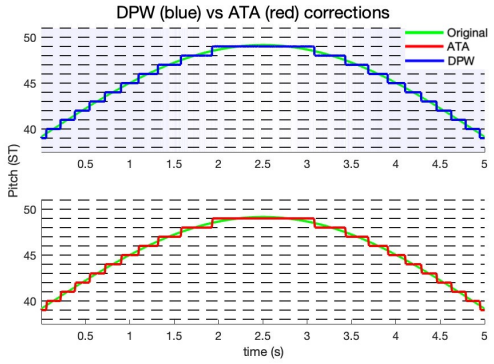


Figure 4: *Extreme correction for a glissando*

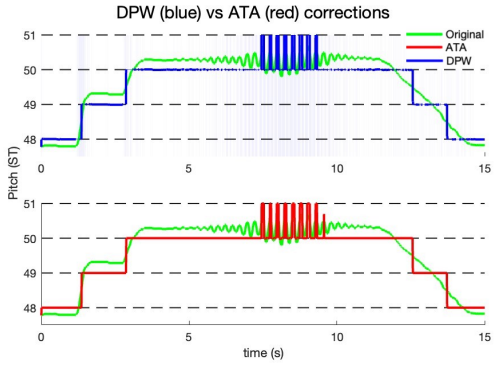


Figure 5: *Extreme correction for an expressive melody*

patent), so we cannot reproduce the exact expected pitch correction of ATA. Therefore, the most effective way to fully understand how the ATA pitch correction algorithm works is to utilize the ATA VST plugin to retune voice samples and extract the corrected f0 from the resulting audio using Praat software [6]. This curves are compared with the DPW correction. To generate the input audio samples, we use Cantor Digitalis (CaD), which is a continuous pitch input synthesizer. CaD takes the trajectory of a wacom stylus, the it generates f_0 and synthesizes a vocal sound. We modified its code to have purposely not-intonated sounds related to the original stylus trajectory. Audio examples can be found in soundcloud [7]. The non-intonated audio samples can be corrected with ATA vist but also with an off-line DPW implementation that we explain later in section 4. Now we proceed to the comparison of both both pitch correction methods.

The simplest case of correction is a shifted note with vibrato. Small vibratos can be effectively corrected with ATA using a retune speed of 50ms. For sustained notes, ATA performs very well and there is no difference with DPW, so we do not present this example here. The difference arises when we have a signal that contains flat notes, free paths, and vibratos. Therefore, it is important to demonstrate how a correction can be performed with ATA using different values of the retune speed parameter, refer to Figure 6. Going up to down we use a retune speed parameter from 0, 15, 50, 100 and 200 ms. The correction is effective at 50ms for the vibrato, but the pitch trajectory after the 12-second mark becomes lost and flattened. Only with a retune speed parameter set to 200ms is it possible to preserve some of the pitch trajectory, but at that configuration, the vibrato is not corrected.

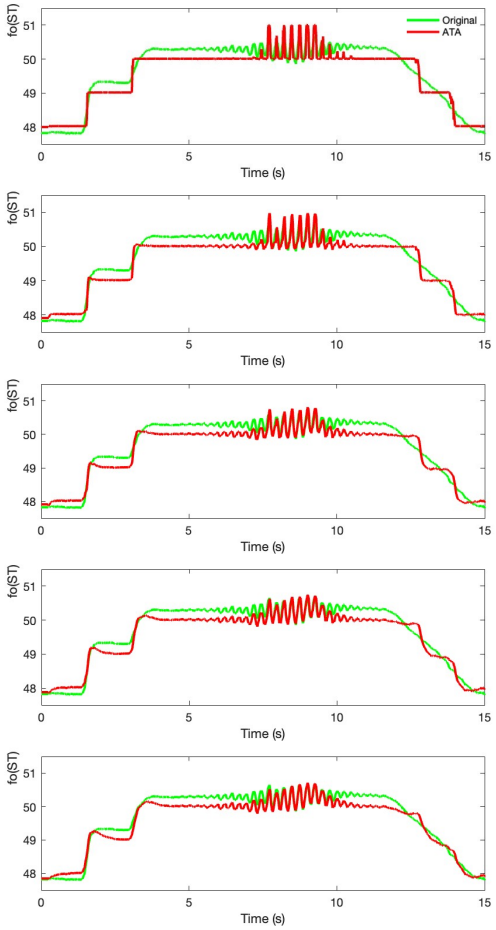


Figure 6: *Correction using different values of retune speed on ATA, RS= 0, 15, 50, 100, 200 ms (up to down)*

Now we will examine the functionality of the ATA flextune parameter. For a more general case, let's now observe what happens when we vary the retune speed while maintaining a specific value for flextune. We have done a configuration with zero retune speed and two values of flextune: zero (red) and 40 cents (violet), figure 7. As we can see, the flextune parameter allows for movement within the range defined by the flextune value after the correction, resulting in the production of smaller ornaments at the output.

In the following example, we will use a non zero value of retune speed, 15ms, and flextune values of zero (red) and 30 cents (violet), as shown in figure 8. As we can see, like the previous example, some ornaments smaller than the flextune value can be preserved at the output.

Now, we present a study with a transition time of 50ms and flextune values of 30 and 60 cents. As we can see in figure 9, a

[6] https://www.fon.hum.uva.nl/praat/

[7] https://on.soundcloud.com/b5NDp last checked: 25 may 2023

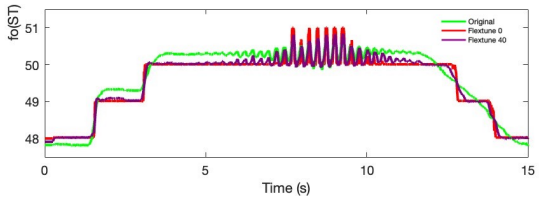


Figure 7: *Correction with ATA, at zero retune speed and flextune: 0 (red) and 40 cents (violet))*

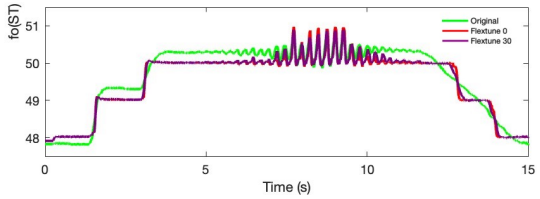


Figure 8: *Correction with ATA, at retune speed equal to 15 ms and flextune: 0 (red) and 30 cents (violet))*

larger value for flextune results in a lack of reactivity. This means the vibrato is not corrected but the path after time equal to 12 s is better preserved than in the other cases. In other words when the notes are well corrected, the general path may be more or less lost depending on the parameters.

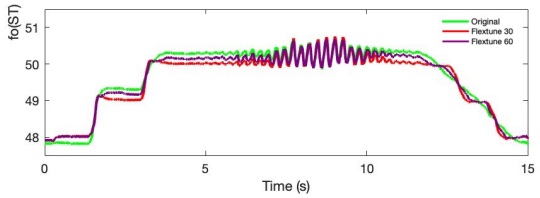


Figure 9: *Correction with ATA, at retune speed equal to 50 ms and flextune: 30 (red) and 60 cents (violet))*

Finally, we show what happens when varying the retune speed for the same flextune parameter. We have chosen a moderate flextune value of 40 cents, while the retune speed varies as follows: 50ms, 100ms, and 200ms. The result can be seen in figure 10. There is always a trade-off between preservation of the main path (free path) and vibrato correction. This means that ATA better preserves the vibrato, but regions such as the one after 12 seconds become staircase-like, resulting in the loss of the original pitch trajectory. In the other hand, parameter values that preserve the shape in that zone, does not correct the vibrato. As we can see in figure 10, the vibrato is not corrected for a retune speed higher than 50ms. On the other hand, when we use flextune at 40 cents and keep zero retune speed (figure 8) the vibrato is corrected but the path after time 12 s is flattened.

3.3. Expressive Correction with DPW

We will show several examples variations of the DPW parameter: critical time and transition time. For the first example, we do choose 100 ms as T_c, then we vary T_t, giving the results in figure

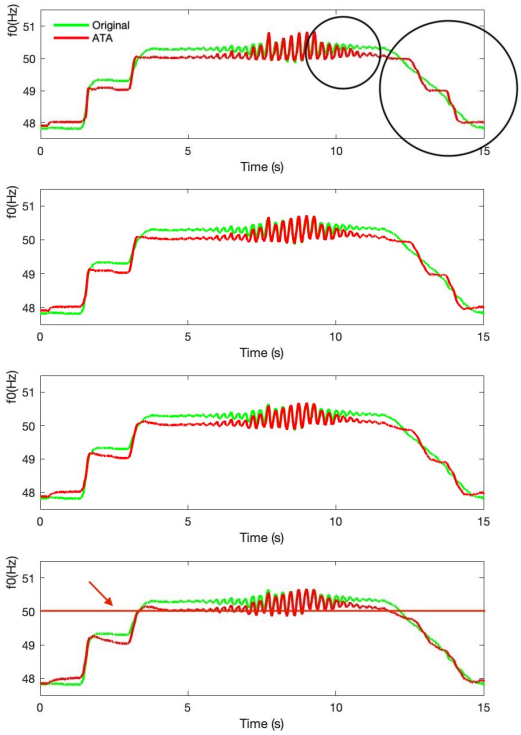


Figure 10: *Correction using different values of retune speed on ATA, RS= 0, 50, 100, 200 ms (up to down) for the same flextune value (40 cents)*

11. As we can see, varying T_t parameter allow us to "smooth" the pitch correction.

Also we have done a correction using a larger critical time equal to 250 ms (optimal according to [11]). It gives the results in figure 12. As we see, the critical time acts as trigger of the correction and the transition time acts as a smoother. The critical time (as parameter) adds an ornament at the beginning of each note step in the staircase region and the transition time modifies the shape of the ornament.

Finally, we have performed a correction using the same transition time (50 ms) while varying the critical time parameter (100 ms, 150 ms, 250 ms). It gives the results in figure 13. As we can see the critical time parameter acts like a trigger for the pitch correction algorithm and the transition time acts as the smoother.

Now we can compare the best configuration for each method. In the case of ATA, it is not possible to achieve good vibrato correction and good preservation of the free path simultaneously. Therefore, we preferred a moderate configuration that performs reasonably well for both purposes. A suitable ATA configuration is a retune speed of 100 ms and flextune of 40 cents (figure 13). For DPW the most suitable correction is done by choosing the critical time as 200 ms (DPW) and then we can choose for example a transition time equal to 50 ms (figure 10) . For simplicity we have put these two cases in the figure 14). This shows that DPW performs a better correction: Firstly the vibrato is well centered in DPW correction while not in ATA; and secondly the DPW preserve better the f_o-path after time 12 s, while ATA flatten it. In contrast, ATA seems visually better in the segment before 5 s while DPW

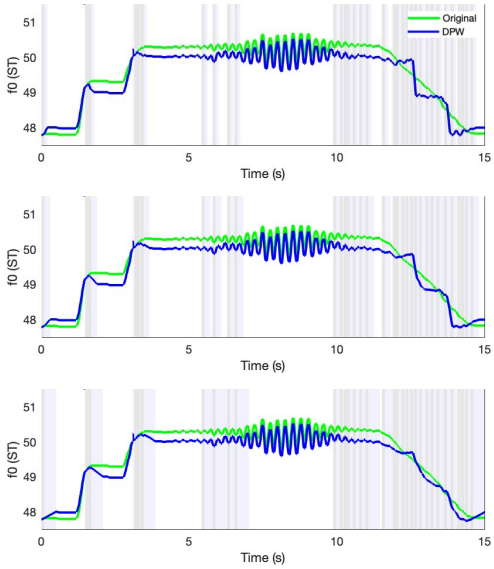


Figure 11: *Correction using different values of transition time in DPW (from up to down: 100,200,400 ms), for the same critical time (100 ms)*

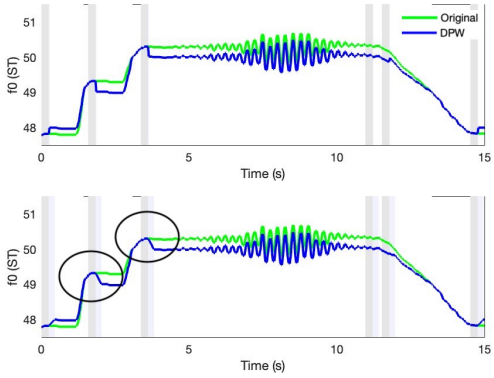


Figure 12: *Correction using different values of T_t in DPW (from up to down: 25,200 ms), for the same T_c (250 ms)*

present an more visible expressive ornament. In the subsequent subsection, we will showcase the measurements that are directly linked to the aforementioned observations, as we will see DPW is closer to the original f_o curve for all the regions.

3.4. Comparison through MSE and MAE

The difference between two curves can be measured in various ways, here we presented two. Firstly, the Mean Squared Error (MSE) that measures the sensitivity to quadratic errors; it is calculated through the difference of squares, which gives larger errors a greater impact on the overall result. MSE also provides a measure of variance between the curves. Secondly, the Mean Absolute Error (MAE) that provides a measure of the average difference in magnitude between the curves, unlike MSE, MAE does not amplify larger errors. We use the follwoing equations:

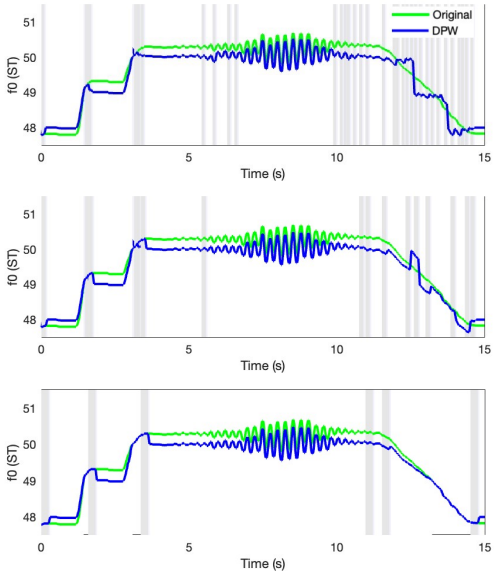


Figure 13: *Correction using different values of T_c in DPW (from up to down: 100,150,250 ms), for the same T_t (50 ms)*

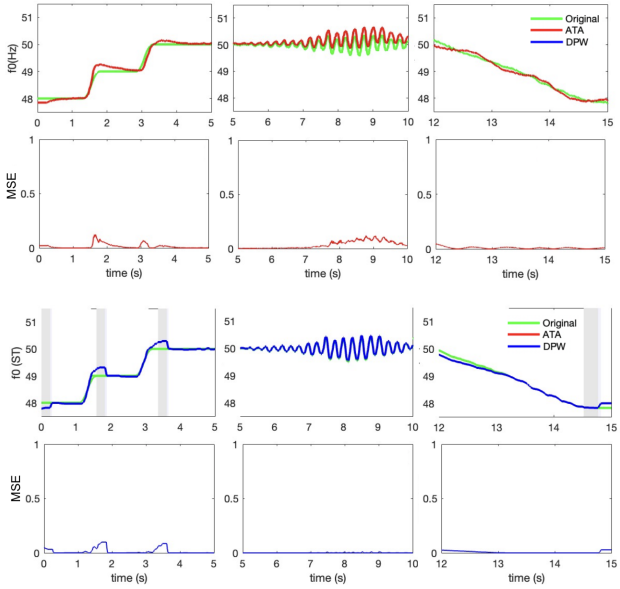


Figure 14: *Correction for the same T_t (50 ms) using flextune at 40 cents for ATA and T_c at 200 ms for DPW and the corresponding MSE.*

$$\text{Mean of MSE} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ij}-\hat{y}_{ij})^2\right) \tag{6}$$

$$\text{Mean of MAE} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}|y_{ij}-\hat{y}_{ij}|\right) \tag{7}$$

Where N represents the number of samples, n_j is equal to 1, cause there always a comparison of one curve with the reference, j represents the curve to compare (ATA or DPW), y_{ij} are the values of the original curve j, and $\hat{y}_{ij}$ are the values of the comparison curve j.

