TUNING MATTERS: ANALYZING MUSICAL TUNING BIAS IN NEURAL VOCODERS

Hans-Ulrich Berendes, Ben Maman, Meinard Müller
International Audio Laboratories Erlangen
{hans-ulrich.berendes, ben.maman, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Vocoders, which reconstruct time-domain waveforms from spectral representations such as mel-spectrograms, are essential in modern music and speech synthesis. Traditional signal-processing techniques like the Griffin-Lim algorithm have largely been replaced by neural vocoders, which leverage generative models to achieve superior audio quality. However, these models can introduce artifacts and biases, potentially affecting their output in unforeseen ways. In this study, we examine how different musical tunings affect neural mel-to-audio vocoders within the context of Western music, where performances do not necessarily adhere to the modern 440 Hz standard tuning. As a key contribution, we evaluate several recent neural vocoders on datasets containing piano, violin, and singing voice recordings. Our results reveal that different vocoders exhibit distinct biases, causing deviation in tuning, and affecting waveform reconstruction quality in case of non-standard tuning. Our work underscores the need for improved vocoder robustness in music synthesis and provides insights for refining future models.

1. INTRODUCTION

Recent advances in speech and music synthesis often follow a two-stage approach: An initial acoustic model generates an intermediate spectral representation, from which a second model, frequently referred to as a vocoder, reconstructs a time-domain waveform [1–4]. A common choice for this intermediate representation is a mel-spectrogram. While traditional signal-processing methods can reconstruct waveforms from mel-spectrograms, their quality depends on the spectral dimensionality. Recent deep learning-based generative models, such as Generative Adversarial Networks (GANs) [5] or Diffusion models [6], achieve high-fidelity reconstruction even from low-dimensional representations. Although historically used only in speech transmission, the term "Vocoder" has recently been adopted for general spectrogram-to-audio models [7, 8].

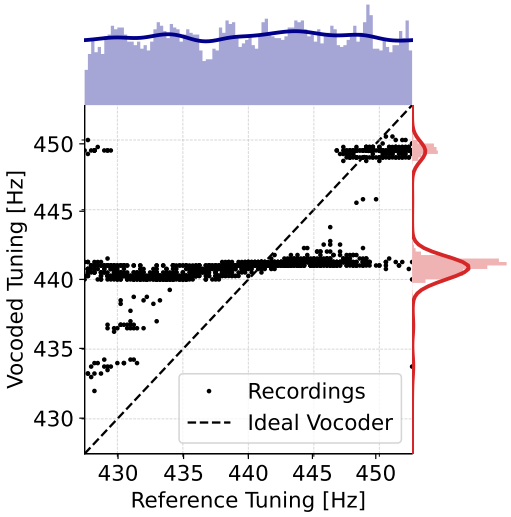


Figure 1: Scatter plot of estimated tuning from vocoded recordings (vertical axis, vocoder from [1]) compared to original audio tuning (horizontal axis). Marginal distributions are shown as histogram plots with a Gaussian kernel density estimation (red and blue line). For an ideal vocoder, the blue and red distribution would align.

This two-stage approach factorizes the problem of audio synthesis, enabling different modeling techniques for each stage. Vocoders can be trained in a self-supervised manner on large amounts of data, enabling researchers in music or speech synthesis to rely on pre-trained vocoders. However, vocoders may introduce artifacts or biases, particularly in case of domain-shift, such as unseen musical instruments. While recent works in music synthesis have moved towards more musically informed spectrogram generation models—such as enabling fine-grained control over timbre and pitch [2, 3]—the impact of vocoders on the musical characteristics of the output remains underexplored. In particular, one musically important but often overlooked aspect is the influence that different musical tunings have on signal reconstruction.

Tuning plays a fundamental role in music and varies across styles and traditions. In Western music, tuning typically refers to the frequency of a reference pitch, from which the frequencies of all other pitches can be derived. While A4 = 440 Hz is the modern standard tuning [9], real-world recordings often exhibit deviations due to historical reasons, or artistic choices [10].

This paper aims to explore how musical tuning affects vocoder performance. A key contribution is our analysis of tuning preservation during waveform reconstruction from mel-spectrograms of real recordings, revealing systematic biases in certain vocoders. Our evaluation compares multiple neural vocoders and a signal-processing baseline across three diverse datasets: piano, violin, and singing with piano accompaniment. Figure 1 highlights one of our findings, showing how a specific vocoder introduces tuning bias, leading to a mismatch between the tuning distributions of the original and reconstructed recordings. As a further contribution, we conduct a listening test to assess how non-standard tuning affects the perceived quality of vocoded audio.

