# Sample-based engine noise synthesis using an enhanced pitch-synchronous overlap-And-Add method

**Article** *in* The Journal of the Acoustical Society of America · November 2012

DOI: 10.1121/1.4754663 · Source: PubMed



**3 authors:**

Jan Jagla
Scientific and Technical Center for Building
**16** PUBLICATIONS **115** CITATIONS
SEE PROFILE

Julien Maillard
Scientific and Technical Center for Building
**51** PUBLICATIONS **509** CITATIONS
SEE PROFILE

Nadine Martin
ASTRIIS
**135** PUBLICATIONS **1,287** CITATIONS
SEE PROFILE

# Sample-based engine noise synthesis using an enhanced pitch-synchronous overlap-and-add method

Jan Jagla[a] and Julien Maillard
*Centre Scientifique et Technique du Bâtiment, Paris-Est University, 24 rue Joseph Fourier, 38400 Saint Martin d'Hères, France*

Nadine Martin
*GIPSA-Lab - UMR 5216, Grenoble Institute of Technology / CNRS, 11 rue des Mathématiques, 38400 Saint Martin d'Hères, France*

(Dated: September 3, 2012)

An algorithm for the real time synthesis of internal combustion engine noise is presented. Through the analysis of a recorded engine noise signal of continuously varying engine speed, a dataset of sound samples is extracted allowing the real time synthesis of the noise induced by arbitrary evolutions of engine speed. The sound samples are extracted from a recording spanning the entire engine speed range. Each sample is delimitated such as to contain the sound emitted during one cycle of the engine plus the necessary overlap to ensure smooth transitions during the synthesis. The proposed approach, an extension of the PSOLA method introduced for speech processing, takes advantage of the specific periodicity of engine noise signals to locate the extraction instants of the sound samples. During the synthesis stage, the sound samples corresponding to the target engine speed evolution are concatenated with an overlap and add algorithm. It is shown that this method produces high quality audio restitution with a low computational load. It is therefore well suited for real time applications.

PACS numbers: 43.50.Lj, 43.50.Rq, 43.60.Ek

## I. INTRODUCTION

The prediction of road traffic noise in urban areas is extensively documented in the literature[1–6]. The majority of these studies focus on the estimation of road traffic noise levels. However, the comfort or discomfort of a listener regarding the sound emitted by a vehicle not only depends on the perceived sound level but also on more complex indicators related to the spectral content and time properties of the sound field. A convenient way to assess such indicators is to synthesize the time domain signal perceived by the listener in the situation of interest and to perform perceptual evaluation of the sound field using an appropriate 3D sound restitution system. This approach is also referred to as auralization of the calculated sound field[7,8]. In the case of traffic noise, the auralization process involves three main tasks: the synthesis of the source signals (such as engine and rolling noise), their filtering to model the sound propagation and the spatial rendering of each individual contribution. This paper focuses on the engine noise signal synthesis. The technique described here is patent pending.

In 2001, Amman and Das[9] presented an engine noise synthesis method based on a deterministic-stochastic signal decomposition widely used in speech processing. The deterministic part is analyzed and re-synthesized using additive synthesis and then subtracted from the input signal to isolate and analyze the stochastic component. In the synthesis stage, framewise additive synthesis in the Fourier domain and a multipulse excited autoregressive model are used to generate the deterministic and the stochastic components of non-stationary engine sounds respectively. A similar signal decomposition is used in the HARTIS (Harmonic Real Time Synthesis) software[10]. In this approach, the harmonic part is also synthesized by additive synthesis and the stochastic part is generated by a smooth overlap granular synthesis method. These two methods involve at least one inverse Fourier transform during synthesis, making them less suited to real time applications.

Another approach also related to speech processing was introduced by Van Rensburg[11] in 2006. It uses the phase vocoder to simulate varying speed engine noise by time scaling and pitch shifting of a single engine sound at a given rotational speed. The main drawback of this method is that the relative amplitudes of engine noise harmonics are supposed to be independent from the rotational speed of the engine. Heitbrink[12] suggested a method that overcomes this problem by using a weighted sum of pitch shifted engine sound signals at different rotational speeds.

Recently, Zhekova[13] proposed a method for the simulation of a diesel engine noise at idle. It is based on a time-frequency analysis using the Gabor Transform and its derivatives to estimate the input parameters of a granular synthesis algorithm. Although this approach allows a precise characterization of the sound produced during an engine cycle at idle, it has not been extended to other engine speeds yet.

Another study[14] discussed the possibility to assemble sequentially samples of engine sounds at different speeds

[a] Author to whom correspondence should be addressed. Electronic mail: jan.jagla@cstb.fr

and loads. However, the approach for constructing the sample dataset is not described. The authors also mention that the dataset must be created out of a large set of recordings.

The method presented in the present paper is based on the principles of the overlap-and-add (OLA) signal processing algorithm and of some of its extensions such as the synchronous OLA method (SOLA)[15],the pitch-synchronous OLA method (PSOLA)[16,17] and the waveform similarity OLA method (WSOLA)[18]. All these methods were originally developed for speech processing and more specifically for pitch-scale and time-scale modifications of speech. However, in these previous works, the sound samples extracted from one input signal always appear in the same order in the synthesized signal. Note that Lent[19] described a pitch synchronous approach for time-scale and pitch shifting of music signals even before the above techniques were introduced for speech.

The present technique creates a dataset of sound samples ordered with respect to their fundamental frequency. The dataset is constructed such that any continuous evolution of the fundamental frequency can be synthesized in real time by overlapping and adding the sound samples. As for the PSOLA or Lent's methods, the instants of extraction in the input signal of the sound samples (also referred to as pitch marks) are chosen in a pitch synchronous way. However, the fundamental frequency component of engine noise signals is time varying and often masked by low frequency noise. Therefore, the pitch mark location process of the PSOLA and Lent's methods cannot be applied directly.

This paper is organized as follows. Section II introduces the structure of engine noise signals and the engine noise model considered in this study. Section III presents an overview of the existing OLA methods and describes the proposed analysis/synthesis scheme. Section IV details the analysis part of the proposed algorithm, including the estimation of the instantaneous fundamental frequency (pitch) of recorded sound signals, the location of pitch marks and the shape of the sample extraction window. Section V describes the synthesis part of the proposed algorithm including the overlap-and-add method, the sample selection and a discussion on the computational efficiency of the algorithm. A validation of the proposed pitch mark selection model together with a comparison with the WSOLA method is presented in Section VI. Finally, conclusions about this work are presented in Section VII.

## II. INPUT SIGNAL

### A. Engine noise characteristics

In the following, the four-stroke internal combustion engine is considered for simplicity as it is used for most production cars, trucks and motor bikes. Two-stroke engines are still in use in low-power motorbikes but their environmental impact forces the manufacturers to promote their four-stroke counterparts.

Independently from the number of cylinders and their disposition, an engine noise is essentially harmonic due to the sequential explosions occurring in the combustion chamber of each cylinder. The explosion frequency $F_{ex}$ for a four-stroke engine is given by

$$F_{ex} = \frac{N_c}{2}\frac{rpm}{60}, \tag{1}$$

where $N_c$ is the number of cylinders in the engine and $rpm$ is the number of revolutions per minute. In the case of a four-cylinder engine, explosions occur periodically in each cylinder with two explosions per engine revolution, while for a six-cylinder engine, there are three explosions per engine revolution. In practice, the harmonics of $F_{ex}$ are the prevailing components of engine noise.

Two successive explosions do not occur in the same cylinder and thus are not identical. Furthermore, two successive engine revolutions do not contain the explosions of the same cylinders. On that account, two other frequencies and their harmonics enrich the spectrum of an engine sound,

$$F_{en} = \frac{rpm}{60}, \tag{2}$$

$$F_c = \frac{1}{2}\frac{rpm}{60}, \tag{3}$$

where $F_{en}$ is the frequency corresponding to the engine revolution and $F_c$ is the fundamental frequency corresponding to a complete cycle of the engine (two revolutions). As an illustration, Figure 1 shows the spectrum of a noise emitted by a four-cylinder engine running at $2,200$ `rpm`. The locations of the different frequencies given by Eqs. 1 to 3 are also depicted in Figure 1.

