Article

An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach

Félix Gontier [1], Mathieu Lagrange [1], Pierre Aumond [2,3], Arnaud Can [2] and Catherine Lavandier [3]

1   LS2N
2   IFSTTAR, AME, LAE
3   ETIS, UMR 8051, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, F-95000 Cergy-Pontoise.
*   Correspondence: mathieu.lagrange@cnrs.fr

1   **Abstract:** The spreading of urban areas and the growth of human population worldwide raised
2   societal and environmental concerns. To better address those, the monitoring of the acoustic
3   environment in urban but also rural or wilderness areas is an important matter. Fostering on
4   the recent development of low cost hardware acoustic sensors, we propose in this paper to consider a
5   sensor grid approach to tackle this issue. In this kind of approach, the crucial question is the nature
6   of the data that is transmitted from the sensors to the processing and archival servers. To this end, we
7   propose an efficient audio coding scheme based on third octave band spectral representation that
8   allows 1) the estimation of standard acoustic indicators and 2) the recognition of acoustic events at
9   state-of-the-art performance rate. The former is useful to provide quantitative information about
10 the acoustic environment, while the latter is useful to gather qualitative information and build
11 perceptually motivated indicators using for example the emergence of a given sound source. The
12 coding scheme is also demonstrated to transmit spectrally encoded data that, reverted to the time
13 domain using state-of-the-art techniques, is not intelligible, thus protecting the privacy of citizens.

14 **Keywords:** acoustic monitoring; audio encoding

15 **1. Introduction**

16   The advent of low cost acoustic sensors together with the need to better monitor and comprehend
17 the acoustic environment of urban and wilderness areas give rise to the deployment of experimental
18 sensor networks such as the SONYC[1] [1] and CENSE[2] [2] projects. To do so, considering the so-called
19 "sensor grid" approach [3,4] has several advantages. The sensor grid approach puts the focus on a) a
20 distributed system of data acquisition with a large set of sensors, b) the production and storage of a
21 large dataset, and c) its availability for intensive and open data analysis computation.

22   In such approach, the number of sensor nodes shall be extended as needed without having to
23 change the hardware and software architecture, *i.e.* the approach shall be scalable. Second, the nodes
24 shall be energy efficient, ideally autonomous, in order to ease the deployment of the sensor grid to the
25 desired topology. Third, the encoding scheme used to transmit the data from the sensor to the storage
26 and processing servers shall be designed with care in order to guarantee the privacy of the citizens.
27 At the same time, the data stored shall remain as rich as possible while maintaining a low bitrate for
28 efficient transmission and storage.

---

1   http://wp.nyu.edu/sonyc
2   http://cense.ifsttar.fr

In the case of urban acoustic monitoring, the data of interest ~~can be quantitative such as standard acoustic indicators (LAeq, ...) or qualitative~~ typically consists of standardized high-quality measurements of energetic indicators such as the ~~presence of sound sources of interest (bird calls, sirens, ...). The latter allows us~~ equivalent sound level L_{eq} in dB SPL or its A-weighted counterpart L_{Aeq} in dBA. However, while these indicators have proven to correlate well with perceptual evaluations for negative impact sound environments [5], they are not sufficient to fully describe urban sound environments [6]. Many other variables can be derived to better account for previously implicit properties [7] including percentile values or time dynamics. Studies have been conducted to select relevant subsets of descriptors in sound environment characterization [8–10].

In addition, considering sound sources allows to better assess the human perception of the urban sound environment in terms of pleasantness and other perceptual indicators [11,12]. ~~This~~ For instance, a bird song and traffic noise with the same energy are perceived very differently. Most acoustic indicators used in current sound maps describe the global sound environment and lack information on its composition. The recognition and the estimation of the level of contribution of sources of interest is thus of great importance to fully model the studied acoustic environments, see for example [13–15]. Other indicators have been found to correlate with specific sound environment content [16] but do not achieve the level of performances of a well-designed source detection scheme in characterizing acoustic environments. This source detection step can respectively be operated online (on the sensors) or offline (on the data servers) leading to different grid designs (Figure 1).

Performing the detection online (case B of Figure 1) is efficient in terms of data storage as only the detected events are transmitted. This approach is thus currently considered in several ~~approaches~~ applications in the literature [1,17,18]. Though, it requires the availability of decent computing resources on the sensors in order to perform the detection step which as of today's efficiency of hardware does not allow them to be autonomous using power sources like ~~decent size~~ solar panels. Furthermore, the detection is done once and cannot be recomputed during offline posterior analyses.