2. BACKGROUND

2.1 Mel-Spectrogram Inversion

Mel-spectrogram computation involves two main stages, both of which can lead to information loss. First, the magnitude short-time Fourier transform (STFT) is computed, discarding phase information. Second, a mel filter bank is applied to the magnitude-STFT, typically reducing frequency resolution.

A signal processing-based approach to mel-spectrogram inversion is to first estimate the magnitude-STFT, often via a pseudo-inverse, using Non-Negative Least Squares (NNLS) [11], and then reconstruct the waveform by estimating the phase, typically using the Griffin-Lim algorithm [12]. The quality of the reconstructed waveform depends heavily on the spectral resolution of both the STFT and mel-spectrogram.

In contrast, neural vocoders are able to synthesize high-quality audio from mel-spectrograms with a lower spectral dimensionality, making them more practical for audio synthesis. Early neural vocoders focused on speech and often failed to generalize to unseen domains such as new speakers or musical instruments. More recently, "universal" neural vocoders have emerged that can robustly handle diverse audio sources, including complex musical signals [5, 13].

For example, Hawthorne et al. [1] train a GAN-based vocoder on 16,000 hours of music data, building upon SoundStream [14] and SEANet [15]. This vocoder is widely used in music synthesis [1–3, 16, 17]. Similarly, BigVGAN [5], originally trained on speech data, has been extended by BigVGAN-V2 [1] with a broader training set including music and environmental sounds, enabling more robust performance across domains. Despite their impressive audio quality, neural vocoders are sensitive to their training data. As a result, they may struggle with out-of-distribution inputs, such as unfamiliar instruments or non-standard musical tunings.

Previous studies have found that many mel spectrogram inversion models produce waveforms with locally unstable pitch when applied to music [18, 19]. However, our study takes a broader perspective by examining tuning as a global statistic over entire recordings, distinguishing it from local pitch fluctuations, as discussed in the next section.

[1] https://github.com/NVIDIA/BigVGAN

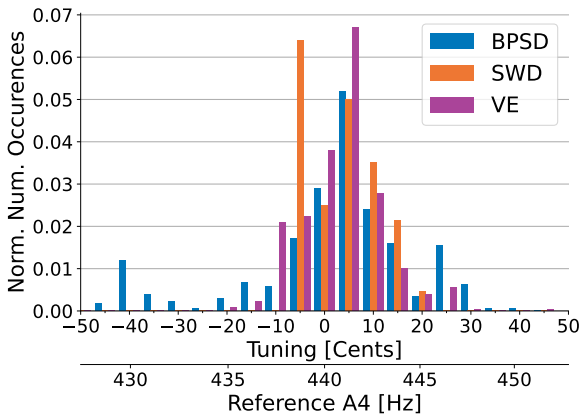


Figure 2: Distribution of tuning (5-Cent resolution) per recording for the three investigated datasets.

2.2 Tuning and Tuning Estimation

In the context of the 12-tone equal temperament system in Western music, tuning can be characterized by the frequency of a given reference pitch, often called *concert pitch*. The modern tuning standard was established in 1975 in ISO16 defining the reference pitch to be 440 Hz for the note A4 [9]. However, this has been subject to change over time, and even today, it is by no means a universally applied standard. To illustrate this, Figure 2 shows the distribution of tuning values per recording for the three datasets used in this paper, which we introduce in Section 3. All three tuning distributions peak around 440 Hz, with a slight tendency toward higher values. However, we can also see considerable variations for some recordings, going as low as 430 Hz. Studies have found that only 50% of Western classical music recordings fall into the tuning range 440–443 Hz [20], underscoring the natural diversity in musical tuning. Qin and Lerch [21] found that tuning can be a confounding variable for music classification algorithms, highlighting the potential impact of tuning on deep-learning-based models.

Tuning can also be expressed as deviation in Cents from A4 = 440 Hz, where one semitone equals 100 Cents, as shown by the two x-axes in Figure 2. In this context, tuning estimation is the task of finding the concert pitch frequency, or equivalently, the deviation of the standard 440 Hz pitch. Due to the importance of tuning in music, many different approaches have been developed for tuning estimation of full performances [20, 22–24]. Most approaches define tuning as a *circular* offset from a reference pitch (typically A4 = 440 Hz) within ± 50 Cents since a deviation of more than ± 50 Cents is indistinguishable from a transposition to the next semitone. For example, an A that is 60 Cents flat (lower) is indistinguishable from a G# that is 40 Cents sharp (higher). For recordings deviating by more than ± 50 Cents, the estimated tuning therefore "wraps around" to the opposite side. To ensure a robust and

unbiased evaluation, we employ two independent tuning estimation methods. This redundancy allows us to cross-validate results and account for potential inaccuracies or method-specific biases in tuning estimation. Both methods operate with a 1-Cent resolution.