Our example is helpful to highlight three types of pitch modification. The first part in $0 < t < 5$ where signal is like a staircase between the notes 48,49 and 50. The second part in $5 < t <$ 0represents the correction of a poorly intonated frequency modulation, similar to the human vibrato. And the third part is a soft path of a f_o trajectory that should not be corrected, the free path represents the case where the singer do not have the intention to play any specific note. Each part must be compared to the desirable pitch curve, which is different for each region. For example for the staircase part, the desired signal is a staircase. For the vibratory part the ideal pitch would be the same vibration but well centered. And for the third part, the original signal would be the ideal pitch, rather than a correction we want to preserve it. These assumptions are illustrated on figure 6, the calculation of MSE is done point by point. The mean over each region is reported in Table 1. As it is shown and mentioned before, DPW perform better correction of vibratos while preserving the free path of the note, and ATA is better for the staircase part while losing more of the vibrato and free path parts.

Table 1: *MSE and MAE between input and corrected f_0 for the different regions*

	MSE		MAE	
Region	DPW	ATA	DPW	ATA
1	0.0146	0.0146	0.0747	0.0914
2	0.0415	0.0642	0.1304	0.2103
3	0.0539	0.0280	0.2015	0.1463

Please note that all the comparison are focused on the pitch correction curves. For DPW we use the pitch correction method that is different than the full algorithm audio. The implementation of the vocoder, described in section 4, is a complex process and the vocoder we have use in making the audio DPW tracks is not as advanced as the vocoder of ATA. As a result, some imprecision may be present in the generated f_0 paths for the DPW audio examples. Despite these limitations, it is worth highlighting the valuable insights gained from this comparison, which shed light on the respective strengths and weaknesses of each method.

4. IMPLEMENTATION OF AN OFF-LINE AUDIO PITCH CORRECTION

This section talks about the off-line implementation of DPW. DPW works in an analogous way to Cantor Digitalis. However, instead of an incoming f_0 given by a table, we use an f_0 value obtained from a pitch tracker on a pre-recorded vocal audio track. The general structure for a autotune system is conformed by: a pitch tracker, a pitch correction algorithm, and a pitch warping algorithm (vocoder). DPW can follow a similar approach using a pitch tracker to acquire f_0.

4.1. Development of the off-line retuner

We developed a methodology for off-line vocal retuning using the DPW method; this process requires obtaining F0 data and a

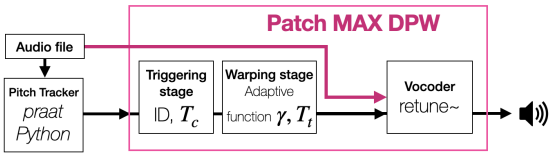


Figure 15: *Configuration of the offline retuner*

transparent vocoder as shown in 15. For pitch tracking, we utilized Praat[8] (software to analyze audio prosody), the *To PitchTier* method allows us to obtain F0 curves for the original audios within a Praat file sampled at Praat time intervals. We did a Python code (with package *wave*) to extract the file's relevant data and to create arrays for time and f_0 information; the arrays were re-sampled at the original audio files sampling rate. The *pathlib* package was employed to process multiple sound library files simultaneously, resulting in a library of the original audios and the f_0 files. The Max/MSP environment was used to process the f_0 information (on Semi tones and Hz) and write retuned audio files using the different vocoders (retune∼, freqshift∼, pitchshift∼, supervp∼, etc). Our goal was to identify the most transparent vocoder that generated a voice signal closest to the input F0, using the original f_0 data, the *retune∼* object was selected as the most transparent modification for the entire library; this ensured that the vocoder avoided introducing sound artifacts that could affect the perception of quality and retuning. However, the overall quality of the presented vocoder, retune∼, is not as precise and good as the ATA vocoder. Therefore, the resulting audio tracks using retune∼ may not be as good as those using the ATA vocoder. Therefore, we dispose of an alternative option, with an wrapper of the *World* [16]vocoder, provided by the research engineers of Lutherie-Acoustique-Musique Group, the audio obtained with *World* is done through a non-real-time transposition through python. The resulting audio has a better quality than the MAX implementation. The sound library for DPW correction using both vocoders can listen on the soundcloud playlist noted in section 3.2.

5. CONCLUSIONS

Through our research, we studied DPW algorithm for audio pitch correction. It is possible to control and trigger a pitch correction thanks to three degrees of freedom that preserves low-amplitude vibratos and ornaments in the neighborhood of the target note. We have also shown how the pitch correction methods are composed of two stages (triggering and warping), and how the modification of the control parameters can lead to equivalent configurations for different systems. We have identified a scenario where ATA and DPW exhibit similarity: extreme correction. Moreover, we have identified three types of correction: staircases, vibratos, and free paths, and have illustrated that DPW performs better for vibratos and free paths, while also being adequate for staircase correction. DPW also exhibits less trade-off between its parameters compared to ATA.

In addition, we have developed an audio application that includes the DPW method. Compared to ATA, its control parameters allow for a smooth pitch trajectory transition towards the nearest

[8] https://www.fon.hum.uva.nl/praat/

notes on a defined scale, minimizing distortion of melodic ornaments between the notes. However, it is important to note that the vocoder used in our application (retune∼) may not provide the same level of quality, precision and accuracy as the ATA vocoder.

We plan to undertake a comprehensive perceptual evaluation of the two systems in a formal setting. This evaluation aims to assess the perceptual salience of the pitch effects introduced by the DPW method, as well as their potential musical relevance.

6. ACKNOWLEDGMENTS

This research was funded through ANR National Research Agency projects: Analysis and Transformation of Singing Style (ANR-19-CE38-0001) and Gepeto: GEsture and PEdagogy of inTOnation (ANR-19-CE28-0018)

7. REFERENCES

[1] C. Vincent, *La voix chanteé*, chapter De l'antipop à l'Autotune, pp. 123–142, N. Henrich & De Boeck Solal, 2013.

[2] P. Boulez and A. Gerzso, "Computers in music," *Scientific Amer.*, vol. 258, no. 4, pp. 44–51, April 1998.

[3] A. Wilson and B. Fazenda, "Perception and Evaluation of Audio Quality in Music Production," in *Proc. of the 16th Int. Conf. on Digit. Audio Effects*, Maynooth, Ireland, September 2013, pp. 68–77.

[4] J.M Chowning, "Digital sound synthesis, acoustics and perception: A rich intersection," in *Proc. of the Int. Conf. on Digit. Audio Effects*, Verona, Italy, December 2000, pp. 1–6.

[5] H. Hildebrand, "Pitch detection and intonation correction apparatus and method," Auburn Audio technologies, Auburn, AL, USA Patent US5973252A, G10H-007/00, Oct. 14, 1992, pp 10-18.

[6] L. Haken, "Position correction for an electronic musical instrument," Champaign, IL, US Patent 76191562009, Int GIOH 1/22, Apr. 19, 2007, pp 6–12.

[7] L. Haken, E. Tellman, and P.Wolfe, "An indiscrete music keyboard," *Comput. Music J.*, vol. 22, no. 1, pp. 30–48, Spring 1992.

[8] L. Feugère, C. d'Alessandro, B. Doval, and O. Perrotin, "Cantor digitalis: chironomic parametric synthesis of singing," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2, pp. 98 1–19, December 2017.

[9] Sebastian Rosenzweig, Simon Schwär, Jonathan Driedger, Meinard MüllerA. Wilson, and B. Fazenda, "Adaptive pitch-shifting with applications to intonation adjustment in a cappella recordings," in *Proc. of the 24th Int. Conf. on Digit. Audio Effects*, Vienne, Austria, September 2021, pp. 121–128.

[10] N. Orio, N. Schnell, and M.M. Wanderley, "Input devices for musical expression: Borrowing tools from hci," in *Proc. of the Int. Conf. on New Interfaces for Musical Expression*, Seattle, USA, April 2001, pp. 62–76.

[11] O. Perrotin and C. D'alessandro, "Target acquisition vs. expressive motion: Dynamic pitch warping for intonation correction," *ACM Transactions on Computer-Human Interaction*, vol. 23, no. 3, pp. 17 1–21, June 2016.

[12] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette, "Vibrato: Detection, estimation, extraction, modification," in *Proc. of the Int. Conf. on Digit. Audio Effects*, Verona, Italy, December 1999, pp. 1–6.

[13] R. Lamb and A.N. Robertson, "Seaboard: a new piano keyboard-related interface combining discrete and continuous control," in *Proc. of the Int. Conf. on Digit. Audio Effects*, Oslo, Norway, June 2011, pp. 503–506.

[14] A.P. McPherson, A. Gierakowski, and A.M. Stark, "The space between the notes: Adding expressive pitch control to the piano keyboard," in *Proc. of the SIGCHI Conf. on Human Factors in Comput. Syst.*, New York, NY, USA, June 2013, p. 2195–2204.

[15] O. Perrotin and C. d'Alessandro, "Quel ajustement de hauteur mélodique pour les instruments de musique numériques?," in *Journées d'Informatique Musicale (JIM 2015)*, Montréal, Canada, May 2015, pp. 186–189.

[16] K. Ozawa M. Morise, F. Yokomori, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, July 2016.

[17] V. Verfaille, U. Zölzer, and D. Arfib, "Adaptive digital audio effects (a-dafx): a new class of sound transformations," *IEEE Transactions on Speech and Audio Processing 14 (5)*, pp. 1817– 1831, 2006.

[18] N. Zacharov, *Sensory Evaluation of Sound*, CRC Press, Boca Raton, FL, USA, first edition, 2019, pp. 60–99, 107-134.

[19] J. Chowning, *Music, Cognition, and Computerized Sound*, MIT Press, Cambridge, Massachusetts, London, England, 1999, pp. 261–276.

[20] S. Bernsee and D. Gökdag, "Methods for extending frequency transforms to resolve features in the spatio-temporal domain," Zynaptiq GmbH , Hannover, Germany, USA Patent 11079418 B2, Aug. 23 , 2018, pp 26–41.

[21] Cycling '74. CA, USA, "Max online documentation," Accessed: 19.09.2022. [Online]. Available: https://docs.cycling74.com/.

[22] Ircam. Paris, France, "Supervp for max max reference pages 11/2012," [Online]. Available: https://forum.ircam.fr/media/uploads/software/SuperVP%20for‰20Max/supervp-for-max.pdf, published Nov-2012, pp 1–14.

Appendix E

Paper: Comparing vocoders for automatic vocal tuning

Daniel Hernan Molina-Villota, Christophe d'Alessandro. Comparing vocoders for automatic vocal tuning. Proc. of 16th International Symposium on Computer Music Multidisciplinary Research, Nov 2023, Tokyo (JP), Japan. pp.756-759, ff10.5281/zenodo.10115215ff. hal-04283705f

A demo of the subjective test was prepared for this presentation using the following website:
http://chorus-digitalis.lam.jussieu.fr/vocoder-comparison-cmmr.html

A poster is also included after the article.

Comparing vocoders for automatic vocal tuning

D. H. Molina Villota[1] and C. D'Alessandro[1] *

Institut Jean Le Rond d'Alembert
Equipe Lutheries-Acoustique-Musique
Sorbonne Université - Centre National de la Recherche Scientifique
Paris, France
daniel.molina_villota@sorbonne-universite.fr

Abstract. We present a compendium of sounds and analyses that support a comprehensive approach to the musical use of the vocoder in automatic vocal tuning correction. Vocoder design has primarily focused on refining the vocoder as a realistic vocal transformer. However, its application within modern music emphasizes its unique sonic identity, adding distinctive coloration to the performer's voice. In this demo, we propose a benchmark that encompasses the vocoder's key elements. The vocoder is considered and analyzed as an audio effect playing an important role in vocal composition, in an approach similar to the study of musical instruments.

Keywords: Vocoder Benchmark Voice Transformation

1 Introduction

The term "vocoder" [1] has two meanings: it can either refer to (i) a software device for transparent voice coding, transmission and natural transformation, or to (ii) a musical device for cross-synthesis and pitch flattening. In this paper, we address the first definition, keeping in mind that this technology may also be used in musical applications, in particular for auto-tuning.

The aim of this work is to establish a parametric benchmark that will facilitate technical discussion of the vocoder, particularly in the case of automatic vocal tuning and audio distortion. In establishing such a benchmark, one should be wary of judging vocoders based on the same criteria as natural voice, whose sound description is extremely challenging [3]. In this demo, we present an audio and graphics repository that supports our benchmark, which can help define the vocoder identity.

2 The Benchmark

Currently, there are no studies that merge musicological and technical approaches to describe the vocoder as a vocal coloring instrument. Acoustically, the vocoder can be

* This Research is funded by National Research Agency: Analysis and Transformation of Singing Style ANR19CE380001 & GEsture and PEdagogy of inTOnation ANR19CE280018

seen as just one more of the many parts that compose the vocal apparatus. The vocoder has its own characteristics and identity which are inherent to its technique. We propose a benchmark that precisely frames the unique characteristics of the vocoder as a vocal coloring instrument. The modern music repertoire evidences two main uses: the distortion due to the technique itself (re-synthesizing with the original F0) and the re-pitching technique (like Autotune).

Methodology: We started with a sample sound which was passed through the Antares autotune software. We framed the two main use cases (presets): one with extreme correction that merely quantizes pitch, and another "transparent" preset that modifies neither pitch nor any other characteristic. The resulting audio files were analyzed with Praat and shaped with Python, generating an f0.wav file as shown in Figure 1. This file, along with the original sound file, was then processed through various vocoders to obtain the sounds with **extreme correction** and the desired **transparent** modification. The samples used come from previous studies at our lab. They can be heard in an online library along with the vocoded tracks(`https://on.soundcloud.com/1d7mx`).

We have used the following vocoders: **Circe** is based on deep learning [4]. The encoder generates a latent code for selected features, and the decoder transforms it back for a given f0 using a bottleneck technique [5]. **Retune** [7] uses frequency and time domain methods such as the Reduced Heisenberg Uncertainty Transform and the Cross-Frequency Phase Coupling . It is used in ZTX, MAX, Digital Performer, and MOTU. **Autotune Antares** (Abbreviated as ATA) [6] serves as an intonation corrector. It is the most commonly used vocoder in contemporary music. **World** [8] is a vocoder based on a custom spectral representation that generates high-quality audio and fast processing . The benchmark descriptors proposal is summarized in table .

