Autotune vs AI Voice Cloning: A Case Study for Automatic Pitch Corrections

A. Haron[1], D. Dzulkifli[1], S. Jibin[2], R. M. Azizul[2], A. Azlan[1] A. Afiq[1], Y. Kaliaperumal[1], F. Norman[1], R. Fauzan[1], A. Halim[3], I. Kamel[4], K. Izam[5] and P. N. 'Aainaa[6]

[1]*Faculty of Creative Multimedia, Multimedia University, Cyberjaya, Malaysia*
[2]*Faculty of Cinematic Arts, Multimedia University, Cyberjaya, Malaysia*
[3]*Network & Intelligent Campus Ecosystems, Multimedia University, Cyberjaya, Malaysia*
[4]*Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia*
[5]*Strategic Marketing Department, Multimedia University, Cyberjaya, Malaysia*
[6]*Student Lifestyle and Experience, Multimedia University, Cyberjaya, Malaysia*

Keywords: Autotune, AI Voice Cloning, Artificial Intelligence, Music Production.

Abstract: This article presents a comparative case study between autotune and AI voice cloning for automatic pitch corrections. A recording of a vocalist with pitch inconsistencies is used as the source signal for pitch correction. Autotune and AI voice cloning results are observed and analysed using melodic range spectrograms to highlight the differences between these approaches. This study's findings revealed each approach's uniqueness, followed by reflections and considerations as recommendations for music production practice.

1 INTRODUCTION

For non-professional singers and hobbyists, pitching inconsistencies are common issues in music recordings. Two approaches that address this issue include autotune and more recently, AI voice cloning. In this article, we present our comparative case study between autotune and AI voice cloning by observing and analysing melodic range spectrograms. The objective of this study is to present a comparison between two different approaches for automatic pitch correction, highlighting the balance between technical refinement and preservation of artistic authenticity.

2 AUTOTUNE AND AI VOICE CLONING

Autotune and AI voice cloning are innovations within music production for the manipulation and correction of the human voice. Autotune, which originates as a corrective tool for pitch imperfections in recordings, has evolved into an aesthetic choice, utilized across various musical genres and production contexts (Danielsen, A., 2018).

Through signal processing, autotune analyses and corrects vocal pitch deviations to align with predetermined pitches. In contrast, AI voice cloning harnesses the power of machine learning and neural networks to replicate the nuances of human speech and singing with realism. By training on extensive datasets of vocal recordings, AI voice cloning systems can synthesize vocal performances that closely emulate the characteristics of specific individuals.

2.1 Autotune

Autotune has significantly influenced the contemporary landscape of popular music. Initially developed as a tool for pitch correction, autotune has evolved into a ubiquitous application employed by recording engineers and artists alike (Tyrangiel, J, 2009). At its core, autotuning involves analysing the incoming audio signal, identifying deviations from the desired pitch and manipulating these deviations to align with predetermined musical scales or notes.

Haron, A., Dzulkifli, D., Jibin, S., Azizul, R. M., Azlan, A., Afiq, A., Kaliaperumal, Y., Norman, F., Fauzan, R., Halim, A., Kamel, I., Izam, K. and Aainaa, P. N.
Autotune vs AI Voice Cloning: A Case Study for Automatic Pitch Corrections.
DOI: 10.5220/0013333500004557
In *Proceedings of the 4th International Conference on Creative Multimedia (ICCM 2024)*, pages 56-60
ISBN: 978-989-758-733-7; ISSN: 3051-6412

This corrective process rectifies off-key or out-of-tune vocal performances.

Fundamentally, autotune operates through a series of signal-processing processes. Upon receiving audio input, autotune employs Fourier analysis to decompose the signal into its constituent frequencies, from which it identifies the dominant pitch of the input signal and compares it to a predefined scale or pitch. This comparison facilitates the detection of discrepancies between the actual pitch and the intended pitch, highlighting areas requiring correction.

Autotune offers varying degrees of correction intensity, allowing granular controls over the extent to which pitch deviations are rectified. This enables users to achieve subtle pitch correction or more pronounced stylistic effects, catering to artistic preferences and production requirements.

Despite its widespread adoption and utility, autotune has raised polarizing responses within the music industry and listeners alike. Critics argue that excessive reliance on autotunes encourages homogeneity in vocal performances, diminishing the uniqueness and authenticity of individual expression. Conversely, proponents argue that autotune serves as an aesthetical choice, for correcting technical imperfections, and fosters creative experimentation.