The first tuning estimation method, implemented in the LibROSA package [25], follows a two-stage process. First, an STFT is computed, and frequency peaks are identified and refined using parabolic interpolation as described in [26]. In the second stage, a frequency histogram corresponding to tuning values is constructed by mapping the interpolated frequencies to the range ± 50 Cents using a modulo operation. The tuning value with the highest count in the histogram is then selected. We denote this approach `FreqHist`. The second tuning estimation method is implemented in the LibFMP package [27] and differs mostly in the first stage. Rather than frequency interpolation, an STFT with a large window size is used to obtain the necessary frequency resolution. The STFT is averaged over time and the resulting frequency is converted to a Cent-scale with 1-Cent resolution using cubic interpolation. The resulting distribution is compared with a set of comb-like template vectors, each representing a specific tuning within ± 50 Cents. The final estimate is given by the template that maximizes the correlation with the distribution. We denote this technique `TempMatch`.

3. EXPERIMENTAL SETUP

3.1 Datasets

We use three datasets with distinct instrumentation, including piano, singing, and violin. **The Beethoven Piano Sonata Dataset (BPSD)** [28] consists of 11 versions of all first movements of Beethoven's Piano Sonatas, totaling 352 recordings, and approximately 40 hours. We choose a piano dataset because the discrete pitch set and stable tuning throughout a piece enable a robust tuning estimate. The BPSD in particular is well-suited for our evaluation for two reasons: First, it contains diverse recordings spanning nearly 90 years (1935—2022) from different performers, acoustic conditions, and pianos. Second, as shown in Figure 2, the dataset exhibits a wide range of tunings. **The Schubert Winterreise Dataset (SWD)** [29] contains nine complete recordings of the "Winterreise" song cycle for singing voice and piano, by nine different performers, totaling approximately 10.5 hours. Unlike the piano, singing voice has a continuous pitch range and the tuning is less stable, making it a valuable addition to our experiments. However, the piano in the SWD provides a stabilizing reference for the voice. The **Violin Etudes (VE) dataset** [30] consists of 925 monophonic violin recordings (approximately 28 hours) from YouTube. Unlike piano, violin tuning estimation can be less reliable due to continuous pitch variation. To ensure robust evaluation, we filter out recordings where the two tuning estimation methods disagree by more than 5 Cents, resulting in a final selection of 651 recordings.

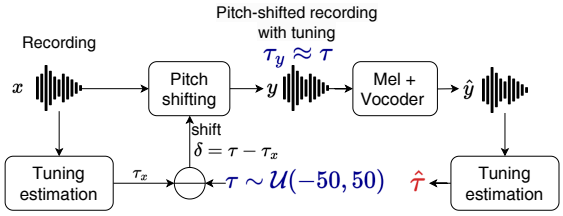


Figure 3: Experimental setup for a single recording x.

3.2 Pitch-shift Augmentation and Vocoding

Since our datasets do not cover the full range of tuning values with a sufficient number of recordings—shown in Figure 2—we use pitch shift augmentation to create a new version of our datasets with a uniform distribution in the tuning space similar to [21], using the Rubber Band Library. [2] Figure 3 shows the experimental setup of the applied pitch shift augmentation. For a given recording x, we estimate the original tuning τ_x using the `FreqHist` estimator. We sample a target tuning $\tau \sim \mathcal{U}(-50, 50)$ and pitch shift x by the difference $\delta = \tau - \tau_x$. This yields a modified recording y with a tuning of $\tau_y = \tau$ (equality holds up to the tuning estimation error). For each recording in our datasets, we generate four pitch-shifted versions, which are subsequently downsampled to 16 kHz. While pitch shifting may introduce minor artifacts, we argue these affect perceived quality but not tuning estimation accuracy.

Next, we calculate a mel-spectrogram from y and subsequently reconstruct the time-domain signal using a vocoder, producing the output $\hat{y}$. We refer to this process as *vocoding* y. The parameters of the mel-spectrogram are always chosen to fit the given vocoder, and after vocoding, each signal is downsampled back to 16 kHz. The tuning of $\hat{y}$ is then estimated, yielding $\hat{\tau}$. By comparing $\hat{\tau}$ with τ, we assess the vocoder's ability to preserve tuning.

3.3 Quantitative Metrics Tuning Preservation

We introduce two metrics to evaluate tuning preservation. A straightforward approach would be to compute the difference $\hat{\tau} - \tau$. However, since our tuning estimation algorithms are circular, large errors can arise from semitone confusion. For example, if a vocoder is applied to a signal with a tuning of $\tau = 45$ Cents and it raises the tuning by 10 Cents, the estimation would return $\hat{\tau} = -45$ Cents (equivalent to +55 Cents). A simple difference would then yield $\hat{\tau} - \tau = -90$ Cents, even though the vocoder only changed the tuning by 10 Cents in this case.