2.1 Descriptors of the benchmark

In this section, we summarize some examples of the benchmark. First, we can identify some descriptors independently of the preset used (transparency or extreme retuning). **Latency** is the first appreciable descriptor: retune has the largest latency and ATA the smallest latency. In addition, vocoding involves changes in spectrum, formants and f0-spreading. For those, the transparent preset allows to test the technique alone, avoiding the f0-jumps collateral effect. If the spectrum and signal shape remain unchanged, the vocoder can be considered "**distortion-free**"; ATA and World exhibit this characteristic. Regarding **formants**, World tends to **deepen** them and Circe/retune to **distort** them. Although Circe is known for performing constant transposition well: it generates a **tremolo aligned to vibrato** when using the transparent preset, we also include this effect as descriptor. Concerning harmony, vocoders can present increasing **harmonic differences** (World) or **residual noise** (Retune); we include these changes as descriptors as well. As discussed later, they also appear with the extreme retuning preset.

The extreme retuning preset also involves latency, changes in signal shape, spectrum and formants. ATA and World show good **preservation of the signal shape** despite the pitch jumps. The extreme retuning preset causes discrete pitch steps; the transitory parts generate spectral changes which manifest as vertical lines on the spectrogram. Those are related to local **f0-spreading** (or f0-loss), which deteriorates pitch perception and vocoder realism on a global scale. On the other hand, f0-spreading adds a particular

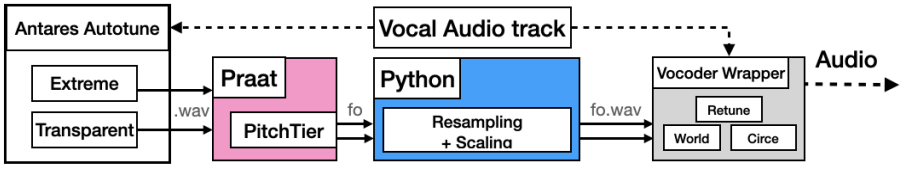


Fig. 1. Flow diagram for the methodology for vocoding with two presets: transparent and extreme retuning (f0 discrete curve).

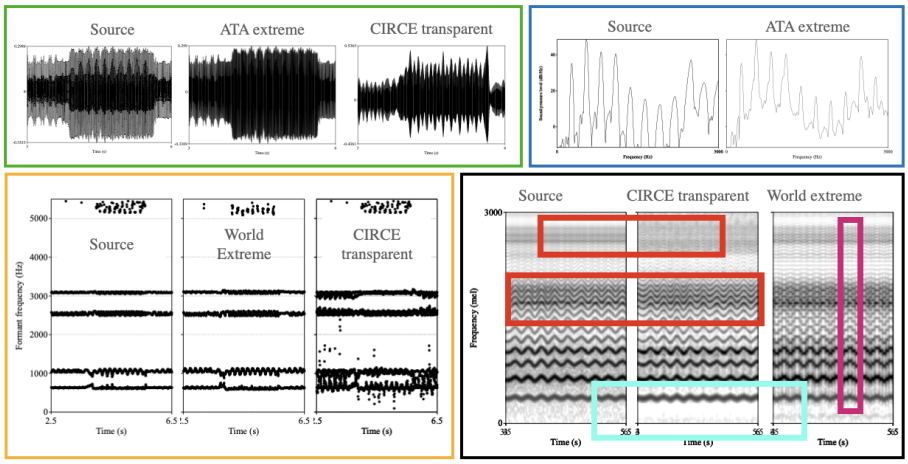


Fig. 2. Green block (Signal Shape): Changes are observed for 2 vocoders. Autotune extreme correction case shows minimal changes while Circe transparent case exhibits significant shape variations. Yellow block (Formants): World shows notable deepening in formant variation and CIRCE exhibits substantial formant alterations. Blue block: (spectral slices): f0-spreading at a given time for original audio and ATA extreme retuning. Black block (spectral changes): In the CIRCE re-synthesis case, upper harmonics appear spread (shown in red), while lower harmonic content seems more prominent in relation to noise (shown in sky blue). In the World retuning case, vertical lines (purple) correspond spectral content spreading at each f0-steps. The audio sample used for all the examples is "real3maleintervals.wav".

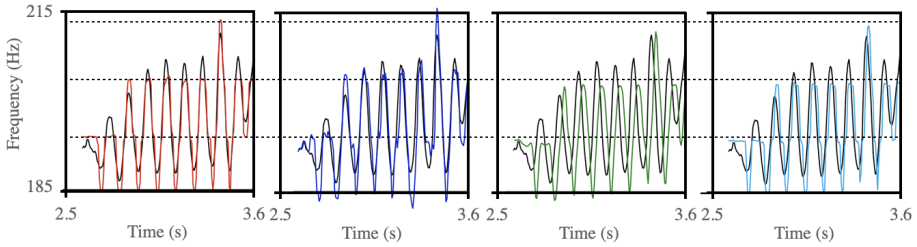


Fig. 3. F0-Path for extreme retuning using (left to right): Autotune, CIRCE, Retune and World. Autotune and World reach exact pitch values more accurately than the others. Retune presents a bigger latency than the other ones.

color due the transient (inherent to the technique) and it contributes to the unique timbre of each vocoder. Each vocoding technique affects harmonics and timbre differently, giving rise to the **harmonic coloration and amplification** descriptors. Circe and Retune are visible examples that alter the harmonic content. Similarly, we observe the **inharmonic coloration** descriptor, which involves residual noise in the low and high-frequency regions of the spectrum. It is notably present in the retune extreme retuning case. Inharmonic coloration affects the presence of noise notably around silences. A summary of the parameters can be seen in Figure 2 and Table 1.

Table 1. Benchmark

Sound Parameter	Latency	Bypass Or resynth Transparency	Formant Deepening	Formant Distorsion	Signal Shape Changing	Tremolo Aligned to Vibrato	F0-Spreading	Upper-Harmonics Modification	Sub-Harmonics Modification	In-harmonic Adding and Residual Noise
Autotune	X	X								
Circe	X		X	X	X	X	X	X	X	X
World		X								
Retune	X		X	X	X			X		X

3 Discusion

Vocoders can introduce changes in timbre properties, like coloration (filter-like action) or discrete pitch variation, while preserving articulation and prosodic content. Our demo provides an audio and visual comparison of the auditory changes introduced by the use of various vocoders. This comparison has been carried out in a systematic way, yielding the benchmark summarized in table 1. Such a benchmark could serve as basis to develop a shared language for technicians and musicians to describe a vocoder's identity.

References

1. Dolson, M.: The phase vocoder:A tutorial. In: Comput. Music J. vol 10 no.4, pp. 14-27 (1986)
2. Lanchantin, P. et al.: Vivos Voco: A survey of recent research on voice transformation at IRCAM. In: Int. Conf. on Digit. Audio Effects, pp.277-285. Paris, France (2011)
3. Castellengo, M.: Perception(s) de la voix chantée. In: La Voix Chantée entre Sciences et Pratiques (N. Henrich),pp. 35-64. De Boeck. Paris, France (2014)
4. Roebel, A. and Bous F.: Neural Vocoding for Singing and Speaking Voices with the Multi-Band Excited WaveNet. In: Information 13(3) 103, pp 1-29 (2022)
5. Bous, F and Roebel.: A. A Bottleneck Auto-Encoder for F0 Transformations on Speech and Singing Voice. In: Information 13(3) 102, pp 1-19 (2022)
6. Hildebrand, H.: Pitch detection and intonation correction apparatus and method. Auburn Audio Technologies, Auburn, AL, USA Patent US5973252A, G10H-007/00, pp 10-18 (1992)
7. Bernsee, S. and Gökdag, D.: Methods for extending freq transforms to resolve feats in the spatio-temporal dom. Zynaptiq GmbH. Hannover(DE). Patent EP3271736B1, pp 1-51 (2016)
8. Morise, M. et al.: WORLD: A Vocoder-Based High-Quality Speech Synthesis Sys. for Real-Time Applications. In: IEICE Transactions on Inf. and Sys., E99.D (7), pp 1877-1884, (2016)

Comparing vocoders for automatic vocal tuning

Daniel Molina Villota, Christophe d'Alessandro

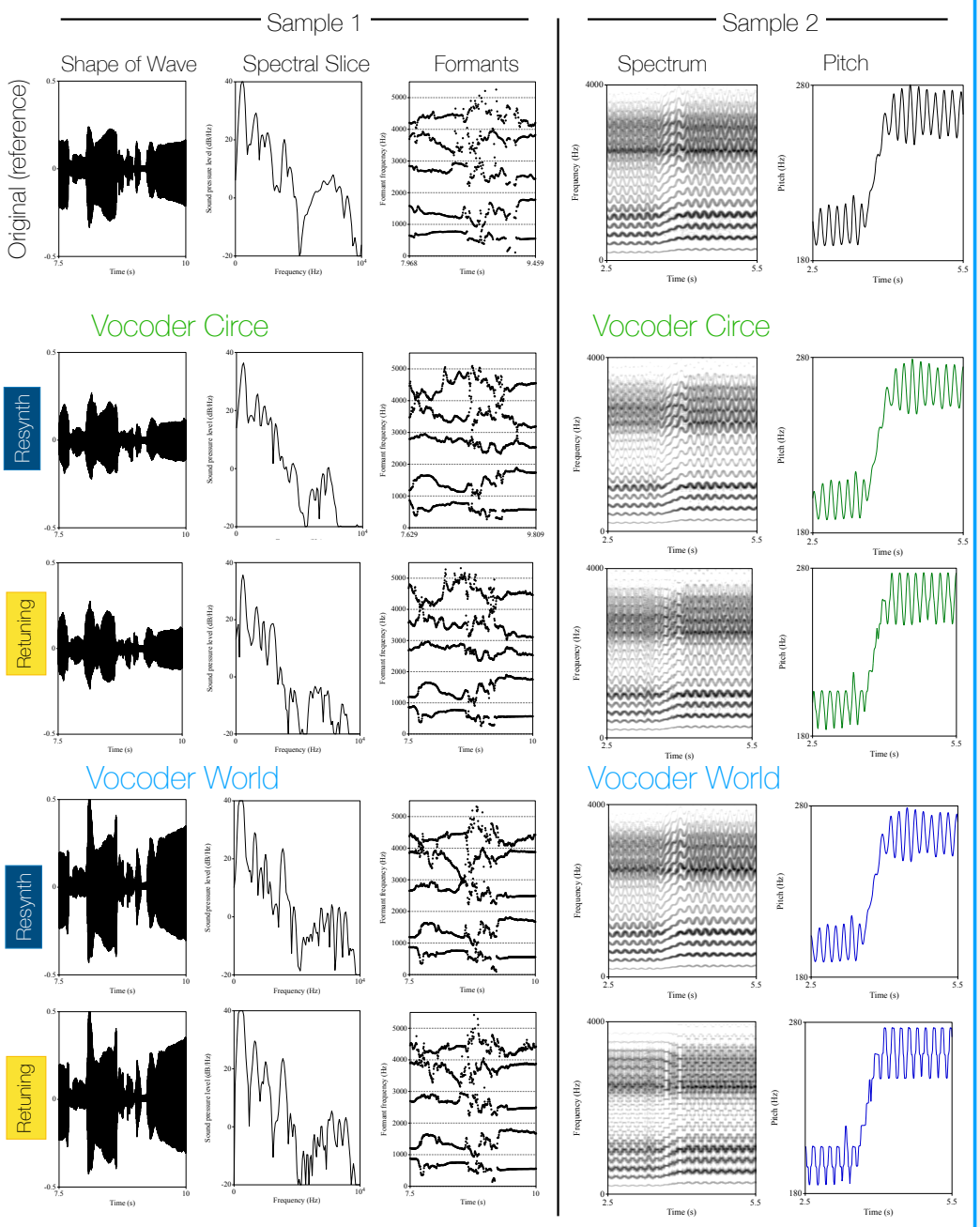


Figure 1. Comparisson for signal shape, spectral slice & formants (sample 1); and for spectrogram and pitch at given time (sample 2), for original and resynthetized audio.

Check the audio test

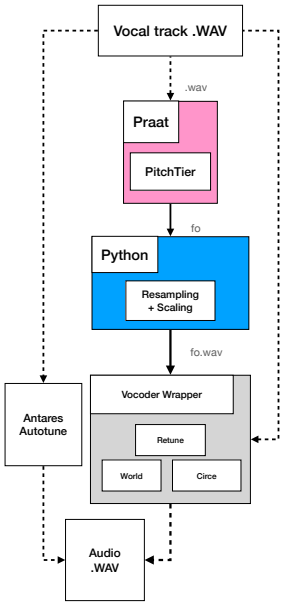


Figure 2. Schema for the resynthesis process

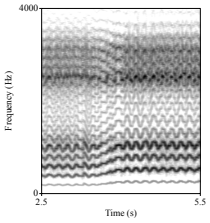


Figure 3. Spectrum for Antares (sample 2)

We present a compendium of sounds and analyses that support a comprehensive approach to the musical use of the vocoder in automatic vocal tuning correction. Vocoder design has primarily focused on refining the vocoder as a realistic vocal transformer. However, its application within modern music emphasizes its unique sonic identity, adding distinctive coloration to the performer's voice. In this demo, we propose a benchmark that encompasses the vocoder's key elements. There are 4 vocoders studied here: Autotune, Circe, Retune and World.

This is a research funded by ANR GEPETO, ANR ARS.

Appendix F

Paper: A Singing Toolkit: Gestural Control of Voice Synthesis

Daniel Hernan Molina-Villota, Christophe d'Alessandro, Grégoire Locqueville, Thomas Lucas. A Singing Toolkit: Gestural Control of Voice Synthesis, Voice Samples and Live Voice. Proc. of 16th International Symposium on Computer Music Multidisciplinary Research, Nov 2023, Tokyo (JP), Japan. pp.704-707, ff10.5281/zenodo.10115215ff. ffhal-04283703f

A demo was prepared with the included devices, a prerecorded performance was shown in a video. The corresponding video file is included in the video library.

A Singing Toolkit: Gestural Control of Voice Synthesis, Voice Samples and Live Voice.

D. H. Molina Villota, C. D'Alessandro, G. Locqueville, and T. Lucas *

Institut Jean Le Rond d'Alembert
Equipe Lutheries-Acoustique-Musique
Sorbonne Université - Centre National de la Recherche Scientifique
Paris, France
daniel.molina_villota@sorbonne-universite.fr

Abstract. The Singing Toolkit demo presents three approaches to real-time gestural control of voice : control of vocal synthesis using the Cantor Digitalis instruments; syllabic re-sequencing and modification of pre-recorded vocal tracks with the Voks instrument; control of real-time vocal performances, using DAFx and inertial devices. These three approaches exemplify the potential of gesture-based control to enhance vocal performances, expand the creative possibilities in vocal music production, and open up new avenues for expressive control and artistic exploration.

Keywords: Gestural Control of Voice, IMU, Theremin, Voks, Chironomic, Gesture, Cantor Digitalis

1 Introduction

The Singing Toolkit demonstrates our recent work in three directions for real-time gesture control and modification of voice signals. The first instrument, Cantor Digitalis, is a formant synthesizer using bimanual (chironomic) gestures for melodic and formantic control with the help of graphic tablet. The second instrument, Voks, allows for syllabic resequencing using tapping gestures and chironomic control of intonation and voice quality. The third approach is real-time voice transformation through gesture-controlled vocal effects using the IMU RiOT-Bitalino inertial measurement units (an Ircam and Bitalino joint project).