Performing the detection offline has several benefits. First, the sensor task is much simpler and can thus be implemented on low energy consumption processors. This means that the sensors can be autonomous in terms of energy, which greatly eases the deployment of the network. Second, it allows researchers to gather large amount of data that can be post-processed and studied further offline. Data can be re-analyzed following newer classification schemes or using new indicators. Though, transmitting the raw audio through the network (case A of Figure 1) has several disadvantages, in terms of required bandwidth, storage capabilities and privacy. Thus, for transmission from the sensors to the storage unit, the data shall be encoded in a more efficient way than lossy audio encoding standards [19]. Indeed, the audio signal is not mandatory for the computation of the acoustic indicators and the features needed for event recognition. Also, as the data is transmitted using the network and stored, one must ensure that the intelligibility of potential speech utterances is lost during the coding process, in order to guarantee the privacy of the citizens.

~~We thus propose~~ The analysis of sources within a sound environment is generally linked with spectral content [20] and high temporal resolutions while acoustic indicators focus on more scarce but highly accurate measurements. We thus present in this paper an audio coding scheme specifically tailored for use in a sensor grid (case C of Figure 1). ~~It has the following features: it is of low bitrate but still allows the computation of most of the standard acoustics indicators with controllable precision. As far as acoustic event detection is concerned, we~~ The proposed method computes and encodes third-octave bands L_{eq} with controllable precision, enabling low bitrate transmissions while allowing both the derivation of other acoustic indicators and the implementation of offline audio event detection schemes. We report equivalent performance of several state-of-the-art classification ~~schemes~~ models from features computed using raw audio data and encoded data at production level rates. Finally, according to a perceptual evaluation, the proposed coding scheme very strongly degrades the intelligibility, thus ensuring citizen privacy.

78   In order to promote reproducible research, the coder as well as the experiments needed to generate
79 the figures is available to the research community[3]. The remaining of the paper is organized as follows:
80 Section ~~?? reviews the types of data that need to be estimated using the transmitted data and Section~~ 2
81 describes the proposed encoding scheme. The experimental protocol designed to validate the proposed
82 approach is described in Section 3 and results are discussed in Section 4.
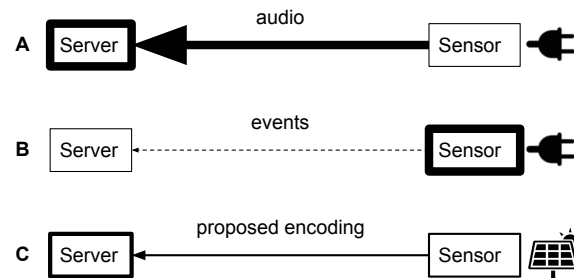

Figure 1. Three alternative implementations of the sensor grid approach for the monitoring of the acoustic environment. In A) the raw audio is transmitted, in B) only the detected events, and in C) a compressed spectral representation. Thickness of arrows indicates bandwidth and thickness of boxes indicates the level of computation or storage required.

83 ## 2. ~~Background~~ Proposed encoding scheme

84   ~~The main contribution of this paper is to present an acoustic data encoding scheme that allows~~
85 ~~both acoustic monitoring and sound event recognition in urban sound environment. Here we briefly~~
86 ~~review state-of-the-art methods in both subjects in order to better motivate our design choices.~~

87 *2.1. Data representation*

88 *2.2. ~~Acoustic monitoring~~*

89   ~~The considered sensor network primarily aims at monitoring urban sound environments, that is,~~
90 ~~continuously assessing their content and impact on the population. In the literature, this is typically~~
91 ~~performed by measuring energetic acoustic indicators such as the equivalent sound pressure level~~
92 ~~L_{eq} in dB SPL or its A-weighted equivalent L_{Aeq} in dBA. However, while these indicators have~~
93 ~~proven to correlate well with perceptual evaluations for negative impact sound environments [5]~~
94 ~~, they are not sufficient to fully describe urban sound environments [6]. Many other variables~~
95 ~~can be derived to better account for previously implicit properties [7] including percentile values~~
96 ~~or time dynamics. Studies have been conducted to select relevant subsets of descriptors in sound~~
97 ~~environment characterization [8–10]. The mentionned L_{eq} features are measured at periods of~~
98 ~~0.125 ms or 1 s, respectively termed as "fast" and "slow" as a convention.~~
99   ~~Most measured or derived indicators describe the global urban acoustic environment and~~
100 ~~lack information on its composition. Sound sources recognition enables a more comprehensive~~
101 ~~understanding of the properties of the acoustic environment. For instance, a bird song and traffic~~
102 ~~noise with the same energy are perceived very differently. The analysis of sources within a~~
103 ~~sound environment is generally linked with spectral content [20] and high temporal measurement~~
104 ~~resolutions. Other indicators can correlate with specific sound environment content [16] but do~~
105 ~~not achieve the level of performances of a good source detection scheme in characterizing acoustic~~
106 ~~environments.~~

