OPEN ACCESS Freely available online
L. Fierro and V. Välimäki, "Enhanced Fuzzy Decomposition of Sound Into Sines, Transients, and Noise"
J. Audio Eng. Soc., vol. 71, no. 7/8, pp. 468–480, (2023 July/August).
DOI: https://doi.org/10.17743/jaes.2022.0077

Enhanced Fuzzy Decomposition of Sound Into Sines, Transients, and Noise

LEONARDO FIERRO,* AND VESA VÄLIMÄKI, *AES Fellow*
(leonardo.fierro@aalto.fi) (vesa.valimaki@aalto.fi)

Acoustics Lab, Department of Information and Communication Engineering, Aalto University, Espoo, Finland

The decomposition of sounds into sines, transients, and noise is a long-standing research problem in audio processing. The current solutions for this three-way separation detect either horizontal and vertical structures or anisotropy and orientations in the spectrogram to identify the properties of each spectral bin and classify it as sinusoidal, transient, or noise. This paper proposes an enhanced three-way decomposition method based on fuzzy logic, enabling soft masking while preserving the perfect reconstruction property. The proposed method allows each spectral bin to simultaneously belong to two classes, sine and noise or transient and noise. Results of a subjective listening test against three other techniques are reported, showing that the proposed decomposition yields a better or comparable quality. The main improvement appears in transient separation, which enjoys little or no loss of energy or leakage from the other components and performs well for test signals presenting strong transients. The audio quality of the separation is shown to depend on the complexity of the input signal for all tested methods. The proposed method helps improve the quality of various audio processing applications. A successful implementation over a state-of-the-art time-scale modification method is reported as an example.

0 INTRODUCTION

Decomposing an audio signal into its sinusoidal, transient, and noise (STN) components has been drawing research interest for over two decades [1–4]. It is a widely used tool in a variety of audio processing applications, ranging from beat tracking [5] and tonality estimation [6] to reduction of spectral complexity in cochlear implants [7] and to virtual bass enhancement [8]. The STN separation is also helpful in time-scale modification (TSM) [1, 9, 10], where it has been combined with the notion of fuzzy logic in order to improve [4] or evaluate the audio quality [11]. In all these audio applications, it is helpful to process sine, transient, and noise components independently of each other. This paper proposes improvements to the fuzzy STN decomposition of audio signals.

The STN separation relies on the assumption that any audio signal can be described as a linear combination of three independent actors: tonal content (sines), impulsive events (transients), and a residual part (noise) that does not belong to either one of the other two classes and adds nuance to the sound. Historically, additive synthesis modeled any sound as a sum of sinusoidal components [12–14]. Serra and Smith expanded the additive synthesis method by introducing the noise class, which was obtained as a residual after a sinusoidal model was subtracted from the original signal [15]. In the resulting method—called spectral modeling synthesis [15]—the frequency, amplitude, and phase of the sinusoidal components were estimated from the short-time Fourier transform (STFT) using a method similar to the McAulay–Quatieri algorithm [16].

The three-way decomposition was first introduced by Verma et al. [1, 17, 18], who showed that including a third component for transients was greatly beneficial in the context of signal analysis and synthesis, as it avoided the smearing of transients, which was a weakness in sines + noise models. Levine and Smith also showed that the adaptiveness of the STN model made it suitable for audio compression and for pitch- and time-scale modification [19].

Fitzgerald discovered that it was possible to decompose an audio signal into its sinusoidal and transient components by using spectral masks extracted via horizontal and vertical median filtering of the STFT [20]. Driedger et al. [2] reintroduced the three-way separation by updating Fitzgerald's method: the noise component could be obtained by

*To whom correspondence should be addressed, e-mail: (leonardo.fierro@aalto.fi).

retrieving spurious information after extracting the other two components with median filtering.

Füg et al. [3] proposed a follow-up method involving the use of structure tensors (ST) to find predominant orientation angles and anisotropy in the time-frequency signal representation, showing an improvement in the separation quality for sounds with vibrato.

Other recent approaches for sines–transients separation include a kernel additive matrix [21], non-negative matrix factorization [22], improved sinusoidal modeling [23, 24], and neural networks [25]. However, these methods do not involve a third class for the noise component; hence, they are not discussed further in this paper.

It should be noted that the STN decomposition does not directly relate to traditional source separation, which usually aims at retrieving musical instruments in a sound mixture [26] or speech from noisy background sources [27]. According to the STN formulation, even strongly percussive sources, such as drums, will have a sinusoidal and noise component—unless they are perfect, synthetic pulses—and, similarly, strongly harmonical sources, such as the violin, will hold, in addition to a sinusoidal part, a transient component in their attack and a noise component to describe the nuances, such as the bowing noise.

Although both Driedger et al. [2] and Füg et al. [3] applied hard binary masks to define the STN classes, Damskägg and Välimäki [4] introduced the concept of fuzzy logic in the context of TSM. The fuzzy classification (FZ) allows spectral bins to simultaneously contribute to the three classes, providing a more refined basis for the three-way separation [4]. This decomposition method was then extended to objective evaluation by Fierro and Välimäki [11] and improved by Moliner et al. [8] to allow perfect reconstruction by ensuring that the three soft spectral masks sum up to unity.

This work proposes a novel way to estimate fuzzy soft masks for STN decomposition of audio signals. The proposed method allows for intermediate classifications of the spectral bins between two components—sines vs. noise and transients vs. noise. This two-stage decomposition is shown to improve the overall sound quality of the separated components, particularly for transients. The masks ensure perfect reconstruction and are optimized for each class to have a large constant region followed by a fast but smooth transition to the adjacent class. The transition slope is refined for both decomposition stages to provide the best separation quality.

The rest of this paper is structured as follows. SEC. 1 discusses previous three-way separation techniques. SEC. 2 introduces the new STN decomposition method, which quasi-optimally extracts the sinusoidal and transient components. SEC. 3 evaluates the proposed method against three previous techniques. SEC. 4 applies the proposed method to TSM, and SEC. 5 concludes.

1 RELATED WORK

This section summarizes three previous STN decomposition methods based on a spectrogram representation of the input signal. A spectrogram X is an M-by-K matrix representing the time-frequency behavior of audio signal x. Each element $X(m, k)$ is computed using the STFT [28, 15]:

$$X(m,k) = \sum_{n=0}^{L-1} x(n+mH)\, w(n)\, e^{-j\omega_k n}, \tag{1}$$

where n is the sample index, $m = 0, 1, 2, \ldots M-1$ is the temporal frame index, $k = 0, 1, 2, \ldots K-1$ is the spectral bin index, w is the analysis window, H is the hop size, L is the window length in samples, which is assumed to be even, j is the imaginary unit, and ω_k is the normalized central frequency of the k^{th} spectral bin.

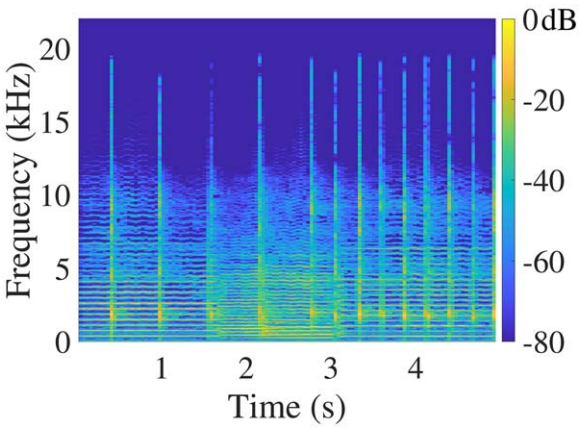

Fig. 1. Spectrogram of a test signal consisting of the castanets and the violin playing simultaneously.

1.1 Harmonic–Percussive–Residual Separation

The Harmonic–Percussive–Residual (HPR) separation [2] builds upon the Harmonic–Percussive (HP) method for sines–transients decomposition [20]. Fitzgerald noted that because sinusoids form flat lines in time direction in the spectrogram and, vice versa, impulsive events appear as flat lines in the frequency direction, they can be detected (suppressed) using a median filter [20]. Fig. 1 shows the spectrogram of a signal consisting of a mixture of violin and castanets, whose time- and frequency-direction ridges are noticeable. The spectrogram was produced using the 2,048-point Fast Fourier Transform (FFT) with a 2,048-sample Hann window and a hop size of 1024, i.e. 50% overlap. The sample rate of the test signal is 44.1 kHz.