### B. Engine noise recording

The analysis-synthesis method proposed in this paper requires a specific engine noise recording as input to the analysis stage. Such a recording can be obtained by placing a microphone near the engine compartment and running the engine with no gear engaged over the available engine speed range. The variation of the engine speed (and thus of the fundamental frequency $F_c$) must be slow enough to ensure its local stationarity on at least one cycle duration. Besides, the recording must be long enough so that the analysis algorithm can extract enough distinct sound samples for the synthesis. This issue will be further discussed in Section IV.C.

### C. Signal model

In this work, the noise emitted by an internal combustion engine is supposed to be fully characterized by its fundamental frequency. An engine noise at constant speed can be modeled as

$$x_{F_0}[n] = \sum_k A_{F_0,k}\sin(2\pi\frac{kF_0}{F_s}n + \phi_{F_0,k}) + w_{F_0}[n], \tag{4}$$

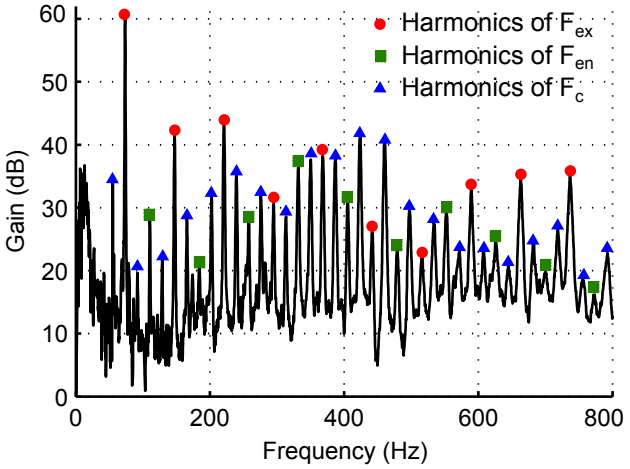


FIG. 1. Spectrum of the low frequency content of the sound signal emitted by a four-cylinder engine running at $2,200$ rpm. The harmonics are separated in three inclusive families (the harmonics of $F_{ex}$ are also harmonics of $F_{en}$ which are harmonics of $F_c$) to emphasize their physical origin. The components $F_c$ and $2F_c$ are masked by noise. (Color online)

where $n$ is the discrete time index, $F_0 = F_c$ is the fundamental frequency, $A_{F_0,k}$ and $\phi_{F_0,k}$ are the amplitudes and the initial phases of the harmonics respectively and $w_{F_0}[n]$ is the stochastic component of the engine noise. It is assumed that the amplitudes and phases of the harmonics and the spectral coloration of the stochastic noise depend solely on the fundamental frequency $F_0$. In the following, the order of the most energetic component is noted $k_{Amax}$. In practice, for a $N_c$-cylinder engine $k_{Amax} = N_c$.

## III. ANALYSIS-SYNTHESIS METHOD

### A. EXISTING OLA METHODS

Unlike engine noise signals, speech signals cannot be defined solely by their fundamental frequency, even the harmonic voiced regions must also be characterized by their formants. However, the principles of various analysis methods of existing OLA algorithms intended for speech processing can also be used to create sample datasets suitable for the synthesis of engine noise signals. As these methods are the basis of our approach, a quick review of their principles is presented below.

Existing OLA methods extract sound samples from the input signal and rearrange them by overlap-and-add to achieve the expected signal modification. In the following, the location and extraction of the sound samples from the input signal is referred to as the analysis stage and the construction of the new signal as the synthesis stage. The positions in the input signal of the extracted sound samples are termed pitch marks, following the PSOLA terminology.

The SOLA system, introduced by Roucos and Wilgus[15] for time scale modifications of speech, asynchronously extracts sound samples from the input signal during the analysis stage. During the synthesis stage, the extracted samples are synchronized with an autocorrelation technique to maintain the pitch period. This algorithm ensures good restitution quality but requires the calculation of correlation functions during the synthesis stage which is not optimal for real time applications.

The WSOLA system presented by Verhelst and Roelands[18] attempts to maximize the waveform similarity between the original and the synthesized speech. The location of each sound sample to be extracted is adjusted so as to maximize its cross-correlation with the natural continuity in the input signal of the last segment added to the synthesized signal. The principles of the WSOLA method can be used to create a dataset of sound samples by maximizing the cross-correlation between the sample to be extracted and already extracted samples with a similar fundamental frequency. However, it will be shown in Section VI that the harmonics of $F_{en}$ and $F_c$ cannot be preserved by this cross-correlation approach.

In the PSOLA system developed by Moulines and Charpentier[16], the pitch marks are placed synchronously with the local pitch period by detecting the glottal impulses of speech signals. This allows the synthesis of pitch shifted signals by overlapping and adding the extracted samples synchronously with a different pitch period. As discussed by Moulines[20], the perceptual effects of OLA operations are difficult to apprehend. In the specific case of the PSOLA system, these effects have been studied by Kortekaas[17,21] on constant fundamental frequency natural speech. For varying fundamental frequency signals, the effects of such methods are even more difficult to predict.

The method introduced by Lent[19] for music signals uses similar time scaling and pitch shifting methods as the PSOLA framework but also provides an interesting method for pitch mark location. It consists in determining the fundamental frequency of the input signal and filtering out this component by means of a band pass digital filter. The zerocrossings of the filtered signal provide information about the periodicity of the input signal. Although this approach is bounded to signals with stationary fundamental frequency, it allows the extraction of single periods of a musical signal.

### B. A NEW PSOLA APPROACH FOR ENGINE NOISE

A new analysis synthesis framework for engine noise signals is proposed in order to account for the specific harmonic structure of these noises. It can be viewed as an extended version of the Lent's method. We propose a pitch mark location method adapted for signals with varying fundamental frequency, including specific sound sample extraction algorithms and synthesis methods.

The purpose of this new analysis synthesis scheme is to synthesize the sound emitted by an engine for arbitrary variations of its rotational speed ($rpm$) provided that recordings spanning the whole $rpm$ range of the engine are available. One important constraint is to minimize the computational load of the synthesis stage to enable real time applications. Also, special attention is

given to phase match between the sound samples that are likely to be concatenated. Provided that the number of cylinders in the engine is known, the proposed pitch mark location enables placing the pitch marks so that all sound samples extracted from the input signal start at the explosion in the same cylinder. It will be shown that this property greatly enhances the fidelity of synthesized signals due to preserved harmonicity.

The analysis stage aims at iteratively collecting sound samples from the input signal at equally spaced fundamental frequencies

$$F_{0_i} = F_0^{min} + i\ \frac{F_0^{max} - F_0^{min}}{M-1} \qquad i = 0, M-1, \quad (5)$$

where $M$ is the total number of sound samples to extract.

Step 1 is to estimate the evolution of the fundamental frequency of the input signal. Note that it may vary arbitrarily as long as it spans the frequency range of interest. As the fundamental is often missing or hard to detect, this step is preferably performed by estimating the evolution of the most energetic harmonic, $k_{Amax}F_0$.

In step 2, this component is extracted from the input signal with a selective time varying passband filter. The central frequency of the filter is updated at every sample to match the local $k_{Amax}F_0$ estimate. Ideally, the resulting signal is a single sinusoid modulated in frequency and amplitude according to the $k_{Amax}F_0$ evolution and remains in phase with the input signal. It provides a convenient representation of the engine noise periodicity. In fact, each period corresponds to the duration between two successive explosions in the cylinders.

