Noise Morphing for Audio Time Stretching

Eloi Moliner, Leonardo Fierro, Alec Wright, Matti S. Hämäläinen, and Vesa Välimäki, *Fellow, IEEE*

***Abstract*—This letter introduces an innovative method to enhance the quality of audio time stretching by precisely decomposing a sound into sines, transients, and noise and by improving the processing of the latter component. While there are established methods for time-stretching sines and transients with high quality, the manipulation of noise or residual components has lacked robust solutions in prior research. The proposed method combines sound decomposition with previous techniques for audio spectral resynthesis. The time-stretched noise component is achieved by morphing its time-interpolated spectral magnitude with a white-noise excitation signal. This method stands out for its simplicity, efficiency, and audio quality. The results of a subjective experiment affirm the superiority of this approach over current state-of-the-art methods across all evaluated stretch factors. The proposed technique notably excels in extreme stretching scenarios, signifying a substantial elevation in performance. The proposed method holds promise for a wide range of applications in slow-motion media content, such as music or sports video production.**

***Index Terms*—Audio systems, interpolation, signal restoration, spectral analysis, timbre.**

I. Introduction

Audio time-scale modification (TSM), a critical process in audio signal processing, involves adjusting the temporal duration of a sound signal without altering its pitch [1]–[4]. This operation is integral in various applications, such as music production [5], sound design [6], [7], and multimedia content manipulation [8], [9]. This task becomes especially challenging with large stretching factors, where conventional methods, such as the phase vocoder, often introduce perceptual artifacts, e.g., transient smearing, loss of presence, and phasiness [3], [4], [10]. The subjective nature of audio time stretching further complicates the problem, as there is no clear objective metric for evaluation [9], [11]. The inherently ill-defined nature of this task, as there is no ideal reference signal, is shaped by subjective expectations and perceptual nuances.

The best performing TSM methods apply the Short-Time Fourier Transform (STFT), manipulate the spectrogram of the signal to change its duration, and then apply the inverse STFT to reconstruct the time-scaled signal [3], [4], [12]. Established TSM methods have predominantly focused on the separation and accurate manipulation of sinusoidal and transient components of sounds [13]–[15]. The noise component describes sound nuances and textures, e.g. plucking or bowing noise from stringed instruments, and is often the main descriptor for environmental sounds [16], [17]. Common TSM approaches, including phase vocoder-based methods, struggle to provide precise descriptions and scaling for such sound nuances, compromising the final time-stretched audio quality [16], [18]. The use of a three-way decomposition to isolate the noise component from sines and transients [19]–[22], in combination with phase randomization [23], [24] in the resynthesis process, showed a first improvement in the quality of the stretched noise component [4], [22]. A solution involving a Wavenet neural synthesizer for the noise component has also proved successful for extreme time stretching of environmental sounds [17].

Previous solutions targeting time-stretching of real-world sounds modeled the stretched noise component via linear interpolation of white Gaussian noise, with the spectral magnitude of the original sound around detected transients [8], [25], or with the residual component of the original sound after the sines were removed [26]. These solution compromise the audio quality when applied to general sounds as they are designed for noisy signals and do not feature a three-way decomposition for transient handling. An alternative technique leveraged generative adversarial networks for TSM of speech signals [27], but its data-driven nature imposes limitations on its application to general audio.

This letter introduces "Noise Morphing" (NM), an approach that combines the core idea behind the aforementioned techniques and the sines-transients-noise decomposition (STN). This involves producing a white-noise excitation signal of equal length to the output signal of the TSM processing. The white-noise signal is morphed with interpolated log-magnitude spectra of the noise component extracted from the target signal. The novelty lies in the application of spectral morphing within the STN framework, which adds a new layer of precision to the TSM processing chain: in the proposed approach, each of the three components is individually processed with the most suitable technique, before being recombined into a time-stretched mixture [22], [28].

This letter is structured as follows. Section II describes the STN decomposition and TSM principles that this work builds upon. Section III details the proposed NM technique. Section IV reports the methods and results of a subjective evaluation conducted against other TSM algorithms to validate the effectiveness of the novel approach, and Section V concludes.