To address this, we introduce a *circular difference*, considering tuning estimates on a circle where $\tau = 50$ and $\tau = -50$ are equivalent. Formally, we define the circular difference between two estimates τ_1 and τ_2 as:

$$\delta_{\text{circ}} = \begin{cases} \delta + 100, & \text{if } \delta < -50 \\ \delta - 100, & \text{if } \delta > 50 \\ \delta & \text{otherwise} \end{cases} \quad (1)$$

[2] `https://github.com/breakfastquay/rubberband`

Vocoder	Short Name	Training Data	# Param.	F_s	# Mel Bands	STFT Win. Len.	Hop Length
Hawthorne et al. [1]	HAWT	Music	15M	16 kHz	128	640	320
BigVGAN [5]	BV	Speech	112M	[22, 24] kHz	[80, 100]	1024	256
BigVGAN-V2 [31]	BV2	Music, Speech, ES	112M	[22, 24, 44] kHz	[80, 100, 128]	1024	256
NNLS & GL [12, 25]	LSGL	—	—	16 kHz	[100, 128, 150]	640	320

Table 1: Overview over investigated vocoders. For vocoders with multiple versions, we show lists with parameters for each version (in order). For example, BigVGAN-V2 with 128 mel bands has a sampling frequency of 44 kHz. "ES" stands for environmental sounds, "NNLS&GL" for Non-Negative Least Squares & Griffin-Lim.

where $\delta = \tau_2 - \tau_1$. This guarantees $\delta_{\text{circ}} \in [-50, 50]$. While the circular difference captures the deviation between τ and $\hat{\tau}$, it does not provide insight into the statistical distribution of $\hat{\tau}$. By comparing the distributions of τ and $\hat{\tau}$ (shown in blue and red for the example in Figure 1, respectively), we quantify how strongly the tuning distribution of the vocoded audio deviates from the input distribution. To this end, we use the Wasserstein Distance (also referred to as the earth mover's distance, or EMD), which is the optimal transport cost between two probability distributions [32]. A lower Wasserstein Distance indicates greater similarity. In particular, given the circular nature of tuning estimation, we compute the *Circular Wasserstein Distance* (CWD), as described in [33]. Thus, in this optimal transport problem, probability mass can flow across the boundaries of the estimation range, as both ends are connected on the considered circle.

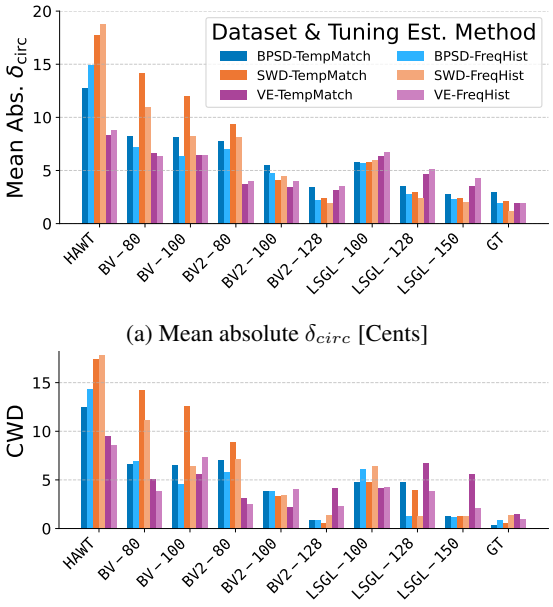
3.4 Vocoders

In Section 2.1, we briefly introduced the main vocoder architectures investigated in this work, for which we will use the following shorthand notations throughout the remainder of the paper: Hawthorne et al. [1] (HAWT), BigVGAN [5] (BV), BigVGAN-V2 [31] (BV2), and the signal-processing-based approach of NNLS [11] followed by Griffin-Lim [12] (LSGL). Note that the vocoder by Hawthorne et al. is ambiguously also referred to as *SoundStream* in the literature [2, 3].
Table 1 gives an overview of the investigated vocoders. HAWT has only a single version, with 128 mel bands. Multiple versions exist of BV and BV2, which differ in the number of mel bands and sampling frequency F_s. For LSGL we use three different numbers of mel bands, with the same underlying STFT properties. In our results, we identify each vocoder by its short name and the number of mel bands. For instance, BV-80 refers to the BigVGAN model with 80 mel bands and a 22kHz sampling rate. This naming convention is unambiguous, as variants of the same vocoder with different sampling rates also have distinct numbers of mel bands.

