Towards Objective Evaluation of Audio Time-Scale Modification Methods

Leonardo Fierro
Acoustics Lab,
Dept. Signal Processing and Acoustics,
Aalto University, Espoo, Finland
leonardo.fierro@aalto.fi

Vesa Välimäki
Acoustics Lab,
Dept. Signal Processing and Acoustics,
Aalto University, Espoo, Finland
vesa.valimaki@aalto.fi

ABSTRACT

The need for high-quality time-scale modification of audio is increasing, as media streaming services are providing new related functionalities to their users. The main goal of a time-stretching method is to preserve the pitch and the subjective quality of the different components of the audio signal, namely transients, noise, and tonal components. Many solutions have been proposed throughout the years, with various results depending on the kind of processed audio input. This paper introduces an evaluation method for audio time-scaling algorithms based on a recent fuzzy time-frequency decomposition, which reveals the energy of the tonal, transient, and noise components in the original and stretched sounds. From the energy curves, typical impairments, such as transient smearing and the loss of tonality, can be observed. This analysis approach is compared with the subjective preferences of different techniques. This leads to suggestions for possible improvements of future algorithms. The ultimate goal is having an objective evaluation method which matches the subjective quality assessment.

1. INTRODUCTION

Time-scale modification (TSM) is the process of changing the duration or the playback speed of a sound without affecting its frequency content, i.e. its pitch, timbre, loudness and brightness [1–3]. TSM allows, for example, to increase or reduce the speed of a speech signal so that the talker seems to be speaking faster or slower, respectively. If an audio signal is simply played at a different speed, i.e. its sample rate is changed, the spectral characteristics are deemed to be altered because the sound formants are moved. This results in an audio output with lower perceived pitch for a slowed-down input or with higher pitch for an input that is sped up. Hence, TSM methods are applied to avoid this phenomenon, retaining the original spectral characteristics of the sound.

TSM has been long used in music, to alter audio signals in an artistic way or to sync recorded sounds during mixing [4]. Time-scaling is applied when slow motion is involved [5] or when media content has to be conformed to a given time slot, as it often happens in TV or radio broadcasting [2]. Lately, streaming services are allowing users to change the video playback speed during reproduction, audio-books are offering "speed reading" functions, and the chance of slowing down words is often requested when learning a new language [6].

Classic time-domain methods for TSM, such as SOLA [7], WSOLA [8], and PSOLA [9] perform well for quasi-harmonic signals, but introduce phase jump artifacts when applied to polyphonic signals and are prone to transient misplacing, skipping or duplication. Following techniques based on the phase vocoder [10, 11] proved useful for signals which can be represented as a sum of slowly varying sinusoids, but also highlighted typical time-scaling artifacts, such as phasiness [12] and transient smearing [3, 13].

Novel techniques aimed at improving the phase vocoder by separating the tonal component of the audio input from the transients either before or during the processing [14]. Damskägg and Välimäki [15] have introduced a modified phase vocoder using a fuzzy classification of the spectral bins in tonal, transient, and noise components. Průša and Holighaus [16] proposed a method for phase correction based on phase gradient estimation that does not require explicit peak picking and tracking. Sharma *et al.* [17] developed a mel-scale based time-varying sinusoidal model for perceptually improved TSM. Recently, Roma *et al.* [18] presented a technique for time stretching involving non-negative matrix factorization, while Roberts and Paliwal [19] proposed a novel fuzzy epoch-synchronous overlap-add method to enhance time-scaling of speech signals.

An effective TSM algorithm should be able to preserve the quality of the tonal, transient, and noise components of the sound after the time scaling [20–23]. In this paper, the fuzzy classification proposed in [15] is implemented to investigate the energy curves of the spectral components of audio samples scaled with multiple techniques, comparing them against the energy curves of the original, non-scaled input. Under the strong hypothesis that perfect time stretching involves conservation of energy for tones, transients, and noise, the aforementioned techniques are evaluated, suggesting where improvements would be required.