In step 3, pitch marks are placed so that every sound sample starts at the explosion in the same cylinder. Recalling $k_{Amax} = N_c$ in practice, these are placed in the input signal at the beginning of every $N_c^{th}$ period of the filtered signal obtained in step 2. For each frequency $F_{0_i}$, a single pitch mark is selected as the one corresponding to the closest fundamental.

In step 4, for each selected pitch mark, a sound sample is extracted from the input signal with an appropriate temporal window. The additive noise $w_{F_0}$ (see Eq. (4)) is naturally extracted with the harmonic part of the signal and requires no additional operations as discussed by Jones and Parks[22]. This four-step process constitutes the analysis stage of the method.

Once the dataset of sound samples is created, any arbitrary evolution of the fundamental frequency can be synthesized. The sample selection and the overlap-and-add constitute the synthesis stage. Detailed information about the analysis and synthesis stages is given in Sections IV and V respectively.

## IV. ANALYSIS

### A. Step 1: Instantaneous fundamental frequency estimation

The estimation of a time varying fundamental frequency is commonly designated as pitch tracking in speech and music signal processing. A variety of robust methods exist in the literature[23–25]. One specific feature introduced in the YAAPT method[25] by Zahorian, the Spectral Harmonics Correlation function ($SHC$), appears to be particularly well suited to the fundamental frequency tracking of engine noise. In the present work, a simplified version of this method based mainly on the $SHC$ function is used to track the $k_{Amax}F_0$ frequency component. More specifically, our method does not make use of the nonlinear signal preprocessing or dynamic programming steps introduced by Zahorian. The signal is analyzed frame by frame with a high overlap rate of about 90%. The frame length is updated for every frame to contain an integer number of periods of the estimated frequency for the previous frame. For our recordings, a value of 20 periods was found appropriate in practice. This choice results from a trade-off between the stationarity assumption of the signal which depends on the $rpm$ variation rate and the precision of the estimation. In fact, the frames must be long enough to achieve good frequency resolution but must also be limited so that the frequency $k_{Amax}F_0$ can be considered stationary over the frame length. The $k_{Amax}F_0$ estimation for frame $t$ is performed by evaluating the $SHC$ function:

$$SHC[t,f] = \sum_{f'=-WL/2}^{WL/2} \prod_{r=1}^{N_H+1} S[t, rf + f'], \qquad (6)$$

where $S[t,f]$ is the magnitude spectrum of frame $t$ at frequency $f$, $WL$, the spectral window length in frequency and $N_H$, the number of harmonics. Note that the values of $f$ and $rf + f'$ are discrete with a spacing depending on the FFT length and the sampling rate. The maxima of the above $SHC[t,f]$ function correspond to the high amplitudes of the frame $t$ spectrum at integer multiples of $f$. Thus, local maxima of $SHC[t,f]$ are the potential values of the frequency $k_{Amax}F_0$ in frame $t$.

In the case of engine noise, an additional assumption can be made to reduce estimation errors. The evolution of the frequency $k_{Amax}F_0$ is continuous and therefore we propose to constraint the estimation at frame $t$ with the estimation at frame $(t-1)$. At each frame, the $SHC$ function is calculated only for a small frequency interval around the last estimated frequency value. Moreover if the fundamental frequency evolution is moderate and if the overlap rate between the frames is sufficient, this interval can easily be reduced to contain only one local maximum. This property ensures no doubling or halving of the estimated value; it is illustrated in Figure 2 for an interval width of 4 `Hz` and ten successive frames.

Different input parameter values provide satisfactory results for engine noises. In this study, $N_H$ was set to 4 and $WL$ to 10 `Hz`. An illustration of the performance of the algorithm using these parameter values is presented in Figure 3 for a four-cylinder engine noise signal.

As the above process is iterative, a specific initialization of the algorithm is necessary to set the initial frequency value and frame length. This value can easily be estimated by calculating the $SHC$ function of the first frame for a larger frequency interval.

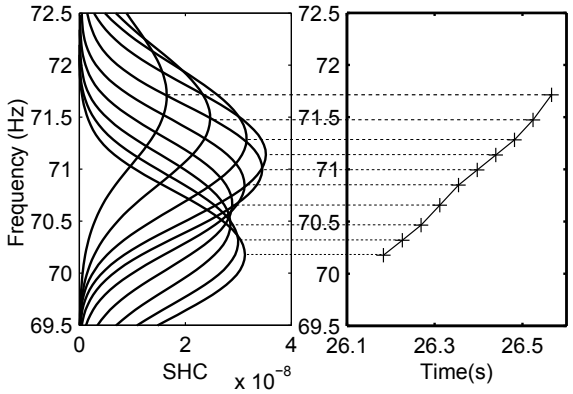


FIG. 2. Estimation of the $k_{Amax}F_0$ frequency on a segment of an engine noise signal. (left) The SHC function for ten successive frames. (right) Corresponding $k_{Amax}F_0$ frequency estimation as a function of time. The search interval is set to $\pm 2$ `Hz` around the estimated value for the previous frame.

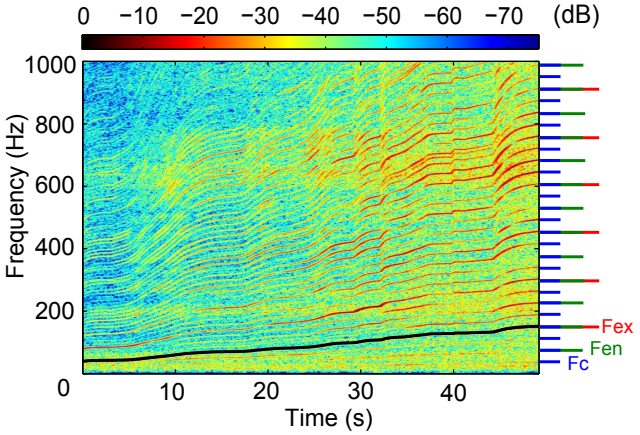


FIG. 3. Spectrogram of a four-cylinder engine noise signal. The solid line is the estimated value of $k_{Amax}F_0$ (explosion frequency) obtained with the simplified YAAPT method. The signal is sampled at $Fs = 44.1$ `kHz` and its spectrogram is calculated using frames of 200 `ms` zeropadded to $32,768$ samples for FFT computations. Hann windows overlapping by 50 % were used. Note that for clarity reasons the spectrogram is showed only for frequencies below $1,000$ `Hz`. On the right of the spectrogram, the colored segments indicate the positions of the harmonics for the last signal frame, as introduced in Figure 1: (red) harmonics of $F_{ex}$, (green) harmonics of $F_{en}$ and (blue) harmonics of $F_c$.

The analysis stage requires the knowledge of the fundamental frequency at every instant of the input signal. The frequency values estimated at each frame are interpolated with a standard cubic interpolation method to obtain an estimation of the instantaneous fundamental frequency at each sample of the input signal.

### B. Step 2: Extraction of the prevailing harmonic

Harmonic extraction methods are mainly used for source separation and speech enhancement systems. The pitch evolution of each speaker is determined and used as input to an adaptive passband comb filter[26]. Such methods have proven their efficiency for the extraction of harmonic signals from noisy environments but they do not fulfill the specific requirements of our application. In the present method, harmonic extraction is performed only on the prevailing harmonic, it has to be selective and not introduce phase delays.

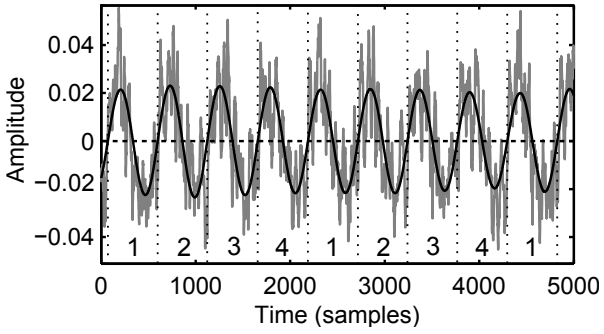


FIG. 4. Segment of a four-cylinder engine noise sampled at 44.1 `kHz` (gray). Extracted prevailing harmonic of frequency $4F_0$ (black). The numbers 1 to 4 indicate the cylinder in which an explosion occurs during the time interval delimited by the dotted vertical lines. Note that the first number is randomly attributed and not associated with a specific cylinder in the engine.