4. RESULTS TUNING PRESERVATION

4.1 Quantitative Results

Figure 4a presents the mean absolute δ_{circ} for all tested vocoders, where distinct colors represent the datasets, and the color shade indicates the tuning estimator. In addition to the tested vocoders, we include metrics for ground truth



(a) Mean absolute δ_{circ} [Cents]

(b) Circular Wasserstein Distance (CWD)

Figure 4: Evaluation metrics for each vocoder, dataset, and tuning estimation method.

audio GT in the figure, where we compare the tuning estimate τ_y of the pitch-shifted audio with the target tuning τ. For GT, we can see that tuning estimation aligns with pitch-shifting, meaning the estimated tuning of a pitch-shifted version differs on average not more than 2 Cents from the target tuning.
As a first and central observation, we see that most neural vocoders introduce tuning deviation, whereas the signal-processing-based LSGL generally shows a lower circular difference. However, BV2-100 and BV2-128 are an exception to this. We also observe that a higher number of mel bands for LSGL and BV2 leads to less tuning deviation, reaching values only marginally above the tuning estimation inconsistencies for a high number of mel bands. When comparing neural vocoders, HAWT exhibits the overall highest circular difference, reaching values up to $\delta_{\text{circ}} = 18.7$ Cents. BV2 shows lower values on average compared to the original BV. From further observing Figure 4a, the different datasets seem to have an impact on the tuning preservation for some vocoders. For instance, BV and HAWT show notably higher tuning deviations for SWD compared to other datasets. The VE dataset is least affected by tuning deviations for all neural vocoders, possibly due to its continuous pitch nature, since it includes

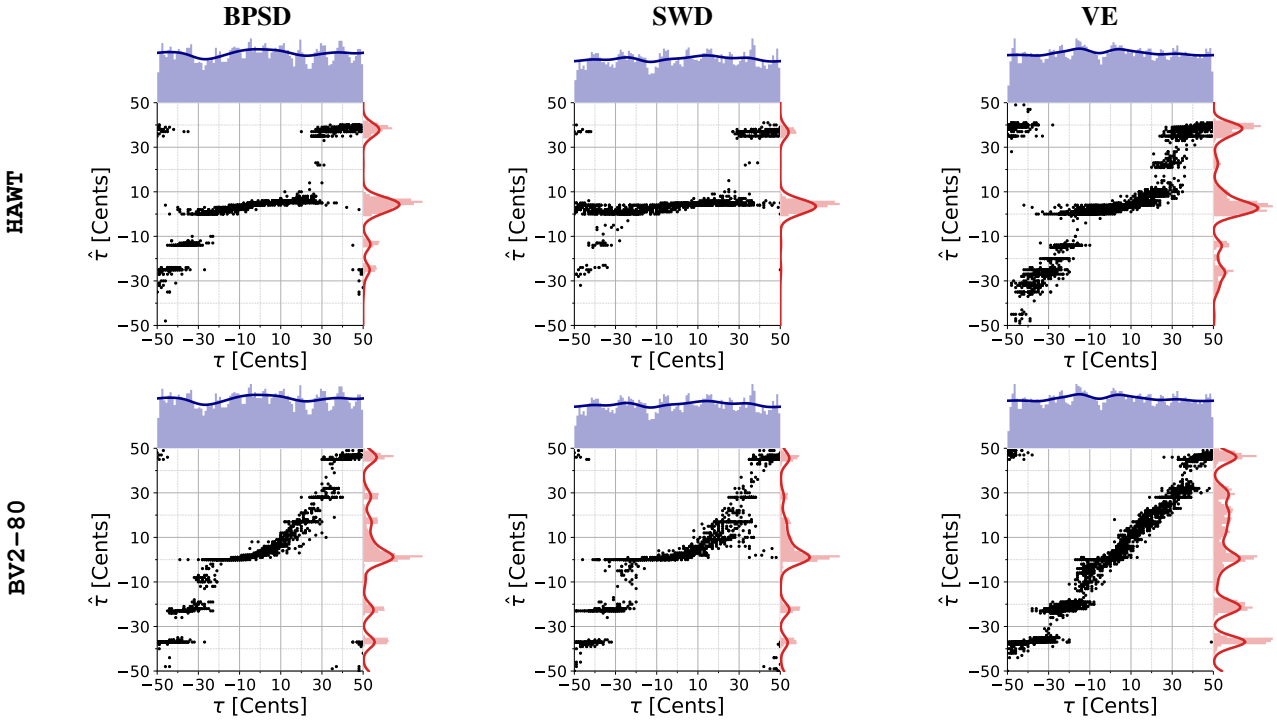
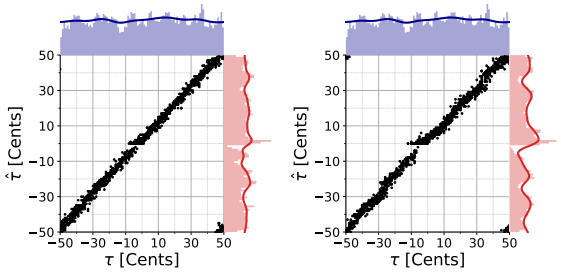