II. Background

According to the STN model [19], [22], any sound can be described as the summation of tonal content (sines), impulsive events (transients), and sound nuances (noise). In this letter, audio signals are decomposed into these three components via

Submitted Dec. 21, 2023; revised Mar. 23, 2024. This work has been financed in part by Nokia Technologies through the DeepSlow 2 project (Aalto University project no. 400610). L. Fierro's work has been supported by the Aalto ELEC Doctoral School. This work is part of the activities of the NordicSMC network (NordForsk project no. 86892). E. Moliner and L. Fierro contributed equally to this work.

E. Moliner, L. Fierro, A. Wright and V. Välimäki are with the Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland (e-mail: name.surname@aalto.fi).

M. S. Hämäläinen is with Nokia Technologies, Tampere, Finland (e-mail: matti.s.hamalainen@nokia.fi).

soft spectral masking of their spectrograms, which leads to a fuzzy decomposition with perfect reconstruction and is the best method to date for this specific task [22].

Given an audio signal $\mathbf{x} \in \mathbb{R}^N$ and its Short-Time Fourier Transform (STFT) $\mathbf{X} \in \mathbb{C}^{M \times K}$, one can obtain a set of class masks following the methodology of Fitzgerald [29]. A median filter is applied to the magnitude spectrogram $|\mathbf{X}|$ in the time and frequency directions, and is used to retrieve the tonalness $\mathbf{R}_\mathrm{s} \in \mathbb{R}^{M \times K}$ and transientness $\mathbf{R}_\mathrm{t} \in \mathbb{R}^{M \times K}$, respectively. Soft masks are then computed as follows [22]:

$$\mathbf{S} = f\left(\mathbf{R}_\mathrm{s}\right), \tag{1}$$

$$\mathbf{T} = f\left(\mathbf{R}_\mathrm{t}\right), \tag{2}$$

$$\mathbf{N} = 1 - \mathbf{S} - \mathbf{T}, \tag{3}$$

where $f(a)$ is an element-wise saturating function [22]:

$$f(a) = \begin{cases} 1, & \text{if } a \geq \beta_\mathrm{U} \\ \sin^2\left(\frac{\pi}{2}\frac{a-\beta_\mathrm{L}}{\beta_\mathrm{U}-\beta_\mathrm{L}}\right), & \text{if } \beta_\mathrm{L} \leq a < \beta_\mathrm{U} \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where β_U and β_L are the upper and lower boundaries of the transition region, respectively.

The masks (1), (2), and (3) are imposed onto the complex spectrogram $\mathbf{X}$ via element-wise multiplication to decompose the three components. The process is repeated for two consecutive stages using different analysis window lengths and separation factors β_U and β_L to improve the decomposition quality [22], [30], [31]. The first stage extracts the sines from the transient and noise residual mixture, using a large analysis window and β_U = 0.80 and β_L = 0.70 for better frequency resolution; the second uses a short analysis window for better temporal resolution, separating the residual into transients and noise [22], using β_U = 0.85 and β_L = 0.75. Thus, three spectrogram representations are obtained, one for each component. As a consequence of the fuzzy classification, each time-frequency bin can belong to two classes simultaneously: to the sine and noise, or to the transient and noise classes [22].

After performing the STN decomposition, different TSM algorithms can be applied for each individual component. The sines are time-stretched using a phase vocoder with identity phase locking [32], as this has been found successful in previous studies [4], [9], [17], [22]. Transients are preserved after extraction by segmenting them into individual events and repositioning each segment in the correct position according to the TSM factor [33].

The noise component has been previously time stretched by randomizing the phase of each signal frame containing noise [4], [12]. However, this leads to an audible disturbance at large time-stretching factors [4]. This letter proposes to use a morphing technique to time-stretch the noise component with an improved perceptual quality, as described next.

III. Noise Morphing

This section introduces NM, a spectral morphing technique designed for the independent stretching of the noise component. A similar concept has been explored in previous works of Moinet et al. [8], [25] and Apel [26], although there were small but significant differences. The core principle of the NM method revolves around applying random phases while maintaining a magnitude consistent with the original audio, in such a way that perfect correlation between successive STFT frames is ensured. The proposed approach is grounded in the assumption that the noise or residual component, being quasi-stochastic, has little perceptual impact from its phase, allowing us to discard it.