As mentioned in Section III.B, the extracted signal is ideally a single sinusoid modulated in frequency and amplitude according to the variations of the $k_{Amax}F_0$ frequency component. Figure 4 shows a segment of a four-cylinder engine noise and its extracted prevailing harmonic $4F_0$. It illustrates the requirements of the filtering process. The later must be sufficiently selective to attenuate the adjacent harmonics $(k_{Amax}-1)F_0$ and $(k_{Amax}+1)F_0$ and must not introduce phase delays so that the filtered signal remains in phase with the input signal. Hereby, no bias is induced in the positioning of the instants of explosion in the cylinders. These concerns oriented the choice of the filter towards the digital Butterworth bandpass filter[27]. During the filtering process, the bandpass center frequency is updated at every sample to match the $k_{Amax}F_0$ estimate. The width of the filter pass band is also updated at every sample to $[(k_{Amax}-0.5)F_0\ (k_{Amax}+0.5)F_0]$ so that adjacent harmonics are filtered out.

At discrete time $n$, the Butterworth filter of order $N$ calculated for the central frequency estimate is defined by its Z-transform as

$$H^n(z) = \sum_{k=0}^{N} \frac{b_k^n . z^{-k}}{a_k^n . z^{-k}}, \tag{7}$$

where $b_k^n$ and $a_k^n$ are respectively the numerator and the denominator coefficients of the filter. They are derived from the analog expression of Butterworth filters using the bilinear transform[27]. As the filter cut-off frequencies $(k_{Amax}-0.5)F_0$ and $(k_{Amax}+0.5)F_0$ are low compared to the sampling frequency 44.1 `kHz`, the poles of $H^n(z)$

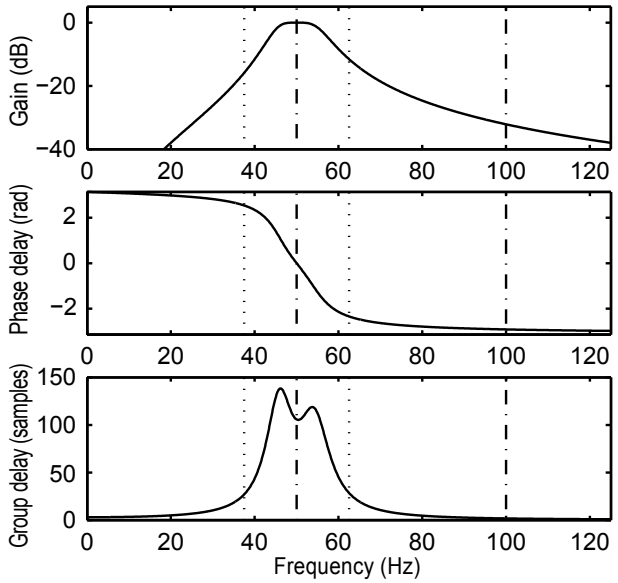


FIG. 5. Response of an order four Butterworth filter with center frequency 50 Hz and pass band [43.75 56.25] Hz: (up) amplitude, (center) phase delay and (bottom) group delay.

are close to the unit circle. To avoid stability issues when increasing the filter order, the input signal is downsampled by a factor 10 to a sampling frequency of 4.41 kHz and the order $N$ is restrained to four. After the filtering process, the signal is upsampled with a cubic interpolation method to restore the original sampling rate of the input signal. We assume that the resampling artifacts are negligible.

If $x[n]$ is the discrete time engine noise signal and $y[n]$ the filtered signal, the filtering process is modeled by the difference equation

$$y[n] = \sum_{k=0}^{N} x[n-k]b_k^n - \sum_{k=1}^{N} y[n-k]a_k^n. \qquad (8)$$

The characteristics of a Butterworth filter designed for a center frequency of 50 Hz are shown in Figure 5. Some of the advantages of the Butterworth bandpass filter is the flat frequency response in the pass band and the null phase delay for the center frequency. Thereby, the output signal is in phase with the input signal. However, the uncertainty on the estimation of the $k_{Amax}F_0$ frequency evolution may induce nonzero phase delay of the extracted harmonic component. This is due to the steep slope of the phase response of the filter close to the center frequency (see Figure 5). Therefore the precision of the frequency estimation described in the previous section is crucial for the extraction of the prevailing harmonic with zero phase delay.

Fourth order Butterworth filters are not very selective. As mentioned above, increasing the order of the filter makes it unstable. Hence, the possibility of filtering the extracted harmonic twice through the same time varying Butterworth filter has been investigated. Such an operation increases the selectivity of the filtering process and hence, the precision on the true periodicity of the prevailing harmonic component but it also doubles both the phase delay induced by the uncertainty on the frequency estimation and the group delay. In the end, it appears that filtering the signal twice does not improve the overall quality of the process.

The group delay induced by the filtering process is not negligible. The energy carried by the prevailing harmonic component is delayed by about one hundred samples (about 23 ms once the output signal is upsampled). However in our application this effect is not critical as long as the filter order is not increased. In fact, as shown in Figure 4, only the phase of the $k_{Amax}F_0$ frequency component is relevant to delimitate its periods.

### C. Step 3: Pitch mark location and selection

As the filtered signal is assumed to be purely sinusoidal, its period can be determined by the locations of the zerocrossings[19]. Every $2N_c^{th}$ zerocrossing is listed as a pitch mark so that each extracted sound sample starts on the explosion in the same cylinder as shown of Figure 6 (up). The correspondence between each pitch mark and the fundamental frequency is given by the of the $k_{Amax}F_0$ frequency estimate. For every frequency $F_0^i$ defined by Eq. (5), the pitch mark corresponding to the closest fundamental frequency is selected. Care is taken to ensure that a given pitch mark is only selected once.

The number of pitch marks to select (noted $M$ in Eq. (5)) in order to create a complete dataset for a given engine depends on the range of the available rpms. Selecting all listed pitch marks (every $2.N_c^{th}$ zerocrossing) results in very large datasets which are not optimal in terms of storage. It is relevant to consider a sufficient number of sound samples per Hz of fundamental, rather than an absolute number of sound samples for an engine. For our purpose, this sound sample density is ideally constant over the entire *rpm* range. This requirement raises the issue of the number of sound samples physically available in the input signal. The input recording must be sufficiently long and the slope of the fundamental frequency evolution sufficiently low to provide enough sound samples for the dataset. In practice, a sample density of 30 sound samples per Hz of fundamental is sufficient. Such datasets contain about $1,000$ sound samples and require about 15 $Mb$ of storage in 32 $bits$ floating point numbers at 44.1 kHz if no signal compression is performed.

