Low-latency Pitch-Shifting with STN decomposition

Maria Polak[1] , Cumhur Erkut[1]
[1] Aalborg University, Copenhagen, Denmark, Email: mpolak23@student.aau.dk

Introduction

Pitch shifting is an effect that modifies the pitch of the input signal while preserving the original tempo and duration. The effect is commonly used for correcting the detuned pitches of a singer, harmonizing voices, key changing, or for creative exploration in sound design and DJing. Manipulating the pitch makes it possible to dramatically change the character of the audio, shape its timber or explore unique sounds. In the light of this, the effect is extensively used both in offline and real-time processing. While offline solutions allow for high-quality pitch modifications on audio recordings, online solutions are still imperfect, often introducing audible artifacts.

A major challenge in pitch-shifting algorithms is the lack of a universal solution that can effectively handle all types of audio and musical content, especially in online processing where latency-quality trade-off plays an important role. Over the years, many algorithms and approaches have been proposed, each suited for a slightly different use case. The whole category of time-domain methods is, in general, suitable for low-latency processing of monophonic signals. On the other hand, frequency-domain (FD) methods, which for years had been considered computationally heavy, can achieve higher quality at the cost of higher latency. Unfortunately, higher quality does not imply flawless. FD methods often suffer from transient smearing and phasiness caused by Fourier Transform nature.

This paper includes fuzzy STN decomposition into the audio processing pipeline to tackle the aforementioned challenges. Separating audio streams allows for using specialized pitch-shifting algorithms tailored to each stream's characteristics. The main contributions of the paper are:

- Fuzzy STN and Noise Morphing in pitch-shifting
- Online implementation of both algorithms

It is important to note that the following chapters often reference work on time scale modification (TSM or time-stretching). This is due to the fact that both problems can be easily mapped to one another. Time-stretching is an effect, reciprocal to pitch shifting, which changes the tempo and duration of the input signal while preserving the audio's pitch. Consequently, pitch-shifting can be viewed as time-stretching and then resampling.

Related Work

The following subsections talk in detail about the works which were used as main references for this project.

Fuzzy STN: Fuzzy STN decomposition is a three-way audio separation method proposed by Fierro and Välimäki [3]. It is a two-stage algorithm which firstly separates sines from the residual component, and then docomposes the residual component into transients and noise. The separation itself builds upon the observation that, in the spectrogram, sinusoids form flat lines in the time direction and impulsive events form lines in the frequency direction. This information is extracted by applying horizontal and vertical median filters on the magnitude spectrum. Resulting spectrograms (X_h and X_v) are later used to calculate tonal-ness R_s and transient-ness R_t as presented in eq (1).

$$R_s = \frac{X_h}{X_h + X_v}, \qquad R_t = 1 - R_s = \frac{X_v}{X_h + X_v} \tag{1}$$

Tonal-ness and transient-ness are used to calculate fuzzy spectral masks S, T, and N as follows:

$$S = f(R_s), \qquad T = f(R_t), \qquad N = 1 - S - T$$

where

$$f(x) = \begin{cases} 1, & \text{if } x \geq \beta_U, \\ \sin^2(\frac{\pi}{2}\frac{x-\beta_L}{\beta_U-\beta_L}), & \text{if } \beta_L \leq x \leq \beta_U, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

and β_U, β_L are control parameters set to values proposed in the original work. Fuzzy masks, while ensuring perfect reconstruction, allow each spectral bin to simultaneously be part of two classes Sines+Noise or Transients+Noise, which yields a natural, SOTA quality separation.

Noise Morphing (NM): Method introduced by Moliner et al. [5] that greatly enhances the quality of the noise component time-stretching. NM is a FD method that applies random phases while maintaining a magnitude equal to that of the original audio. The algorithm consists of two steps - analysis and synthesis. In the analysis step, a log-magnitude spectrum $\mathrm{N}_{orig} \in \mathbb{R}^{M \times K}$ of the original noise component is calculated as

$$\mathrm{N}_{orig} = 10 \log_{10}(|STFT(n_{orig}|) \tag{3}$$

The resulting frames are then linearly interpolated to match the desired length according to factor α, using the two spectra surrounding the interpolation point, according to

$$\mathrm{N}^{\alpha} = \mathrm{lerp}(\mathrm{N}_{orig}, \alpha) \tag{4}$$

where lerp() is the linear interpolation function.

In the synthesis step, a white noise signal of the desired length is generated, transformed to the frequency domain, and then its magnitudes are normalized. Next,

the resulting frames are modulated by one another via element-wise multiplication:

$$\mathrm{N}^{\alpha}_{\mathrm{NM}} = E \odot 10^{\mathrm{N}^{\alpha}/10} \tag{5}$$

This method was found highly effective in time stretching by extreme factors reaching SOTA results.