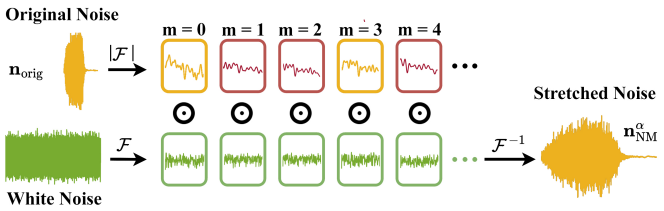

Fig. 1: Conceptualization of noise morphing, for $\alpha = 3$. The original noise log-magnitude spectra (yellow) are time-interpolated (red) and used to modulate the white-noise spectra (green) to produce the time-stretched output.

The proposed algorithm, depicted in Fig. 1, follows a structured analysis and synthesis procedure. The original noise component $\mathbf{n}_\mathrm{orig} \in \mathbb{R}^N$ is first processed with the STFT, using a Hann window of 2048 samples (46 ms) and a hop size of 1024 samples (23 ms) at a sample rate $f_\mathrm{s} = 44.1$ kHz. The log-magnitude spectrum of each STFT frame $\mathbf{N}_\mathrm{orig} \in \mathbb{R}^{M \times K}$ is computed as

$$\mathbf{N}_\mathrm{orig} = 10 \log_{10}(|\mathcal{F}(\mathbf{n}_\mathrm{orig})|), \tag{5}$$

where $\mathcal{F}()$ represents the STFT operator. The log-magnitude spectrum is then linearly interpolated according to the stretching factor α based on the two neighboring spectra, occurring before and after the interpolation point, following

$$\mathbf{N}^{\alpha} = \mathrm{lerp}(\mathbf{N}_\mathrm{orig}, \alpha), \tag{6}$$

where $\mathrm{lerp}(\cdot)$ is the linear interpolation function and α is the stretching factor. In the time dimension, the length of the spectrogram $\mathbf{N}^{\alpha} \in \mathbb{R}^{\lceil \alpha M \rceil \times K}$ is α times that of the spectrogram $\mathbf{N}_\mathrm{orig} \in \mathbb{R}^{M \times K}$ rounded up to the nearest integer.

In the synthesis phase, a white-noise excitation signal $\boldsymbol{\epsilon} \in \mathbb{R}^{\lceil \alpha N \rceil}$ is first generated matching the length of the output signal after time stretching, as shown in Fig. 1. According to our experiments, the perceptual impact of the noise sequence's distribution is negligible, provided its spectrum is white, and the sequence is standardized with zero mean and unit variance. Consequently, uniformly or Gaussian distributed noises, when normalized, are both viable options. In this work, the noise signal is sampled from a standard Gaussian distribution.

As shown in Fig. 1, the STFT is also applied to the white noise, using the same window and hop size as above. The resulting complex time-frequency signal $\boldsymbol{\mathcal{E}} \in \mathbb{C}^{\lceil \alpha M \rceil \times K}$ must be normalized by the window energy to ensure that the flat spectral magnitude equals one. Subsequently, the noise spectral frames are modulated by the interpolated magnitude spectra via element-wise multiplication:

$$\mathbf{N}^{\alpha}_{\mathrm{NM}} = \boldsymbol{\mathcal{E}} \odot 10^{\mathbf{N}^{\alpha}/10}. \tag{7}$$