### D. Step 4: Sound sample extraction

Once a pitch mark has been selected from the pitch mark list, an appropriate temporal window centered on the pitch mark is applied to the input signal to extract the corresponding sound sample (see Figure 6 (bottom)). The standard choice in existing OLA methods for the sample extraction envelope is a Hann window containing an integer number of pitch periods. In the case of engine noise, each sound sample contains information on one engine cycle plus some overlap to ensure smooth transitions

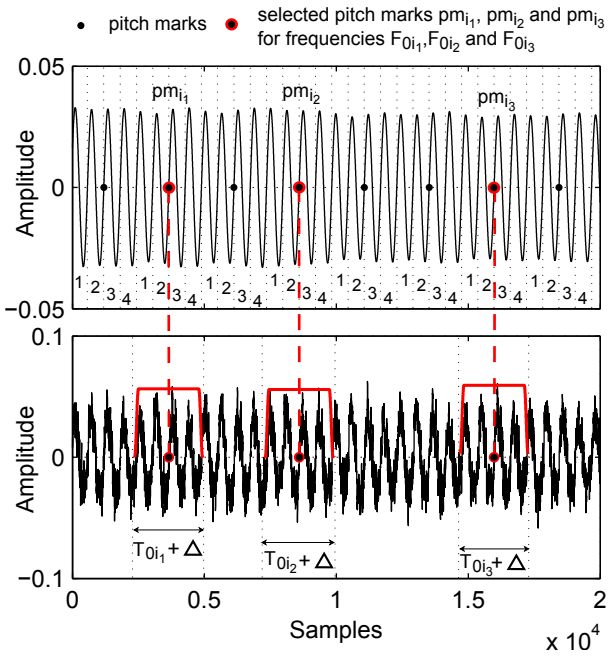


FIG. 6. (up) Segment of a filtered four cylinder engine noise signal with the positions of the pitch marks. The numbers 1 to 4 indicate the cylinder in which the explosions occur. (bottom) Segment of the input signal corresponding to the filtered segment above. The selected pitch marks and the positions and lengths of the extraction windows are also depicted. $T_{0_{i_1}}$,$T_{0_{i_2}}$ and $T_{0_{i_3}}$ are the periods corresponding to the frequencies $F_{0_{i_1}}$, $F_{0_{i_2}}$ and $F_{0_{i_3}}$. $\Delta$ is the overlap length of the sound samples envelope (see Section IV.D). (Color online)

between sound samples during the synthesis. To achieve this, a straightforward solution is to use Hann windows of two periods of the local fundamental frequency $F_{0_i}$. However, when overlapping sound samples with fundamental frequencies lower than the threshold of human hearing (20 `Hz`), a beating phenomenon with a period equal to the Hann window length occurs. This beating is not observed in standard PSOLA methods since the concerned fundamental frequency values in speech or music signal processing never reach 20 `Hz`. It is induced by the summation of slightly different harmonic components in the overlapping sections of successive sound samples. It appears that lowering the overlap rate of the sound samples reduces significantly this beating. For this purpose, Tukey windows with a low and constant overlap of $\Delta$ samples are used instead of Hann windows. Extracted sound samples now contain exactly one period of the fundamental frequency $F_{0_i}$ plus $\Delta$ samples to ensure a smooth overlap during the synthesis. The length of the overlap $\Delta$ needs to be sufficiently long to mask the amplitude and frequency differences between sound samples during the synthesis but needs to be short enough to attenuate the beating occurring in the low *rpm* range. A value of $\Delta = 128$ samples or 3 `ms` at 44.1 `kHz` represents a good compromise.

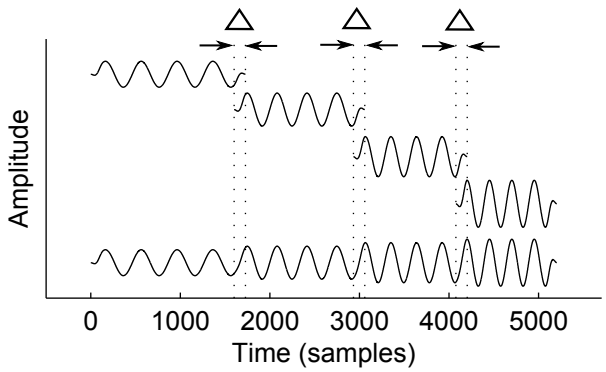


FIG. 7. Schematic representation of the OLA method applied to increasing fundamental frequency and increasing amplitudes sound samples. $\Delta$ is the constant overlap between sound samples set to 128 samples at 44.1 `kHz`.

## V. SYNTHESIS

### A. Overlap and add processing

The product of the analysis stage is a dataset of sound samples corresponding to equally sampled fundamental frequencies over the frequency range covered by the input signal (see Eq. 5). The samples in the dataset are sorted in an ascending order with respect to their fundamental frequency. The dataset allows for the synthesis of any fundamental frequency evolution by overlapping and adding the sound samples (see Figure 7).

The Tukey window requires an overlap of $\Delta$ samples during the synthesis stage to satisfy the constant overlap and add property[28]. This overlap between successive sound samples is sufficient to suppress the perceptive artifacts caused by slight differences in frequency and amplitude. Moreover, the phases of the harmonics of $F_c$ and $F_{en}$ at the boundaries of the sound samples are guaranteed to be continuous since the overlapping signal segments correspond to the explosions in the same cylinders.

### B. Sample selection algorithm

A specific sample selection algorithm is implemented to ensure high quality synthesis. In the case of a slowly varying fundamental frequency target, looping on the same sound sample should be avoided as this would induce audible artifacts. In the context of time scaling of speech signals, Shnell and Peeters[29] proposed to increase the number of available sound samples for the synthesis by using time domain or frequency domain interpolation between extracted samples (TDI-PSOLA or FDI-PSOLA). However, this solution does not solve the problem of a strictly stationary fundamental frequency target. Therefore, the algorithm proposed here acts on the sample selection process rather than the number of available samples. To avoid selecting successively the same sound sample, a pseudo-random oscillation is added to the selected sample index. This oscillation has zero mean to ensure that the average fundamental of the synthesized signal

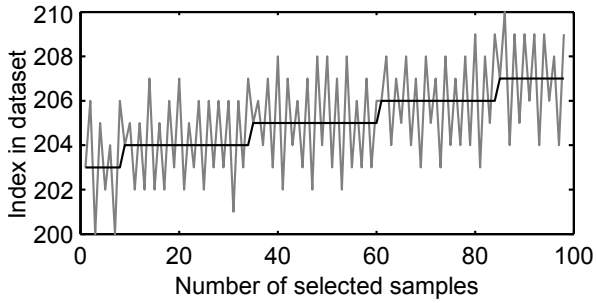


FIG. 8. Sample index matching the fundamental frequency target (black). Selected sample index with a maximum amplitude oscillation of 3 (gray). In the case of stationary *rpm* targets, the sample index is alternatively selected over and under the expected index in the dataset.

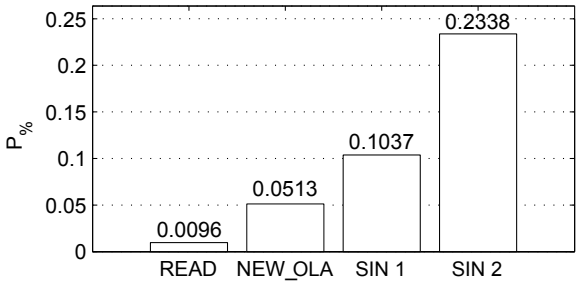


FIG. 9. Percentage of processor time spent for the real time generation of a sound signal. READ corresponds to a simple signal reading of a stored signal in memory. NEW_OLA corresponds to an engine sound synthesis with our overlap-and-add algorithm. SIN 1 and SIN 2 correspond to the generation of a single sinusoid respectively by the direct-form second-order digital resonator method[30] and with standard C library sine function.

matches the stationary target value. This randomized selection algorithm is illustrated in Figure 8.

The choice of the maximum amplitude of the random oscillation depends on the sound sample density in the dataset. In fact, the overlap-and-add algorithm is not intended to handle important spectral variations from one sound sample to another due the low overlap between successive sound samples. In practice, when the dataset contains about 30 sound samples per `Hz` of fundamental, the maximum amplitude of the random oscillation can be set to 3.

### C. Computational cost

The main advantage of this analysis-synthesis scheme is that most of the algorithmic complexity is kept in the analysis stage which is not performed in real time. The synthesis stage is limited to the selection of sound samples in the dataset and $\Delta$ additions per sound sample (where $\Delta$ is the overlap length between two samples) .