This paper is structured as follows. In Section 2, the fuzzy classification described in [15] is briefly summarized. In Section 3, energy curves for the different spectral components are derived, leading to the evaluation method pro-

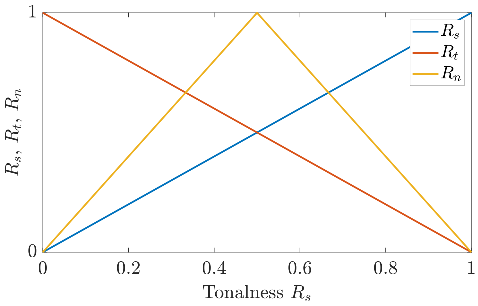


Figure 1: Relationship between individual fuzzy membership functions and tonalness R_s [15]. Transientness R_t is the opposite of tonalness whereas noisiness R_n peaks when the other two components become equal.

posed in Section 4 to compare performances of recent TSM techniques. Finally, a simple objective evaluation score is tested and some reflections and suggestions are discussed in Section 5.

2. FUZZY CLASSIFICATION OF SPECTRAL BINS

A fuzzy classification allows spectral bins to be described by their simultaneous contribution to tones, transients, and noise in a time-frequency representation of the signal [15]. Consider the Short–Time Fourier Transform (STFT):

$$X(m,k) = \sum_{n=-N/2}^{N/2} x(n + m\, H_a)\, w(n)\, e^{-j\omega_k n} \qquad (1)$$

where $x(n)$ is the input signal, $w(n)$ is the analysis window, m is the frame index, k is the spectral bin, H_a is the analysis hop size, N is the frame length in samples, and ω_k is the normalized central frequency of the k^{th} spectral bin.

Tones appear as time-direction flat lines in the spectrogram; conversely, transients appear as frequency-direction flat lines. Tonal and transient spectrograms are computed using median filters [15, 24], which highlight the desired component and suppress the opposite one:

$$X_s(m,k) = \\ \text{median}\Big[|X(m - \frac{L_t}{2} + 1, k)|, ..., |X(m + \frac{L_t}{2}, k)|\Big] \qquad (2)$$

$$X_t(m,k) = \\ \text{median}\Big[|X(m, k - \frac{L_f}{2} + 1)|, ..., |X(m, k + \frac{L_f}{2})|\Big] \qquad (3)$$

where L_t and L_f are the lengths (in samples) of the median filters in time and frequency directions, respectively.

The median-filtered STFTs are then used to compute tonalness R_s, transientness R_t, and noisiness R_n for each bin:

$$R_s(m,k) = \frac{X_s(m,k)}{X_s(m,k) + X_t(m,k)}, \qquad (4)$$

$$R_t(m,k) = 1 - R_s(m,k), \qquad (5)$$

and

$$R_n(m,k) = 1 - |R_s(m,k) - R_t(m,k)|. \qquad (6)$$

The relationship between the three spectral components is visualized in Fig. 1 as a function of tonalness R_s.

3. DERIVATION OF ENERGY CURVES

Membership functions R_s, R_t, and R_n can be used as soft masks for the STFT of an audio signal to individually evaluate the behavior of the three spectral components:

$$X_i(m,k) = X(m,k)\; R_i(m,k), \qquad i = s,t,n. \qquad (7)$$

The decomposition does not allow for perfect reconstruction as it is, i.e.:

$$R_s(m,k) + R_t(m,k) + R_n(m,k) \neq 1, \qquad (8)$$

and thus:

$$X_s(m,k) + X_t(m,k) + X_n(m,k) \neq X(m,k). \qquad (9)$$

A temporal energy curve for each component can now by easily computed:

$$\mathrm{E}_{X_i}(m) = \sum_k |X_i(m,k)|^2 \qquad i = s,t,n. \qquad (10)$$

Each curve reflects the spectral behavior of the associated component: the tonal curve resembles a slowly varying event, while the transient curve presents quick energy bursts interspersed by gaps of low energy.