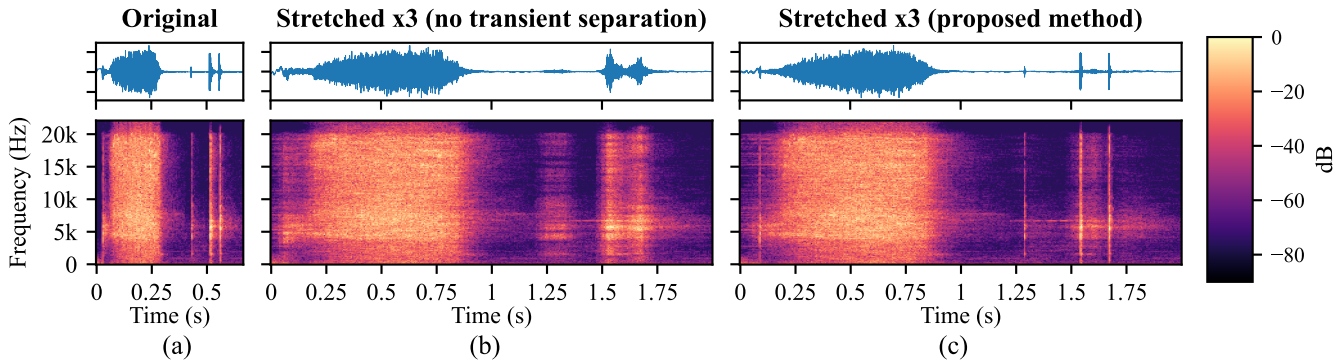

Fig. 2: A can opening sound (a) at normal speed and stretched with $\alpha = 3$ (b) without transient separation, which leads to transient smearing, and (c) with the proposed method, which preserves transients with apt handling of the noise component.

Finally, the morphed noise signal in the time-domain $\mathbf{n}^{\alpha}_{\mathrm{NM}} \in \mathbb{R}^{\alpha N}$ is obtained by applying the inverse STFT using the same parameters as in the analysis (see also Fig. 1):

$$\mathbf{n}^{\alpha}_{\mathrm{NM}} = \mathcal{F}^{-1}(\mathbf{N}^{\alpha}_{\mathrm{NM}}). \tag{8}$$

A notable difference between our method above and the work of Moinet et al. [25] is that the latter directly replaces the magnitude of the time-frequency signal $\boldsymbol{\mathcal{E}}$ with the interpolated magnitudes through polar coordinates, neglecting the white-noise magnitude spectra. Our observations suggest that the modulation approach of (7) yields a more organic effect, as the stochastic variations in the magnitude of the white-noise signal contribute to a perceptually smoother and less artifact-prone sound. Apel [26] combines the white-noise spectra and the interpolated magnitude spectra in the same way as here, but in his work, the residual component contains a mixture of noise and transients, which leads to the need for additional spectral smoothing techniques to enhance the sound quality.

A crucial parameter shaping the quality of the synthesized time-stretched audio is the window length. A long window introduces a smoother signal, akin to noise, but comes at the expense of diminished temporal detail in the output signal, and rapidly changing nuances tend to get smeared. On the contrary, a short window captures finer nuances of the sound, enhancing overall clarity, but has the potential of introducing musical noise artifacts, which may compromise the quality of the synthesized sound. Moinet made similar observations regarding the window length [8]. However, the challenges associated with long windows become more pronounced when transients are not separated. Moreover, our approach of multiplying the noise spectral frames with the interpolated magnitude spectra achieves more natural results with a short window, compared to replacing the magnitudes as Moinet et al. suggested [25].

A. Audio Time-Stretching Example

A comprehensive insight into the efficacy of the proposed TSM method is offered by the example visualized in Fig. 2. The waveform and spectrogram of the unprocessed signal, featuring hisses and clicks from the opening of a soda can, are shown in Fig. 2(a). The stretched noise is highlighted in Fig. 2(b), as well as the need for transient preservation: when the signal is stretched by a factor of 3, transients between 1.5 and 1.75 s are clearly smeared over time, resulting in a characteristic undesirable effect. In striking contrast, Fig. 2(c) showcases the proposed method's performance by preserving the transients between 1.5 and 1.75 s during the time-stretching process. Notably, the method adeptly manages the stretching of the noise component appearing around 5 kHz starting at about 1.5 s. when transients are separated, emphasizing its ability to achieve desirable audio TSM outcomes.

IV. Evaluation

The proposed method has been evaluated against a set of relevant baselines by means of a formal blind listening test. The evaluation process and results are reported in this section.