* This Research is funded by ANR National Research Agency: Analysis and Transformation of Singing Style ANR-19-CE38-0001 & Gepeto: GEsture and PEdagogy of inTOnation ANR-19-CE28-0018

2 Cantor Digitalis : Chironomic control of synthesized voice

Cantor Digitalis [1] is a vowel and semi-vowel singing instrument controlled by chironomic gestures [2]. It translates manual gestures into formant synthesis parameters based on the linear model speech production [1], allowing musicians to control the pitch, vocal effort, and vowel of a synthetic voice in real time. The primary gesture interface used for controlling Cantor Digitalis is the Wacom graphic tablet. Writing or drawing gestures by the preferred hand are controlling pitch and vocal effort, while the other hand control the vowel space using a 2D (2 formants) surface, as shown in Figure 1a.

The pen's low latency (5 ms) makes sound produced by Cantor Digitalis seem to exhibit a direct causality similar to that of acoustic instruments. A visual cue is also printed on the tablet to enhance usability. The graphic tablet has proven effective for controlling voice intonation and singing with Cantor Digitalis. Cantor Digitalis can also be controlled with other continuous interfaces, e.g. the Roli Seaboard RISE Multi-dimensional Polyphonic Expression interface (MPE) [6]. In this case, pitch is controlled using a chromatic keyboard, and vocal effort is controlled by pressure on the touch surface. MPE allows for continuous transitions between notes and pressure levels. Cantor Digitalis [7] [3] won the first prize in the Margaret Guthman Musical Instrument Competition (2015). Cantor Digitalis is limited to vowels or vocalic sounds, to the exclusion of most consonants.

3 Voks: Syllabic sequencing of a prerecorded voice

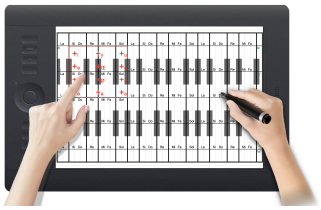
The Voks singing instrument [4] makes it possible to control any voice utterance, including consonants. As it appeared impossible to control each individual articulatory parameter in real time, the syllable is chosen as rhythmic control unit. In practice, the user first loads a sample recording of the desired text being uttered, together with a syllabic annotation of said recording. The loaded sample needs not have any particular rhythm or melody. Then, during the performance, the system resequences the loaded sample, with a rhythm, pitch and vocal quality controlled in real time by the performer's manual gestures.

Syllabic sequencing: Syllabic rhythm control is performed using a cyclic tapping gesture. Several interfaces can capture such gesture data, including buttons, keys, pads, and pressure sensors. Upon tapping/pressing or releasing one's finger on the interface, a one-time signal is sent to the system, triggering advancement of a virtual playhead to the next frame timestamp.

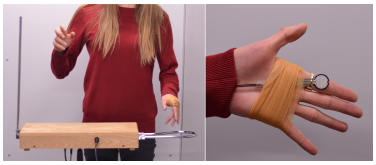
Other gestures: In addition to rhythm sequencing, other parameters are to be controlled by the performer: pitch, vocal effort, vocal tract stretching factor. Some of those parameters are common to Cantor Digitalis, although they are not implemented in the same way — in Cantor, synthesis parameters are controlled directly, whereas in Voks, a prerecorded sample is modified in real time based on control values.

Following Cantor Digitalis, the graphic tablet and MPE interfaces are used to control pitch and vocal effort in Voks. In addition, the theremin has been used as a control

[1] https://github.com/CantorDigitalis



(a) A Wacom graphic tablet.



(b) Theremin and pressure sensor

Fig. 1: Two interfaces that can be used for gestural control of vocal synthesis. (a) The Wacom tablet has been used with Cantor Digitalis (pen and finger) and Voks (pen only) (b) The theremin and pressure sensor have been used to control Voks.

interface, with one antenna controlling pitch and the other controlling vocal effort, and an added pressure sensor placed in between the thumb and index of the performer for rhythm control. T-Voks (i.e. Voks played by a Theremin and a rhythm control button) won second place in the 2022 Guthman musical instrument competition.

4 Gesture Control of Digital Audio Effects with IMU

The third tool in the Singing Workshop is interactive real-time gestural control of digital audio effects (DAFx) for voice. The the BITalino R-IoT (abbreviated as R-IoT)[5] is chosen because of its lightness and powerfullness. It is a 9-axis digital IMU sensor (LSM9DS1) that provides absolute orientation in space with low latency over the OSC protocol. The data flow follows the structure indicated in Figure 2. First, R-IoT data is carried to the computer by a router through wifi. Then, data from R-IoT (orientation, quaternions, and acceleration) is received in MAX using the dedicated Bitalino object and Mubu package (by IRCAM). For each DAFx, a selection of parameters, mapping, limit conditions, and appropriate scaling must be made. The data is then sent from Max to the TouchOSC object in Ableton Live using the OSC protocol. There, another mapping is performed to assign those OSC values to different controls in the effects used.

Now we will describe briefly some effects that have been implemented. We have mapped hand rotation to panning: visually, the performer can make an opening gesture, which allows capturing an appropriate range of orientation values for the axis of rotation. Body limitations help define the scaling limits in MAX so that the movement adequately covers the maximum, minimum, and center of stereo panning. Figure 3 a) illustrates this gesture simply. The second effect is an overdrive effect. Within the specific musical piece for which it has been developed, this effect involves distortion applied to all vocal tracks, which gradually increases towards the end of the song. The backward movement of the hand, as shown in Figure 3 b), relates to the incremental distortion by tilting the arm. Finally, another performer triggers a delay effect momentarily using the same gesture. In this case, the sudden movement launches the delay effect based on the speed of the motion, making the control of the delay much more efficient than with a traditional knob. This movement can be seen in Figure 3 c).

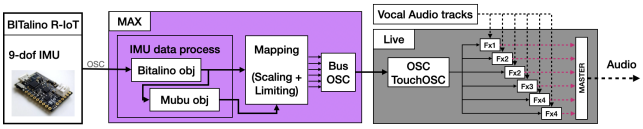


Fig. 2: Flow diagram for Interactive Vocal DAFx with R-devices using MAX and Ableton LIVE.

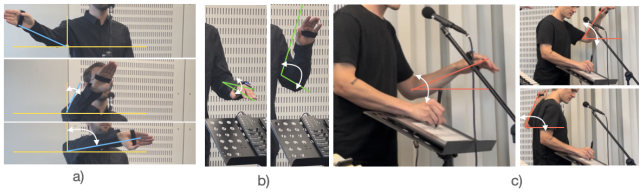


Fig. 3: Schema for the configuration of a) Panning, b)distorsion, c) delay using the R-IoT devices.

5 The Demo

The Singing Workshop the demo consists of a room with the three devices set up, each with its corresponding interfaces and computers. Additionally, there will a poster and three assessors who will explain how the three devices work using musical pieces as examples, within there are also included some tracks of the Chorus Digitalis project, including Cantor Digitalis, Voks and real voices.

References

1. L. Feugère and C. d'Alessandro. Contrôle gestuel de la synthèse vocale. les instruments cantor digitalis et digitartic. *Traitement du Signal*, 32:417–442, 12 2015.
2. L. Feugère, C. d'Alessandro, B. Doval, and O. Perrotin. Cantor digitalis: chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017.
3. L. Feugère, S. Le Beux, and C. d'Alessandro. Chorus digitalis: polyphonic gestural singing. 03 2011.
4. G. Locqueville, C. d'Alessandro, S. Delalez, B. Doval, and X. Xiao. Voks: Digital instruments for chironomic control of voice samples. *Speech Communication*, 125:97–113, 2020.
5. Plux Wireless Biosignals and IRCAM. *User Manual R-IoT Bitalino*. `https://www.bitalino.com/storage/uploads/media/manual-riot-v12.pdf`.
6. Roli. Zimphony: The seaboard rise with cantor digitalis, 2016. `https://youtu.be/mC4pmokMwRo`.
7. S. S.Le Beux, L. Feugère, and C. d'Alessandro. Chorus digitalis: Experiments in chironomic choir singing. pages 2005–2008, 08 2011.

Appendix G

Paper: Correction dynamique et adaptative de la justesse en voix chantée

Daniel Hernan Molina Villota, Christophe d'Alessandro, Olivier Perrotin. Correction dynamique et adaptative de la justesse en voix chantée. CFA 2022 - 16ème Congrès Français d'Acoustique, Société Française d'Acoustique; Laboratoire de Mécanique et d'Acoustique, Apr 2022, Marseille, France. ffhal03848052f

This study presents the application of Pitch Warping and Tuning (PWT) method to correct pitch from audio signals on real time. PWT use the dynamic pitch warping (DPW) method over f_o of an incoming audio signal. Through a pitch velocity condition, the correction is released and adapted to the constantly changing pitch values, in that way the curve of pitch is continuously adapted to the entry. At the end f_o a soft curve on real time is generated and the audio from incoming signal is forced to follow this curve. Several tests have been done using sinusoidal signals and singing voice audio on real time, these results show a significant difference with the traditional pitch correction systems. Our proposal shows a soft variation of pitch that is released over a pitch's velocity condition with an adaptive continuous modification

1 Introduction

Nous menons un projet d'étude systématique des effets audionumériques appliqués au style vocal. Dans ce cadre, il nous a semblé important de travailler sur les effets d'intonation. Le propos est de rendre ces effets dynamiques, sous le contrôle du chanteur, pour qu'il puisse changer au cours de la performance les paramètres des effets. De façon analogue au guitariste électrique, qui utilise depuis longtemps des effets audionumériques dynamiques, le chanteur pourrait ainsi changer le son de sa voix ou son style mélodique et s'emparer ainsi de nouveaux moyens d'expression.

La correction de justesse occupe une place de choix parmi les effets audio-numériques appliqués à la voix. Le succès considérable d'AutoTune de la société Antares (abrégée ATA dans cet article) ne se dément pas depuis plus de 25 ans. De façon inattendue, les artefacts sonores générés par des changements brutaux d'intonation ont très vite attiré l'attention des musiciens de façon positive. En effet, la qualité électronique particulière obtenue a été appréciée dans les musiques populaires, au point de devenir une marque de style musical, et d'être utilisée de façon quasiment systématique. L'effet change en effet la dynamique et l'identité de l'instrument vocal, permettant des transitions de notes beaucoup plus rapides que celles de la voix naturelle. Ces transitions vocales rapides donne un peu à la voix le caractère d'un instrument trous de jeu comme la flûte ou le saxophone, voire d'un instrument à clavier. Comme la rapidité des transitions viole les lois de la physique de la production vocale, le son prend une coloration électronique. ATA modifie l'intonation à l'aide d'un vocodeur, sans prendre en compte les variations naturelles du conduit vocal. Ces transitions rapides s'accompagnent alors de transitoires spectraux, qui sont devenus une signature acoustique de la méthode. Ainsi, ATA conçue au départ comme une méthode transparente du point de vue sonore, a acquis par ses artefacts le statut d'effet audionumérique bien identifié.

Dans cet article une nouvelle méthode pour la correction d'intonation vocale, Dynamic Pitch Warping, ou déformation mélodique dynamique (abrégée DPW dans cet article), qui peut fonctionner en temps-réel, est proposée. Cette méthode est l'adaptation à la voix d'une méthode de correction de trajectoire chironomique pour la synthèse vocale performative [3]. Après la présentation de la méthode DPW, les questions d'analyse de la fréquence de voisement et de modification par un vocodeur de l'intonation sont étudiées. Une évaluation comparative avec Autotune est ensuite proposée. Les perspectives ouvertes pour le développement d'effets audionumériques vocaux dynamiques et contrôlés par le chanteurs sont évoquées.

2 Méthodes de correction mélodique

2.1 Correction mélodique ATA

ATA est une méthode audionumérique de correction mélodique, destinée au début, aux traitements de studio. Cette méthode a été développée par H. Hildebrand et brevetée [10] par Antares Audio Technologies en 1997 sous le nom d'Autotune.

ATA est aujourd'hui utilisé de façon massive dans l'industrie musicale. Lors de son apparition le principal défaut d'ATA a été la transition prononcée entre les notes. Mais ce défaut a été aussi la raison de son succès. L'utilisation comme correction transparente, en temps différé (studio), est presque systématique dans les productions commerciales [11]. L'exploitation des distorsions d'ATA a trouvé un "niche" dans des genres comme le rap, l'hyperpop et la musique électronique. ATA possède au départ une seul paramètre de contrôle : *retune speed* qui défini le temps de transition jusqu'à la note correcte visée. Le logiciel actuellement commercialisé possède un second paramètre de contrôle, *flex tune*, qui est l'intervalle en Cents MIDI dont la fréquence fondamentale peut s'écarter sans déclencher la correction. En plus de ces deux paramètres principaux, le logiciel offre plusieurs paramètres pour changer la qualité sonore. On peut par exemple choisir différentes échelles musicales ou changer la tonalité (*mineur, majeur*), conserver les variations subtiles de hauteur dans les notes longues et stables avec *humanize*, désactiver la modification des formants (*formants*) et modifier le modèle du conduit vocal, modifier la fréquence de référence pour l' échelle musical (*transpose*,*detune*), et ajouter du vibrato (réglable) aux notes tenues (*vibrato*). Pour l'analyse de fréquence fondamentale, plusieurs type de voix sont proposées, ce qui facilite les calculs en spécifiant l'ambitus moyen et en évitant ainsi les erreurs d'octave.

ATA n'est pas le seul correcteur de hauteur. D'autres dispositifs pour la modification de hauteur, avec des fonctionnalités similaires sont par exemple des dispositifs numériques comme metaTune, WavesTune et Ztx, et des dispositifs électroniques comme Model BOSS Vocal Performer, Boss VE-20, TC Helicon, Tascam TA-1VP et Roland VT-4. Mais les méthodes de correction de hauteur ne s'appliquent pas qu'aux effets vocaux. On trouve des

méthodes de corrections également dans les instruments musicaux numériques. Dans ce cas, la correction améliore la précision de la capture du geste et la facilité de jeu musical pour les utilisateurs. Parmi ces instruments musicaux on peut citer [15] Continuum Fingerboard2 [16], Seaboard3 [17], Garageband, TouchKeys et Cantor Digitalis. Le Cantor Digitalis [2] comprend une méthode de correction de hauteur du tracé obtenue par interface tactile. Cette méthode [3] est basée sur une fonction de déformation dynamique dénommé *Dynamic Pitch Warping*.

Dans cet article la méthode DPW est mise en œuvre non pour corriger une trajectoire graphique, mais la courbe de hauteur d'un signal vocal d'entrée.

2.2 Méthode de correction DWP

La fonction proposé par DPW est nommé *fonction élastique* [4]. Cette méthode possède trois degrés de liberté, l'intervalle de détection, le temps critique, et le temps de transition. L'intervalle de détection et temps critiques se combinent pour calculer la condition indirecte de vitesse mélodique, qui détermine l'activation de la correction. La vitesse mélodique permet de définir un seuil de temps critique de détection dans un intervalle de détection. Cela permet d'inférer si la note est assez stable,intentionnelle, et donc mérite d'être corrigée. Si la variation mélodique est trop rapide, la note n'est pas stable et il ne faut pas déclencher de correction. L'avantage des trois degrés de liberté est de permettre la préservation des variations mélodiques expressives, comme le vibrato, en ne corrigeant pas systématiquement la hauteur.