"Vase-Phocoder": An algorithm introduced by Geraint Luff at ADC22 conference[1]. The idea behind the algorithm is rooted in Phase Vocoder (PV), but unlike the traditional PV, it introduces a multi-step prediction method for phase changes. These predictions include horizontal (time-based) and vertical (frequency-based) relationships within the STFT representation. The first iteration makes an initial prediction based solely on horizontal phase-observations, and then the second iteration uses that result to make vertical predictions, ending up with a design that resembles predicting with diagonal phase-observations. This technique helps to maintain clarity and reduces de-tuning effects. [9]

Methods

This work adapts findings from the Moliner's paper [5] to implement a polyphonic low-latency pitch-shifting method. The algorithm was adapted to operate within real-time processing context and implemented as a VST audio plug-in.

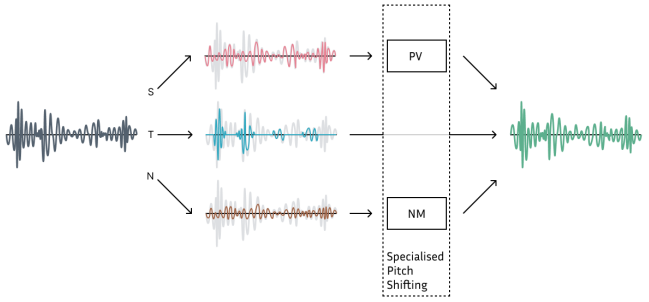


Figure 1: System overview

Algorithm

The core idea behind this project is to incorporate STN decomposition into the pitch-shifting pipeline and then apply specialized processing to each of the extracted components. Separated audio streams, namely sines, transients and noise, are processed in the following ways:

Sines: From the definition of the STN decomposition, the extracted sines component is transient-less. This removes the main cause of audible artifacts in PV-based processing and makes it a fitting technique for pitch-shifting. PV-based processing was also found successful in previous studies [5]. The sines are shifted by "Vase-Phocoder" provided by the Signalsmith library.

Transients: In the related TSM studies, transients are usually repositioned in time to match the TSM factor [5]. In pitch-shifting no additional processing is needed as transients carry no pitch information. Similarly to work of Knesebeck [7] transients get reapplied after sines and noise processing, ensuring their preservation.

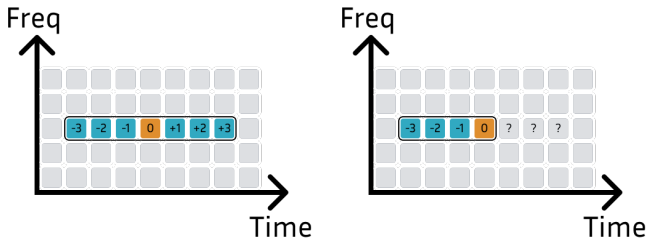
Noise: The noise component is pitch-shifted through a two-step process: time scaling followed by sampling rate conversion. First, TSM is performed by NM from [5]. Then, the resulting signal is resampled back to the original sample rate yielding a pitch-shifted version of the input.

Real-Time

Algorithms operating in the FD always introduce delays. Since all main algorithms in the processing pipeline (STN, NM, and "Vase-Phocoder") are FD-based, latency was a major consideration. The acceptable latency was derived from two commercial pitch-shifting audio plug-ins, which introduce delays of 6946 (157.5ms) and 7170 (162.6ms) samples at a 44.1 kHz sample rate, respectively. The final latency of the developed solution is at most 7325 (166.1ms) samples when using the highest FFT size setting.

Porting the internal algorithms from offline to online processing required two modifications due to the lack of future knowledge in online processing.

The first change is in the STN decomposition, specifically in the horizontal median filters. These filters are applied to the STFT frames in both stages of the decomposition algorithm to filter spectral content on the time axis. By default, median filters are centered around the currently processed point, implying that the filter looks at $filterlength/2$ frames forward in time. For the plug-in implementation, the horizontal median filters were modified to only consider the frames available at the time of calculation—previous frames, as presented in Fig 2.



(a) Original **(b)** Real-Time

Figure 2: Horizontal Medial Filter Modification

The second modification involves frame interpolation in the NM algorithm. The challenge was handling time-stretching by a fractional factor. The offline implementation assumed re-indexing of the original frames and then interpolating to fill the missing indices, which is not feasible in frame-by-frame processing. Instead, the real-time implementation always generates a k number of evenly spaced frames between the previous and currently processed frame:

[1] Geraint Luff ADC22 - `https://youtu.be/fJUmmcGKZMI?si=qDBEBHEfGeEAUKop`

$$k = \lceil \alpha \rceil = \left\lceil 2^{st/12} \right\rceil \tag{6}$$

where st is the number of semitones of pitch shift. To achieve the correct length of a time-stretched signal, the STFT overlap R_s is adjusted according to α:

$$R_s = \left\lceil \frac{R_{org} * \alpha}{k} \right\rceil \tag{7}$$

Finally, a gain correction is applied to compensate for the changing overlap factor.