A. Compared Methods

We considered several baseline methods to provide a comprehensive benchmark for our proposed approach (NM). To establish a lower performance threshold, we included a standard phase vocoder [18], [34] as anchor (AN). As additional baselines, we incorporated the fuzzy phase vocoder [4] (FZ) and its enhanced version with transient preservation [22] (FT). Furthermore, we integrated a prior method in which the stretching of the noise component was achieved using a neural synthesizer [17] (WN).

In addition to these baselines, we conducted two ablation studies aimed at elucidating crucial factors influencing the time-stretching quality of the proposed method. One variant of our approach involved applying noise morphing without prior decomposition and transient separation (ND), resembling previous works by Moinet [8] and Apel [26]. Lastly, we included a version of our proposed method in which the noise morphing employs spectral magnitude replacement instead of multiplication (NI), as suggested by Moinet [8].

B. Listening Test Design

Our test approach, a variation of the standard MUSHRA test [35], has been used earlier in TSM studies [4], [17], and employs a multiple-stimuli method with the original, unprocessed sound as the reference. Across 15 trials, we presented sets of 7 stimuli, with 5 trials being conducted for

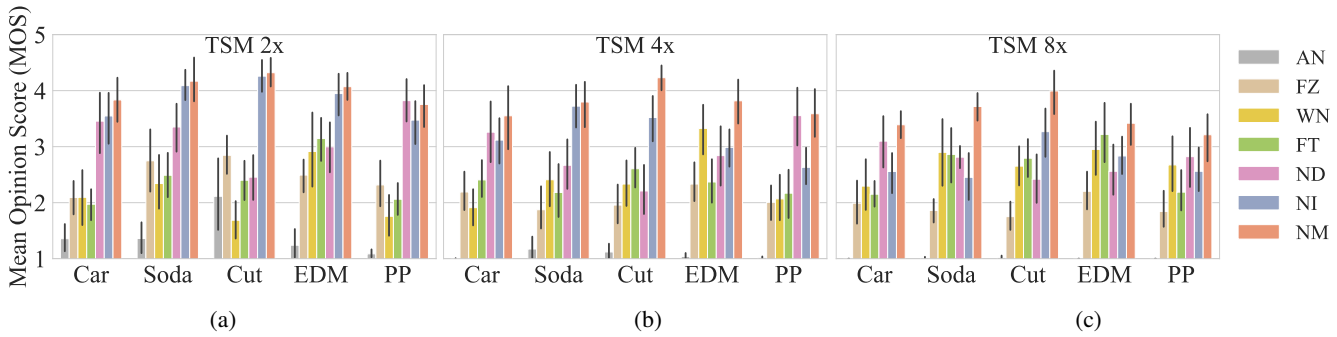

Fig. 3: Listening test results, showing MOS with 95% confidence intervals for (a) $\alpha = 2$, (b) $\alpha = 4$, and (c) $\alpha = 8$.

TABLE I: Audio excerpts used in the listening test

Item name	Description
Car	Live recording of a rally car passing by
Soda	Hiss and click sounds from a can opening
Cut	A knife cutting food on a cutting board
EDM (music)	Electronic music sample
PP (Ping Pong)	Sounds from an amateur ping pong game

each TSM factor $\alpha = 2, 4$, and 8. Each set included stimuli representing the proposed method and the 6 baseline methods outlined in Section IV-A. A set of 5 representative mono audio excerpts were included in the experiment. While we would have preferred to include more examples, we deemed it impractical as it would have resulted in a lengthy and tiring listening test for participants. The audio samples under test are listed in Table I and are available on the companion webpage for this letter[1].

To accommodate the extreme stretching factors involved in the test, each audio sample's duration was kept very short (approximately 2 s). This ensured that the longest time-stretched sounds remained below 18 s in duration [35].

A total of 13 volunteers participated in the experiment, ranging from 26 to 35 years of age. None of the participants had hearing impairments. The participants were instructed to rate each presented stimulus on a scale from 0 to 100, indicating the degree to which the sample met their own subjective expectations for a time-stretched version of the reference, together with the overall audio quality. The participants were not obligated to use the full scale, since ideal examples of best nor worst quality do not exist.

The test software was a customized version of WebMushra [36]. The audio items were played through a single pair of Sennheiser HD 650 headphones within a soundproof listening booth at the Aalto Acoustics Lab in Espoo, Finland.