Horizontal (time-oriented) and vertical (frequency-oriented) median filtering can be applied to the spectrogram $X(m, k)$ to highlight the desired component and suppress the other [20]:

$$X_{\text{h}}(m,k) = \text{med}\Big[|X(m-\frac{L_{\text{h}}}{2}+1,k)|, \ldots, |X(m+\frac{L_{\text{h}}}{2},k)|\Big] \tag{2}$$

and

$$X_{\text{v}}(m,k) = \text{med}\Big[|X(m,k-\frac{L_{\text{v}}}{2}+1)|, \ldots, |X(m,k+\frac{L_{\text{v}}}{2})|\Big], \tag{3}$$

where med[·] is the median function, and X_h and X_v are the resulting horizontally and vertically enhanced magnitude spectrograms, respectively. Parameters L_h and L_v are the median filter lengths (in samples) in the time and frequency directions, respectively.

Matrices X_h and X_v are then used to extract the tonalness R_s and transientness R_t matrices with the following elements [20]:

$$R_s(m,k) = \frac{X_h(m,k)}{X_h(m,k) + X_v(m,k)} \tag{4}$$

and

$$R_t(m,k) = 1 - R_s(m,k) = \frac{X_v(m,k)}{X_h(m,k) + X_v(m,k)}, \tag{5}$$

respectively. Fitzgerald [20] used R_s and R_t directly as spectral masks, whereas Driedger et al. [2] later introduced a controllable separation factor β and a third class (noise) to describe those parts of the sound that are neither sines nor transients.

From Eqs. (4) and (5), a set of hard spectral masks S (sinusoidal), T (transient), and N (noise) can be derived as follows [2]:

$$S(m,k) = \begin{cases} 1, & \text{if } R_s(m,k)/R_t(m,k) > \beta \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

$$T(m,k) = \begin{cases} 1, & \text{if } R_t(m,k)/R_s(m,k) > \beta \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

and

$$N(m,k) = 1 - S(m,k) - T(m,k). \tag{8}$$

Their relationship for a chosen β is shown in Fig. 2. The spectral masks are then imposed on $X(m,k)$ to retrieve the three desired spectral components:

$$X_s = S \odot X, \qquad X_t = T \odot X, \qquad X_n = N \odot X, \tag{9}$$

where ⊙ represents the Hadamard product, or element-wise multiplication.

It has been observed that the quality of the HPR separation largely varies for the sinusoidal and the transient components depending on the choice of the analysis window length L [29, 30, 2]. A large window length L for the STFT, ensuring sufficient frequency resolution but poor time resolution, results in a faultless extraction of sines but a low-quality transient output; conversely, a smaller value of L leads to a better extraction of the transient component but a worse description of sines.

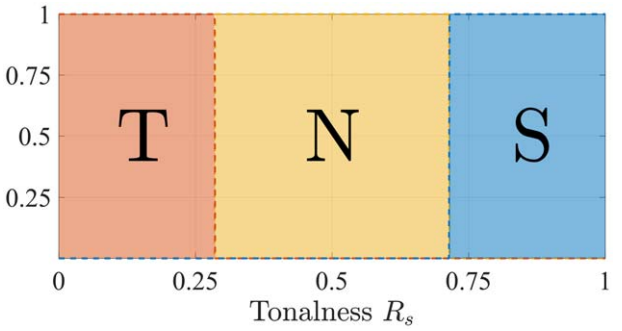

Fig. 2. Hard masks for transients, noise, and sines, as used in the HPR method [2], for separation factor β = 2.5.

To overcome the time-frequency limitation, Driedger et al. [2] divided the decomposition process into two cascaded iterations [2]. In the first stage, a longer analysis window is applied to extract the sinusoidal component [2]

$$x_s = \text{ISTFT}\left[S_1 \odot X\right], \tag{10}$$

while transients and noise remain mixed together in the residual

$$x_{res} = \text{ISTFT}[(T_1 + N_1) \odot X], \tag{11}$$

where ISTFT is the Inverse STFT. Subsequently, the residual x_{res} from the first stage is separated again with shorter windowing, leading to the final decomposition [2]:

$$x_t = \text{ISTFT}[T_2 \odot X_{res}], \tag{12}$$

$$x_n = \text{ISTFT}[(S_2 + N_2) \odot X_{res}]. \tag{13}$$

The noise signal x_n will also contain residuals of sine components, unless they were perfectly separated on the first stage. Fig. 3 shows the separated STN components of the example audio signal used here obtained with the HPR method.

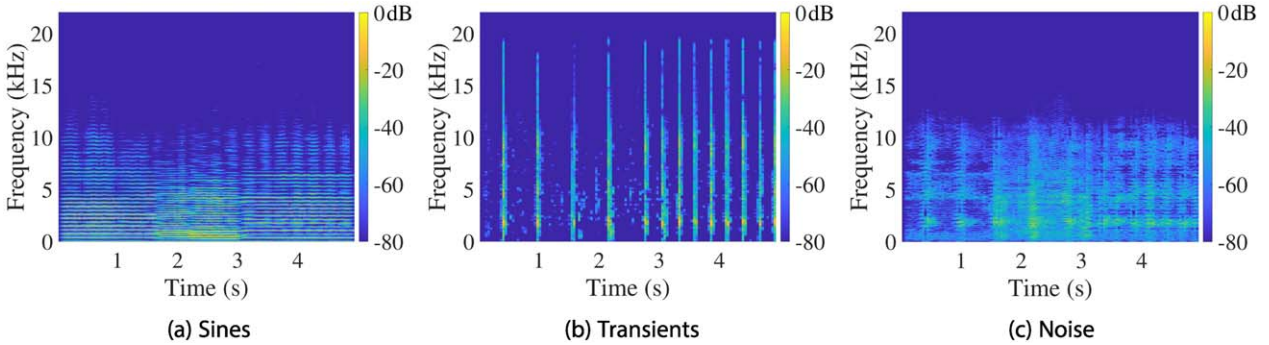

Fig. 3. STN separation performed over the mixture of castanets and violin, cf. Fig. 1, using the HPR method: (a) sines, (b) transients, and (c) noise.

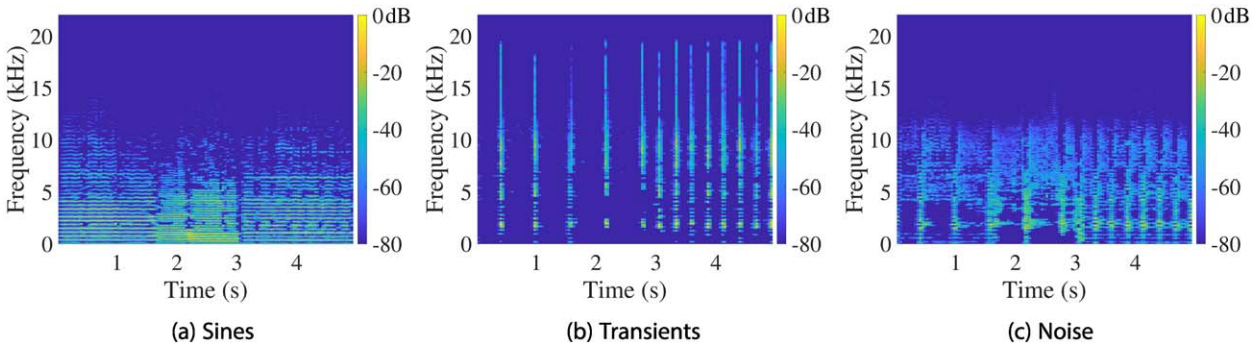

Fig. 4. STN decomposition obtained using ST method, cf. Fig. 3: (a) sines, (b) transients, and (c) noise.

1.2 Sinusoidal, Transient, and Noise Separation Based on Structure Tensor

Füg et al. [3] noted that sounds exhibiting vibrato, which carry tonal information and are perceived as sines, do not present strictly horizontal structures in the spectrogram. This results in a leakage of energy in between different spectral components. The strictness of the median filtering can be overcome using an ST, a widely used tool in image processing, to obtain a measure of the frequency change rate and local anisotropies in the spectrogram, which will then be used as features to define the spectral masks [3].