To illustrate the performance of the synthesis algorithm, a comparison between different simple methods to produce a sound signal is presented in Figure 9. Our OLA approach is compared with a simple signal reading from memory and two methods of generating a single pure tone using the standard C library sine function and the direct-form digital resonator method[30]. The single pure tone generation methods are introduced here to point out the benefit of our OLA synthesis scheme over additive synthesis approaches for harmonic signals. The comparison criterion estimates the percentage of processor time spent on the generation of the signal. These percentages are calculated as

$$P_{\%} = \frac{N_{CPU}}{F_{CPU}} \times 100, \tag{9}$$

where $N_{CPU}$ the number of processor cycles $N_{CPU}$ spent to generate one second of a sound signal sampled at 44.1 `kHz` and $F_{CPU}$ is the processor frequency $F_{CPU}$.

It appears that the synthesis of an engine sound is nearly twice as fast as the generation of a single sinusoid with the direct-form digital resonator method. Thus it is much faster than any time domain or frequency domain additive synthesis method. The direct form digital resonator method requires one multiplication and two additions per sample. We can deduce that the total complexity of our method is approximately equivalent to half a multiplication and one addition per synthesized sample. It appears that more than $1,500$ engine sounds can be synthesized simultaneously in real time on a standard personal computer (Intel(R) Xeon(R) CPU at 2.33 `GHz`) using a single thread implementation. As such, this method is particularly well suited to synthesize engine sounds of multiple vehicles and can therefore be applied to the simulation of the sound field perceived by a listener in urban areas under dense traffic conditions.

## VI. VALIDATION

The analysis of the spectral content of a synthesized signal provides valuable material to assess the performance of the analysis-synthesis algorithm. Important psychoacoustic clues such as the A-weighting or the logarithmic sensitivity to loudness of human hearing can be considered in a spectral analysis. In the following, a novel method to evaluate the similarity between two signals through the analysis of their spectral content is proposed.

In practice four engines will be considered, three low power four-cylinder engines used in light vehicles (noted *engine 1* to *3*) and one six-cylinder engine used in a heavy duty vehicle (*engine 4*). The performance of our method is assessed by synthesizing engine signals with different methods and comparing them to a recorded signal of the same engine. The different methods include the WSOLA approach[18] and different ways of taking into account the phases of the harmonics using the framework described in this paper.

### A. Method

For each vehicle, two engine noise signals are recorded with different variations of the fundamental frequency. The first signal, referred to as *input 1*, is used for the creation of a sample dataset. The fundamental frequency of the second signal (referred to as *input 2*) is used as the target value to synthesize engine noise signals with different methods as introduced above. We propose a criterion to evaluate the similarity between the *input 2* signal and the synthesized signal (with the same *rpm* evolution). The computation of the similarity criterion between the two signals is performed through the following steps:

1. The Short Time Fourier Transforms of both signals ($STFT_{input2}[m,f]$ and $STFT_{synth}[m,f]$) are computed using frames at time $m$ of 16,384 samples (371 `ms` at 44.1 `kHz`) windowed with Hann windows and a 90 % overlap.

2. The squared error $D[m,f]$ in dB and the power spectrum $P_{input2}[m,f]$ in dB of the reference *input 2* are:

$$D[m,f] = 10\log_{10}\left(\left(|STFT_{input2}[m,f]| - |STFT_{synth}[m,f]|\right)^2\right), \quad (10)$$

$$P_{input2}[m,f] = 10\log_{10}\left(\left(|STFT_{input2}[m,f]|\right)^2\right), \quad (11)$$

   where $m$ is the frame index, $f$, the discrete frequency value and $|.|$, the absolute value operator.

3. A-weighting, $A_w[f]$, is added to account for the sensitivity of human hearing:

$$D^A[m,f] = D[m,f] + A_w[f], \quad (12)$$

$$P^A_{input2}[m,f] = P_{input2}[m,f] + A_w[f]. \quad (13)$$

4. The mean over time and frequency with a $1/f$ weighting (giving equal importance to each octave) is:

$$\overline{D^A} = mean_{t,f}\left(D^A[m,f].\frac{f_0}{f}\right), \quad (14)$$

$$\overline{P^A_{input2}} = mean_{t,f}\left(P^A_{input2}[m,f].\frac{f_0}{f}\right), \quad (15)$$

   where $f_0$ is a frequency reference for the $1/f$ weighting, set to 1000 `Hz` by analogy to the A-weighting. The mean is performed on the logarithmic decibel scale to follow the properties of human perception.

5. Finally, the proposed similarity criterion is calculated on a linear scale as the mean error $\overline{D^A}$ normalized by the energy $\overline{P^A_{input2}}$.

$$C = \frac{10^{\overline{D^A}/10}}{10^{\overline{P^A_{input2}}/10}}. \quad (16)$$

   Although, the value $\overline{D^A}$ in dBA already appraises the similarity between a recorded and a synthesized signal, the normalization makes it easier to interpret by providing a value less sensitive to the signal energy.

This criterion is used as a first estimate of the perceptive difference between the signal *input 2* and signals of the same *rpm* evolution synthesized with different datasets created out of *input 1*. If it is close to one, the energy of the error is close to the energy of the input signal pointing out a poor quality synthesis. If it tends to zero, the perceptually important part of the spectral structure is correctly reproduced. The next section describes the different sound sample datasets created out of *input 1*. These differ only by the pitch mark locations in order to emphasize the improvement brought by our pitch mark location method.

### B. Tested datasets

First, to show the importance of placing the pitch marks in a pitch synchronous way, a dataset is created without taking the phase of the fundamental into account. The sound samples still contain information on one complete engine cycle but start at an arbitrary instant of the cycle. The signal synthesized from this dataset, following the fundamental frequency evolution of *input 2*, is noted $x_{asyn}[n]$ (for asynchronous signal).

Then, the principles of an existing speech processing technique (WSOLA) are applied to the creation of a dataset. The pitch marks are placed so as to maximize the cross-correlation between the sample to be extracted and already extracted samples with a similar fundamental frequency. The synthesized engine signal is $x_{wsola}[n]$.

Finally, to assess the importance of the order of the explosions in each sound sample, three different sample datasets are created. For the first, the order of the explosions in the cylinders is not taken into account. Each sound sample can start at the explosion in any cylinder, this is performed by taking every second zerocrossing of the $k_{Amax}F_0$ frequency component as a pitch mark (see Section IV.C). For the second, every $N_c^{th}$ zerocrossing of the $k_{Amax}F_0$ frequency component is listed as a pitch mark. For the third, every $2N_c^{th}$ zerocrossing is considered so that all sound samples start at the explosion in the same cylinder. The synthesized signals for these three datasets are noted respectively $x_2[n]$, $x_{N_c}[n]$ and $x_{2N_c}[n]$.

The spectrograms of the signals are computed and the similarity criterion $C$ is evaluated. As mentioned previously, this study is carried out on three different light vehicle with four-cylinder engines (*engine 1* to *3*) and one heavy duty vehicle with a six-cylinder engine (*engine 4*).

### C. Results

The spectrograms of the recorded and the synthesized signals of *engine 2* are presented in Figure 10 over the low frequency range. First, one can observe that the phase matching between the sound samples is crucial for engine

noise synthesis. In $x_{Asyn}[n]$, the harmonicity of *input 2* is not correctly preserved. The energy of the harmonics is spread over nearby frequencies. Moreover, the short transitions ($\Delta = 128$ samples) between sound samples introduce a comb structure around prevailing harmonics which in practice results in audible artifacts. Second, the spectrograms of $x_{wsola}[n]$ and $x_2[n]$ are very similar. The harmonics of $F_c$ and $F_{en}$ are missing. This suggests that the order of the explosions in the sound samples has a considerable impact on the harmonicity of the synthesized signals. One can also assume that in the cross-correlation calculations of the WSOLA-like method, the $k_{Amax}F_0$ harmonic component has a prevailing influence and induces a high error rate on the differentiation of the order of the explosions in the sound samples.