---

3   https://github.com/felixgontier/cense-coder

2.2. ~~Event detection~~

~~The recognition and the estimation of the level of contribution of sources of interest is thus of great importance to fully model the studied acoustic environments, see for example [13–15]. As far as the encoding scheme is concerned, it is therefore important that the encoded data allows us to compute the above cited indicators but also preserve a sufficient amount of information to reliably detect the different sound sources of interest.~~

~~The identification of those~~ The identification of sources from audio streams has been the subject of extensive research in the past for speech data [21], music data [22], and lately environmental data in which the current work falls. Studied classification methods are diverse, ranging from time-dependant modeling with hidden ~~markov~~ Markov models [23] to "bag-of-frames" approaches [24,25]. Common architectures include learning-based classifiers such as support vector machines [26], Gaussian mixture models [27] or neural networks [28,29]. Also, several type of features are considered to feed those classifiers. The most used are certainly spectral [30] or cepstral [31] representations of the signal. Among them, Mel spectrograms and their cepstral-domain derivation, the Mel-frequency cepstrum coefficients (MFCC), are the most widely used as of today. These representations effectively model the human cochlear response to sounds by grouping frequency components around critical bands in a logarithmic scale. ~~The resulting representations are generally more densely sampled in time and frequency than in acoustic monitoring applications since acoustic indicators focus on more scarce but very accurate measurements.~~

~~They may also be exploited together with features~~ Features can also be computed in other domains to better model signal properties. For instance, [32] adds spectral features related to harmonicity and salient frequencies, and [33] uses a matching pursuit (MP) algorithm to deduce time-domain features. Alternative data representations such as the scattering transform [34] show good results in environmental sound classification tasks [35]. Another promising solution is feature engineering via unsupervised learning, which [36] implements with a clustering technique. ~~Alternative data representations such as the scattering transform [34] show good results in environmental sound classification tasks [35].~~ More complex systems using multiple-domain data representation as well as unsupervised models have been found to handle unrestricted environments suche as urban acoustic scenes efficiently [37]. A more detailed review of used methods is available in [38].

3. ~~Encoding scheme~~

~~We thus need to provide data that allows both the computation of previously described indicators and the recognition of sound events. To do so, we propose to consider the calculation of spectral energetic indicators such as the 31~~ In the acoustic monitoring field, the measurement of energetic indicators is mandatory to the development of noise maps. The third-octave ~~bands within the human audition range 20 Hz - 20 kHz. Slow or fast third-octave~~ bands ~~measurement~~ L_{eq} appears suitable for this work's purpose: in addition to being a relevant descriptor [39] for urban sound environment assessment, it allows for the computation of most cited indicators while representing reasonably small, fixed amounts of data to be transmitted. ~~It additionnaly contains spectral information that can be used for source recognition tasks as it is close to~~ Interestingly, third-octave bands are closely related to MFCC as they both are logarithmic-scaled spectral representations of the audio signal and diverge only in their motivation, mathematical and perceptive respectively.

We therefore consider the calculation of spectral energetic indicators such as the ~~Mel spectral representation, being just another filterbank transform with logarithmic frequency scaling~~ 31 third-octave bands within the human audition range 20 Hz - 20 kHz. "Fast" measurements of third-octave bands with an integration time of 125 ms are preferred to suit the temporal resolution needs of AED schemes, although "slow", 1 s measurements are also considered to further study the effect on our application's constraints.