The ST matrix is obtained from the partial derivatives of the spectrogram with respect to time and frequency, and the orientation angles α and the anisotropy C of the spectral bins are computed from the eigenvalues and the eigenvectors of such a matrix, as described in [3]. The instantaneous frequency change rate R is computed for each bin from the orientation angles:

$$R(m,k) = \frac{{f_s}^2}{HM} \tan[\alpha(m,k)], \tag{14}$$

where f_s is the sample rate. The spectral masks are then obtained as follows:

$$S(m,k) = \begin{cases} 1, & \text{if } |R(m,k)| \le r_s \ \wedge \ C(m,k) > c \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

$$T(m,k) = \begin{cases} 1, & \text{if } |R(m,k)| \ge r_t \ \wedge \ C(m,k) > c \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

where c is the anisotropy threshold, and r_s and r_t are the frequency rate thresholds for the sinusoidal and the transient component, respectively. The noise mask is computed as described in Eq. (8), and the spectral components are then derived as in Eq. (9).

Fig. 4 shows the separated STN components of the example audio signal using the ST method. Some differences can be observed in comparison to the separation results of the HPR method in Fig. 3, such as some holes in the transient events at low and middle frequencies in Fig. 4(b).

1.3 Fuzzy Separation

Damskägg and Välimäki [4] introduced the concept of fuzzy classification of the spectral bins, which corresponds to a nonbinary classification using continuous values between 0 and 1. This method was later extended by Moliner et al. [8] to ensure perfect reconstruction, i.e., all masks summing up to unity. In [8], a third membership function for noisiness R_n is derived from Eqs. (4) and (5):

$$R_n(m,k) = 1 - \sqrt{|R_s(m,k) - R_t(m,k)|}. \tag{17}$$

The soft spectral masks are computed as

$$S(m,k) = R_s(m,k) - \frac{1}{2} R_n(m,k), \tag{18}$$

$$T(m,k) = R_t(m,k) - \frac{1}{2} R_n(m,k), \tag{19}$$

and

$$N(m,k) = 1 - S(m,k) - T(m,k) = R_n(m,k). \tag{20}$$

Their relationship is shown in Fig. 5. The spectral masks are once again imposed on $X(m, k)$ to obtain the spectral components using the Hadamard product, as in Eq. (9).

Fig. 6 shows the separated STN components of the example signal using the fuzzy masks. The results are again slightly different from those obtained with the two previous techniques, presented in Figs. 3 and 4. One apparent feature is the leakage of energy from the other components to the transient component at frequencies below about 5 kHz, shown in Fig. 6(b).

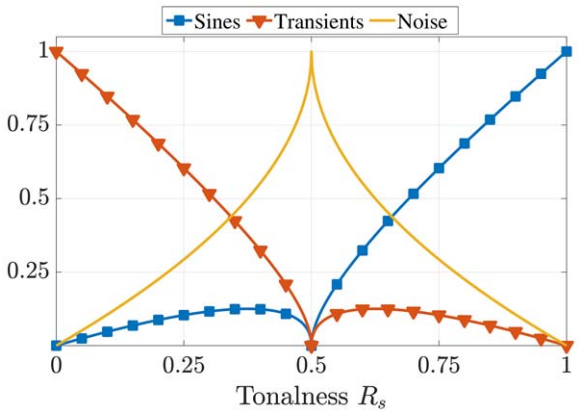

Fig. 5. Transient, noise, and sinusoidal masks, as used in the FZ method, which ensures perfect reconstruction [8].

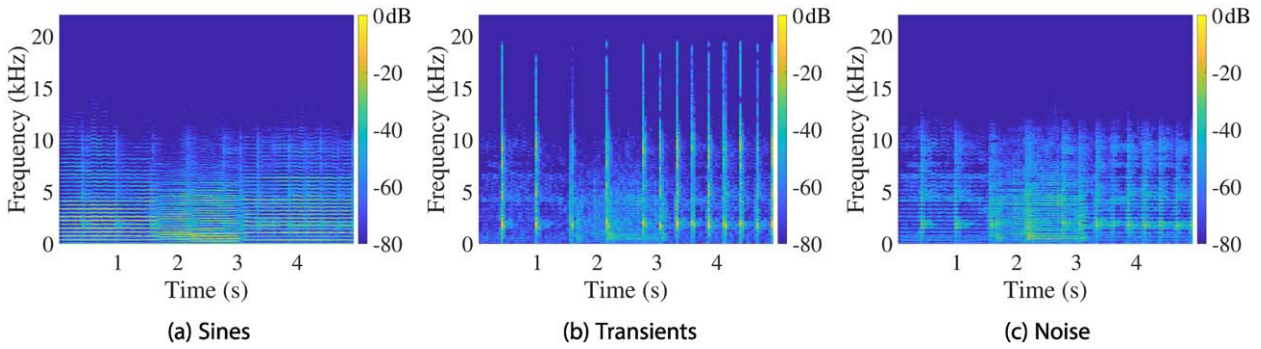

Fig. 6. STN decomposition obtained using the FZ method, cf. Figs. 3 and 4: (a) sines, (b) transients, and (c) noise.

2 PROPOSED METHOD

Fuzzy logic has proved useful in TSM, in which the separation of the STN components using a soft mask leads to the state-of-the-art performance [4]. However, the mask for each component must be designed carefully to obtain the best performance. For example, the FZ separation proposed in [8] is suboptimal for this task because of the leakage caused by the secondary lobes of the S and T masks and the peaky behavior of the N mask, which can be seen in Fig. 5. Ideally, a good fuzzy masking approach guarantees perfect reconstruction, smooth transitions between classes, and a well-defined dominant region per each class. In this section, a novel method comprising an extension to the HPR concept of clustered STN regions with soft masks resulting from fuzzy classification is proposed to fulfill this target.

2.1 Prototype Soft Masking

A prototype function meeting all the aforementioned requirements is the raised–cosine function, also known as the Hann window:

$$w(n) = \sin^2(\pi\, n/L), \qquad 0 \le n < L. \tag{21}$$

It is possible to take advantage of the symmetry of the raised–cosine function, using only its one wing (appropriately shifted) to describe the different transitions. The spectral masks for sines and transients can then be obtained as follows:

$$S(m,k) = \begin{cases} \sin^2\left[\pi(R_s(m,k) + \frac{1}{2})\right], & \text{if } R_s(m,k) \ge \frac{1}{2} \\ 0, & \text{otherwise,} \end{cases}$$

$$T(m,k) = \begin{cases} \sin^2\left[\pi(R_s(m,k) - \frac{1}{2})\right], & \text{if } R_s(m,k) \le \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

with $N(m, k)$ being computed according to Eq. (8). Their relationship is shown in Fig. 7. Although the masks defined in Eq. (22) already provide an audible improvement over the FZ masks, they are affected by a strong leakage of sines and transients into the noise component, suggesting that transitions between adjacent masks should be stricter.

2.2 Improving Noise Classification

Damskägg and Välimäki [4] suggested that the tonalness distribution of pure noise, e.g. white or pink noise, can be used to verify the shape of the noise mask $N(m, k)$. In the following, a description of the noise distribution over the tonalness and, consequently, sines-to-noise and transients-to-noise transitions is experimentally identified.

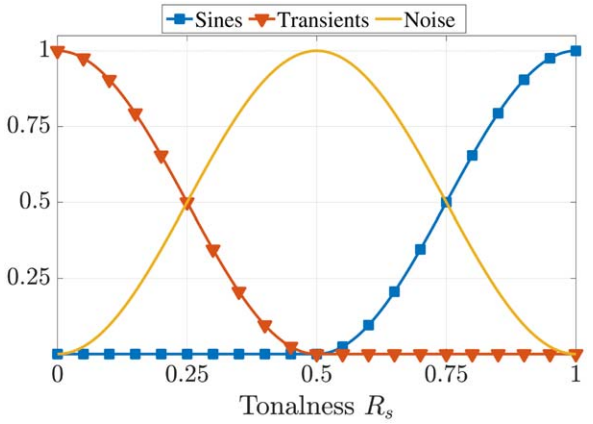

Fig. 7. Prototype soft masks for transients, noise, and sines, as computed from Eq. (22).