In $x_{N_c}[n]$, the sound samples start on the explosion either in the first or third cylinder. The order of the explosion is respected for every engine rotation but not for complete engine cycles (two engine revolutions). The harmonics of $F_{en}$ are therefore present while the harmonics of $F_c$ are still missing. When the order of the explosions in the cylinders is respected, all harmonic components of the engine noise signal are rendered properly. This result can be interpreted in terms of phase matching between the sound samples in the dataset. The instantaneous phases of the harmonics of $F_{en}$ and $F_c$ are continuous at the sound sample boundaries only if the order of the physical events occurring during an engine revolution and an engine cycle is respected.

Figure 11 shows the similarity criterion $C$ values calculated for the signals synthesized with the five datasets for each engine. These results confirm the visual observations in the low frequency range [0 1,000 Hz] (Figure 10). The relative values of the criterion $C$ confirm that $x_{2N_c}[n]$ has the closest spectral structure to the recorded engine noise for all tested engines. The results for the six-cylinder engine (*engine 4*) are similar to those obtained for the four-cylinders engines. This attests that the approach described in this paper is general and can be applied to different internal combustion engines. The variance of the $C$ criterion for each engine is related to the energy ratios between the different harmonic families $F_c$, $F_{en}$ and $F_{ex}$ and the stochastic noise component. As all the considered methods for pitch mark location benefit to different harmonic families, they have variable influence on the $C$ criterion depending on the tested engine.

### D. Preliminary perceptive analysis

To form its own opinion on the quality of the synthesized signals, the reader is invited to listen to the audio samples of the signals presented in Figure 10 and provided with the supplementary material.

For all considered cases, when synthesizing engine noise signals with stationary *rpm*, the random sample selection algorithm presented in Section V.B is efficient. There is no audible sample repetition and the perceived pitch sounds stationary when the sound sample density is sufficient in the dataset. More thorough listening tests based on psychoacoustic test procedures will be carried out to assess the similarity between recorded and synthesized engine noise signals. These psychoacoustic tests will also be used as a first validation of the correlation between the perceptive difference and the $C$ criterion.

## VII. CONCLUSION

The present paper introduces a method to synthesize internal combustion engine noise. Based on the PSOLA framework and the Lent's method, the proposed method does not require knowledge of the pitch. Instead, it extracts information about the prevalent harmonic component of the signal to determine accurate pitch mark locations and ensure a high quality synthesis. The output of the analysis stage is a dataset of sound samples corresponding to equally-sampled fundamental frequency values. The sound samples are concatenated using appropriate sample selection and OLA algorithms to simulate any fundamental frequency evolution. The main feature of this modified OLA method is the optimization of the pitch mark locations in the input signal so that every extracted sound sample contains the same physical events, in the same order but for different engine speeds. It has been shown that this ensures the continuity of the phases of all the harmonics and hence, better preserves the harmonicity of the engine noise signals in comparison to standard OLA methods.

As a future work, the influence of the load applied to the engine on the emitted noise will be considered. The sample selection algorithm can also be improved by taking advantage of other speech processing methods. Particularly, the unit selection algorithms introduced by Hunt and Black[31,32] could be adapted to our sample selection process. These selection algorithms estimate transition costs between sound samples from large speech databases and optimize their selection during the synthesis stage to ensure the natural continuity of the synthesized signal. In the case of engine noise sample datasets, the cost could be estimated by evaluating the differences in frequency and amplitude of the harmonics of the sound samples. This optimization would improve the quality of the synthesized signals but would also increase the computational load of the synthesis stage.

The low complexity of the proposed synthesis scheme makes it particularly relevant for real time synthesis applications. The real time synthesis of engine noise has numerous applications in the field of road traffic auralization, as well as in other areas such as driving simulators, evaluation of exterior and interior vehicle sound quality or the emission of internal combustion engine noise for the detection of silent electric vehicles. In the specific case of traffic noise in urban areas, the auralization system requires very high quality and low complexity sound source synthesis. Indeed, such simulators involve the synthesis of numerous sound sources, the calculation of all significant acoustic propagation paths and the filtering and spatialization of each contribution. All these steps are time consuming and therefore require optimizations. The proposed OLA method applied to engine sound synthesis clearly benefits traffic noise auralization.

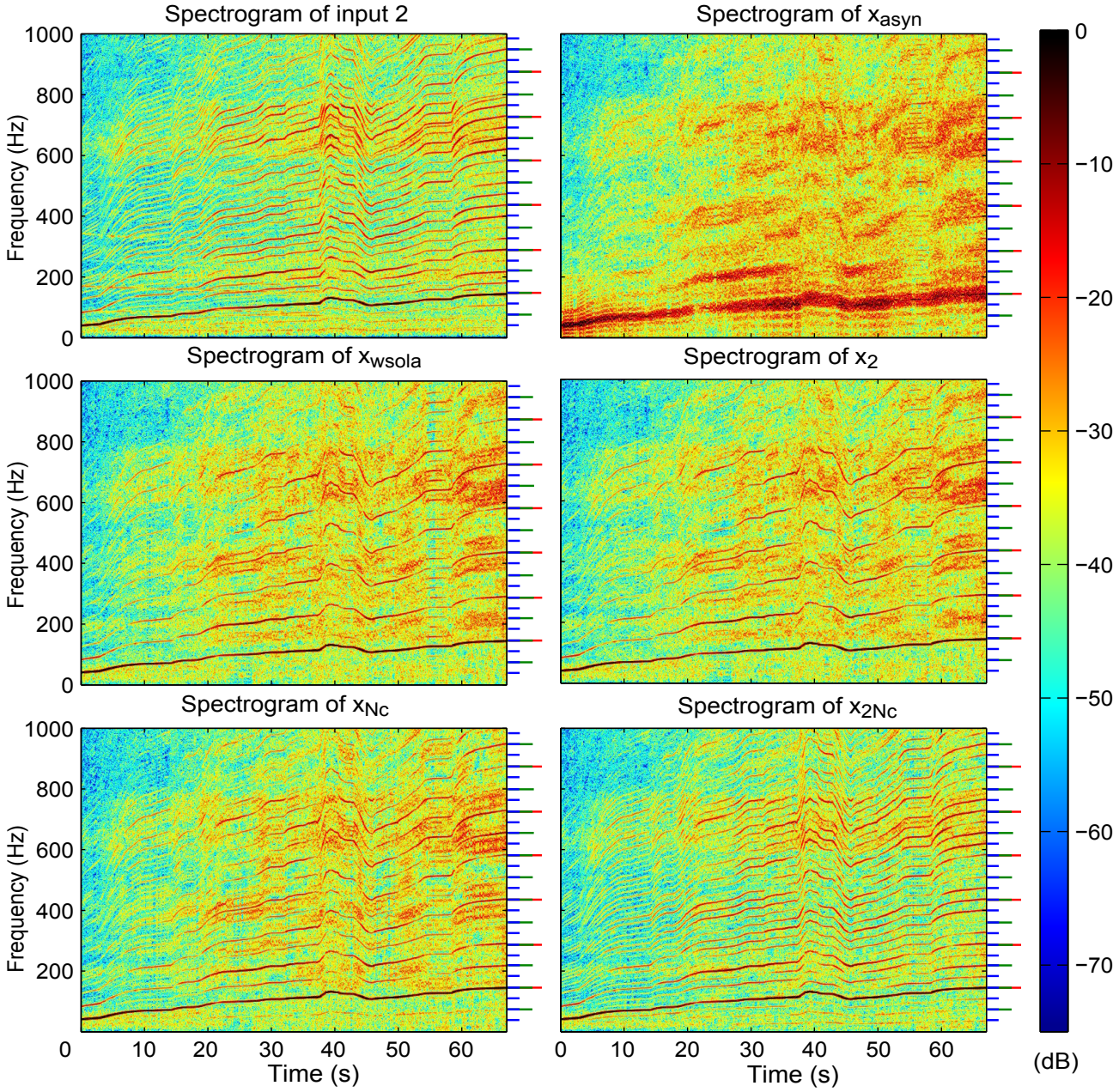


FIG. 10. Spectrograms of the recorded signal *input 2* and the synthesized signals calculated using frames of 200 `ms` zeropadded to 32,768 samples (at a sampling rate of 44.1 `kHz`) for FFT computations. Hann windows overlapping by 50 % were used. The frequency range is restrained to [0 1,000 `Hz`] for clarity reasons. On the right of the spectrogram, the colored segments indicate the positions of the harmonics for the last signal frame, as introduced in Figure 1: (red) harmonics of $F_{ex}$, (green) harmonics of $F_{en}$ and (blue) harmonics of $F_c$.