2.1. *Signal analysis*

A common way to implement third-octave analysis is to first design the highest octave three filters, and use them on progressively time-decimated versions of the input signal [40]. ~~Antoni et al.~~ [41] presents an alternative analysis method based on direct frequency weighting, which yields a lower computational complexity as opposed to time filtering. The weights matrix design procedure complies with both ANSI S1.1-1986 [42] and IEC 61260-1:2014 [43] standards. It also meets the partition of unity principle over all frequency bins in the relevant range. Resulting filters for different parameter l values are compared with Couvreur's implementation [44] with 2nd order Butterworth time-domain filters, as shown in Figure 2. The major difference is that gains at cutoff frequencies are fixed at the optimal -3dB value by design.
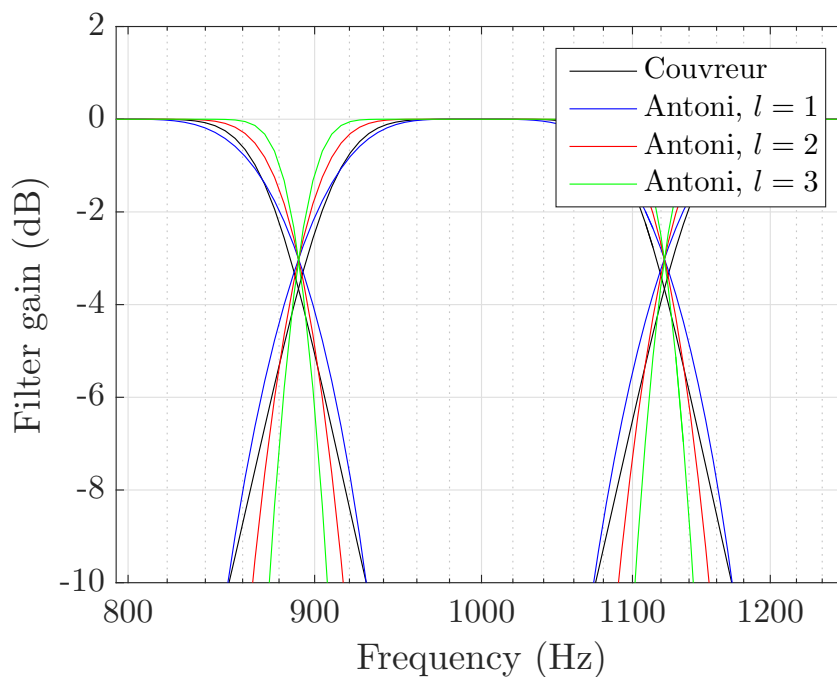

Figure 2. Comparison of Couvreur's and Antoni's implementations of third-octave filters. Frequency-weighting allows for arbitrary transfer functions and thus more accurate gains as standards impose.

The first step is to represent the signal in the frequency domain using a short-term Fourier transform (STFT). The continuously sampled audio is segmented into either 125 ms or 1 s frames to provide "fast" or "slow" acoustic measurements. The data is zero-padded to the next power of two to improve computational efficiency. The other parameters such as windowing functions and frame overlap are dependant on the specific constraints, and discussed in section 3. The phase is discarded as it is not meaningful for third-octave analysis. Similarly, negative frequency components contain the same information as positive ones because the input signal is real. We then compute third-octave bands from the squared spectral magnitude by matrix multiplication with the frequency weights proposed in [41] for $l = 2$ [4]. We include analysis for bands i from -17 to 13, ie. center frequencies f_i ranging from 20 Hz to 20 kHz with $f_0 = 1\ kHz$. The representation is thus composed of 31 values.

---

[4] When implementing said method, we found that squaring the $\phi_l(p), l \geq 1$ intermediate function (p. 887) was the correct approach to meet cutoff frequencies requirements, and we believe this was the author's original intention.

2.2. Data Encoding

A Huffman coding scheme [45] is an efficient method to further reduce data rate while keeping the underlying information. At this point of the processing chain, the data takes the form of a spectrogram with amplitude values represented on an arbitrary data type. It is therefore assimilable to a discrete distribution. By defining symbols as the possible values taken by the data, a discrete probability density function (PDF) can be expressed and estimated. The Shannon entropy then specifies the minimum average achievable output data rate for this given distribution, or average information content. The Huffman algorithm is an entropy coding scheme, as it uses the PDF to assign a code to each symbol so that the average output rate is the closest to the entropy. The resulting mapping function is called a Huffman dictionary.

The efficiency of the coding step thus directly depends on the entropy of the transmitted data. It is defined as $H = -\sum_i p_i log_n(p_i)$, where p_i is the probability of appearance of a given symbol i and n the numerical base in which the output code is represented. A lower entropy can be obtained with two factors. First, the dictionary size *i.e.* the number of symbols must be kept low. This implies performing a quantization on the data resulting in a loss of precision. Second, the data distribution itself is important. Entropy decreases when very few symbols have a very high probability of appearance and is maximum for a uniform distribution.