Dans la version initiale de DPW, les problèmes pratiques de détection et de modification de la fréquence fondamentale ne se posent pas : l'entrée donne exactement la fréquence fondamentale mesurée sur le capteur, et ensuite le synthétiseur vocal la calcule directement.

3 Système de correction mélodique

3.1 Architecture

Le système de correction mélodique utilisant la fonction DPW travaille de la façon suivante. L'interprète chante; on capture le signal sonore et on calcule la fréquence fondamentale. Si la hauteur reste stable dedans un intervalle dans le voisinage d'une note (intervalle de détection, I) pendant un seuil temporel (temps critique, t_c), le système déclenche la correction. Cette condition peut être traduite comme une condition de vitesse, cela signifie que les changements rapides de notes tels que les trajectoires naturelles de la voix ne vont pas être corrigés et que seulement les notes stables vont être corrigées. La correction est faite de manière douce pendant une période de transition et de façon adaptative, ainsi les notes suivantes seront aussi accordées grâce à la fonction élastique.

La figure 1 montre la fonction de notes étendues utilisée dans certaines méthodes de correction et la fonction élastique développée par Perrotin et D'Alessandro. Le zéro représente

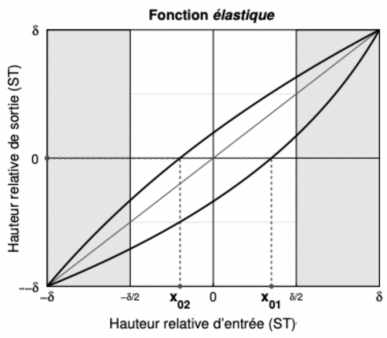


FIGURE 1 – Méthode pour la correction de hauteur [3]

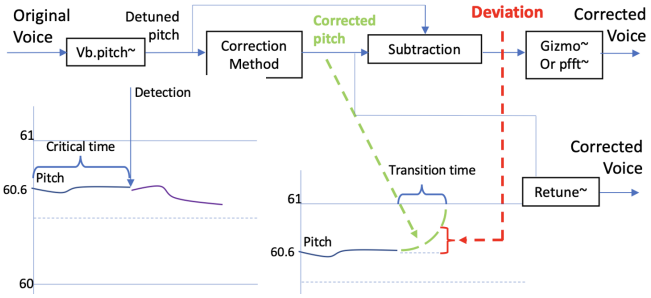


FIGURE 2 – Schéma du DPW

le point de l'échelle musicale de référence. $\delta = 1$ représente une différence d'un degré (un demi-ton le plus souvent) avec la note suivante et précédente. On utilise les conditions limites $y(\pm\delta) = \pm\delta$. et le fonction élastique suit :

$$g(y) = Ae^{\gamma(y+B)+C} \tag{1}$$

Pour calculer la solution, C est remplacée dans $g(y)$ et on efface ainsi la dépendance en B. Avec $y(\pm\delta) = \pm\delta$ on obtienne :

$$y_E(x) = \begin{cases} \frac{1}{\gamma}\left[log\left[(e^{2\gamma}-1)(\frac{x}{\delta}+1)\frac{1}{2} \right.\right. \\ \qquad\qquad \left.\left.+1\right]\right] - \delta, \text{if } \gamma \neq 0 \\ x, \text{if } \gamma = 0 \end{cases} \tag{2}$$

x est la hauteur d'entrée et y_E est la hauteur de sortie. Il ne doit pas y avoir correction pour les valeurs entières, donc $y_E(x_o, \gamma_o) = 0$, et on peut obtenir la courbure :

$$\gamma = \frac{1}{\delta} log\left(\frac{\delta - x_o}{\delta + x_o}\right) \tag{3}$$

Pour fzero donné, la valeur de γ est déduite a partir de. La sortie est accordée seulement si la hauteur d'entrée est stable. Quand l'entrée prend les valeurs $x_{integer} \pm 1$, on a $y(\pm\delta) = \pm\delta$. Quand la correction est désactivée, γ revient graduellement à 0, et peut s'activer encore dès que la condition de vitesse est remplie de nouveau. Un schéma du fonctionnement est donné figure 2.

3.2 Adaptation a la voix

3.2.1 Analyse de la fréquence fondamentale d'entrée

Le premier pas pour adapter la méthode à la voix, est d'obtenir la courbe fzero d'entrée. Pour cela les méthodes suivantes ont été testées : yin [14], vb.pitch (V. Bohm), sigmund and fzero [8]. Les meilleurs résultats sont obtenus avec l'objet fzero, avec la configuration suivante : @onsetamp 0.0001 @onsetpitch 0.001, les seuils de détection d'amplitude et de hauteur. La sortie (fzero) est ensuite dirigée vers le patch de correction DPW qu'on a crée. Les paramètres contrôlables dan ce patch sont les suivants :

- Fzero,
- Temps de detection t_c,
- Temps de transition
- Liste de notes activées/désactivées

Une fois obtenue la courbe de fzero, elles est passée dans le filtre des notes activées. Dans le patch assigné a ce filtre, on localise fzero dans une octave. Dans l'octave on connaît les notes activées ou désactivées. Si fzero est dans un intervalle activé, il passe a la sortie du filtre. Sinon on calcule la note activée la plus proche et on obtient le valeur extrême de l'intervalle en restant dans le voisinage d'une note activée. Quand la fréquence évolue entre régions activées en passant par des régions désactivées, on utilise les valeurs extrêmes des régions désactivées, cela permet d'utiliser uniquement les régions activées et de contrôler la vitesse de transition entre les zones interdites. Ce temps de transition entre notes désactivées peut être considéré comme un 4ème degré de liberté. L'utilisation d'un objet $gate$ permet de bloquer le passage de fzero quand on est sur les zones interdites. Dans les régions activées fzero passe directement à la sortie.

Le filtre de notes activées et désactivées est important car il permet de contrôler l'échelle mélodique de l'interprète. La différence avec ATA provient des degrés de liberté. ATA propose seulement le contrôle du temps de transition, identique pour les intervalles activés et désactivés, alors qu'avec DPW on a le temps de détection, le temps de transition, les intervalles de détection et le temps de transition entre notes interdites.

3.2.2 Correction de l'intonation par DPW

La condition pour déclencher la correction est basée intrinsèquement sur l'idée de stabilité. Seulement les notes d'une durée supérieure à t_c et comprises dans un intervalle I déclencheront la correction de l'algorithme. Toutes les autres entrées sont interprétées comme des trajectoires de hauteur naturelle et ne seront pas corrigées. Comment noté précédemment, la fonction élastique fait que la sortie et l'entrée convergent dans les extrêmes vers les valeurs entières. Si aucune correction n'est nécessaire (la condition de déclenchement n'est pas remplie), alors l'entrée passe directement à la sortie.

Pour la mise en ouvre de la méthode DPW, nous avons besoin des paramètres des 3 degrés de liberté plus le temps de transition pour les notes désactivées. Avec l'intervalle de détection, nous divisons l'échelle musicale en micro-tons de 0,1 demi-tons (ST). Cela signifie que l'intervalle $(\delta, -\delta)$ est découpé en tranches de 0.1 ST. Si la hauteur d'entrée reste dans l'une des sections de micro-tons pendant le temps de détection, la correction est déclenchée pendant le temps de transition. Le résultat est une courbe de hauteur déformée où les notes suffisamment stables sont corrigées. L'autre dégré de liberté est le temps transition (t_t) est comparable a *retune speed*. On peut considérer aussi un quatrième degré de liberté qui es le temps de transition entre notes désactivées.

3.2.3 Resynthèse de l'intonation modifiée

Une fois la courbe mélodique corrigée calculée, il faut utiliser un vocodeur pour la resynthèse du signal correspondant. Le premier vocodeur a été breveté par Homer Dudley aux Bell Labs en 1935 pour reconstruire une voix à l'aide d'un algorithme d'analyse-synthèse, il avait été créé pour réduire la bande passante d'un signal vocal $v(t)$ pour les télécommunications, mais finalement appliqué et la musique et au bandes son de films [11]. Aujourd'hui de nombreux types de vocodeurs existent : channel vocoder qui utilise la somme pondéré des filtres passe-bande, le LPC qui utilise un filtre IIR, et le cepstrum qui utilise une convolution circulaire d'un filtre FIR. [13]

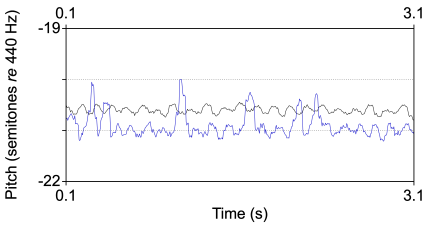


FIGURE 3 – Courbe intonative avec vibrato (noir) et sa version corrigée avec DPW (bleu).

Les vocodeurs testés pour notre système utilisent les objets suivants : pitchshift , freqshift , psych , supervp.trans , fbinshift, transposer, hilbert, gizmo et retune. Tous sauf $retune$ reçoivent en entrée la valeur en midicents à transposer. On ne peut généralement pas accéder à la détection de seuil de cette entrée, et les objets ne sont pas assez réactifs. Avec supervp et gizmo, la réponse est meilleure mais n'est pas assez précise pour notre objectif. Le meilleur résultat a été obtenu avec $retune$ un objet Max MSP basé sur zynaptiq ztx (Precision Time Stretching and Pitch Shifting) [12] une méthode brevetée, comme ATA. L'objet retune est un vocodeur qui sert de base pour les vocodeurs natifs d'Ableton tels que mono vocoder, poly vocoder et autotuna. Cet objet peut recevoir une entrée (signal) pour forcer le pas de sortie souhaité. Nous utilisons cette possibilité pour déformer la hauteur de sortie avec notre correction. Actuellement notre système ne dispose pas

des fonctions supplémentaire d'ATA telles que la sélection du type de voix, vibrato et humanisation.

4 Exemples de corrections et évaluation comparative

L'évaluation de la méthode est conduite en deux temps. Dans une premier temps, il s'agit d'analyser le comportement de la méthode, et ainsi de vérifier si elle répond à nos attentes. Dans un deuxième temps, la nouvelle méthodes est comparée à ATA dans divers cas de correction, afin de mettre en évidence les différences entre les deux méthodes et les avantages possibles de la nouvelle méthode.

Pour la phase d'évaluation initiale, les sons de référence sont généré par un synthétiseur vocal performatif, le Cantor Digitalis [2]. Ce système permet un contrôle chironomique très précis de l'intonation, et donc d'obtenir des sons vocaux de synthèse avec davantage de précision et de justesse que la voix naturelle [1]. L'objectif est de vérifier la réactivité de la méthode et son comportement pour des situation de transitions de notes et d'ornement. Les cas typique du vibrato, du glissando et du portamento sont étudiés.

4.1 Vibrato

Le vibrato est un ornement musical courant, qui est perçu de façon complexe [9]. C'est une modulation de fréquence (mais aussi d'intensité et de timbre) d'une amplitude de l'ordre de quelques dizaines de Cents, et avec des fréquences inférieures de 4 à 6 Hz environ. Le vibrato ajoute de l'expression, de l'emphase, de la variété de timbre, renforce la présence des chanteurs [7]. La figure 3 montre une correction de hauteur qui préserve le vibrato. Pour un exemple similaire, dans le cas d'ATA sans les nouvelles caractéristiques, le vibrato est entièrement supprimé, et la hauteur corrigée saute d'un demi-ton à l'autre, en fonction de la vitesse de correction. Ici, le vibrato est assez rapide, et son apmlitude de l'ordre de 25 Cents en valeur absolue. Avec le DPW, l'intonation moyenne est correctement modifiée, mais le vibrato est bien conservé, ce qui est un des buts de la méthode. Cependant, lorsque l'intonation est trop proche de la note (demi-ton) supérieure, la correction "accroche" cette note. Pour préserver le vibrato, il faut donc soigneusement régler les paramètres de la méthode. L'exploitation musicale de cette source d'artefact reste à explorer.

4.2 Glissando

Les glissando, même s'il sont rarement employés musicalement, sont très utiles pour analyser le comportement des méthode de correction d'intonation. Les glissando sont réalisés avec le Cantor Digitalis, sur une octave, avec une pente plus ou moins rapides. L'objectif du test est de vérifier si le système réagit correctement. la figure 4 montre les résultats pour différentes réglages des paramètres de correction et différentes pentes de glissando. Pour l'image du haut, la pente est de 3 demi-tons/s. La correction est conforme aux attentes, avec la transformation du glissando

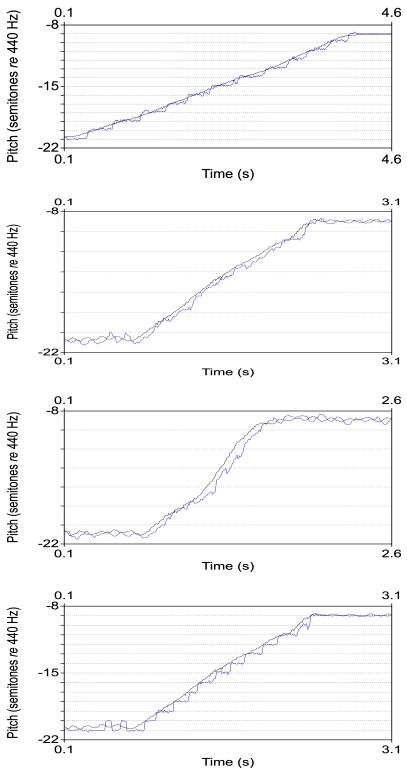


FIGURE 4 – De haut en le bas : Glissando sans vibrato de C3 a C4 pendant 4000ms. Glissando lent avec vibrato de C3 a C4. Glissando rapide avec vibrato de C3 a C4. Glissando lent avec vibrato de C3 a C4 avec une correction instantanée

en escalier mélodique, qui suit l'échelle des demi-tons, tout en préservant par ailleurs les micro-variations mélodiques. Notre correction apporte donc une certaine variabilité au son.

Des variation de vélocité du glissando, en présence de vibrato, sont analysées sur les deux images centrales. Dans ces exemples, la vitesse de déclenchement de la correction en fonction de la vitesse de variation mélodique permet de préserver la pente du glissando, tout en corrigeant le début et la fin des notes. Il est important de mentionner que la correction mélodique n'est pas immédiate ici. Le but de ces deux exemples est de vérifier que le glissando corrigé suit assez bien la forme initiale, mais avec une correction dans les notes extrêmes du son.

L'image du bas de la Figure 4 montre l'effet d'une correction instantanée de hauteur (c'est-à-dire avec un temps de détection et un temps de détection nuls). Le résultat est un escalier mélodique, avec des marches quasiment plates et des contremarches quasiment verticales. Ce résultat est très similaire à celui attendu avec ATA pour un temps de transition nul. La partie plate initiale du son est corrigée vers la note la plus proche, ensuite l'allure du glissando est conservée, mais en escalier, puis, la partie finale du son est corrigée. Dans cet exemple, il y a moins de variation de hauteur à l'intérieur d'une note stable, à cause du temps de détection utilisé qui est de 0.