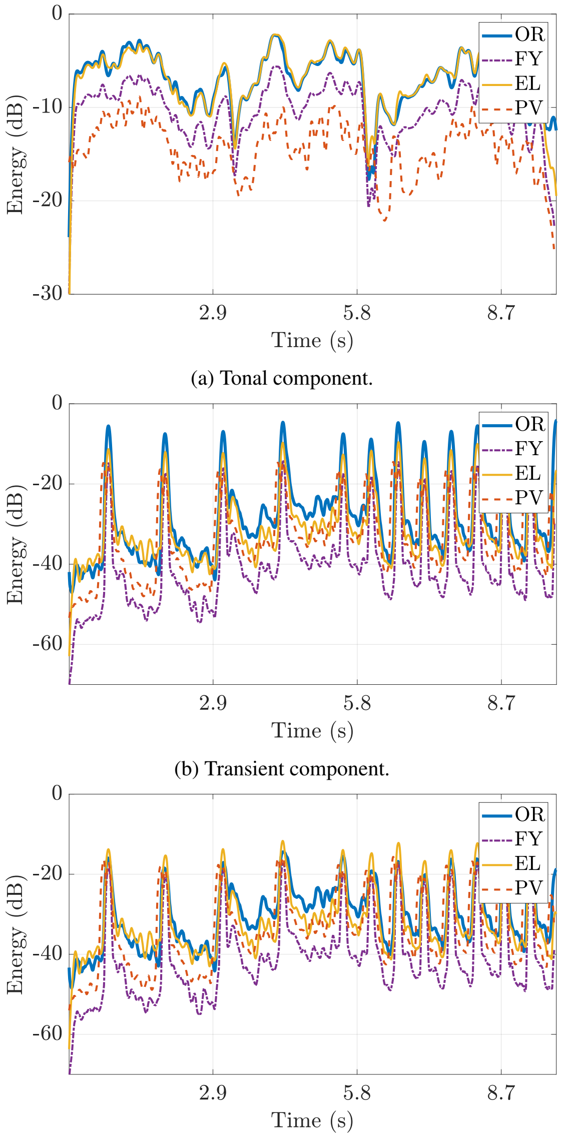
The following step is to confront an input signal $x(n)$ with its time-scaled version $y(n)$ by means of their energy curves E_{X_i} and E_{Y_i}. However, this comparison is not straightforward.

3.1 Time-axis interpolation

The time-scaled output has a different time axis with respect to the original signal. The amount of stretching (or compression) performed by a TSM algorithm is defined by the TSM factor α, which is usually related to the ratio between the analysis hop size and the synthesis hop size used for the processing.

The same number of time points (on different scales) for the energy curves of both the input and the output is needed for the comparison, so it is necessary to interpolate the input energy curves by a factor α. If α is a fractional number that can be represented as $\alpha = L/M$, where L and M are integers, the curves can be first interpolated by a factor L and then decimated by a factor M.

Furthermore, the input and output tracks need to be synchronized if some delay has been introduced during the time-scaling process. This is simply achieved by finding the first point of maximum in the cross-correlation function.

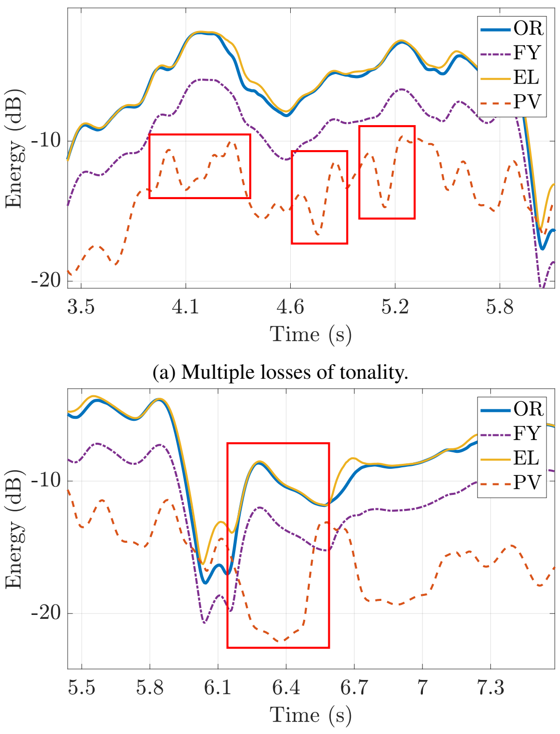


(a) Tonal component.