2.2 AI Voice Cloning

AI voice cloning represents a recent groundbreaking advancement in the field of audio synthesis and digital signal processing. At its core, AI voice cloning uses machine learning algorithms, to analyse and emulate the nuances of human speech and singing (Sisman, B., et al. 2020) (Ren, Z., 2024).

The training process for AI voice cloning involves processing large quantities of audio data. Through training, it learns to discern subtle nuances in pitch, timbre, intonation, and articulation, encoding the unique traits of a vocalist.

AI voice cloning finds applications across diverse domains, including entertainment, multimedia production, accessibility services, and interactive systems such as video games. Nevertheless, ethical considerations and societal implications accompany the proliferation of AI voice cloning technology. Concerns regarding privacy, consent, identity manipulation, and the potential for malicious misuse highlight the need for responsible practices within music production.

3 COMPARISON OF MELODIC RANGE SPECTROGRAMS

In this section, we present melodic range spectrograms (Gohari, M., et al. 2024) of four distinct audio clips. Each clip is a snippet taken from the same phrase and location in the music, each with equal duration. All plots are generated using Sonic Visualizer using the same parameter values. In our study, we used Reaper's ReaTune VST plug-in with automatic pitch correction for autotune, and Retrieval-Based Voice Conversion (RVC) WebUI for AI voice cloning.

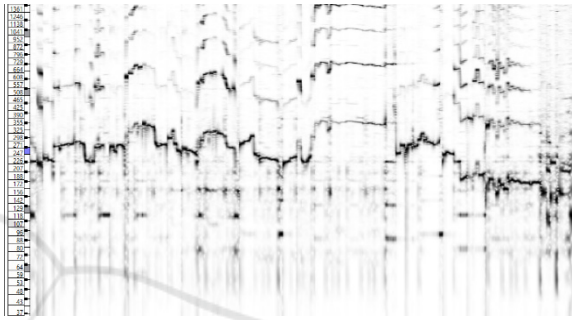

Figure 1: Melodic range spectrogram of unaltered recording.

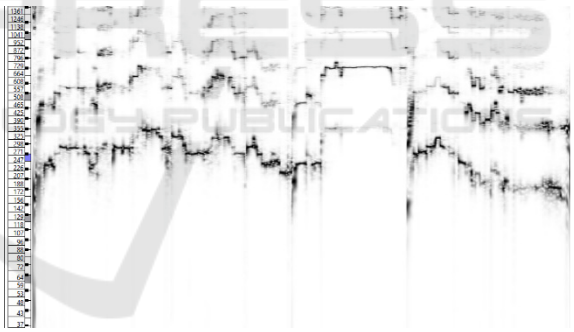

Figure 2: Melodic range spectrogram of original singer.

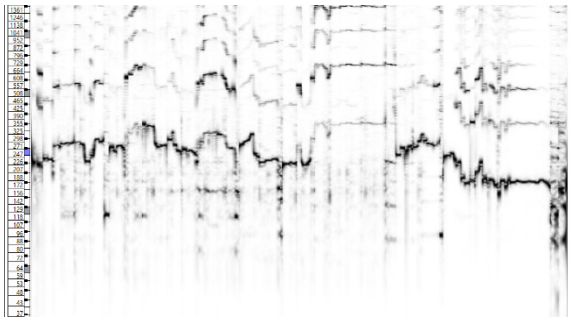

Figure 3: Melodic range spectrogram of recording with autotune.

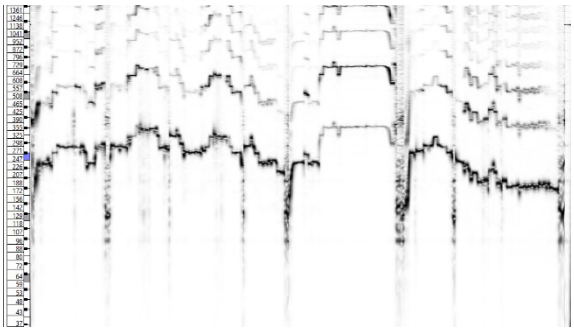

Figure 4: Melodic range spectrogram of recording with AI voice cloning.

3.1 Initial Observations

Figure 1 shows a spectrogram created from the actual recording session. The recording session was a multitrack recording with the full band. As such, there is some signal bleeding into the singer's microphone from other instruments in the recording studio. This signal bleed can be observed in Figure 1 and highlighted in Figure 5.