4.3 Portamento

La figure 5 montre l'analyse dans le cas du portamento, ou transition "portée" entre deux notes.

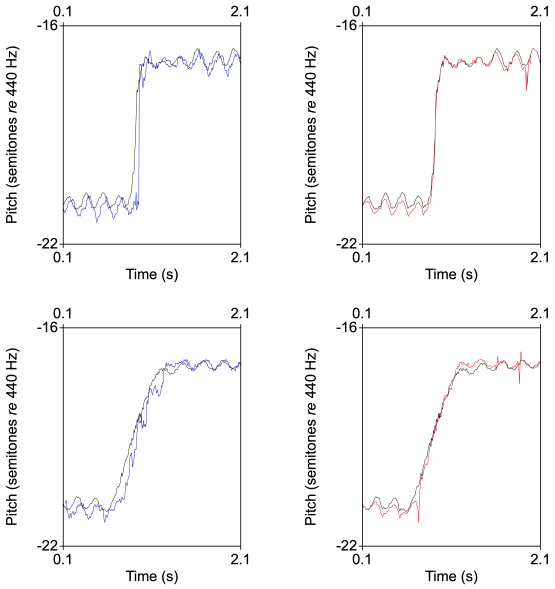


FIGURE 5 – En haut : DPW (15ms de détection et 3Oms de transition) (bleu) et ATA (45ms)(rouge) pour un portamento rapide. En bas : DPW (15ms de détection et 3Oms de transition) (bleu) and ATA (45ms)(rouge) pour un portamento plus lent.

Les résultat sont donnés sur la figure 5. La méthode donne une bonne correction, meilleure pour le portamento plus lent. Dans les deux cas la forme générale de la transition est conservée, et il y a une correction moyenne de hauteur qui préserve le vibrato pour les notes satbles avant et après la transition. Les images du bas montrent le même type de correction avec Autotune.

Nous pouvons observer quelques différences dues à un meilleur contrôle du grain de correction avec notre méthode. Pour une correction avec une contante de 45ms ATA ne corrige presque pas, alors que notre méthode apporte quelques modifications. Elle est plus rapide dans la correction qu'ATA. Notre méthode permet un meilleur contrôle du grain de correction comme on le voit sur la figure.

La figure 6 montre l'effet des variations de paramètres dans notre méthode. Un glissando est corrigé selon trois conditions différentes (temps de détection 7 ms, temps de transition 15ms, temps de détection 15 ms, temps de transition 15ms,temps de détection 30 ms, temps de transition 60ms) en magenta, bleu et marron respectivement, l'échantillon original est en noir. Comme nous pouvons le voir (et l'entendre) la modification devient moins audible, plus douce pour des temps de détection et de transition plus longs.

L'échelle mélodique de correction peut être spécifiée à volonté. Dans ce cadre, nous avons fait un test avec une échelle mélodique chromatique pour laquelle certaines

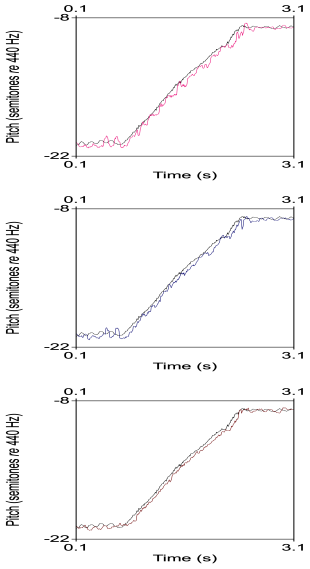


FIGURE 6 – Correction par DPW avec 3 réglages différents.

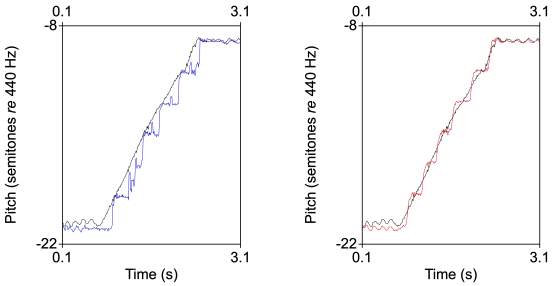


FIGURE 7 – Test avec la moitie des notes désactivées

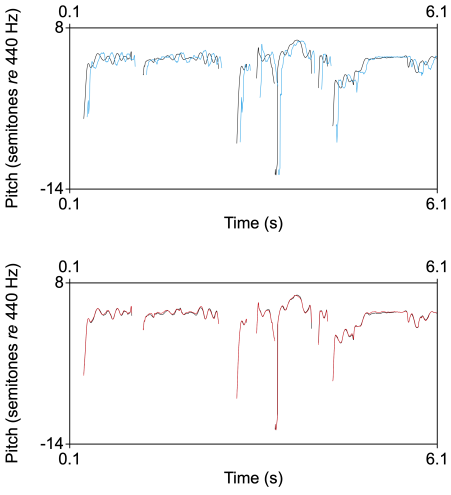


FIGURE 8 – Test avec un son réel corrige avec DPW (bleu) et Antares (rouge)

notes sont interdites ou désactivées (notes désactivées $= 1, 3, 5, 7, 9, 11$; notes activées $= 0, 2, 4, 6, 8, 10$) avec un temps de détection de 7 ms et un temps de transition de

15ms. Le résultat est porté sur la partie gauche de la figure 7. La partie droite montre le résultat d'ATA pour un temps de 21 ms. Cette dernière correction semble de meilleure qualité.

Finalement une chanson est analysée avec notre méthode et ATA dans la figure 8. La correction est plus claire avec notre méthode pour une configuration de 45 ms de durée (15 ms pour le temps de déclenchement et 30 ms pour le temps de transition). Il faut remarquer que notre méthode n'est actuellement pass plus rapide qu'ATA, mais ne pose pas de problèmes de rendu sonore.

5 Discussion

Le premier objectif de cette recherche était l'application à la voix chantée, et en temps réel d'un nouvel algotithme de correction de hauteur, qui avait été développé pour la correction de cibles sur une tablette graphique. Nous avons ainsi montré que l'algorithme DPW, en lui adjoignant une détection de fréquence fondamentale en entrée et le vocodeur ZTX en sortie peut s'appliquer à la voix en temps réel. Il est ainsi possible d'ajouter des degrés de liberté pour le contrôle de l'effet. De plus le logiciel est ouvert, et donne des résultats sonores comparables à ceux d'ATA pour le temps réel. Les paramètres de contrôle permettent de ne corriger que les changement de note sur l'échelle et non les ornements mélodiques entre les notes. Le vocodeur lui même semble introduire des imprécision, et il sera nécessaire de travailler encore cet aspect. Ces imprécisions qui apparaissent visuellement sur les tracés semblent avoir peu d'effet auditif. Le fait d'avoir des contrôles différents introduit des artefacts différents, dont l'exploitation musicale doit être étudiée.

L'utilisation d'un contrôle dynamique de l'effet de correction est un de nos objectifs. Nous avons démontré que le système fonctionne de manière robuste. Notons que le système est ouvert, via MAX MSP. Ainsi la méthode peut s'appliquer dans divers systèmes ou modules d'effet.

L'utilisation de nouveaux paramètres permet de nouvelles possibilités pour la création et la performance musicale. Pour la suite du travail, une évaluation perceptive formelle doit être menée. L'étude de différents vocodeurs et des artefacts qui en résultent doit également être conduite. En effet les variation de timbre résultant des modifications des transitions naturelles sont de la plus haute importance pour la perception de l'effet sonore et pour son utilisation artistique.

Remerciements

Cette recherche a bénéficié d'un financement à travers les projets de l'Agence Nationale de la Recherche ARS : Analysis and Transformation of Singing Style (ANR-19-CE38-0001) et Gepeto : GEsture and PEdagogy of inTOnation (ANR-19-CE28-0018).

Références

[1] C. d'Alessandro, L. Feugere, S. Le Beux, O. Perrotin, A. Rilliard, Drawing melodies : Evaluation of chironomic singing synthesis, JASA 135 (6) (2014) 3601–3612.

[2] L. Feugère, C. d'Alessandro, B. Doval, O. Perrotin, Cantor digitalis : chironomic parametric synthesis of singing, EURASIP Journal on Audio, Speech, and Music Processing 2017 (1) 2.

[3] O. Perrotin, C. D'Alessandro, Target Acquisition vs Expressive Motion : Dynamic Pitch Warping for Intonation Correction, *ACM Transactions on Computer Human Interaction 23*, 1-21 (2016).

[4] O. Perrotin, Chanter avec les mains : interfaces chironomiques pour les instruments de musique numériques *Thèse de doctorat Informatique, U.Paris Saclay, Ch4*, 93-123 (2015).

[5] V. Verfaille, M. Wanderley, P. Depalle, Mapping Strategies for Gesturaland Adaptive Control of Digital Audio Effects, *Journal of New Music Research* **35-1**, 71-93 (2006).

[6] V. Verfaille, C. Guastavino, C. Traube, An interdisciplinary approach to audio effect classification, *Proceedings 9th International Conference on DAFx*, 106–113 (2006)

[7] L. Regnier, PhD Thesis - Localization, Characterization and Recognition of Singing Voices, *Université Pierre et Marie Curie - Paris VI*, Paris (2012).

[8] M. Zbyszynski, D. Zicarelli, R. Collecchia, fzero Fundamental estimation for Max 6 *International Computer Music Conference*, Perth, Australia (2013).

[9] C. D'Alessandro, M. Castellengo, The pitch of short-duration vibrato tones, *Journal of the Acoustical Society of America*, **93(3)** 1617-1630(1994)

[10] H. Hildebrand, Autotune Antares Patent, Pitch detection and intonation correction apparatus and method, *Auburn Audio Technologies, CA*, (1998)

[11] C. Vincent, De l'antipop à l'Autotune, *La voix chantée, N. Henrich, Boeck solal* **Ch 8** 123-137(2014)

[12] S. Bernsee, D. Gökdag, ZTX Patent - Frequency transform extension methods to solve characteristics in the space-time domain, *Zynaptiq Gmbh*, (2015)

[13] U. Zôlzer, DAFX : Digital Audio Effects, *Wiley*, 315 (2002)

[14] A. de Cheveigné, H. YIN, A fundamental frequency estimator for speech and music, *J. Acoust Soc Am*, 1917-1930 (2002)

[15] Vincent Goudard, Hugues Genevois et Lionel Feugre : On the playing of monodic pitch in digital music instruments. In Anastasia Georgaki et Giorgos Kouroupetroglou, diteurs : Proceedings of the International Computer Music Conference (ICMC), pages 1418–1425, Athens, Greece, September 2014. National and Kapodistrian University of Athens.

[16] Lippold Haken : Position correction for an electronic musical instrument, novembre 17 2009. US Patent 7,619,156.

[17] Roland Lamb et Andrew N. Robertson : Seabord : a new piano keyboardrelated interface combining discrete and continuous control. In Proceedings of the International Conference on New Interfaces for Musical Expression (NIME), NIME '11, pages 503–506, Oslo, Norway, May 30 - June 1 2011.

[18] IRCAM, SuperVP-TRaX, http ://forumnet.ircam.fr/373.html ?L=1.

Bibliography

[Abe et al., 2008] Abe, T., Nakamura, Y., Kawahara, H., and Shikano, K. (2008). Analysis-and-manipulation approach to pitch and duration of musical instrument sounds without distorting timbral characteristics. In *Proceedings of the 11th International Conference on Digital Audio Effects*, pages 1–8, Espoo, Finland.

[Apel, 1969] Apel, W. (1969). *Harvard Dictionary of Music*. Belknap Press of Harvard University Press, Cambridge, Massachusetts, 2nd edition.

[Ardaillon, 2017] Ardaillon, L. (2017). *Synthesis and expressive transformation of singing voice*. PhD thesis, Université Pierre et Marie Curie - Paris VI, Signal and Image processing. English. ffNNT : 2017PA066511ff. fftel-01710926v2f.

[Arfib et al., 2002] Arfib, D., Couturier, J., Kessous, L., and Verfaille, V. (2002). Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces. *Organised Sound*, 7(2):127–144.

[Babacan et al., 2013] Babacan, O., Drugman, T., d'Alessandro, N., Henrich Bernardoni, N., and Dutoit, T. (2013). A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *ICASSP 2013 - 38th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, Vancouver, Canada.

[Bernsee and Gökdag, 2016] Bernsee, S. and Gökdag, D. (2016). *Methods for extending frequency transforms to resolve features in the spatio-temporal domain*. Zynaptiq GmbH, Hannover, Germany.

[Bevilacqua et al., 2016] Bevilacqua, F., Caramiaux, B., and Françoise, J. (2016). Perspectives on real-time computation of movement coarticulation. In *3rd International Symposium on Movement and Computing*, pages 1–5, Thessaloniki, Greece.

[Bevilacqua et al., 2013] Bevilacqua, F., Schnell, N., Rasamimanana, N., Bloit, J., Flety, E., et al. (2013). De-mo: Designing action-sound relationships with the mo interfaces. In *CHI Conference on Human Factors in Computing Systems '13 Extended Abstracts on Human Factors in Computing Systems*, pages 2907–2910, Paris, France.

[Bohm, 2022] Bohm, V. (2022). vb.pitch - a pitch tracker. https://vboehm.net/downloads. Last access 19-09-2022.

[Bonada, 2004] Bonada, J. (2004). High quality voice transformations based on modeling radiated voice pulses in frequency domain. In *Proceedings of the 7th International Conference on Digital Audio Effects*, pages 1–4, Naples, Italy.

[Boulez and Gerzso, 1998] Boulez, P. and Gerzso, A. (1998). Computers in music. *Scientific American*, 258(4).

[Bous, 2023] Bous, F. (2023). *A Neural Voice Transformation Framework for Modification of Pitch and Intensity*. Phd thesis, ED130 - École doctorale Informatique, Télécommunications et Électronique (Paris). Sciences et Technologie de la Musique et du Son (UMR 9912 STMS). Institut de Recherche et de Coordination Acoustique Musique (IRCAM) Équipe Analyse/Synthèse des Sons. Sorbonne Université.

[Bous and Roebel, 2022] Bous, F. and Roebel, A. (2022). A bottleneck auto-encoder for f0 transformations on speech and singing voice. *Information*, 13(3-102):1–19.

[Bouty and Sabine, 1901] Bouty, E. and Sabine, W. (1901). Architectural acoustics. part. i. reverberation (acoustiquearchitecturale. 1re partie. réverbération). *J. Phys.Theor. Appl.*, 10(1):38–48.

[Bowler et al., 1990] Bowler, I., Purvis, A., Manning, P., and Bailey, N. (1990). On mapping n articulation onto m synthesiser-control parameters. In *Proceedings of the International Computer Music Conference*, pages 181–184. Cited on page 10.

[Castellengo, 2014] Castellengo, M. (2014). Perception(s) de la voix chantée. In Henrich, N., editor, *La Voix Chantée entre Sciences et Pratiques*, pages 35–64. De Boeck, Paris, France.

[Chowning, 1999] Chowning, J. (1999). Digital sound synthesis, acoustics and perception: A rich intersection. In *Proceedings of the International Conference on Digital Audio Effects (DAFX-00)*, pages 1–6, Verona, Italy.