Figure 5: Tuning estimates of vocoded audio ($\hat{\tau}$) over input audio (τ) on all three datasets for vocoders `HAWT` and `BV2-80`. Marginal distributions are shown as histograms with a moving average smoothing of width five applied and a Gaussian kernel density estimate (line). Tunings were estimated with `TempMatch` estimator.

violin only. However, this dataset seems to be difficult for `LSGL`. Comparing the results for the two tuning estimation methods, we observe similar trends despite some variations for specific vocoder-dataset combinations.

Figure 4b shows the CWD for all vocoders, datasets, and tuning estimation methods. We observe a strong correlation with δ_{circ} from Figure 4a. A high CWD indicates that the tuning of the vocoder output follows a distribution different from that of the uniformly distributed pitch-shift augmented datasets, suggesting that, in general, when a vocoder introduces tuning deviations, these deviations follow a non-uniform distribution.

4.2 Qualitative Results



(a) `BV2-128` (b) `LSGL-150`

Figure 6: Tuning estimates of vocoded audio ($\hat{\tau}$) over input audio (τ) for `BV2-128` and `LSGL-150` on the SWD.

To better understand the tuning changes introduced by the vocoders, we analyze two example vocoders in more detail: `HAWT` and `BV2-80`, as these are among the most commonly used for music data in the literature [1–3, 17, 34, 35]. Figure 5 shows a scatter plot of all vocoded tunings $\hat{\tau}$ over input tunings τ for all three datasets, alongside marginal distributions with a Gaussian kernel density estimation (KDE). Note that we account for the circular continuity of the tuning estimates to calculate the KDE.

As indicated by our quantitative analysis, the output distribution of $\hat{\tau}$ is heavily altered from the uniform input distribution in almost all cases. For both vocoders, a cluster around $\hat{\tau} = 0$ is evident, corresponding to $\mathrm{A4} = 440\,\mathrm{Hz}$, though the actual peak is slightly above $\hat{\tau} = 0$. This could reflect a bias in the training data, as many recordings are tuned between 440 and 443 Hz, as discussed in Section 2.2. This effect is similar for the BPSD and the SWD, but less pronounced for the VE. Figure 6 shows the same scatter plot for `BV2-128` and `LSGL-150` on SWD, where output tuning closely follows input tuning, consistent with the low tuning deviation metrics discussed earlier. The bias observed in Figure 5 appears to vanish for the higher resolution `BV2` model (which has more mel bands and higher sampling rate). This is potentially due to the reduced information loss in the mel spectrograms, making accurate reconstruction an easier task. Additional figures for all datasets, vocoders, and tuning estimator combinations are available on our website. [3]

[3] https://www.audiolabs-erlangen.de/resources/2025-ISMIR-VocoderTuningEstimation

5. LISTENING TEST

While we showed that vocoders may not preserve tuning, we did not examine whether this affects perceived quality for non-standard tuning. In principle, tuning deviation and output quality could be orthogonal; a vocoder might alter tuning yet still produce high-quality audio.

To investigate this question, we conducted a listening test, focusing on the BPSD due to its diverse original tuning. Instead of comparing samples from different vocoders, we compare only vocoded samples of the same excerpt under different tunings from a single vocoder. This goal introduces two challenges for the listening test design. First, we require test items with identical original quality but different tunings. A potential solution is pitch shifting, which can robustly replicate a specific tuning (Section 4.1, Figure 4a), but here, the introduction of small artifacts might impact the perceived quality. Second, the commonly used MUSHRA test [36] is unsuitable for comparing items that differ in tuning from the reference, as pitch differences would directly influence the listeners' judgment. We address both issues in our test design.

The test follows an AB format without a reference: Participants compare two excerpts and choose the one with the better-perceived quality, having a "no preference" option as well. As a first step, we select four BPSD recordings with diverse original tunings of -42, -11, 0, and 34 Cents and pitch-shift each one three times to replicate the other three tunings, yielding 16 test items (four per tuning).