Considering an estimated PDF for our data as shown in Figure 3 (top-left), immediately applying a linear quantization is clearly suboptimal, as the data mostly consists of very low values and a direct quantization results in most of the information being lost. We therefore flatten the PDF by taking the logarithm of the representation (top-right). This allows to better use the available range of symbols. A linear quantization is then applied with $2^{q-1} - 1$ output values to obtain Figure 3 (bottom-left). Finally, we use a Δ-encoding algorithm along the time dimension to reduce redundancies between consecutive frames. This effectively concentrates higher probabilities on symbols around zero (Figure 3 bottom-right), yielding a higher amount of symbols at $2^q - 1$ but nevertheless a smaller entropy. In the example shown here, the entropy before Δ-compression is $H_{log} = 6.24\ sh$ which is reduced to $H_\Delta = 3.54\ sh$.

Huffman encoding is then computed with either a frame-specific symbol-code dictionary or a fixed dictionary generated using a learning dataset. A comparison of both methods is shown in Figure 4. For most short texture frame durations, the local Huffman algorithm is found to be much faster. In fact, the encoding computational complexity is a function of the number of dictionary elements. A frame-specific dictionary only contains the symbols found in a given data packet, while a fixed dictionary must contain every possible symbol as it does not depend on the present data. The output bitrate is found to be close for both methods, as sending an optimal local dictionary balances the suboptimality of globally generated symbol-code pairs.

The decoding process is quite straightforward, as both Huffman and Δ compressions are lossless and directly invertible. The entire coder scheme is summarized in Figure 5.

3. Validation protocol

A set of metrics is computed to assess the efficiency of the proposed coder scheme and to determine the impact of the degrees of freedom given by the parameters of the proposed algorithm. Those parameters are as follow: 1) the desired word size prior to Huffman encoding, set by quantization for both signal representation choices, 2) the number of considered bands, and 3) the STFT parameters including frame duration, overlap and windowing function usage. This section details experimental protocols for the validation of bitrate, measurement error, audio event recognition and intelligibility developped to best suit the corresponding constraints.
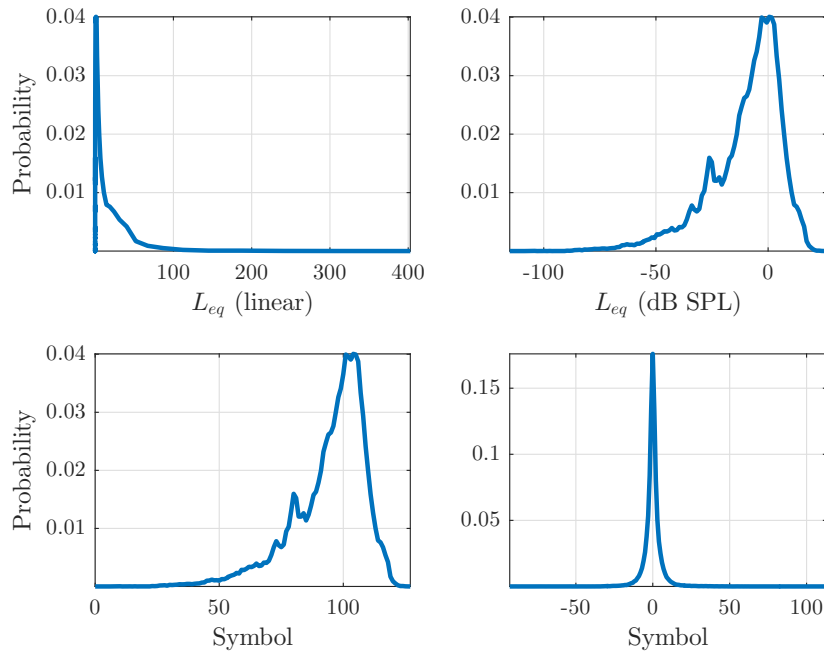
Figure 3. Example of estimated probability density functions of the data throughout the encoding step. Unchanged output of the representation step (top-left), concentrated towards very low values. PDF "flattening" effect induced by logarithm application (top-right). Here values are mapped to the range $[0, 2^7 - 1](q = 8)$ and rounded to perform quantization (bottom-left). Output of the Δ compression (bottom-right), with desirable probabilities as the input to a Huffman algorithm.
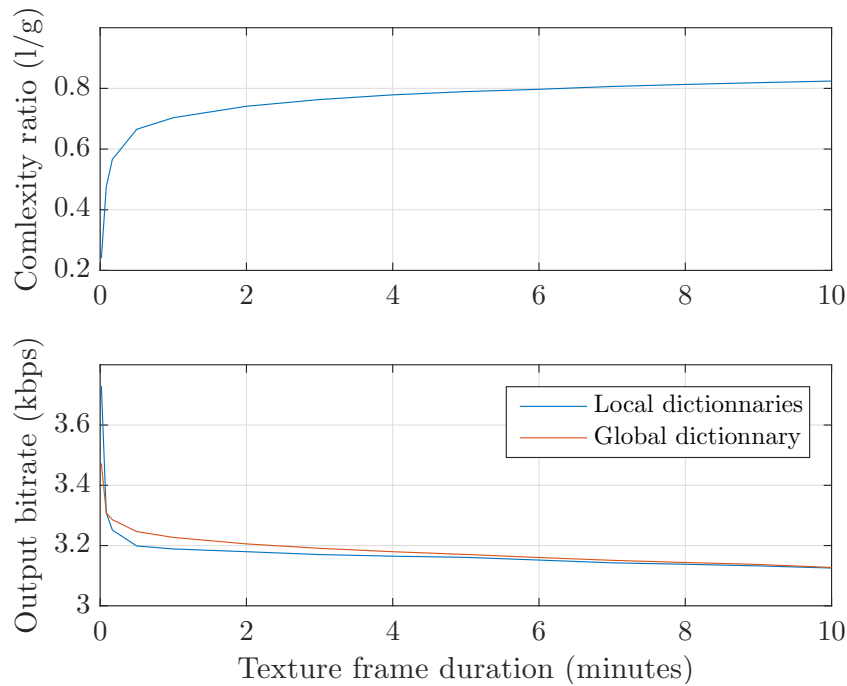