[Cook, 1998] Cook, P. (1998). Towards the perfect audio morph? - singing voice synthesis and processing. In *Proceedings of the 1st International Conference on Ditial Audio Effects*, pages 1–8, Barcelona, Spain.

[Cook, 1999] Cook, P. (1999). Perceptual fusion and auditory perspective. In Cook, P., editor, *Music, Cognition, and Computerized Sound*. MIT Press, Cambridge, Mass.

[Dattoro, 1997] Dattoro, J. (1997). Effect design, part 2: Delay-line modulation and chorus. *Journal of the Audio Engineering Society*, 45(10):764–788.

[de Cheveigné and Kawahara, 2002] de Cheveigné, A. and Kawahara, J. (2002). Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930.

[De Götzen et al., 2001] De Götzen, A. et al. (2001). Traditional implementations of a phase-vocoder: the tricks of the trade. In *Proceedings of the 3rd International Conference on Digital Audio Effects*, pages 1–6, Verona, Italy.

[Delle Monache et al., 2018] Delle Monache, S., Rocchesso, D., Bevilacqua, F., Lemaitre, G., Baldan, S., et al. (2018). Embodied sound design. *International Journal of Human-Computer Studies*, 118:47–59.

[Disch and Edler, 2010] Disch, S. and Edler, B. (2010). An amplitude- and frequency-modulation vocoder for audio signal processing. In *Proceedings of the 13th International Conference on Digital Audio Effects*, pages 1–8, Graz, Austria.

[Dolson, 1986] Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27.

[Donahue et al., 2019] Donahue, C., McAuley, J., and Puckette, M. (2019). Adversarial audio synthesis. In *International Conference on Learning Representations*, pages 1–16, New Orleans, Louisiana.

[Doval et al., 2003] Doval, B., D'Alessandro, C., and Henrich, N. (2003). The voice source as a causal/anticausal linear filter. VOQUAL'03, Geneva, Switzerland, August 27-29, ISCA International Speech Communication Association Archive. [Online]. Available: https://www.lam.jussieu.fr/cantordigitalis/media/Dov03.pdf.

[Dudley, 1937] Dudley, H. W. (1937). *System for the Artificial Production of Vocal or Other Sounds*. Bell Telephone Laboratories Incorporated, Garden City, N.Y. US Patent 2121142S erial No. 35,466.

[Dudley, 1939] Dudley, H. W. (1939). *The vocoder*. Bell Labs Record. Reprinted in R. W. Schafer and J. D. Markel, eds., Speech Analysis, New York: IEEE Press, 1979, pp. 347-351.

[Dutilleux, 1998] Dutilleux, P. (1998). Filters, delays, modulations and demodulations: A tutorial. In *Proceedings of the International Conference on Digital Audio Effects (DAFx-98)*, Barcelona.

[D'Alessandro, 2003] D'Alessandro, C. (2003). Voice material for the workshop voice quality: Functions, analysis and synthesis. VOQUAL'03, Geneva, Switzerland, August 27-29, 2003, ISCA Archive. [Online]. Available: http://www.isca-speech.org/archive/voqual03/voq3_audio.html.

[D'Alessandro and Castellengo, 1994] D'Alessandro, C. and Castellengo, M. (1994). The pitch of short-duration vibrato tones. *Journal of the Acoustical Society of America*, 95(3):1617–1630.

[Engel et al., 2019] Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019). GANSynth: Adversarial neural audio synthesis. In *7th International Conference on Learning Representations*, New Orleans, Louisiana.

[Favreau, 2001] Favreau, E. (2001). Phase vocoder applications in grm tools environment. In *Proceedings of the 4th International Conference on Digital Audio Effects*, pages 1–4, Limerick, Ireland.

[Feugère, 2013] Feugère, L. (2013). *Synthèse par règles de la voix chantée contrôlée par le geste et applications musicales*. PhD thesis, Groupe Audio et Acoustique (LIMSI-CNRS), ED 391 - Sciences mécaniques, acoustique, électronique et robotique de Paris (SMAER), Université Pierre et Marie Curie - Paris VI, Paris, France.

[Feugère et al., 2017] Feugère, L., d'Alessandro, C., Doval, B., and Perrotin, O. (2017). Cantor digitalis: Chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(2):1–19.

[Feugère et al., 2016] Feugère, L., O.Perrotin, d'Alessandro, C., Beux, S. L., and Doval, B. (2016). Samples performed by m. schaff and r. treiner in demo -chant à partir du texte- of anr project chanter cha(nt)n(umérique)te(mps)r(éel). (accessed: 01.09.2024).

[Flanagan and Golden, 1966] Flanagan, J. and Golden, R. (1966). Phase vocoder. *Bell Systems Tech Journal*, 45:1493–1509.

[Françoise, 2015] Françoise, J. (2015). *Motion-Sound Mapping by Demonstration*. PhD thesis, ED 130 L'EDITE : École doctorale Informatique, Télécommunications et Électronique, Institut de Recherche et Coordination Acoustique/Musique IRCAM, Université Pierre et Marie Curie.

[Fulford and Gingsborg, 2013] Fulford, R. and Gingsborg, J. (2013). The sign language of music: Musical shaping gestures (msgs) in rehearsal talk by performers with hearing impairments. *Empirical Musicology Review*, 8(1):53–67.

[Garnier, 2007] Garnier, M. (2007). *Communiquer en environnement bruyant :de l'adaptation jusqu'au forçage vocal*. PhD thesis, Université Pierre et Marie Curie - Paris VI, Paris, France.

[Garnier et al., 2007] Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., and Dubois, D. (2007). Characterisation of voice quality in western lyrical singing: from teachers' judgements to acoustic descriptions. *Journal of interdisciplinary music studies*, 1(2):62–91.

[Garnier et al., 2005] Garnier, M., Henrich, N., Dubois, D., Castellengo, M., Poitevineau, J., and Sotiropoulos, D. (2005). Etude de la qualité vocale

dans le chant lyrique. *Scolia [sciences cognitives, linguistique et intelligence artificielle / revue de linguistique]*, 20(1):151–169.

[Geslin, 1998] Geslin, Y. (1998). Sound and music transformation environments: A twenty-year experiment at the "groupe de recherches musicales". In *Proceedings DAFX-98 Digital Audio Effects Workshop*, pages 241–248.

[Godøy and Leman, 2009] Godøy, R. I. and Leman, M. (2009). *Musical Gestures: Sound, Movement, and Meaning*. Routledge.

[Haken, 2009] Haken, L. (2009). *Position Correction for an electronic musical instrument*. Champaign, IL. U.S. Patent US7619156B2.

[Haken et al., 1998] Haken, L., Tellman, E., and Wolfe, P. (1998). An indiscrete music keyboard. *Computer Music Journal*, 22(1):30–48.

[Hallam et al., 2016] Hallam, S., Cross, I., and Thaut, M., editors (2016). *The Oxford Handbook of Music Psychology*. Oxford University Press, 2nd edition.

[Henrich Bernardoni, 2001] Henrich Bernardoni, N. (2001). *Etude de la source glottique en voix parlée et chantée: modélisation et estimation, mesures acoustiques et électroglottographiques, perception*. Phd thesis, Université Pierre et Marie Curie - Paris VI. ⟨tel-00123133⟩.

[Henrich Bernardoni, 2014] Henrich Bernardoni, N. (2014). *La voix chantée*. Voix parole langage. De Boeck Superieur.

[Henrich Bernardoni, 2015] Henrich Bernardoni, N. (2015). *HDR Habilitation à diriger des recherches, La voix humaine : vibrations, résonances, interactions pneumo-phonorésonantielles*. Laboratoire Grenoble Images Parole Signal Automatique UMR 5216 GIPSA-lab, Université Grenoble Alpes.

[Henrich Bernardoni et al., 2008] Henrich Bernardoni, N., Bezard, P., Garnier, M., Expert, R., Guerin, C., et al. (2008). Towards a common terminology to describe voice quality in western lyrical singing: Contribution of a multidisciplinary research group. *Journal of Interdisciplinary Music Studies*, 2(1&2):71–93.

[Hijleh, 2012] Hijleh, M. (2012). *Towards a Global Music Theory: Practical Concepts and Methods for the Analysis of Music Across Human Cultures*. Routledge, 1st edition.

[Hildebrand, 1996] Hildebrand, H. (1996). *Method and Apparatus for Digital Filtering Signals*. Auburn Audio Technologies, Auburn, AL. U.S. Patent US5727074A.

[Hildebrand, 1998] Hildebrand, H. (1998). *Pitch detection and intonation correction apparatus and method*. Auburn Audio Technologies, Auburn, AL. U.S. Patent, Patent US5973252A.

[Hildebrand, 2014] Hildebrand, H. (2014). *Virtual tuning of a string instrument*. Auburn Audio Technologies, Auburn, AL. U.S. Patent US8648240B2.

[Hoffman and Cook, 2008] Hoffman, M. and Cook, P. (2008). Real-time dissonancizers: Two dissonance-augmenting audio effects. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland.

[Hunt and Wanderley, 2002] Hunt, A. and Wanderley, M. M. (2002). Mapping performer parameters to synthesis engines. *Organised Sound*, 7(2):97–108.

[ITU-R-BS.1116-3, 2015] ITU-R-BS.1116-3 (2015). Methods for the subjective assessment of small impairments in audio systems. Technical report, International Telecommunication Union Radiocommunication Sector Recommendation ITU.

[ITU-R-BS.1534-3, 2015] ITU-R-BS.1534-3 (2015). Method for the subjective assessment of intermediate quality level of audio systems. Technical report, International Telecommunication Union Radiocommunication Sector Recommendation ITU.

[Kameoka and Kuriyagawa, 1969] Kameoka, A. and Kuriyagawa, M. (1969). Consonance theory, part ii: Consonance of complex tones and its computation method. *Journal of the Acoustical Society of America*, 45(6):1460–1469.

[Keiler et al., 2001] Keiler, F., Arfib, D., and Zölzer, U. (2001). Efficient linear prediction for digital audio effects. In *Proceedings of the 15th International Conference on Digital Audio Effects*, pages 277–285, Verona, Italy.

[Klatt and Klatt, 1990] Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857.

[Lamb and Robertson, 2011] Lamb, R. and Robertson, A. N. (2011). Seaboard: a new piano keyboard-related interface combining discrete and continuous control. In *Proceedings of the 11th International Conference on Digital Audio Effects*, pages 503–506, Oslo, Norway.

[Lanchantin et al., 2011] Lanchantin, P., Farner, S., Veaux, C., Degottex, G., Obin, N., Beller, G., Villavicencio, F., Huber, S., Peeters, G., Roebel, A., and Rodet, X. (2011). Vivos voco: A survey of recent research on voice transformation at ircam. In *International Conference on Digital Audio Effects*, pages 277–285, Paris, France.

[Laroche and Dolson, 1999] Laroche, J. and Dolson, M. (1999). New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications. *Journal of the Audio Engineering Society*, 47(11):928–936.

[Loscos and Bonada, 2004] Loscos, A. and Bonada, J. (2004). Emulating rough and growl voice in spectral domain. In *Proceedings of the 7th International Conference on Digital Audio Effects*, pages 1–4, Naples, Italy.

[Lucas et al., 2021] Lucas, T., d'Alessandro, C., and De Laubier, S. (2021). Mono-replay: a software tool for digitized sound animation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Shanghai, China.

[Makhoul, 1975] Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

[Makhoul, 1977] Makhoul, J. (1977). Stable and efficient lattice methods for linear prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-25(5):423–428.

[Makhoul and El-Jaroudi, 1986] Makhoul, J. and El-Jaroudi, A. (1986). Time-scale modification in medium to low rate speech coding. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1705–1708, Tokyo, Japan.

[Martinez Ramírez, 2020] Martinez Ramírez, M. (2020). *Deep Learning for AUdio Effects Modeling*. PhD thesis, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK.

[McPherson et al., 2013] McPherson, A. P., Gierakowski, A., and Stark, A. M. (2013). The space between the notes: Adding expressive pitch control to the piano keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2195–2204, New York, NY, USA. ACM.

[Mitchell and Heap, 2011] Mitchell, T. J. and Heap, I. (2011). Soundgrasp: A gestural interface for the performance of live music. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, pages 465–468, Oslo, Norway.

[Mitchell et al., 2012] Mitchell, T. J., Madgwick, S. O. H., and Heap, I. (2012). Musical interaction with hand posture and orientation: A toolbox of gestural control mechanisms. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, pages 499–502, Ann Arbor, MI, USA.

[Moinet and Dutoit, 2011] Moinet, A. and Dutoit, T. (2011). Pvsola: A phase vocoder with synchronized overlap-add. In *Proceedings of the 14th International Conference on Digital Audio Effects*, pages 269–275, Paris, France.

[Moorer, 1979a] Moorer, J. (1979a). The use of linear prediction of speech in computer music applications. *Journal of the Audio Engineering Society*, 27:134–140.

[Moorer, 1979b] Moorer, J. A. (1979b). About this reverberation business. *Computer Music Journal*, 3(2):13–18.

[Morise, 2015] Morise, M. (2015). Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67:1–7.

[Morise, 2016] Morise, M. (2016). D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, 84:57–65.

[Morise et al., 2016] Morise, M., Yokomori, F., and Ozawa, K. (2016). World: A vocoder-based high-quality speech synthesis system for real-time applications. *Institute of Electronics, Information and Communication Engineers (IEICE) Transactions on Information and Systems*, E99.D(7):1877–1884.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467.

[Moulines and Laroche, 1995] Moulines, E. and Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16(2):175–205. Voice Conversion: State of the Art and Perspectives.

[Neubacker, 2011] Neubacker, P. (2011). *Sound object oriented analysis and note object oriented Processing of polyphonic sound recordings*. Celemony, Munich, Germany. U.S. Patent US8022286B2.

[Orfanidis, 1990] Orfanidis, S. J. (1990). *Optimum Signal Processing, An Introduction*. McGraw-Hill, 2nd edition.

[Orio et al., 2001] Orio, N., Schnell, N., and Wanderley, M. (2001). Input devices for musical expression: Borrowing tools from hci. In *Proceedings Workshop on New Interfaces for Musical Expression (NIME-01)*, pages 15–18, Seattle, USA.

[Peeters and Richard, 2021] Peeters, G. and Richard, G. (2021). Deep learning for audio and music. In *Multi-faceted Deep Learning: Models and Data*. Springer, M.

[Perrotin and D'Alessandro, 2016] Perrotin, O. and D'Alessandro, C. (2016). Target acquisition vs. expressive motion: Dynamic pitch warping for intonation correction. *Association for Computing Machinery (ACM). ACM Transactions on Computer-Human Interaction Journal (TOCHI)*, 23(3):1–21.

[Politis et al., 2012] Politis, A., Pihlajamaki, T., and Pulkki, V. (2012). Parametric spatial audio effects. In *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, pages 1–8, York, UK.

[Portnoff, 1976] Portnoff, M. (1976). Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):243–248.

[Portnoff, 1980] Portnoff, M. R. (1980). Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, Signal Processing*, ASSP-28:55–69.

[Rasamimanana et al., 2011] Rasamimanana, N., Bevilacqua, F., Schnell, N., Guedy, F., Maestracci, E. F., Zamborlin, B., and Petrevsky, U. (2011). Modular musical objects towards embodied control of digital music. In *Proceedings of the 5th International Conference on Tangible, Embedded, and Embodied Interaction (TEI '11)*, pages 9–12, Funchal, Portugal. Association for Computing Machinery (ACM).