Evaluation

The proposed solution (PS) was perceptually evaluated using a formal blind listening test, comparing it to a set of relevant baselines. This section details the evaluation process and presents the results.

Methods

In total three pitch shifting real-time solutions were chosen as baselines for evaluation. Two commercial solutions: Elastique Pitch V2[2] (EL) and SoundShfiter[3] (WS), and a pitch shifting C++ library Signalsmith Stretch (SS). Additionally, a simple resampling algorithm was used as a low-quality anchor (AN). Three sound samples of length 6-8 seconds were processed by all selected algorithms: full mix, drum loop, and guitar loop. Presented test samples were pitch-shifted either by a couple of semitones or by a round trip up (shifting the signal up and then down an octave).

The experiment has been set up accordingly to the work of Moliner et al. [5]. The evaluation consisted of a blind listening test using a modified version of webMUSHRA followed by a short interview. Each participant first completed a training session and then proceeded through six test sessions. During each listening session, participants were presented with reference audio and five test samples processed by different pitch-shifting algorithms. After the listening session participants were asked to answer three interview questions.

The experiment was conducted on the premises of Aalborg University Copenhagen. All participants used the same computer and a pair of Beyerdynamic DT 990 Pro headphones.

Participants were instructed to rate each condition on a scale from 0 to 100 reflecting how well the condition met their expectation for a pitch-shifted version of the reference, also considering the overall audio quality. The participants were not required to use the full scale, as there are no ideal examples of the best or worst quality.

In total eight people took part in the experiment. All participants specified that they have a professional and/or educational background in music.

[2]Elastique Pitch V2 - `https://products.zplane.de/products/elastiquepitch/`

[3]SoundShifter by Waves - `https://www.waves.com/plugins/soundshifter`

Results

The results of the listening experiment are shown in Fig 3. Similarly to the reference studies [4, 3, 5] Wilcoxon signed-rank test [8], followed by Cliff's Delta and its effect size calculation [1], was chosen as a data analysis method. As expected, the low-quality anchor achieved a significantly lower score than the proposed method (Wilcoxon signed-rank test: $p < 0.01$, Cliff's $\delta = 0.85$). The analysis shows, with great statistical significance, that the proposed solution has not achieved the quality of SOTA methods (details of the test results can be found in online resources). The system, in terms of quality, is the closest to the Signalsmith library. Surprisingly, the proposed solution was rated lower than the library even though the library is internally used by the system. This might suggest an incompatibility between NM and the library, either caused by conflicting designs or potential implementation issues.

A thematic analysis [2] of the interview comments highlighted a few prominent themes. Half of the participants stated that some of the stimuli sounded very similar to each other. Moreover, few of them even suggested that the same software was used twice in each test set. Most of the participants (63%) expressed that they enjoyed the experiment, either stating that looking for the differences was fun, or that they liked the sounds of the effect/stimuli itself. Interestingly, 38% of participants stated that they liked the sound of the artifacts present in some stimuli. Notably, two participants suggested that if the audio quality was not to be rated, they would put a higher rating on "bad" sounding conditions, as they "developed a taste to it". Another participant shared that, over the last years, in the music production industry (e.g. in the hyperpop genre) "aliasing/phasing" artifacts have become more popular and accepted as means of artistic expression.

Discussion and Future Work

The evaluation results indicate several areas for improvement. This section discusses the shortcomings of the proposed solution and suggests future work to address these issues.

Audio Quality

The audio quality of the system needs enhancement. The evaluation results show that the perceived quality is lower than commercial alternatives. One way to improve this is by enhancing the STN decomposition quality. The horizontal median filter, which processes STFT frames on the time axis, currently only considers prior frames due to real-time processing constraints. Introducing a one-frame look-ahead or fine-tuning the filter length could improve results.