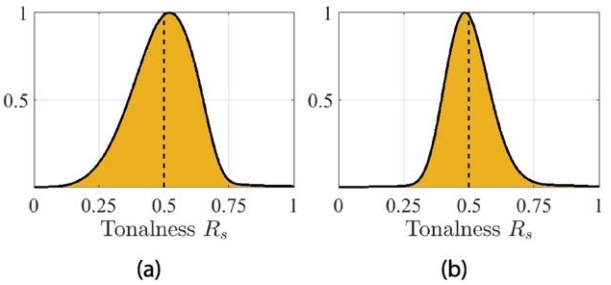

Fig. 8. Normalized tonalness distribution for median-filtered white noise with (a) long and (b) short analysis windows.

A set of 100 instances of random white noise were generated, and their tonalness was computed, independently, with a long window (185 ms, or $L = 8192$ samples at 44.1 kHz) and a short window (11 ms or $L = 512$ samples). Normalized histograms for tonalness values are shown, respectively, in Figs. 8a and 8b. A visual inspection indicates that the noise component remains relevant for a large range of tonalness values around 0.5 before quickly decaying in both directions. This suggests that mask transitions should be much steeper than in FZ (cf. Fig. 5). The success of hard

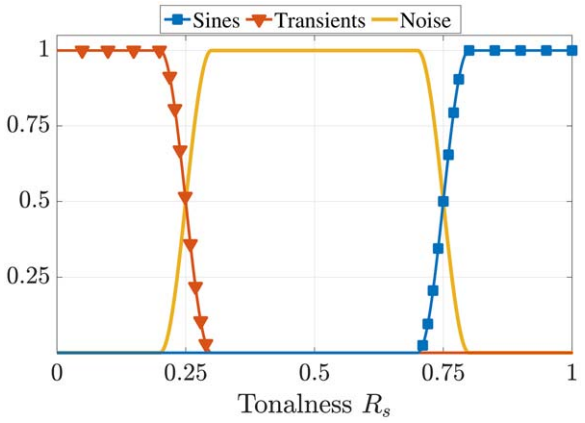

Fig. 9. Proposed enhanced transient, noise, and sines fuzzy masks for STN decomposition, $\beta_U = 0.8$, $\beta_L = 0.7$.

masking of HPR also indicates that a single-classification dominant region should be included in each mask.

It is also noted in Figs. 8a and 8b that the shape of the tonalness distribution of noise is asymmetric for different window lengths. The peak value is not centered at $R_s = 0.5$ but shifts towards one side or another depending on L. The two wings of the distribution have different degrees of steepness: a sharper "sinusoidal" (right) side and a more relaxed "transient" (left) side can be identified for longer L (Fig. 8a); the opposite behavior is exhibited for shorter L (Fig. 8b). As for HPR, this could lead to a two–stage decomposition featuring masks with different transition regions.

2.3 Enhanced Soft Masking with Fuzzy Logic

Following the considerations discussed here, a new set of masks can be derived by altering Eq. (22) to include dominant and cutoff regions for each mask while retaining the smoothness of the raised cosine function for the transition region. Parameters β_U and β_L are introduced to control the limits of the transition region and the bounds for, respectively, the dominant (upper) region and the cutoff (lower) region. The enhanced fuzzy masks are obtained as follows:

$$S(m,k) = f\left(R_s(m,k)\right), \quad T(m,k) = f\left(R_t(m,k)\right) \tag{23}$$

where

$$f(a) = \begin{cases} 1, & \text{if } a \geq \beta_U \\ \sin^2\left(\frac{\pi}{2}\frac{a-\beta_L}{\beta_U-\beta_L}\right), & \text{if } \beta_L \leq a < \beta_U \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

and $N(m,k)$ is computed using Eq. (8). The relationship of the proposed soft masks is shown in Fig. 9, when $\beta_U = 0.8$ and $\beta_L = 0.7$.

It has been shown earlier that, with hard masks, the two–stage STN decomposition yields better results than a single-stage separation [2, 30]. The same concept can be extended to fuzzy masks. Following Eqs. (10) and (13), two sets are obtained: $\{S_1, T_1, N_1, \beta_{U,1}, \beta_{L,1}\}$ for sines extraction with a large L, and $\{S_2, T_2, N_2, \beta_{U,2}, \beta_{L,2}\}$ for the residual transient and noise separation using a small L. The analysis process is summarized in Fig. 10.

Table 1. Transition area bounds for each decomposition stage used in this study.

Stage	Decomposition	β_U	β_L
1	Sines vs. Residual	0.8	0.7
2	Transients vs. Noise	0.85	0.75

2.4 Choosing the Transition Area

In order to find suitable values for the β_U and β_L parameters of the two decomposition stages, an optimization algorithm was run over the STN decomposition of a mixture of synthetic sounds, each belonging almost perfectly to a single class: a sum of sinusoids (for S), a short Gaussian monopulse[1] (for T), and a white noise sequence (for N). As the original sources are known, the decomposition error can be evaluated for each class. The goal is not to find a single optimal pair of values for the interval, as it is known that the quality of STN decomposition greatly varies with different audio inputs [2, 4]. Instead, a range of tonalness values for each bound yielding a small-enough decomposition error can be identified; the following analysis was conducted in order to find a pair of quasi-optimal values that suits multiple audio inputs.

The genetic algorithm [31] was chosen for the optimization process, which is divided in two stages. The first optimization run narrows down a set of paired bounds $B_1 = \{(\beta_{U,1}, \beta_{L,1})_i\}$ over the S_1 mask by minimizing the decomposition error over the sinusoidal part. Following that, different optimizations can be run by fixing a single pair $\{\beta_{U,1}, \beta_{L,1}\}$ from B_1 and finding its optimal pair $\{\beta_{U,2}, \beta_{L,2}\}$ over the T_2 mask. Finally, an audible comparison over multiple separations using the obtained sets determines the final quasi-optimal set $\bar{B} = \{(\beta_{U,1}, \beta_{L,1}), (\beta_{U,2}, \beta_{L,2})\}$, which ensures that the decomposition quality remains similar for different audio inputs. The results of this quasi-optimal choosing process are reported in Table 1.

Fig. 11 shows the separated STN components of the example audio signal using the proposed method with the chosen set $\bar{B}$. The separation results differ somewhat from the previous separation examples. In particular, the transients in Fig. 11(b) are unbroken, and the pauses between them are practically free from leakage from the other components, as desired.

3 EVALUATION

The audio quality of STN decomposition is typically degraded by intercomponent leakage, loss of tonality, loss of presence, or other artifacts, e.g., musical noise [32]. In previous works, the separation quality was evaluated by means of audio blind source separation performance assessment metrics, such as Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifacts Ratio

[1] $x_{GP}(t) = \sqrt{e}2\pi f_c t e^{-2(\pi f_c t)^2}$, where f_c is the center frequency.

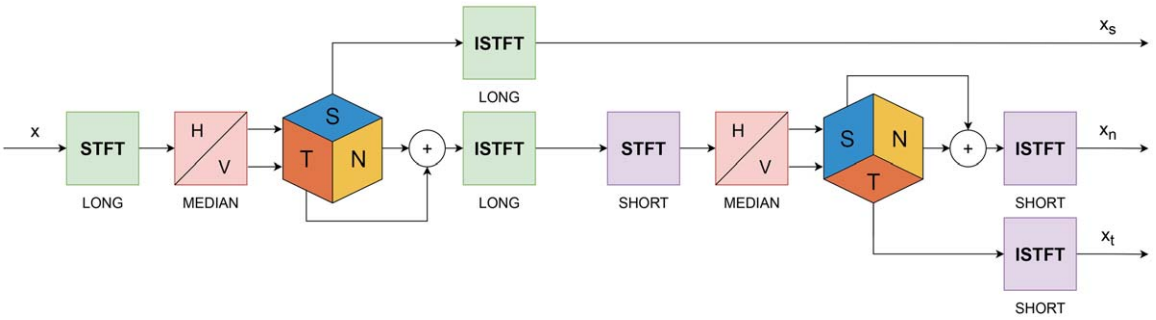

Fig. 10. Block diagram of the proposed two-stage STN decomposition method.

(SAR) [2, 3, 33]. However, those metrics require a mixture of three independent sources, one per class, that is subsequently decomposed again for the separation quality assessment. This prevents nonsynthetic audio inputs, e.g., music or speech, from being tested: unless dealing with perfectly tonal, impulsive, or noisy sources, the input sounds themselves are composed of a mixture of unknown STN parts.