[Rehding and Rings, 2020] Rehding, A. and Rings, S. (2020). *The Oxford Handbook of Critical Concepts in Music Theory*. Oxford Handbooks. Oxford University Press. Online edition, Oxford Academic, 6 Jan. 2015.

[Roche, 2016] Roche, F. (2016). *Music sound synthesis using machine learning: Towards a perceptually relevant control space*. PhD thesis, École Doctorale Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS), Laboratoire Grenoble Images Parole Signal Automatique (GIPSA-Lab), Université Grenoble-Alpes, Grenoble, France.

[Roebel, 2003] Roebel, A. (2003). A new approach to transient processing in the phase vocoder. In *Proceedings of the 6th International Conference on Digital Audio Effects*, pages 344–349, London, United Kingdom.

[Roebel, 2010] Roebel, A. (2010). A shape-invariant phase vocoder for speech transformation. In *Proceedings of the 13th International Conference on Digital Audio Effects*, pages 1–8, Graz, Austria.

[Roebel and Bous, 2022] Roebel, A. and Bous, F. (2022). Neural vocoding for singing and speaking voices with the multi-band excited wavenet. *Information*, 3-103(13):1–29.

[Roebel and Rodet, 2005] Roebel, A. and Rodet, X. (2005). Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proceedings of the 8th International Conference on Digital Audio Effects*, pages 30–35, Madrid, Spain.

[Rogers, 2017] Rogers, L. (2017). *Discrimination Testing in Sensory Science: A Practical Handbook*. Elsevier Science.

[Roucos and Wilgus, 1985] Roucos, S. and Wilgus, A. M. (1985). High quality time-scale modification for speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 493–496.

[Rovan et al., 1997] Rovan, J., Wanderley, M. M., Dubnov, S., and Depalle, P. (1997). Instrumental gestural mapping strategies as expressivity determinants in computer music performance. In *Proceedings of the AIMI International Workshop*, pages 68–73. Cited on pages 7, 8, and 20.

[Schaeffer, 1966] Schaeffer, P. (1966). *Traité des Objets Musicaux*. Seuil, Paris, France.

[Schnell et al., 2011] Schnell, N., Bevilacqua, F., Rasamimanana, N., Bloit, J., Guedy, F., and Flety, E. (2011). Playing the mo – gestural control and re-embodiment of recorded sound and music. In *Proceedings of the 11th International Conference on New Interfaces for Musical Expression (NIME '11)*, pages 1–4, Oslo, Norway. University of Oslo and Norwegian Academy of Music.

[Schoeffler et al., 2018] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J. (2018). webmushra — a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1):1–6.

[Schroeder, 1962] Schroeder, M. (1962). Natural-sounding artificial reverberation. *Journal of the Audio Engineering Society*, 10(3):219–223.

[Schwarz and Rodet, 1999] Schwarz, D. and Rodet, X. (1999). Spectral envelope estimation, representation, and morphing for sound analysis, transformation, and synthesis. In *ICMC: International Computer Music Conference*, pages 1–4, Pekin, China.

[Schörkhuber et al., 2012] Schörkhuber, C., Klapuri, A., and Sontacchi, A. (2012). Pitch shifting of audio signals using the constant-q transform. In *Proceedings of the 15th International Conference on Digital Audio Effects*, pages 269–275, York, UK.

[Seashore, 1931] Seashore, C. (1931). The natural history of the vibrato. *Proceedings of the National Academy of Sciences*, 17(12):623–626.

[Strange, 1983] Strange, A. (1983). *Electronic Music, Systems, Techniques and Controls*. W. C. Brown.

[Sundberg, 1994] Sundberg, J. (1994). Acoustic and psychoacoustic aspects of vocal vibrato. *Speech, Music and Hearing Quarterly Progress and Status Report 2-3, Swedish Royal Institute of Technologies*, 35(2-3):45–68.

[Verfaille, 2003] Verfaille, V. (2003). *Effets audionumériques adaptatifs : théorie, mise en œuvre et usage en création musicale numérique*. PhD thesis, Ecole Doctorale 353 – Mécanique, Physique et Modélisation, Laboratoire de Mécanique et d'Acoustique — UPR 7051, Université de la Méditerranée - Aix-Marseille II, Marseille, France.

[Verfaille and Arfib, 2002] Verfaille, V. and Arfib, D. (2002). Implementation strategies for adaptive digital audio effects. In *Proceedings of the 5th International Conference on Digital Audio Effects*, pages 21–26, Hamburg, Germany.

[Verfaille et al., 2006a] Verfaille, V., Guastavino, C., and Traube, C. (2006a). An interdisciplinary approach to audio effect classification. In *Proceedings of the 9th International Conference on Digital Audio Effects*, pages 107–113, Montreal, Canada.

[Verfaille et al., 2006b] Verfaille, V., Wanderley, M., and Depalle, P. (2006b). Mapping strategies for gestural and adaptive control of digital audio effects. *Journal of New Music Research*, 35(1):71–93.

[Verfaille et al., 2006c] Verfaille, V., Zölzer, U., and Arfib, D. (2006c). Adaptive digital audio effects (a-dafx): A new class of sound transformations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1817–1831.

[Wanderley, 2001] Wanderley, M. M. (2001). Gestural control of music. In *International Workshop on Human Supervision and Control in Engineering and Music*. Cited on page 8.

[Wilmering et al., 2020] Wilmering, T., Alessia, M., and Sandler, M. (2020). A history of audio effects. *Applied Sciences*, 10(3):791.

[Wilson and Fazenda, 2013] Wilson, A. and Fazenda, B. (2013). Perception and evaluation of audio quality in music production. In *Proceedings of the 16th International Conference on Digital Audio Effects*, pages 68–77, Maynooth, Ireland.

[Xiao et al., 2019] Xiao, X., Locqueville, G., d'Alessandro, C., and Doval, B. (2019). T-voks: the singing and speaking theremin. In *NIME 2019 International Conference on New Interfaces for Musical Expression*, pages 110–115, Porto Alegre, Brazil.

[Zacharov et al., 2018] Zacharov, N., Worch, T., and Ramsgaard, J. (2018). *Sensory Evaluation of Sound*. CRC Press, Boca Raton, Florida.

[Zbyszynski et al., 2013] Zbyszynski, M., Zicarelli, D., and Collecchia, R. (2013). fzero fundamental estimation for max 6. In *Proceedings of the 39th International Computer Music Conference (ICMC)*, pages 1–6, Perth, Australia.

[Zölzer, 2008] Zölzer, U. (2008). Equalizers. In Zölzer, U., editor, *DAFX: Digital Audio Effects*, pages 115–189. John Wiley & Sons, Ltd, Chichester, UK.

[Zölzer, 2011] Zölzer, U. (2011). *Digital Audio Effects*. John Wiley and Sons, Ltd, West Sussex, United Kingdom, second edition.

[Zölzer and Boltze, 1995] Zölzer, U. and Boltze, T. (1995). Parametric digital filter structures. Technical Report Paper 4099, Hamburg University of Technology, Hamburg, Germany.

Musical Recordings References

[Adele, 2011] Adele (2011). Rolling in the deep. In: 21. XL Recordings and Columbia.

[Beatles, 1966] Beatles, T. (1966). Tomorrow never knows. In: Revolver. Published by Parlophone.

[Billie Eilish, 2019] Billie Eilish (2019). bad guy, strange addiction, xanny, ilo milo, bury a friend. In: WHEN WE ALL FALL ASLEEP, WHERE DO WE GO? Published by Darkroom/Interscope Records.

[Billie Eilish, 2021] Billie Eilish (2021). Nda. Happier Than Ever. Published by Darkroom/Interscope Records.

[Björk, 2004] Björk (2004). Who is it, the pleasure is all mine. In: Medúlla. Published by One Little Independent.

[Björk, 2015] Björk (2015). Stonemilker. In: Vulnicura. Published by One Little Independent Records.

[Björk, 2022] Björk (2022). Mycelia. In: Fossora. Published by One Little Independent.

[Bon Jovi, 2013] Bon Jovi (2013). What about now. Published by Island Mercury.

[Charli XCX, 2020a] Charli XCX (2020a). Forever, claws, c2.0. how i'm feeling now. Published by Atlantic Records, Asylum Records.

[Charli XCX, 2020b] Charli XCX (2020b). Von dutch. In: Von Dutch - Single. Published by Atlantic Records.

[Cher, 1998] Cher (1998). Believe. In: Believe. Published by Warner Bros. Records.

[Color Me Badd, 1991] Color Me Badd (1991). I wanna sex you up. In: C.M.B. Published by Giant Records.

[Daft Punk, 1997] Daft Punk (1997). Around the world. In: Homework. Published by Daft Life Ltd, ADA France.

[Daft Punk, 2000] Daft Punk (2000). Single. Published by Virgin Records.

[Daft Punk, 2001] Daft Punk (2001). Harder better faster stronger, one more time. In: Discovery. Published by Virgin Records.

[Daft Punk, 2013] Daft Punk (2013). Instant crush. In: Random Access Memories. Published by Columbia Records.

[Drake, 2011] Drake (2011). Headlines. Single. Published by Young Money Entertainment, Cash Money Records.

[Dua Lipa, 2020] Dua Lipa (2020). Don't start now. In: Future Nostalgia. Published by Warner Records, Prescription Songs, Warner Chappell Music.

[Dua Lipa et al., 2020] Dua Lipa, Madonna, and Missy Elliott (2020). Levitating (the blessed madonna remix). In: Club Future Nostalgia. Published by We Still Believe, Warner.

[Ellie Goulding, 2012] Ellie Goulding (2012). Anything could happen. In: Halcyon. Published by Polydor Records.

[Foo Fighters, 2011] Foo Fighters (2011). White limo. In: Wasting Light. Published by Roswell Records.

[Garbage, 1995] Garbage (1995). As heaven is wide. In: Garbage. Published by Almo Sounds.

[Garbage, 2012] Garbage (2012). Blood for puppies. In: Not your Kind of People. Published by Vibecrusher Music (BMI) and Deadarm Music (ASCAP), administered by Kobalt Music Group. (p)(c) Garbage Unlimited.

[Gotye, 2011] Gotye (2011). Somebody that i used to know. In: Making Mirrors. Published by Eleven, Universal,V2 - Universal Music Australia.

[Jimi Hendrix, 1967] Jimi Hendrix (1967). Third stone from the sun. Published by Legacy Recordings, Sony Music Entertainment.

[Justin Bieber, 2015] Justin Bieber (2015). What do you mean? In: Purpose. Published by Def Jam Recordings.

[Kygo and Selena Gomez, 2017] Kygo and Selena Gomez (2017). It ain't me. Single from the EP Stargazing. Published by Interscope, Sony, Ultra.

[Lady Gaga, 2009] Lady Gaga (2009). Monster. In: The Fame Monster. Published by Interscope.

[Lana Del Rey, 2012a] Lana Del Rey (2012a). Doin' time. In: Norman Fucking Rowell!! Published by Stranger Records, Interscope Records, Polydor Records.

[Lana Del Rey, 2012b] Lana Del Rey (2012b). Summertime sadness. In: Born to Die. Published by Stranger Records, Interscope Records, Polydor Records.

[Lorde, 2013] Lorde (2013). Royals, team. Pure Heroine. Published by Universal Music.

[Loreen, 2022] Loreen (2022). Neon lights - single. Published by Universal Music AB.

[Loreen, 2023] Loreen (2023). Tattoo (acoustic) - single. Published by Universal Music AB.

[Madonna, 2000] Madonna (2000). Music. In: Music. Published by Warner Records.

[Madonna, 2005] Madonna (2005). Forbidden love. In: Confessions on a Dance Floor. Published by Warner Bros. Records.

[Madonna, 2005] Madonna (2005). Sorry. In: Confessions on A Dance Floor. Published by Warner Records.

[Madonna and Sickick, 2021] Madonna and Sickick (2021). Frozen. In: Frozen (Sickick Remix). Published by Artist Partner Group Inc/Warner Records.

[Maroon 5, 2019] Maroon 5 (2019). Memories. In: Jordi. Published by 222 Records, Interscope Records.

[Melanie Martinez, 2023a] Melanie Martinez (2023a). The contortionist, void. In: Portals. Published by Atlantic Records.

[Melanie Martinez, 2023b] Melanie Martinez (2023b). Death. In: Portals. Published by Atlantic Records.

[Mike Posner, 2015] Mike Posner (2015). I took a pill in ibiza. In: At Night, Alone. Published by Island Records.

[Mon Laferte, 2023] Mon Laferte (2023). Tenochtitlan, no+sad, casta diva, te juro que volveré. In: Autopoietica. Published by Universal.

[Mon Laferte, 2024] Mon Laferte (2024). Obra de dios. In: Obra de Dios - Single. Published by Mon Laferte Inc.

[Pharrell Williams, 2013] Pharrell Williams (2013). Happy. In: Girl. Published by Back Lot Music, i am OTHER, Columbia Records.

[Prince, 1998] Prince (1998). Good love. In: Crystal Ball. Published by Legacy Records.

[Raye, 2024] Raye (2024). Black mascara. In: My 21st Century Blues. Label: Human Re Sources.

[Robyn and Zhala, 2018] Robyn and Zhala (2018). Human being. Published by Konichiwa, Island, Interscope.

[Robyn & la bagatelle magique, 2015] Robyn & la bagatelle magique (2015). Got to work it out, set me free. In: Love is free. Published by Cherrytree Records/Kiersznbaum.

[Rosalía, 2018] Rosalía (2018). De aquí no sales, di mi nombre. In: El Mal Querer. Published by Columbia, Universal Music Publishing, Warner Chappell Music.

[Rosalía, 2022] Rosalía (2022). Como un g, diablo, cuuuuuuuuuute,. In: motomami. Published by Columbia, Concord Music Publishing, Sony Music Publishing, Universal Music Publishing, Warner Chappell Music.

[T-Pain, 2005] T-Pain (2005). I'm sprung. Single. Published by Konvict Muzik, Jive Records.

[Taylor Swift, 2017] Taylor Swift (2017). Delicate, gorgeous. In: Reputation. Published by Big Machine Records LLC, MXM Music.

[Taylor Swift, 2019] Taylor Swift (2019). Cruel summer, me!, after glow. In: Lover. Published by Taylor Swift, Hipgnosis Songs Group, Universal Music Publishing.

[Taylor Swift, 2020] Taylor Swift (2020). cardigan, cardigan (the long pond studio sessions). In: Folklore - The Long Pond Studio Sessions version). Published by Republic Records.

[The Carters (Beyoncé, Jay-Z), 2018] The Carters (Beyoncé, Jay-Z) (2018). Apeshit. In: Everything is Love. Published by Parkwood Entertainment, Roc Nation, Sony Music Publishing.

[The Chainsmokers, 2016] The Chainsmokers (2016). Closer. Single. Published by Disruptor Records, Columbia Records.

[The Weeknd, 2016] The Weeknd (2016). Starboy. In: Starboy. Published by Republic Records.

[Tove Lo, 2014] Tove Lo (2014). Habits (stay high) - hippie sabotage remix. In: Queen of the Clouds. Published by Universal Music.