In order to analyze tuning bias within each vocoder, we do not compare across vocoders but rather select one set of item pairs that is then vocoded and presented independently for each vocoder. Each original (non-pitch-shifted) recording is paired with its three pitch-shifted versions, ensuring each tuning is tested against the original. For example, the item with an original tuning of -42 Cents is paired with its pitch-shifted versions with tunings -11, 0, and 34 Cents. In this example, if a listener prefers the 0-Cent tuning over the original -42-Cent tuning, even though the 0-Cent was obtained through pitch-shifting, this can indicate that tuning affects quality stronger than pitch shifting. In total, this yields 12 item pairs per vocoder, and we test four vocoders: `HAWT`, `BV2-80`, `BV2-128`, and `LSGL-150`. We split the items into two separate listening tests with 24 pairs each, to limit the test duration per listener. Additionally, we include four control pairs for both subgroups with identical items, where attentive listeners should indicate "no preference". We exclude listeners who indicate more than once a preference for control item pairs.

5.1 Results

In total, 25 participants took part in our listening test, 19 male and 6 female, with a median age of 27, ranging from 21 to 58. Among them, 20 participants had some prior experience with listening tests. A total of 5 listeners did not meet the post-screening criterion, leaving 20 listeners distributed evenly among the two item subsets. For each tuning value, we aggregate the number of times it was preferred. If tuning had no impact on quality, we would expect either a tendency towards "no preference" votes, or an approximately equal distribution of preferences across tuning values.

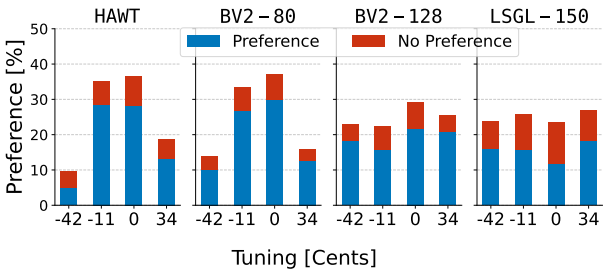


Figure 7: Listening test results: listeners' preference towards tuning values as a percentage of total votes for each vocoder individually. "No preference" votes are split evenly between both excerpts in a pair, e.g., a vote between 0 and 11 counts as half a vote in the red bar for each, meaning that bars sum up to 100% for each vocoder.

Figure 7 illustrates this percentage of preference votes for each tuning value and tested vocoder. For `HAWT` and `BV2-80` we can see a trend: The tunings -42 and $+34$ Cents receive fewer preference votes compared to the middle values. In contrast, `BV2-128` does not exhibit a strong trend, while `LSGL-150` shows a generally lower number of preference votes, suggesting that listeners perceived fewer quality differences compared to the neural vocoders. When aggregating preferences into groups of "original" and "pitch-shifted", listeners show a slight preference towards the original items for the neural vocoders, indicating that pitch-shifting also has a negative influence on quality (see supplementary website). Therefore, fully disentangling the effects of pitch shifting and tuning in the listening test remains challenging. However, due to our test design, always comparing the vocoded original recordings with their vocoded pitch-shifted counterpart, Figure 7 still shows a meaningful trend.

Overall, the results indicate that vocoders which show a bias in tuning preservation (`HAWT` and `BV2-80`) also show a decrease in quality when reconstructing signals with out-of-distribution tuning.

6. CONCLUSIONS

In this study, we investigated how musical tuning affects neural vocoders, focusing mainly on tuning preservation. Our findings reveal that vocoders can significantly alter both individual tunings and overall tuning distributions, with some exhibiting a bias towards modern standard tuning. Additionally, our listening test suggests a decline in reconstruction quality for signals with non-standard tunings when processed by a vocoder with tuning bias.

Our work underscores the importance of tuning in music generation and vocoder design. Future work should focus on mitigating tuning biases during vocoder training. Our evaluation approach, based on pitch shifting and quantitative evaluation metrics, gives researchers a straightforward yet effective method for assessing tuning robustness in music vocoders.

7. ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 350953655 (MU 2686/11-2) and Grant No. 500643750 (MU 2686/15-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

8. REFERENCES

[1] C. Hawthorne, I. Simon, A. Roberts, N. Zeghidour, J. Gardner, E. Manilow, and J. H. Engel, "Multi-instrument music synthesis with spectrogram diffusion," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 598–607.

[2] B. Maman, J. Zeitler, M. Müller, and A. H. Bermano, "Performance conditioning for diffusion-based multi-instrument music synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, South Korea, 2024, pp. 5045–5049.

[3] D. Kim, H.-W. Dong, and D. Jeong, "Violindiff: Enhancing expressive violin synthesis with pitch bend conditioning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1–5.

[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R.-S. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 4779–4783.

[5] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.

[6] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *Proceedings of the International Conference on Learning Representations, ICLR*, Virtual, 2021.