**Acknowledgments**

This research has been undertaken in the frame of the European project number 234306 HOSANNA.

[1] C. Steele, "A critical review of some traffic noise prediction models", Appl. Acoust. **62**(3), 271–287 (2001).

[2] B. Gauvreau, M. Bérengier, P. Blanc-Benon, and C. Depollier, "Traffic noise prediction with the parabolic equation method: Validation of a split-step padé approach in complex environments", J. Acoust. Soc. Am. **112**(6), 2680–2687 (2002).

[3] H. Jonasson, U. Sandberg, G. V. Blokland, J. Ejsmont, G. Watts, and M. Luminari, "Source modelling of road vehicles", Technical Report, Deliverable 9 of the Harmonoise Project (2004).

[4] G. Dutilleux, J. Defrance, B. Gauvreau, and F. Besnard, "The revision of the french method for road traffic noise prediction", J. Acoust. Soc. Am. **123**(5), 3150 (2008), full paper available on the CD-ROM of the Proceedings of the Joint ASA-EAA Meeting in Paris, 2008 (Acoustics'08).

[5] R. Makarewicz and M. Zóltkowski, "Variations of road traffic noise in residentials areas", J. Acoust. Soc. Am. **124**(6),

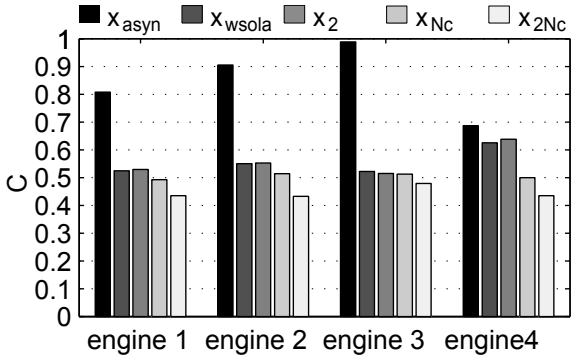


FIG. 11. Criterion $C$ values for the five considered synthesis datasets of the four test engines. *Engine 1*, *engine 2* and *engine 3* are the three four-cylinder engines while *engine 4* is the six-cylinder engine.


3568–3575 (2008).

[6] E. Salomons, H. Zhou, and W. Lohman, "Efficient numerical modeling of traffic noise", J. Acoust. Soc. Am. **127**(2), 796–803 (2010).

[7] J. Forssén, T. Kaczmarek, J. Alvarsson, P. Lundén, and M. E. Nilsson, "Auralization of traffic noise within the LISTEN project - preliminary results for passenger car passby", in *Proc. of the European Conference on Noise and Control*, EuroNoise-09, Edinburgh, Scotland, (2009).

[8] J. Maillard, "Prediction and auralization of construction site noise", in *Proc. of the European Conference on Noise and Control*, EuroNoise-09, Edinburgh, Scotland, (2009).

[9] S. Amman and M. Das, "An efficient technique for modeling and synthesis of automotive engine sounds", IEEE Transactions on Industrial Electronics **48**, 225–234 (2001).

[10] J.-F. Sciabica, M.-C. Bézat, V. Roussarie, R. Kronland-Martinet, and S. Ystad, "Towards timbre modeling of sounds inside accelerating cars", in *Proc. of Auditory Display : 6th International Symposium*, Copenhagen, Denmark, 377–391 (2009).

[11] J. V. Rensburg, "Phase vocoder technology for the simulation of engine sound", International Journal of Modern Physics C **17**, 721–731 (2006).

[12] D. Heitbrink and S. Cable, "Design of a driving simulation sound engine", in *Proc. of the Driving Simulation Conference 2007*, Iowa City, IA (2007).

[13] I. Zhekova, "Analyse temps-fréquence et synthèse granulaire des bruits moteur diesel au ralenti: Application pour l'étude perceptive dans le contexte des scènes auditives. (Time-frequency analysis and granular sythesis of diesel engine noise at idle: Application for a perceptive study in the context of the auditory scene analysis)", Ph.D. thesis, Université de la Méditerranée - Aix-Marseille II (2010), URL `http://tel.archives-ouvertes.fr/tel-00491452/en/` (date last viewed : 05/02/2012).

[14] K. Genuit and W. R. Bray, "Prediction of sound and vibration in a virtual automobile", Sound and Vibration **36**, 12–19 (2002).

[15] S. Roucos and A. M. Wilgus, "High quality time-scale modifications for speech", in *Proc. of International Conference on Acoustics Speech and Signal Processing*, Tampa, FL, **10**, 493–496 (1985).

[16] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication **9**(5/6), 453–467 (1990).

[17] R. W. L. Kortekaas and A. Lohlrausch, "Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli", J. Acoust. Soc. Am. **101**(4), 2202–2213 (1997).

[18] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modifications of speech", in *Proc. of International Conference on Acoustics Speech and Signal Processing*, Minneapolis, MN, **2**, 554–557 (1993).

[19] K. Lent, "An efficient method for pitch shifting digitally sampled sounds", Computer Music Journal **13**(4), 65–71 (1989).

[20] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modifications of speech", Speech Communication **16**, 175–205 (1995).

[21] R. W. L. Kortekaas, "Psychoacoustical evaluation of PSOLA. II. double-formant stimuli and the role of vocal perturbation", J. Acoust. Soc. Am. **105**(1), 522–535 (1999).

[22] D. L. Jones and T. W. Parks, "Generation and combination of grains for music synthesis", Computer Music Journal **12**, 27–34 (1988).

[23] D. Talkin, *Speech Coding and Synthesis*, chapter 14. A robust algorithm for pitch tracking (RAPT), 495–518 (Elsevier Science Inc., New York) (1995).

[24] A. de Cheveigné, "YIN, a fundamental frequency estimator for speech and coding", J. Acoust. Soc. Am. **111**(4), 1917–1930 (2002).

[25] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking", J. Acous. Soc. Am. **123**(6), 4559–4571 (2008).

[26] M. Gainza, B. Lawlor, and E. Coyle, "Harmonic sound source separation using FIR comb filters", in *Proc. of the 117th Audio Engineering Society Convention*, San Francisco, CA (2004).

[27] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Prentice-Hall, New Jersey), 219–224, 227–228 (1975).

[28] J. O. Smith, *Spectral Audio Signal Processing* (W3K Pulishing), page 188 (2009), URL `https://ccrma.stanford.edu/~jos/sasp/` (date last viewed : 05/02/2012).

[29] N. Schnell and G. Peeters, "Synthesizing a choir in real-time using pitch synchronous overlap add (PSOLA)", in *Proc. of the International Computer Music Conference*, Berlin, Germany, (2000).

[30] J. O. Smith and P. R. Cook, "The second-order digital waveguide oscillator", in *Proc. of the International Computer Music Conference*, San Jose, Costa Rica, 150–153 (1992).

[31] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", in *Proc. of the International Conference on Acoustics Speech and Signal Processing*, Atlanta, GA, **1**, 373–376 (1996).

[32] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis", in *Proc. of the Fifth European Conference on Speech Communication and Technology*, Rhodes, Greece, **2**, 601–604 (1997).