Figure 4. Performance comparison of locally and globally generated Huffman dictionaries. (top) The mean execution time ratio favors the use of frame-specific dictionaries for tested texture frame lengths. (bottom) The mean output bitrate is close for both algorithms, showing that the necessity of sending symbol-code pairs mostly compensates for their optimality.
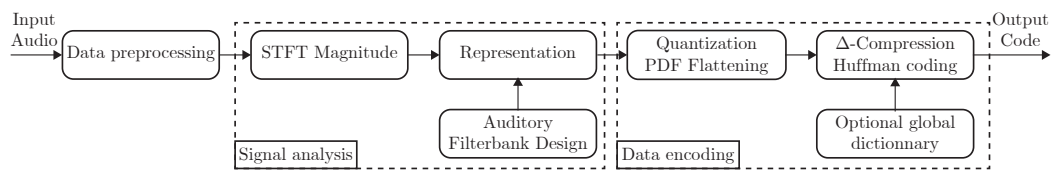
Figure 5. Overview of the ~~coder~~ encoding process.

3.1. Efficiency

The coder's output bitrate as well as measurement error is computed for both data representations. This metric is composed of the spectral third-octave band analysis error and the additional loss yielded by quantization before the encoding process.

~~To evaluate the analysis error, we consider~~ In order to assess the quality of the propoposed scheme's spectral measurement precision, we compare it to a known reference: the Matlab *ita_toolbox* implementation [46]~~as a reference~~. This toolbox's implementation shows less than 0.1 *dB* error on all bands when compared to a class 1 sound level meter software on 2 s third-octave bands measurements. We study both "fast" computation with 125 ms windows and "slow" 1 s analysis to assess the impact of time integration. Third-octave bands are computed using the proposed method as well as the *ita_toolbox* in the same experimental conditions, then averaged over time and compared to the toolbox's measurements on longer ~~extracts. In all experiments the sound dataset is the same and is composed of~~ 2 s extracts. The experiment is run for Gaussian white noise as well as urban environmental recordings: the former allows direct estimation bias evaluation, while the latter provides a simulation of real operating conditions. As the analysis frames are of short duration, poor frequency resolution and spectral leakage are expected when performing the STFT, particularly in the lower frequencies bands. For this reason we evaluate the ~~effect of the use of~~ impact of windowing functions on each band measurement error in order to potentially mitigate this loss of precision.

To provide relevant statistics, these quantities are estimated with data from the 4500 audio recordings of the UrbanSound8k dataset [47] described in the next section and 4500 *iid.* Gaussian noise sequences.