[7] H. Dudley, "Remaking Speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[8] A. Mustafa, N. Pia, and G. Fuchs, "StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Canada, 2021.

[9] ISO, "Acoustics – standard tuning frequency (standard musical pitch)," *ISO16:1975*, 1975.

[10] F. Gribenski, *Tuning the World: The Rise of 440 Hertz in Music, Science, and Politics, 1859–1955*, ser. New Material Histories of Music. University of Chicago Press, 2023. [Online]. Available: https://books.google.de/books?id=VJKpEAAAQBAJ

[11] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. Society for Industrial and Applied Mathematics, 1995.

[12] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[13] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., virtual, 2020.

[14] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 30, pp. 495–507, 2022.

[15] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "SEAnet: A multi-modal speech enhancement network," in *Proceedings of the Annual Conference of the International Speech Communication Association, (Interspeech)*, H. Meng, B. Xu, and T. F. Zheng, Eds., Shanghai, China, 2020, pp. 1126–1130.

[16] H. Kim, S. Choi, and J. Nam, "Expressive acoustic guitar sound synthesis with an instrument-specific input representation and diffusion outpainting," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Seoul, Republic of Korea, 2024, pp. 7620–7624.

[17] B. Maman, J. Zeitler, M. Müller, and A. H. Bermano, "Multi-aspect conditioning for diffusion-based music synthesis: Enhancing realism and acoustic control," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 68–81, 2025.

[18] B. D. Giorgi, M. Levy, and R. Sharp, "Mel spectrogram inversion with stable pitch," in *Proceedings of the International Society for Music Information Retrieval Conference, ISMIR*, Bengaluru, India, 2022, pp. 233–239.

[19] J. H. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," in *International Conference on Learning Representations, ICLR*, New Orleans, LA, USA, 2019.

[20] A. Lerch, "On the requirement of automatic tuning frequency estimation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Victoria, Canada, 2006, pp. 212–215.

[21] Y. Qin and A. Lerch, "Tuning frequency dependency in music classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 401–405.

[22] A. Degani, M. Dalai, R. Leonardi, and P. Migliorati, "Comparison of tuning frequency estimation methods," *Multimedia Tools and Applications*, vol. 74, no. 15, pp. 5917–5934, Aug. 2015.

[23] K. Dressler and S. Streich, "Tuning frequency estimation using circular statistics," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, 2007, pp. 357–360.

[24] V. Gnann, M. Kitza, J. Becker, and M. Spiertz, "Least-squares local tuning frequency estimation for choir music," in *Proceedings of the Audio Engineering Society (AES) Convention*, New York City, New York, USA, 2011.

[25] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proceedings the Python Science Conference*, Austin, Texas, USA, 2015, pp. 18–25.

[26] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music." *Journal of the Acoustical Society of America (JASA)*, vol. 111, no. 4, pp. 1917–1930, 2002.

[27] M. Müller and F. Zalkow, "libfmp: A Python package for fundamentals of music processing," *Journal of Open Source Software (JOSS)*, vol. 6, no. 63, pp. 3326:1–5, 2021.

[28] J. Zeitler, C. Weiß, V. Arifi-Müller, and M. Müller, "BPSD: A coherent multi-version dataset for analyzing the first movements of beethoven's piano sonatas," *Transaction of the International Society of Music Information Retrieval*, vol. 7, no. 1, pp. 195–212, 2024.

[29] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, and H. Grohganz, "Schubert Winterreise dataset: A multimodal scenario for music analysis," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 14, no. 2, pp. 25:1–18, 2021.

[30] N. C. Tamer, P. Ramoneda, and X. Serra, "Violin etudes: A comprehensive dataset for f0 estimation and performance analysis," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 2022, pp. 517–524.

[31] S.-G. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN GitHub repository," 2024. [Online]. Available: https://github.com/NVIDIA/BigVGAN

[32] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Foundations and Trends in Machine Learning*, vol. 11, no. 5–6, pp. 355–607, 2019.

[33] J. Delon, J. Salomon, and A. Sobolevski, "Fast Transport Optimization for Monge Costs on the Circle," *Society for Industrial and Applied Mathematics Journal on Applied Mathematics*, vol. 70, no. 7, pp. 2239–2258, 2010.

[34] H. Kim, S. Choi, and J. Nam, "Expressive acoustic guitar sound synthesis with an instrument-specific input representation and diffusion outpainting," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7620–7624.

[35] S. Dai, M.-Y. Liu, R. Valle, and S. Gururani, "Expressivesinger: Multilingual and multi-style score-based singing voice synthesis with expressive performance control," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3229–3238.

[36] International Telecommunications Union, "ITU-R Rec. BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems," 2015.