Informal subjective listening tests on STN separation have previously been conducted by asking the participants whether the sinusoidal and transient components extracted by the method under test met the expectation of representing the sines and transients of the audio reference [2]. The STN decomposition algorithm proposed in this work is evaluated against other techniques by extending the same idea to a formal blind listening test, involving experienced listeners as participants and asking them to rate the quality of the sines and transients extraction for different STN methods.

3.1 Listening Test Design

A formal blind listening test was conducted on a selection of 19 experienced listeners, 17 of which reported previous experience in test design. No participant reported any hearing impairments or relevant medical conditions. The test software was run on a machine running MacOS 10.14.6, using a single pair of Sennheiser HD 650 headphones, inside a soundproof listening booth at the Aalto Acoustics Lab, Espoo, Finland.

A set of nine audio samples of short duration (4 to 6 s) was selected, consisting of two synthetic sounds and seven musical excerpts from various genres, featuring different spectral contents. The test samples are listed in Table 2.

Table 2. Audio samples used in the listening test.

Name	Description
Synth	Synthetic mix of tones, pulses, and white noise
CastViol	Solo violin and castanets, from [34]
ICanSee	Excerpt from *I Can See Clearly*, by *Holly Cole Trio*
Eddie	Excerpt from *Early in the Morning*, by *E. Rabbit*
Jazz	Mix of trumpet, piano, bass, and drums, from [34]
Vocals	Excerpt from *Tom's Diner*, by *Suzanne Vega*
Vibrato	Synthetic mix of vibrato, pulses, and pink noise
Billie	Intro of *Billie Jean*, by *Michael Jackson*
Drum	Solo performed on a drum set, from [34]

In each trial of the test, subjects were presented with one of the audio excerpts, referred to as *reference*, and were asked to blindly rate the quality of the extraction of the sound component under test (sines or transients, respectively) from such a reference, for four different STN decomposition methods: HPR, ST, FZ, and the proposed one (PROP). The following settings were used: the sample rate was 44.1 kHz, and the window and FFT lengths were $L_1 = 8{,}192$ samples for the first round and $L_2 = 512$ samples for the second round, with 75% overlap and Hann windowing. The length of the median filters was 500 Hz (93 bins in the first round and 6 in the second) in the frequency direction and 200 ms (4 bins in the first round and 69 in the second) in the time direction.

The original reference was also included among the samples under test, to provide a lower bound. Subjects were

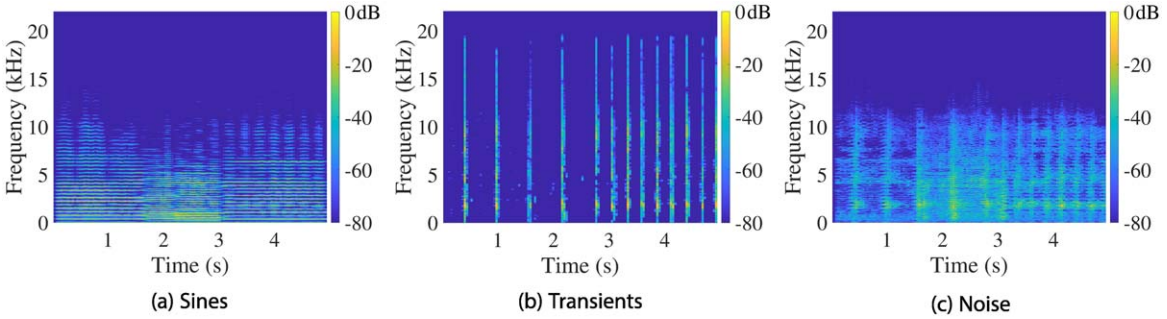

Fig. 11. STN decomposition of the castanets and violin using the proposed method, cf. Figs. 3, 4, and 6: (a) sines, (b) transients, and (c) noise.

asked to rate each sample on a scale from 0 to 100, with the specific request of assigning 0 to the sample identified as the reference. As there was no ground truth separation for any of the synthetic sounds, it is anticipated that none of the samples was perceived to be an ideal decomposition: hence, there was no obligation to rate any of the samples as 100 (full scale).

The test was divided in two parts consisting of nine trials each, the first focusing on sines extraction and the second on transients extraction, for a total of 18 trials and 90 audio samples under investigation. Listeners were allowed a short training session before starting the actual test to get acquainted with the interface, the keyboard shortcuts, and the task itself. The results of the training session were not included in the statistical analysis. Prior to the test, familiarity of the subjects with the concepts of sound separation and sines, transients, and noise was asserted. The experiment was conducted over WebMushra [35], although the designed test did not follow the MUSHRA recommendation. The processed audio excerpts and the test software are available at the companion webpage [36].

3.2 Results

Mean Opinion Scores (MOS) were computed from the ratings given by the subjects to estimate the quality of the decomposition methods under test. Boxplots displaying the distribution of the data for the sines and transients extraction are shown, respectively, in Figs. 12(a) and 12(b). Data distribution can also be observed via histograms, available at the companion webpage [36].

Overall, the proposed method consistently performed better than or as well as HPR for every audio excerpt for both sinusoidal and transient decompositions. The proposed methods achieved a median MOS score between 3.0 and 5.0 ("good" to "excellent") in all but one test case; see Figs. 12(a) and 12(b). The HPR method was below 3.0 in two cases. The ST method generally scored intermediate values, with the exception of *Vibrato*, in which ST outperformed the other separation methods: this was expected, considering that ST was designed purposely for sound mixtures presenting vibrato. FZ was the lowest-ranked algorithm with a significant difference in the subjective ratings from the other three methods, as appears from Figs. 12(a) and 12(b).

Considering the sines extraction only, the proposed method performed quite similarly to HPR in Fig. 12(a) and significantly better only for the *Vocals* and *Drum* excerpts, in which the fuzzy soft masking helped in the preservation of the tonality variations. However, the median value for the proposed method was consistently higher than that of the HPR method.

A larger improvement was observed in the transient decomposition in Fig. 12(b). Considering the median of the distributions, the proposed method surpassed HPR for all excerpts but *Icansee* and *Jazz*. ST proved to be a competitive algorithm for *Vocals*. The large amount of variance in the data came from the absence of a proper separation "reference," i.e., an upper limit for the subjective grading in each trial. The subjects had to apply their own scale during the grading process, which consequently lead to data that are distributed in a non-Gaussian fashion. This was also confirmed by an inspection of data skewness and kurtosis, which is reported in the companion website [36].

Further analysis was conducted on the results to assess statistical significance in the data distribution, i.e., the difference between the distribution of data for different methods had statistical significance. In this case, the observation of non-Gaussian distributions called for a nonparametric paired difference test. The Wilcoxon signed-rank test was chosen for the task. For a 95% confidence interval, statistical significance was achieved if the signed-rank test returned a p value below the threshold $\alpha = 0.05$. Thresholded resulting p values for sinusoidal and transient separation data are shown, respectively, in Figs. 13 and 14.

The results showed that statistical significance for the difference in data distribution from the proposed and the competitive methods was achieved for at least one of the two components (sinusoidal or transient separation) for seven excerpts out of nine, with the transient separation being the discriminant factor in six cases out of seven. *Icansee* and *Jazz* were the two most complex mixtures among the collected audio samples: this suggested that the more complex the sound mixture was, the harder it became to discriminate the separation performance. The traditional statistical analysis via ANOVA and paired t test, carrying similar results, is reported in the companion website [36].

4 APPLICATION TO TIME STRETCHING

The proposed method is adapted to audio TSM, which is a suitable application for the STN decompositions [4, 11, 34]. For this purpose, the fuzzy phase vocoder (FPV) developed by Damskägg and Välimäki [4], which received the highest average score in a recent comparison of audio time-stretching methods [37], is modified to include the proposed decomposition method. The refined separation allows for the transients to be preserved and repositioned onto the stretched time axis [38], while the sinusoidal and noise components are processed via the phase vocoder with identity phase locking and phase randomization, respectively.

The enhanced TSM processing for a section of the *CastViol* excerpt that has been slowed down to half speed is shown in Fig. 15. The sound processed with the original FPV visibly suffers from transients smearing, a recurring phenomenon in phase-vocoder-based TSM [39], With the STN-enhanced version, the transients appear much sharper and better resemble the ones visible in the original signal.