(b) Transient component.

(c) Noise component.

Figure 2: Comparison between the energy curves of the fuzzy spectral components, $\alpha = 2$, sample: *cast-viol*.



(a) Multiple losses of tonality.

(b) Great loss of tonality.

Figure 3: Loss of tonality in PV after time-stretching.

3.2 Weighting and normalization

The overall loudness of the time-strecthed audio output may largely differ from the input loudness, depending on the involved TSM technique. In order to suppress possible differences related to this matter during the evaluation step, the output spectrograms $Y_i(m, k)$ can be A-weighted and normalized to fit the range $[-1, 1]$ prior to the fuzzy classification process.

4. VISUAL EVALUATION OF TSM METHODS

A primary evaluation can be conducted by studying input and output audio files from the listening test in [15], namely samples *drumsolo* (a noisy solo of drums), *vocals* (a short a cappella from Suzanne Vega's "Tom's Diner"), *cast-viol* (a sequence of castanet sounds over a violin) and *techno* (a short dance music sample).

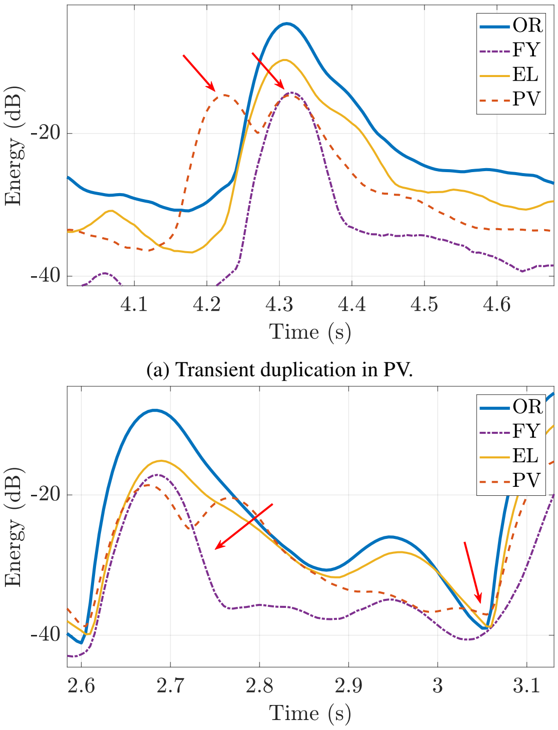
The following analysis parameters are set: $f_s = 44.1\,\text{kHz}$, $H_a = 512$ samples, $L_t = 500\,\text{ms}$, $L_f = 200\,\text{Hz}$ and $\alpha = 2$. Using an Hamming window for the STFT, the window length is $N = 4096$ samples ($\approx 100\,\text{ms}$) for the input and αN for the time-scaled output.

The hypothesis of conservation of the energy of the separate spectral components appears to be valid, as it can be seen from Fig. 2. The energy curves for the original *cast-viol* signal (OR, blue) are compared with the output energy curves for the standard phase vocoder (PV, orange), the fuzzy phase vocoder (FY, purple), and the popular commercial algorithm *Élastique* (EL, yellow) [25]. All the curves have been normalized with respect to the maximum energy value of the non-decomposed input signal.

As expected, the curves clearly show how the standard phase vocoder performs visibly worse than novel methods: it heavily suffers from tonality loss (Fig. 3), transient duplication and smearing (Fig. 4a).

Another example of transient smearing can be seen in Fig. 4b for the *drumsolo* sample, where FY fails to accurately follow the offset of the first transient and consequently misses the steep transition into the second transient. This is confirmed by listening to the time-stretched track, where the smearing is clearly audible.

A further analysis of Fig. 2 also reveals that novel techniques nicely follow the spectral behavior of tones and



(a) Transient duplication in PV.