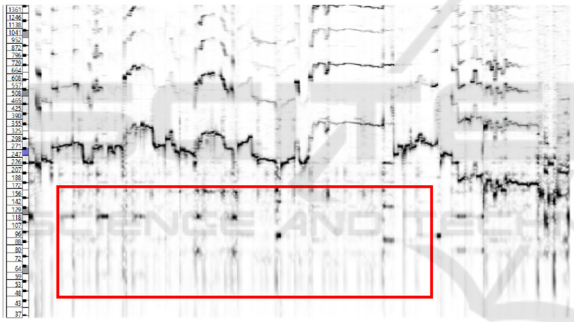

Figure 5: Plot of Figure 1 with a red box to highlight signal bleed.

Figure 2 shows a spectrogram of the original singer of the song we covered in the recording session. This vocal track was created through source separation from the original track using multi-scale multi-band densenet (Takahashi, N., & Mitsufuji, Y., 2017). The darkest areas in Figure 2 form a blockish contour that exhibits distinct steps and clear horizontal lines compared to the smoother contours and slanted horizontal lines observed in Figure 1. These steps indicate exact pitch changes, and the clear horizontal lines indicate a steady pitch, as one would expect from a highly acclaimed professional singer. Figure 6 shows this difference side by side.

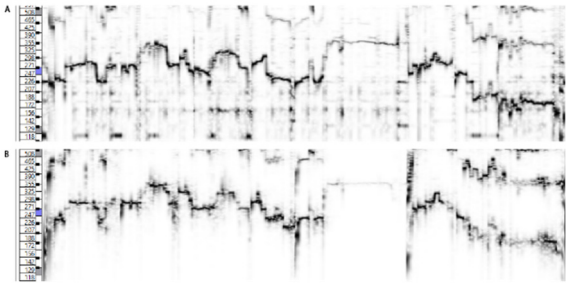

Figure 6: (A, top) Unaltered recording, (B, bottom) Original singer.

3.2 Comparing Unaltered Recording and Autotuned Recording

In this section, we present our observations when comparing the unaltered recording against the autotuned output.

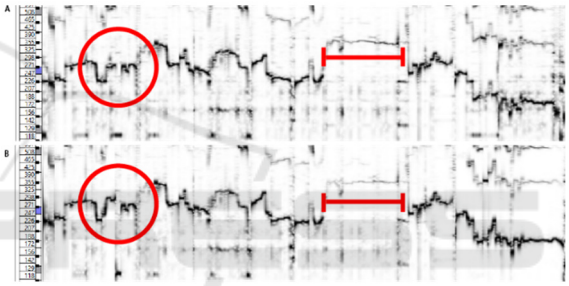

Figure 7: (A, top) Unaltered recording, (B, bottom) Autotuned version.

Figure 7 illustrates spectrograms from both the unaltered recording and the autotuned version of the recording. The autotuned version clearly shows improvement in pitch as highlighted with circles and lines in Figure 7.

Melody contour in the circled region of the unaltered recording and autotuned version illustrates pitch correction as a result of the autotune process. The contour in the autotuned version exhibits steps that are not observed in the unaltered recording. While the melody contour in the autotuned version above the highlighted line illustrates a steadier pitch compared to the unaltered recording. These two examples are not exclusive, as many smaller and more nuanced differences can be observed between the two plots. These two instances are highlighted as they exhibit clear differences between before and after autotune. It can also be observed that the signal bleed remains after applying autotune to the unaltered recording.

One major advantage of autotune is the minimal computational resources required compared to AI

voice cloning. Autotune is commonly used in real-time during live performances, with a small delay that is manageable with the right parameters and hardware.

3.3 Comparing Unaltered Recording and with AI Voice Cloning

In this section, we present our observations when comparing the unaltered recording against AI voice cloning results. Figure 8 illustrates these signals with an additional melodic range spectrogram of the original singer as a reference.

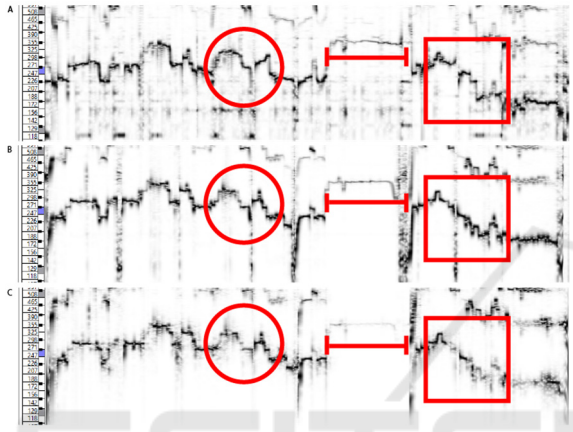

Figure 8: (A, top) Unaltered recording, (B, middle) AI voice cloned version, (C, bottom) Original singer.