4.1 Comparison

A preference test was conducted to compare the original FPV [4] with the proposed STN-modified one (PROP), in order to evaluate the enhancement brought by STN decomposition to a suitable method. Eleven experienced listeners participated in the test, which was realized on the same hardware and with similar modalities as the one described in SEC. 3.1. Subjects were asked to listen to a reference

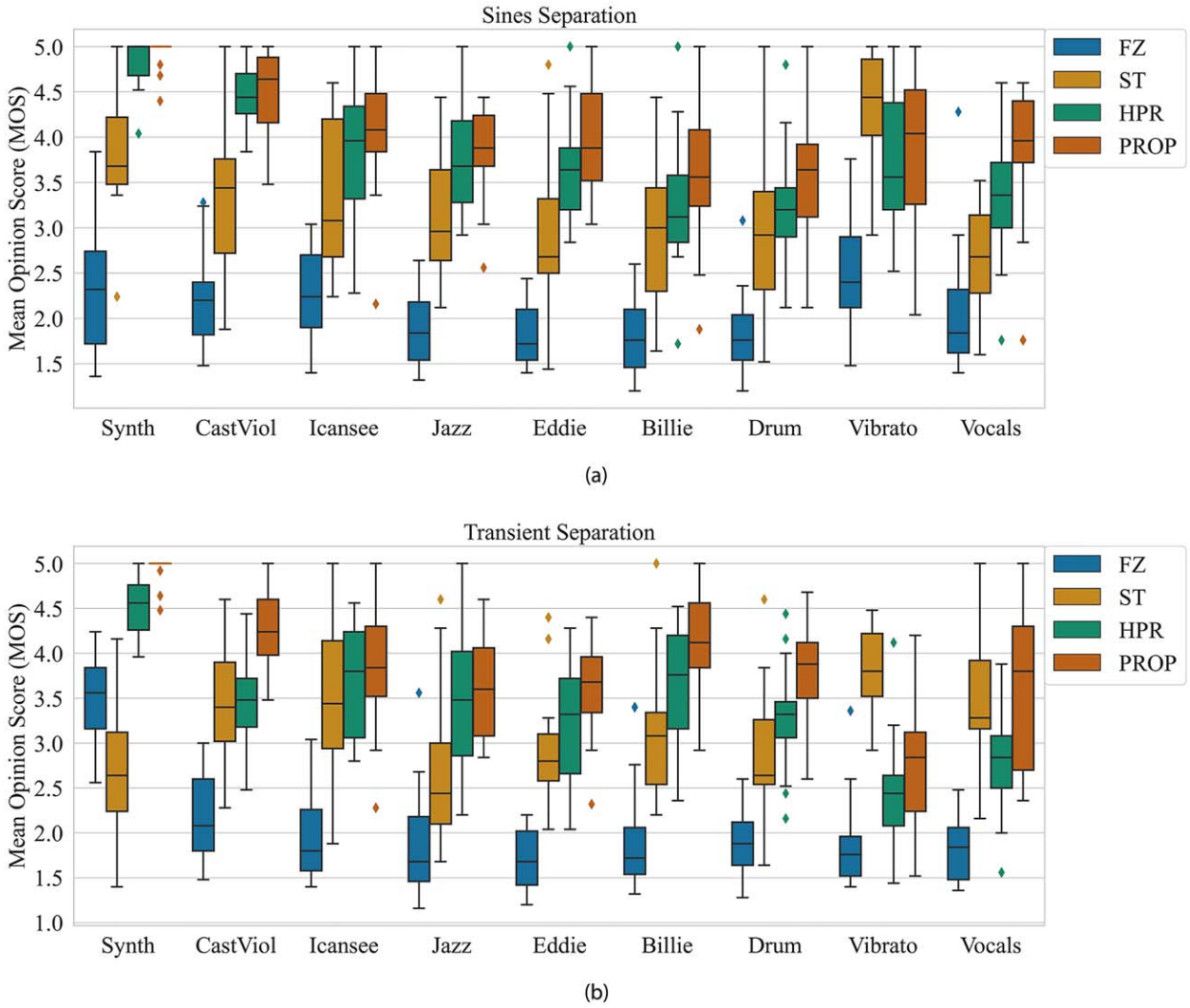

Fig. 12. MOS and confidence intervals for (a) sines and (b) transients separation of nine audio samples, showing that the proposed method is the winner or among the best methods in almost all cases.

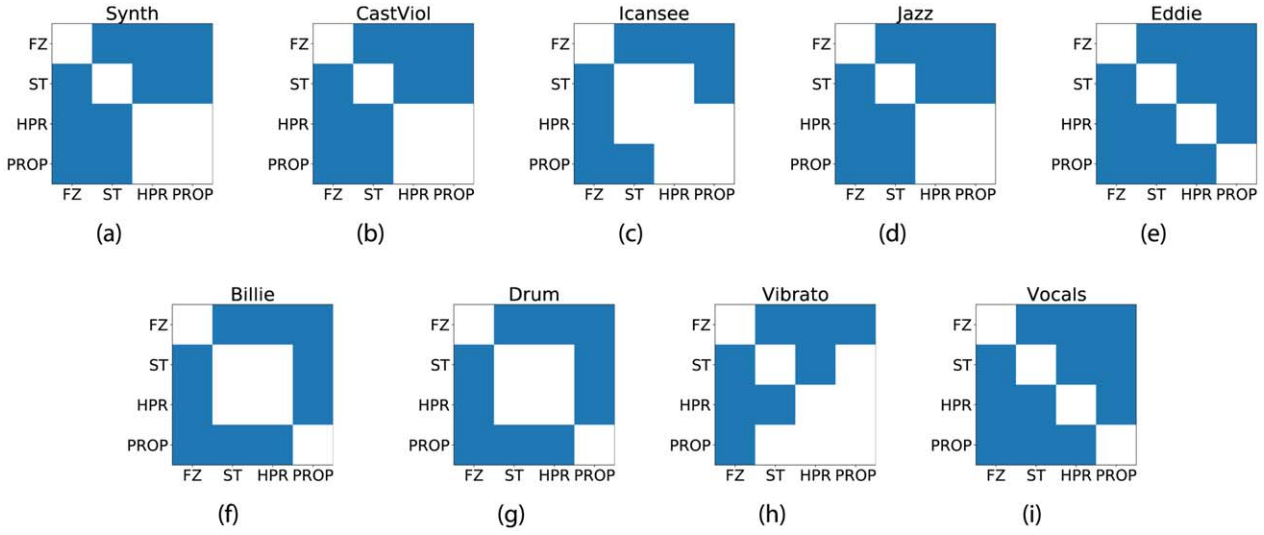

Fig. 13. Thresholded p values resulting from the Wilcoxon signed-rank test over sinusoidal separation data. Statistical significance ($p \leq \alpha$, $\alpha = 0.05$) is highlighted by coloring the cell. (a) *Synth*, (b) *CastViol*, (c) *Icansee*, (d) *Jazz*, (e) *Eddie*, (f) *Billie*, (g) *Drum*, (h) *Vibrato*, and (i) *Vocals*.