C. Results

The results of the listening test are presented in Fig. 3. Notably, the proposed Noise Morphing method consistently emerged with the highest Mean Opinion Scores (MOS) across all examples and TSM factors except one, underscoring its efficacy in delivering perceptually superior time-stretched audio. The recommended Wilcoxon signed-rank test [37] shows a general trend of statistical significance in the data distributions, despite occasional overlap in some distributions. Results are reported in the companion website[1]. In this section, our analysis centers on comparing situations where confidence intervals occasionally overlap.

A comparative analysis between NM and NI reveals interesting dynamics. For α = 2, NM and NI exhibited similar performance. However, as the stretching factor increased to α = 4 and α = 8, NI received significantly lower scores in most examples. This reinforces our suggestion that the modulation of the magnitude spectra produces a more realistic noise output than simple magnitude replacement. Our results indicate that noise morphing without transient decomposition (ND) performs poorly on examples containing clear and frequent transients, such as Cut and EDM. This observation highlights the beneficial contribution of the STN decomposition in the time-stretching framework. Interestingly, WN (α = 4) and FT (α = 8) show comparable performance in the EDM example, while NI and ND experience a quality drop. This is most likely due to the nature of the sound, suggesting that WN and FT are more suited for time-stretching music signals.

Qualitative comparisons with Élastique, a renowned piece of commercial software for audio TSM, are not directly addressed here; instead, readers are directed to audio examples available on the accompanying website[1] due to the need for third-party software. This limitation precluded a direct quantitative comparison within our controlled testing environment.

To provide an overview of NM capabilities wider than what is shown in the results, a larger subset of processed examples is also available for listening on the companion website[1].

V. Conclusions

This letter introduces a method to improve the time-stretching of the noise component of an audio signal, which is obtained by separating tonal and transient components. The proposed Noise Morphing method exhibits consistent superiority in audio quality across various stretch factors when compared to baseline methods. The suggested approach shows potential for extensive use in various slow-motion media productions, including music processing or sports videos. Future work involves exploring how to expand the method for stereo and multichannel audio signals.

[1]http://research.spa.aalto.fi/publications/papers/ieee-spl-noisemorphing

REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 453–467, Dec. 1990.

[2] J. Bonada, "Automatic technique in frequency domain for near-lossless time-scale modification of audio," in *Proc. Int. Computer Music Conf.*, (Berlin, Germany), p. 396–399, Aug. 2000.

[3] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Appl. Sci.*, vol. 6, no. 2, p. 57, 2016.

[4] E.-P. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Appl. Sci.*, vol. 7, p. 1293, Dec. 2017.

[5] D. Cliff, "Hang the DJ: Automatic sequencing and seamless mixing of dance-music tracks," *HP Lab. Tech. Rep.*, vol. 104, 2000.

[6] V. Välimäki, J. Rämö, and F. Esqueda, "Creating endless sounds," in *Proc. 21st Int. Conf. Digital Audio Effects (DAFx)*, (Aveiro, Portugal), pp. 32–39, Sep. 2018.

[7] C. Malloy, "Timbral effects: The Paulstretch audio time-stretching algorithm," *J. Acoust. Soc. Am.*, vol. 151, pp. A158–A158, Apr. 2022.

[8] A. Moinet, *Slowdio: Audio Time-Scaling for Slow Motion Sports Videos*. PhD thesis, University of Mons, Mons, Belgium, 2013.

[9] T. Roberts, A. Nicolson, and K. K. Paliwal, "Deep learning-based single-ended quality prediction for time-scale modified audio," *J. Audio Eng. Soc.*, vol. 69, pp. 644–655, Sept. 2021.

[10] J. Laroche and M. Dolson, "Phase-vocoder: About this phasiness business," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, (New Paltz, NY), Oct. 1997.

[11] L. Fierro and V. Välimäki, "Towards objective evaluation of audio time-scale modification methods," in *Proc. Sound Music Comp. Conf. (SMC)*, (Torino, Italy), pp. 457–462, Jun. 2020.