3.2. ~~Event~~ Preserving event recognition *performance at low bitrate*

In order to ensure that the proposed scheme allows the recognition of events using ~~state of the art~~ state-of-the-art methods, we consider the results achieved by several recognition systems on the UrbanSound8k dataset[47]. It features about 9 hours of urban environmental sounds recordings separated in 8732 audio files ranging from 1 to 4 seconds each. The recordings are labeled in 10 classes (air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music), and distributed in 10 independent folds. A method and baseline results are also provided for the classification task. ~~It is used in all the following experiments with the exception of intelligibility computation for which a small dataset of high-quality voice recordings is preferred. Audio~~ In this study, audio files are resampled at 44.1 kHz, normalized and only the left channel is considered for further analysis.

~~Mel spectrograms are computed with the rastamat library [49] to assess the relative performance of third-octave bands. STFT analysis frames are obtained by applying a Hann window on 23.2 ms of signal with 50% (11.6 ms) overlap to allow for efficient phase recovery in further tests.~~

~~We~~ The aim of this experiment is two-fold: to evaluate the global loss of information induced by the ~~coder by implementing the four classificationmodels~~ coding process, and to evaluate the adequacy of third-octave bands indicators for classification. As studying the best performing classical models [48], deep learning approaches [28] or unsupervised feature learning is beyond the scope of this paper, we choose to replicate the baseline models and results proposed in [47].
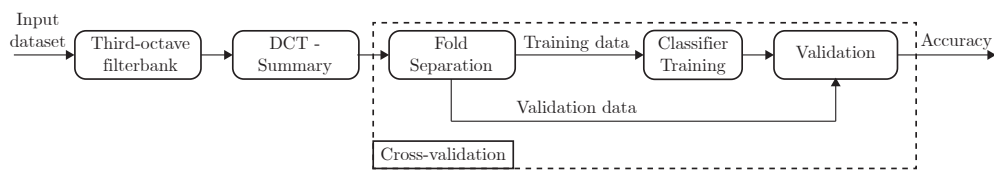
Figure 6. Classification process flowchart. Here third-octave bands are obtained by reversing the operations in Section 2.2 and cepstral coefficients are computed before training the model. Cross-validation is achieved by independantly training and testing the models with each of the 10 combination of a 90%-10% training-validation dataset separation.

The four classification models are as follow: a support vector machine with a radial basis function kernel, a random forest classifier with 500 trees, a decision tree and a k-nearest neighbors classifier with $k = 5$. Values for the SVM parameter C and RBF kernel variance σ^2 are determined using a grid search. We optionally apply a discrete cosine transform to the critical band signal representation to obtain cepstra, known as Mel Frequency Cepstrum Coefficients or MFCC in the case of Mel spectrograms. The 25 first coefficients are kept, and summarized along time with several statistics such as the mean, variance, skewness, kurtosis, minimum, maximum, median, derivative mean and variance, second order derivative mean and variance operators. The feature vector is thus comprised of 275 values to reproduce available results and compare with the data produced by the proposed coding scheme. Models are trained for each setup using the 10-fold cross validation method provided by the authors of [47], that is, every combination of testing one fold on models trained with the other nine. A summary of the considered classification process is shown in Figure 6.

Mel spectrograms are also computed with the *rastamat* library [49] to assess the relative performance of third-octave bands. STFT analysis frames are obtained by applying a Hann window on 23.2 ms of signal with 50% (11.6 ms) overlap to allow for efficient phase recovery.

3.3. Inintelligibility

It is important that the proposed scheme ensures a high level of recognition of acoustic events but, as the data is transmitted over the network and potentially stored, the level of intelligibility of speech utterances potentially captured in the decoded stream should be as low as possible. Thus, intelligibility in decoded and reconstructed audio is also an important concern of this study.

The use of an audio recording as an evidence is considered by the forensic phonetics [50] field. The readers shall be aware that even recorded with a good quality, such recording remains a weak biometrical indicator [51]. Still, even if the identity of the speaker cannot be asserted, one can attempt to transcribe the spoken utterances. As such, we believe that it is important to demonstrate that the proposed encoding scheme do not allow such thing.

In order to decrease the intelligibility, several approaches can be considered. First, a speech versus non speech detector can be implemented on the sensor. If some speech is detected, the data is not transmitted. This approach is however prone to failure of the detector and will lead to unavailability of data during speech time periods which might compromise the relevance of some indicators.