The AI voice cloning process requires ample training data for it to deliver the desirable results. Training data in this case refers to voice recording samples of the individual who will be the inferencing voice. The inferencing voice will be mimicking the voice from the signal to be processed. In our case, the inferencing voice will be the singer of the unaltered recording, while the signal to be processed is the voice recording of the original singer.

In the Retrieval-Based Voice Conversion (RVC) WebUI, a few samples of inference voices are provided. These inference voices were trained with nearly 50 hours of training data. However, for an inferencing voice with acceptable quality, around 10 minutes of high-quality training data will be sufficient. We used roughly 10 minutes of training data to create an inferencing voice. Training an inferencing voice is computationally expensive compared to autotune. It took us around 20 minutes to train an inference voice using 10 minutes of training data using a Windows PC with an 11th Gen Intel Core i5 and an Intel Iris Xe GPU. Training can be expected to perform quicker with better and faster hardware. Currently, a real-time implementation for AI voice cloning exists, using pre-trained inference voices.

In Figure 8, three regions of interest are highlighted, each with a different shape. Melody contour within the highlighted circle for the AI voice cloned output (Figure 8, middle) exhibits similarities from both the unaltered recording (Figure 8, top) and the original singer (Figure 8, bottom). The contour for AI voice cloned output in this region shows more distinctive steps, which closely resemble the original singer yet also shows significant dips or valleys in the middle of the contour within the highlighted circle, which closely resembles the unaltered recording. Similar can be observed for the regions above the highlighted line and regions within the highlighted square. The shape of the melody contour for the AI voice cloning results closely resembles the contour of the original singer, while the intensity or loudness of the AI voice cloning results closely resembles the intensity of the unaltered recording.

4 REFLECTIONS AND CONSIDERATIONS

Autotune is portrayed as a transformative tool in audio engineering, originally developed for pitch correction but widely used across various musical genres as an aesthetic choice. However, we need to acknowledge the polarizing reception of autotune within the music industry, with critics raising concerns about its potential to homogenize vocal performances and detract from authentic expression.

The emergence of AI voice cloning is a groundbreaking advancement, enabling the replication of human speech and singing with unprecedented realism. AI voice cloning raises important ethical considerations surrounding privacy, consent, and identity manipulation, underscoring the need for responsible practices in music production.

This study provided a comparative analysis of autotune and AI voice cloning through observation and analysis of melodic range spectrograms. These spectrograms offer insights into the output of each technology on automatic pitch correction, highlighting differences in melodic contours and intensity between the original recordings and processed versions. By comparing the processed versions to the original recordings and the performances of professional singers, this study offers evaluations of the output from each technology.

We are of the opinion that for most use cases, a mixture of unaltered recording, autotuned and AI voice cloned output should be used in the final mix. Using only autotune without audibly hearing the artefacts doesn't yield an acceptable result for pitch correction while using only the AI voice cloning is unsuitable as the nuances in singing such as vibrato and tremolo are taken from the original singer. The authors advocate for the use of such technologies as tools to correct rather than to improve singing quality.

5 CONCLUSIONS

This study highlighted the complex interplay between technological innovation and artistic integrity in music production. By examining autotune and AI voice cloning as solutions to correcting vocal pitch imperfections, this study discussed functionalities and ethical considerations for such tools. While autotune offers reliable pitch correction capabilities, its polarized reception raises debate around artistic authenticity. Conversely, AI voice cloning presents a leap in vocal synthesis realism, yet also raises concerns, particularly regarding privacy and identity manipulation. A comparative analysis through melodic range spectrograms was presented and provides insights into each technology's output. Finally, the authors of the study advocate for a balanced integration of these tools, prioritizing artistic authenticity while leveraging technological advancements to correct, rather than to improve.

REFERENCES

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An open-source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1467-1468).

Danielsen, A. (2018) Music, media and technological creativity in the digital age. Nordic Research in Music Education Yearbook Vol. 18, 9.

Gohari, M., Bestagini, P., Benini, S., & Adami, N. (2024). Spectrogram-Based Detection of Auto-Tuned Vocals in Music Recordings. arXiv preprint arXiv:2403.05380.

Ren, Z. (2024). Selection of Optimal Solution for Example and Model of Retrieval-Based Voice Conversion. In 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023) (pp. 468-475). Atlantis Press.

Sisman, B., Yamagishi, J., King, S., & Li, H. (2020). An overview of voice conversion and its challenges: From statistical modelling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 132-157.

Takahashi, N., & Mitsufuji, Y. (2017). Multi-scale multi-band densenets for audio source separation. In 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 21-25). IEEE.

Tyrangiel, J. (2009). Auto-tune: Why pop music sounds perfect. Time Magazine, 1877372-3.