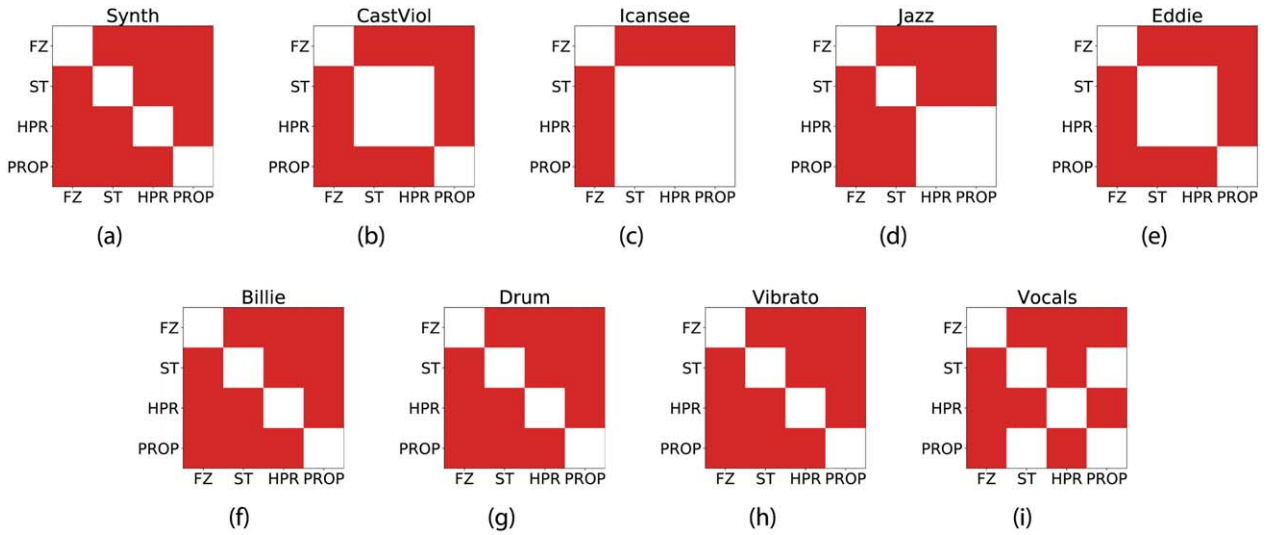

Fig. 14. Thresholded p values of transient separation data, cf. Fig. 13. Statistical significance ($p \leq \alpha$, $\alpha = 0.05$) is highlighted by coloring the cell. (a) *Synth*, (b) *CastViol*, (c) *Icansee*, (d) *Jazz*, (e) *Eddie*, (f) *Billie*, (g) *Drum*, (h) *Vibrato*, and (i) *Vocals*.

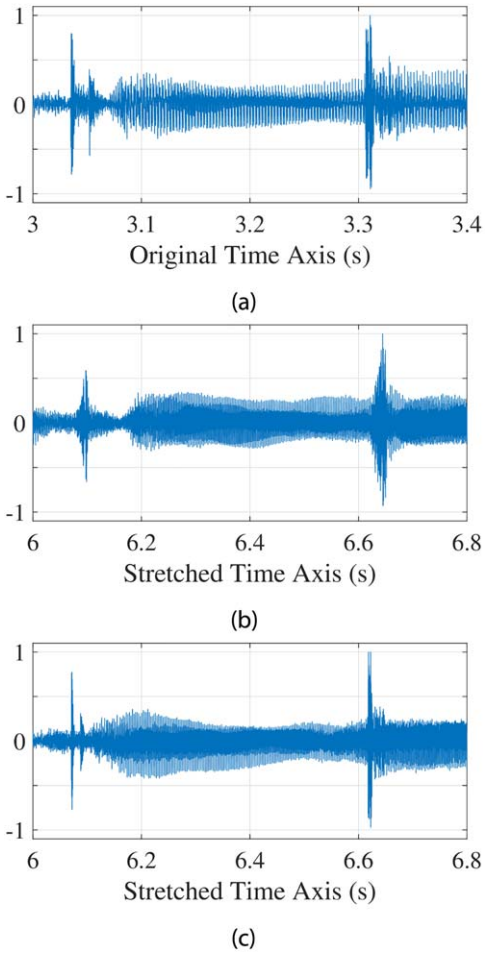

Fig. 15. Comparison between (a) a section of *CastViol* and its modifications via (b) fuzzy phase vocoder (FPV) [4] and (c) its STN enhancement (PROP), for a TSM factor of 2. Note the different time scales in the subfigures.

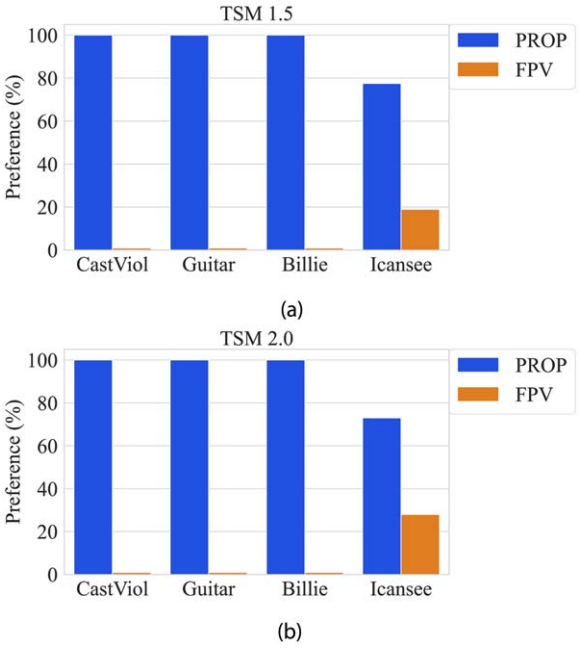

Fig. 16. Subject preferences between original FPV [4] and the STN modification (PROP), for TSM factors (a) 1.5 and (b) 2.

sound and to select their preferred time-stretched version in terms of sound quality from two options, processed respectively with FPV and PROP. Four audio excerpts (*CastViol, Billie*, and *Icansee* from the previous test, plus a *Guitar* plucking sound) were time-stretched with factors 1.5 and 2. Loudness normalization was applied to compensate for the nonperfect reconstruction of FPV. Results are visualized in Fig. 16, and all test sounds are available on the companion webpage [36].

Test subjects showed a strong preference for PROP over all samples for both time-stretching factors, as can be seen in Figs. 16(a) and (b). A minority expressed a preference for

FPV for *Icansee*, mentioning that although the transients' "punch" was well retained by the other method (PROP), it created a dissonance with the noisy part of the transient, which goes through the phase vocoder and phase randomization and is heavily smeared. A different processing for the noise component, which can now be isolated through the STN decomposition, could further improve the audio time-stretching performance.

5 CONCLUSION

In this paper, the three-way sound decomposition into sines, transients, and noise using fuzzy logic was enhanced. A set of soft spectral masks was derived to fulfill the task while preserving the perfect reconstruction property. Using such soft masks, the novel two-stage STN decomposition method proposed in this paper allows a single spectral bin to be simultaneously classified either as sine and noise, or as transient and noise. Soft masking positively affects the decomposition by attenuating or removing common artifacts, e.g., musical noise or loss of transient presence.

The results of a subjective listening test against three other methods showed that the proposed decomposition method typically improves the separation quality in terms of transient extraction, with a comparable performance for sines extraction with the previous best method. It was also shown how the complexity of the audio signal affects the quality of the decomposition. For instance, the proposed separation method struggles when the sinusoidal part contains vibrato, as does the competing previous method.

The proposed method can help improve sound quality in many audio processing tasks. A successful application to audio time stretching was shown to improve the performance of the state-of-the-art algorithm.

6 ACKNOWLEDGMENTS

This work belongs to 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. The authors are grateful to Dennis Bontempi for the helpful discussions and to Alec Wright for proofreading.

7 REFERENCES

[1] T. S. Verma and T. H. Y. Meng, "Extending Spectral Modeling Synthesis With Transient Modeling Synthesis," *Comput. Music J.*, vol. 24, no. 2, pp. 47–59 (2000 Jun.). https://doi.org/10.1162/014892600559317.

[2] J. Driedger, M. Müller, and S. Disch, "Extending Harmonic-Percussive Separation of Audio Signals," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pp. 611–616 (Taipei, Taiwan) (2014 Oct.).

[3] R. Füg, A. Niedermeier, J. Driedger, S. Disch, and M. Müller, "Harmonic-Percussive-Residual Sound Separation Using the Structure Tensor on Spectrograms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 445–449 (Shanghai, China) (2016 Mar.). https://doi.org/10.1109/ICASSP.2016.7471714.

[4] E.-P. Damskägg and V. Välimäki, "Audio Time Stretching Using Fuzzy Classification of Spectral Bins," *Appl. Sci.*, vol. 7, no. 12, paper 1293 (2017 Dec.), https://doi.org/10.3390/app7121293.

[5] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stajylakis, "Music Tempo Estimation and Beat Tracking by Applying Source Separation and Metrical Relations," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 421–424 (Kyoto, Japan) (2012 Mar.). https://doi.org/10.1109/ICASSP.2012.6287906.

[6] Á. Faraldo Pérez, *Tonality Estimation in Electronic Dance Music: A Computational and Musically Informed Examination*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain (2018 Mar.).