The quality was also affected by issues with audio volume normalization, breaking the perfect reconstruction normally assured by fuzzy STN and NM. The Signalsmith library used for sines pitch-shifting inconsistently attenuates audio levels. Replacing it with an alterna-

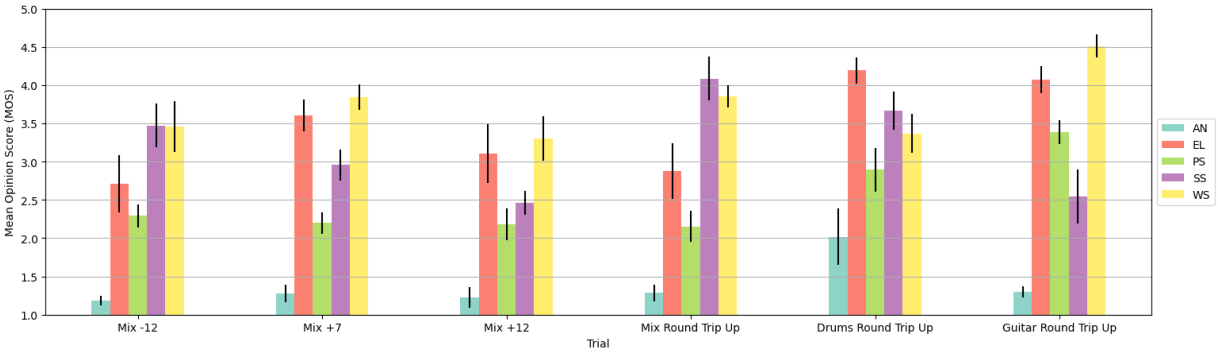


Figure 3: Listening test results - MOS with 95% confidence intervals

tive solution (e.g., zynaptiq ZTX[4]) could improve quality. Additionally, the vague description of white noise normalization in the NM paper led to unexpected results. A more precise normalization equation based on empirical evaluation could help.

Leveraging a non-linear frequency map could also enhance audio quality. Lower harmonics are crucial for pitch perception, while higher harmonics mainly affect timbre. Introducing a tonality limit, where frequency scaling follows the pitch-shifting factor up to a corner frequency and remains 1:1 beyond that, could reduce artifacts and better preserve the input timbre [9].

Performance

The plug-in's performance can be optimized by parallelizing the calculations of sines and noise components, as they become independent after decomposing the input signal. Additionally, merging the calculations of STN decomposition and pitch shifting could reduce the computational load. Currently, each algorithm calculates its own FFT and IFFT, increasing CPU-intensive operations.

Evaluation

The evaluation should be repeated and potentially expanded. A loudness bias [6] might be present in the results, as not all audio excerpts were normalized during exporting. This issue is only magnified by the loudness attenuation in the Signalsmith library.

Future evaluations should cover high pitch-shifting factors. NM has shown superior performance to its competitors in high TSM factors. Consequently, the proposed system should outperform alternatives in similar scenarios. This was not addressed in the current evaluation, as commercial solutions do not support pitch shifting by more than an octave up or down.

Conclusion

This paper details the design, implementation and evaluation of a low-latency pitch-shifting plug-in. The implementation centers around a pitch-shifting algorithm based on findings from previous TSM-related studies, with additional adaptations for real-time usage. The evaluation shows that, according to the quantitative data, the proposed solution falls short compared to commercial SOTA solutions. On the contrary, qualitative data suggest alternative use cases originating from the recent music trends. Future work involves improving audio quality, increasing plug-in performance, and altering evaluation strategy.

[4] zynaptiq ZTX `https://www.zynaptiq.com/ztx/`

References

[1] L. Alchieri, N. Abdalazim, L. Alecci, S. Gashi, M. Gjoreski, and S. Santini: Lateralization effects in electrodermal activity data collected using wearable devices. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8 (2024), 1–30.

[2] V. Braun and V. Clarke: Using thematic analysis in psychology. Qualitative Research in Psychology 3 (2023), 77–101

[3] L. Fierro and V. Välimäki: Enhanced fuzzy decomposition of sound into sines, transients, and noise. J. Audio Eng. Soc. 71 (2023), 468-480

[4] C. Mendonça and S. Delikaris-Manias: Statistical tests with mushra data. In Proc. Audio Eng. Convention Milan, Italy, 05 (2018)

[5] E. Moliner, L. Fierro, A. Wright, M. S. Hämäläinen, and V. Välimäki: Noise morphing for audio time stretching. IEEE Signal Process. Lett. 31 (2024), 1144-1148

[6] E. C. Poulton: Models for biases in judging sensory magnitude. Psychological Bulletin 86 (1979), 777-803

[7] A. von dem Knesebeck, P. Ziraksaz, and U. Zölzer: High quality time-domain pitch shifting using PSOLA and transient preservation. In Proc. Audio Eng. Convention, number 8202, San Francisco, CA, USA (2010).

[8] F. Wilcoxon: Individual comparisons by ranking methods. Biometrics Bulletin 1 (1945), 80-83

[9] G. Luff. The Design of Signalsmith Stretch, URL: `https://signalsmith-audio.co.uk/writing/2023/stretch-design`