(b) Transient smearing in FY.

Figure 4: Typical transient artifacts emerging from the comparison of the fuzzy energy curves.

transients of the input signal and that the noise component gives an important contribution to the overall energy. It is also peculiar how its spectral energy curves strongly resembles the transient ones, hinting that transients in the audio input probably have a strong noise component.

In order to have a better understanding of the TSM performances, we can evaluate the energy deviation of the TSM outputs with respect to the non-scaled input (Fig. 5), considering the energy curves for the non-decomposed tracks. The goodness of EL is immediately visible, as its deviation is almost always around 0 dB. FY appears to suffer from a constant energy loss, but that might be a consequence of a loudness mismatch between the input and the output tracks. For this reason, energy deviation curves for each fuzzy component can be normalized by matching the input and output mean deviations:

$$\Delta \mathrm{E}_i = L_{Y,i} - L_{X,i} - (\bar{L}_{Y,i} - \bar{L}_{X,i}), \qquad (11)$$

where

$$L_{Y,i} = 10 \log_{10}(\mathrm{E}_{Y_i}),$$
$$\bar{L}_{Y,i} = \mathrm{mean}[L_{Y,i}], \quad i = s, t, n.$$

The normalized deviation curves for the *cast-viol* sample are reported in Fig. 6. With this visual representation, FY appears to be better overall for the aforementioned sample with respect to EL, as reflected by the listening test conducted in [15]. In particular, Fig. 6a shows that the tonal component is quite well preserved, while Fig. 6b and 6c show that, using FY, a lot of the transient energy appears to be converted into noise energy after the time-stretching, providing an explanation for the poor performances of FY with transient-dominant sounds.

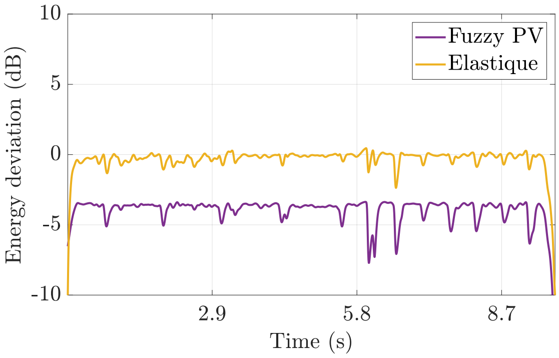


Figure 5: Energy deviation with respect to the original input, $\alpha = 2$, sample: *cast-viol*.

5. PROPOSED OBJECTIVE SCORING

A visual evaluation method for TSM techniques has been described in the previous Section, but it may be of interest to also produce a performance "score" for TSM algorithms using the information and the knowledge provided by the energy deviation curves. Ideally, the objective score should closely match the subjective MOS (Mean Opinion Score) that would result from a formal listening test.

A simple way of using the energy curves to synthesize a final score is to compute the MSE (Mean Squared Error) for every spectral component and then use linear regression to model a relationship between the energy deviations and the MOS given in [15]:

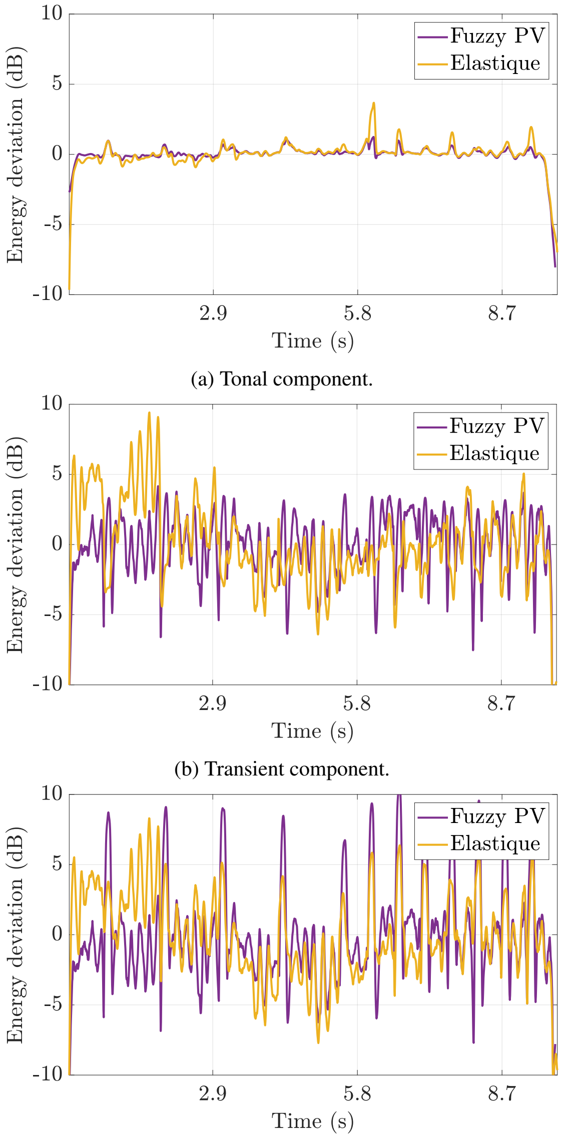
$$e_i = \mathrm{mean}[\Delta \mathrm{E}_i^2] \qquad i = s, t, n. \qquad (12)$$

Different regression models have been generated, first using data from a single TSM factor ($\alpha = 1.5$ and $\alpha = 2.0$) and then combining the datasets. Coefficients for each model are reported in Table 1: as it can be seen, all the models correctly try to set a bias $b_0 = 3$ (the central value in the MOS scale). The estimated objective evaluation score $\hat{S}$ is finally obtained as:

$$\hat{S} = b_0 + \sum_i b_i e_i \qquad i = s, t, n. \qquad (13)$$

Real MOS and estimated objective scores using the combined model are reported in Table 2 for some samples and different stretching factors and TSM algorithms: FY, EL and HP[1] (Harmonic-Percussive Separation, [14]). The best real and predicted scores are highlighted. The amount of data available from the listening test in [15] is clearly not enough to generate an accurate model and it is unlikely that a linear function alone is entirely capable of describing the

[1] Energy curves for Harmonic-Percussive separation were not displayed in Section 4 to avoid an overcrowded visualization.



(a) Tonal component.

(b) Transient component.

(c) Noise component.

Figure 6: Energy deviation curves for the individual spectral components, $\alpha = 2$, sample: *cast-viol*.

relationship between the energy curves and the subjective MOS.

6. CONCLUSION

In this paper, the application of energy curves for the tonal, transient, and noise spectral components of sounds is proposed to analyze time-scaled audio signals. The separate components were extracted from the spectrogram of the signal with median filters, operating separately in the time and in the frequency directions. Typical impairments appearing in time-stretched audio could be clearly identified from the energy curves by visual inspection, as demonstrated by several example figures.

	$\alpha = 1.5$	$\alpha = 2.0$	**Combined**
b_0	2.957	2.787	2.996
b_s	−0.355	−0.013	−0.025
b_t	0.034	−0.137	−0.111
b_n	0.055	0.138	0.104

Table 1: LR coefficients for generated models.

Audio Sample	α	**FY**		**EL**		**HP**	
		True	Pred	True	Pred	True	Pred
Drumsolo	*1.5*	2.3	2.7	3.2	2.8	**3.5**	**3.0**
	2.0	1.8	2.7	**2.5**	2.6	2.4	**2.7**
Cast-viol	*1.5*	**4.1**	2.6	3.6	2.7	3.8	**3.4**
	2.0	**4.1**	**3.7**	3.3	3.0	3.6	3.7
Vocals	*1.5*	3.4	2.7	2.9	2.8	**3.5**	**2.9**
	2.0	3.1	2.8	2.7	2.8	**3.3**	**3.0**

Table 2: Real MOS scores (taken from [15]) and estimated objective scores for some samples and different TSM algorithms. The best results on each row are highlighted to ease comparison.