[7] B. Lentz, A. Nagathil, J. Gauer, and R. Martin, "Harmonic/Percussive Sound Separation and Spectral Complexity Reduction of Music Signals for Cochlear Implant Listeners," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8713–8717 (Barcelona, Spain) (2020 May). https://doi.org/10.1109/ICASSP40776.2020.9052920.

[8] E. Moliner, J. Rämö, and V. Välimäki, "Virtual Bass System With Fuzzy Separation of Tones and Transients," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx)*, pp. 86–93 (Vienna, Austria) (2020 Sep.).

[9] F. X. Nsabimana and U. Zölzer, "Audio Signal Decomposition for Pitch and Time Scaling," in *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 1285–1290 (St. Julians, Malta) (2008 Mar.). https://doi.org/10.1109/ISCCSP.2008.4537424.

[10] J. Driedger, M. Müller, and S. Ewert, "Improving Time-Scale Modification of Music Signals Using Harmonic-Percussive Separation," *IEEE Signal Process. Letters*, vol. 21, no. 1, pp. 105–109 (2013 Jan.). https://doi.org/10.1109/LSP.2013.2294023.

[11] L. Fierro and V. Välimäki, "Towards Objective Evaluation of Audio Time-Scale Modification Methods," in *Proceedings of the 17th Sound and Music Computing Conference (SMC)*, pp. 457–462 (Malaga, Spain) (2020 Jun.).

[12] J. W. Beauchamp, "Additive Synthesis of Harmonic Musical Tones," *J. Audio Eng. Soc.*, vol. 14, no. 4, pp. 332–342 (1966 Oct.).

[13] J.-C. Risset and M. V. Mathews, "Analysis of Musical-Instrument Tones," *Phys. Today*, vol. 2, no. 22, pp. 23–30 (1969 Feb.). https://doi.org/10.1063/1.3035399.

[14] J. A. Moorer, "Signal Processing Aspects of Computer Music: A Survey," *Proc. IEEE*,

vol. 65, no. 8, pp. 1108–1137 (1977 Aug.). https://doi.org/10.1109/PROC.1977.10660.

[15] X. Serra and J.III Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Comput. Music J.*, vol. 14, no. 4, pp. 12–24 (1990 Winter). https://doi.org/10.2307/3680788.

[16] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 4, pp. 744–754 (1986 Aug.). https://doi.org/10.1109/TASSP.1986.1164910.

[17] T. S. Verma, S. N. Levine, and T. H. Y. Meng, "Transient Modeling Synthesis: A Flexible Analysis/Synthesis Tool for Transient Signals," in *Proceedings of the International Computer Music Conference (ICMC)*, pp. 164–167 (Thessaloniki, Greece) (1997 Sep.).

[18] T. S. Verma and T. H. Y. Meng, "An Analysis/Synthesis Tool for Transient Signals That Allows a Flexible Sines+Transients+Noise Model for Audio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6, pp. 3573–3576 (Seattle, WA, USA) (1998 May). https://doi.org/10.1109/ICASSP.1998.679647.

[19] S. N. Levine and J. O.III Smith, "A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications," presented at the *105th Convention of the Audio Engineering Society* (1998 Sep.), paper 4781.

[20] D. Fitzgerald, "Harmonic/Percussive Separation Using Median Filtering," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)* (Graz, Austria) (2010 Sep.).

[21] D. Fitzgerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/Percussive Separation Using Kernel Additive Modelling," in *Proceedings of the IET Irish Signals and Systems Conference (ISSC)* (Limerick, Ireland) (2014 Jun.). https://doi.org/10.1049/cp.2014.0655.

[22] F. J. Canadas-Quesada, D. Fitzgerald, P. Vera-Candeas, and N. Ruiz-Reyes, "Harmonic-Percussive Sound Separation Using Rhythmic Information From Non-negative Matrix Factorization in Single-Channel Music Recordings," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx)*, pp. 276–282 (Edinburgh, UK) (2017 Sep.).

[23] J. Neri and P. Depalle, "Fast Partial Tracking of Audio With Real-Time Capability Through Linear Programming," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx)*, pp. 326–333 (Aveiro, Portugal) (2018 Sep.).

[24] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Phase-Aware Harmonic/Percussive Source Separation via Convex Optimization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 985–989 (Brighton, UK) (2019 May). https://doi.org/10.1109/ICASSP.2019.8683821.

[25] K. Drossos, P. Magron, S. I. Mimilakis, and T. Virtanen, "Harmonic-Percussive Source Separation With Deep Neural Networks and Phase Recovery," in *Proceedings of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 421–425 (Tokyo, Japan) (2018 Sep.). https://doi.org/10.1109/IWAENC.2018.8521371.

[26] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. Fitzgerald, and B. Pardo, "An Overview of Lead and Accompaniment Separation in Music," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 8, pp. 1307–1335 (2018 Aug.). https://doi.org/10.1109/TASLP.2018.2825440.

[27] S. Makino, T.-W. Lee, and H. Sawada, eds., *Blind Speech Separation* (Springer, Dordretcht, the Netherlands, 2007).

[28] J. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 25, no. 3, pp. 235–238 (1977 Jun.). https://doi.org/10.1109/TASSP.1977.1162950.

[29] J. Bonada, "Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio," in *Proceedings of the International Computer Music Conference (ICMC)* (Berlin, Germany) (2000 Aug.).

[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, no. 1, pp. 228–237 (2014 Jan.). https://doi.org/10.1109/TASLP.2013.2287052.

[31] S. Mirjalili, *Evolutionary Algorithms and Neural Networks* (Springer, Cham, Switzerland, 2019).

[32] L. Fierro and V. Välimäki, "SiTraNo: A MATLAB App for Sines-Transient-Noise Decomposition of Audio Signals," in *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx)*, pp. 73–80 (Vienna, Austria) (2021 Sep.).

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469 (2006 Jul.). https://doi.org/10.1109/TSA.2005.858005.

[34] J. Driedger and M. Müller, "TSM Toolbox: MATLAB Implementations of Time-Scale Modification Algorithms," in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx)*, pp. 249–256 (Erlangen, Germany) (2014 Sep.).

[35] M. Schoeffler, S. Bartoschek, F.-R. Stöter, et al., "WebMUSHRA–A Comprehensive Framework for Web-Based Listening Tests," *J. Open Res. Softw.*, vol. 6, no. 1, paper 8 (2018 Feb.). http://doi.org/10.5334/jors.187.

[36] L. Fierro and V. Välimäki, "Enhanced Fuzzy Decomposition of Sound into Sines, Transients and Noise: Companion Web Page," http://research.spa.aalto.fi/publications/papers/jaes-stn (accessed Oct. 19, 2022).

[37] 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, no. 9, pp. 644–655 (2021 Sep.). https://doi.org/10.17743/jaes.2021.0031.

[38] F. Nagel and A. Walther, "A Novel Transient Handling Scheme for Time Stretching Algorithms," presented at the *127th Convention of the Audio Engineering Society* (2009 Oct.), paper 7926.

[39] A. Röbel, "A New Approach to Transient Processing in the Phase Vocoder," in *Proceedings of the Sixth International Conference on Digital Audio Effects (DAFx)*, pp. 344–349 (London, UK) (2003 Sep.).

THE AUTHORS

Leonardo Fierro

Vesa Välimäki

Leonardo Fierro is a doctoral researcher at Aalto University, Espoo, Finland. He received his M.Sc. degree in Communication Technologies and Multimedia from the University of Brescia, Italy, in 2019. He has been part of the Audio Signal Processing group of the Aalto Acoustics Lab since 2019. His research interests involve audio time–scale modification, transient processing, and loudness equalization.

•

Vesa Välimäki is a Full Professor of Audio Signal Processing and Vice Dean for Research at Aalto University, Espoo, Finland. He received his M.Sc. and D.Sc. degrees from the Helsinki University of Technology in 1992 and 1995, respectively. In 1996, he was a postdoctoral researcher at the University of Westminster, London, UK. In 2008–2009, he was a visiting scholar at the Stanford University Center for Computer Research in Music and Acoustics (CCRMA). His research interests are in signal processing and machine learning applied to audio and music technology. Prof. Välimäki is a Fellow of the AES and a Fellow of the IEEE. He was the General Chair of the 11th International Conference on Digital Audio Effects DAFx in 2008 and of the 14th International Sound and Music Computing Conference SMC in 2017. He is the Editor-in-Chief of the *Journal of the Audio Engineering Society*.