[12] A. Röbel, "A shape-invariant phase vocoder for speech transformation," in *Proc. 13th Int. Conf. Digital Audio Effects (DAFx)*, (Graz, Austria), p. 298–305, Sep. 2010.

[13] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Process. Lett.*, vol. 21, pp. 105–109, Jan. 2014.

[14] J. Driedger and M. Müller, "TSM Toolbox: MATLAB implementations of time-scale modification algorithms," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, (Erlangen, Germany), pp. 249–256, Sep. 2014.

[15] G. Roma, O. Green, and P. A. Tremblay, "Time scale modification of audio using non-negative matrix factorization," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, (Birmingham, UK), Sep. 2019.

[16] W.-H. Liao, A. Roebel, and A. W. Y. Su, "On stretching Gaussian noises with the phase vocoder," in *Proc. 15th Int. Conf. Digital Audio Effects (DAFx)*, (York, UK), pp. 131–134, Sep. 2012.

[17] L. Fierro, A. Wright, V. Välimäki, and M. Hämäläinen, "Extreme audio time stretching using neural synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, (Rhodes Island, Greece), pp. 1–5, Jun. 2023.

[18] D. Arfib, F. Keiler, U. Zölzer, V. Verfaille, and J. Bonada, "Time-frequency processing," in *DAFX: Digital Audio Effects* (U. Zölzer, ed.), pp. 219–278, Chichester, UK: Wiley, 2nd ed., 2011.

[19] T. S. Verma and T. H. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 6, (Seattle, WA), pp. 3573–3576, May 1998.

[20] S. N. Levine and J. O. Smith III, "A sines+transients+noise audio representation for data compression and time/pitch scale modifications," in *Proc. Audio Eng. Soc. 105th Conv.*, (San Francisco, CA), Sep. 1998.

[21] T. S. Verma and T. H. Y. Meng, "Time scale modification using a sines+transients+noise signal model," in *Proc. Digital Audio Effects Workshop (DAFX)*, (Barcelona, Spain), pp. 49–52, Nov. 1998.

[22] L. Fierro and V. Välimäki, "Enhanced fuzzy decomposition of sound into sines, transients, and noise," *J. Audio Eng. Soc.*, vol. 71, pp. 468–480, Jul. 2023.

[23] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.*, vol. 14, no. 4, pp. 12–24, 1990.

[24] P. Hanna and M. Desainte-Catherine, "Time scale modification of noises using a spectral and statistical model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 6, (Hong Kong, China), pp. 181–184, Apr. 2003.

[25] A. Moinet, T. Dutoit, and P. Latour, "Audio time-scaling for slow motion sports videos," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, (Maynooth, Ireland), pp. 2–5, Sep. 2013.

[26] T. Apel, "Sinusoidality analysis and noise synthesis in phase vocoder based timestretching," in *Proc. Australasian Computer Music Conf.*, (Melbourne, Australia), pp. 7–12, Jul. 2014.

[27] E. Cohen, F. Kreuk, and J. Keshet, "Speech time-scale modification with GANs," *IEEE Signal Process. Lett.*, vol. 29, pp. 1067–1071, Apr. 2022.

[28] T. S. Verma and T. H. Y. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music J.*, vol. 24, no. 2, pp. 47–59, 2000.

[29] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, (Graz, Austria), p. 217–220, Sep. 2010.

[30] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, pp. 228–237, Jan. 2014.

[31] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, (Taipei, Taiwan), pp. 611–616, Oct. 2014.

[32] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 323–332, May 1999.

[33] F. Nagel and A. Walther, "A novel transient handling scheme for time stretching algorithms," in *Proc. Audio Eng. Soc. 127th Conv.*, (New York, NY), Oct. 2009.

[34] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, Feb. 1995.

[35] IET, "BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation ITU-R BS.1534-1, International Telecommunication Union, Geneva, Switzerland, 2015.

[36] M. Schoeffler, S. Bartoschek, F.-R. Stöter, *et al.*, "WebMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Research Software*, vol. 6, Feb. 2018.

[37] C. Mendonça and S. Delikaris-Manias, "Statistical tests with MUSHRA data," in *Proc. 144th Audio Eng. Soc. Conv.*, (Milan, Italy), May 2018.