Furthermore, this paper investigated the possible use of the energy curves for objectively evaluating the sound quality of time-stretched signals. A linear regression model was fit to available listening test data to predict mean opinion scores using the three energy components. A regression model generated using the combined data of two different time-stretch factors gives a superior prediction in comparison to models produced from a single stretch factor. The accuracy of the proposed prediction method is still insufficient to replace the listening test. Further investigations, possibly including more listening tests, will be required for devising an accurate objective evaluation method for audio time-scale modification.

Acknowledgments

The authors are grateful to Eero-Pekka Damskägg for helpful discussions and for providing his TSM programs. The help of Nils Meyer-Kahlen is gracefully acknowledged. This work is part of the activities of the “Nordic Sound and Music Computing Network (NordicSMC)”, NordForsk project number 86892. The work of Leonardo Fierro was funded by the Aalto ELEC Doctoral School.

7. REFERENCES

[1] E. Moulines and J. Laroche, “Non-parametric techniques for pitch-scale and time-scale modification of speech,” *Speech Communication*, vol. 16, no. 2, pp. 175–205, 1995.

[2] P. Dutilleux, G. De Poli, A. von dem Knesebeck, and U. Zölzer, “Time-segment processing (chapter 6),” *DAFX: Digital Audio Effects, Second Edition; Zölzer, U., Ed*, pp. 185–217, 2011.

[3] J. Driedger and M. Müller, “A review of time-scale

modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.

[4] D. Cliff, "Hang the DJ: Automatic sequencing and seamless mixing of dance-music tracks," *Hp Laboratories Technical Report Hpl*, vol. 104, 2000.

[5] A. Moinet, "Slowdio: Audio time-scaling for slow motion sports videos," Ph.D. dissertation, University of Mons, Mons, Belgium, 2013.

[6] O. Donnellan, E. Jung, and E. Coyle, "Speech-adaptive time-scale modification for computer assisted language-learning," in *Proc. 3rd IEEE Int. Conf. Advanced Learning Technologies*, Athens, Greece, 2003, pp. 165–169.

[7] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP-85)*, 1985, pp. 493–496.

[8] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP-93)*, vol. 2, 1993, pp. 554–557.

[9] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[10] M. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 3, pp. 374–390, 1981.

[11] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 3, pp. 323–332, 1999.

[12] ——, "Phase-vocoder: About this phasiness business," in *Proc. Workshop Appl. Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 1997.

[13] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. Int. Conf. Digital Audio Effects (DAFx-03)*, London, UK, 2003, pp. 344–349.

[14] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 105–109, 2013.

[15] E.-P. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Applied Sciences*, vol. 7, no. 12, p. 1293, 2017.

[16] Z. Průša and N. Holighaus, "Phase vocoder done right," in *Proc. European Signal Process. Conf. (EUSIPCO)*, 2017, pp. 976–980.

[17] N. Sharma, S. Potadar, S. R. Chetupalli, and T. Sreenivas, "Mel-scale sub-band modelling for perceptually improved time-scale modification of speech and audio signals," in *Proc. National Conf. Communications (NCC)*, 2017, pp. 1–5.

[18] 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, 2019.

[19] T. Roberts and K. K. Paliwal, "Time-scale modification using fuzzy epoch-synchronous overlap-add (FESOLA)," in *Proc. IEEE Workshop Appl. Signal Process. Audio and Acoustics (WASPAA-19)*, New Paltz, NY, USA, 2019, pp. 31–34.

[20] 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'98)*, vol. 6, 1998, pp. 3573–3576.

[21] 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.*, 1998.

[22] T. S. Verma and T. H. Meng, "Time scale modification using a sines+transients+noise signal model," in *Proc. Digital Audio Effects Workshop (DAFX'98)*, Barcelona, Spain, 1998, pp. 49–52.

[23] ——, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music J.*, vol. 24, no. 2, pp. 47–59, 2000.

[24] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. Int. Conf. Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010.

[25] Zplane Development. Élastique: industry-leading time stretching pitch shifting. Accessed: 03.06.2020. [Online]. Available: http://licensing.zplane.de/technology#elastique