Comparing vocoders for automatic vocal tuning

Daniel Hernan Molina-Villota, Christophe d'Alessandro

▶ **To cite this version:**

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, 10.5281/zenodo.10115215 . hal-04283705

HAL Id: hal-04283705
https://hal.science/hal-04283705v1

Submitted on 14 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

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

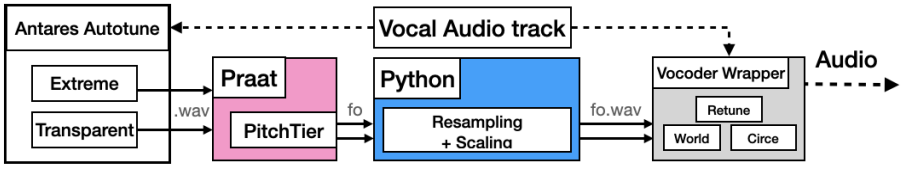

Fig. 1. Flow diagram for the methodology for vocoding with two presets: transparent and extreme retuning (f0 discrete curve).

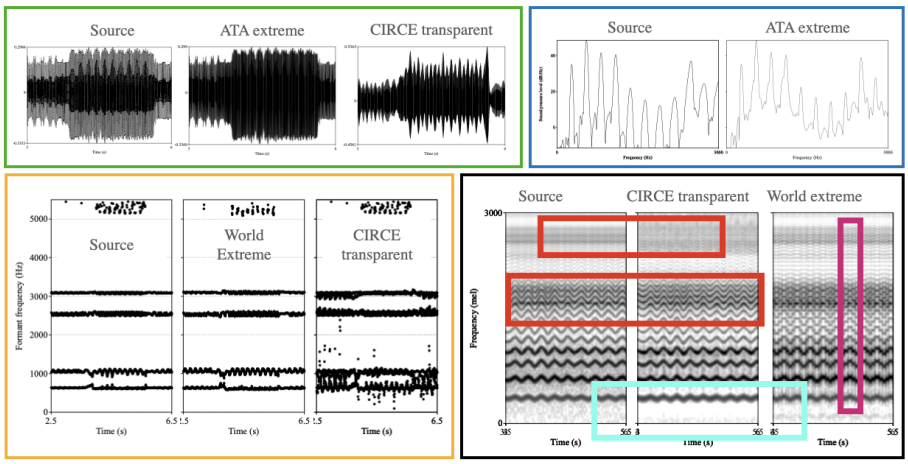

Fig. 2. Green block (Signal Shape): Changes are observed for 2 vocoders. Autotune extreme correction case shows minimal changes while Circe transparent case exhibits significant shape variations. Yellow block (Formants): World shows notable deepening in formant variation and CIRCE exhibits substantial 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".

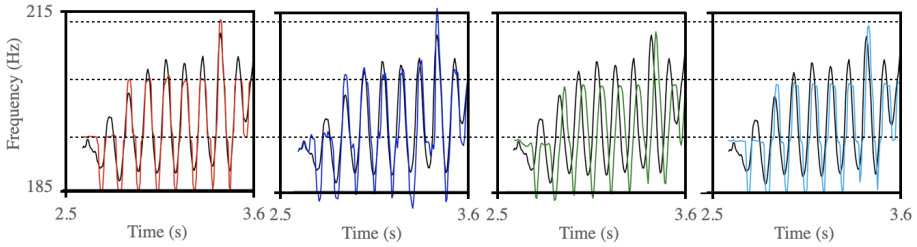

Fig. 3. F0-Path for extreme retuning using (left to right): Autotune, CIRCE, Retune and World. Autotune and World reach exact pitch values more accurately than the others. Retune presents a bigger latency than the other ones.

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)