A second approach is to wisely choose the frame rate. Since a phoneme is about 100ms to 200ms long [52] [53], a frame rate lower than 4 Hz should therefore dramatically reduce intelligibility.

Thus, the impact of the frame rate parameter of the proposed encoding scheme on intelligibility is evaluated. To do so, a perceptual test is conducted. For the evaluation, a dataset of clean speech utterances recorded in studio conditions is used. The dataset is ~~made~~ composed of 9 french sentences enunciated by 3 male and 3 female speakers. The sentences have ~~an~~ a mean duration of 2 *s* and are composed of 5 to 12 words, averaging at 8. The 54 extracts are encoded using the proposed approach with 4 analysis frame rate parameters of respectively 50 Hz, 16 Hz, 8 Hz and 4 Hz leading to 216 audio excerpts being tested in total. However, as the recording conditions are the same for all extracts the

speaker identity is considered as a source of variability rather than as a necessary factor in the test statistics. As a result the objective is primarily to evaluate the 9 sentences with 4 encoding frame rates, *i.e.* 36 parameters.

Several subjective intelligibility metrics are standardized in the literature such as the system proposed in [54]. However, metrics regarding the signal only (SIG) or background noise only (BAK) which are often used to evaluate speech enhancement schemes [55], are not suitable for this test due to the background and signal being mixed with an uncontrollable SNR by the encoding process. Given these conditions a subjective score on the whole signal is used similarly to the OVRL score in [54]. As decoding the transmitted audio induces a framing effect unevenly distorting the input utterance, a second metric is designed to assess the comprehension of words. For the perceptual test, ~~each subject is presented with sentences and asked after listening to each sentence to both transcribe the utterances and rate the global~~ 18 different sentences evenly distributed over the 4 frame rate parameters are selected for each subject. For each sentence the participant first listens to the extract and is then asked to transcribe the utterance and to rate the perceived intelligibility from 1 (not intelligible) to 5 (intelligible). ~~Due to the low number of sentences, each one is tested twice by the same subject but under different encoding setting and speaker identity. As a result each subject evaluates 18 extracts evenly distributed over the 4 frame rate parameters.~~ The first metric is computed as the ratio of correctly transcribed words and will be referred as the intelligibility ratio (IR) in this paper. The second metric is directly the score attributed by the listener, referred as the intelligibility score (IS).

12 subjects of age ranging from 17 to 60 which reported normal hearing participated to the listening test. The test was conducted using a Matlab interface displayed with a desktop computer and audio was played through *Beyerdynamics DT 770* headphones. The output level was set by the experimenter at the same level for all subjects. The mean duration of the test was about 8 minutes.

This test requires an implementation of the inverse operation of the coding process to provide the listener with an estimate of the time-domain signal. To do so, a linearly-scaled spectrogram from the band-focused representations is estimated. This estimation is achieved by multiplying the spectral representation by the scaled transpose of the forward transformation matrix. A loss of resolution increasing with frequency is induced due to the shape of the transformation, as illustrated by Figure 7. ~~The signal~~ Each white square can be interpreted as the filter applied to obtain a single third-octave band. The filtering process will average the energy content of the signal over the corresponding band. As a result the energy located at a specific frequency is impossible to recover and the best inverse approximation is to split the third-octave band's observed energy into all the corresponding frequency bins with the same proportions as during analysis. This leads to the phenomenon seen on Figure 7 where part of the energy of a frequency bin in the original signal is found in all other bins of the same band at reconstruction, in almost equal proportions. The signal phase is recovered with the Griffin & Lim algorithm [56] if analysis overlap is used and white noise spectrogram scaling if not. The time-domain signal is then retrieved using the overlap-add method. When using overlap to compute third-octave bands, framing effects produced by rectangular windowing are reduced by convoluting the signal with a Hann windowing function prior to inverse-STFT computation.

4. Results

4.1. Coder efficiency

The main indicator of performance is the bitrate obtained at the output of the coder. The three varying factors studied are the data word size q, the number of bands and the effect of reducing time-resolution by averaging analysis frames. Figure 8 shows estimations of the output bitrate for different values of q. To match third-octave bands computation principle where a 125 ms ~~analysis is mandatory~~(fast) or 1 s (slow) analysis is standardized, Mel frames are averaged over time. Because we use the most common parameters, namely 23.2 ms window with 50% overlap, the closest achievable rate considering simple averaging is 7.74 frames per second. Third-octave representation on 31 bands
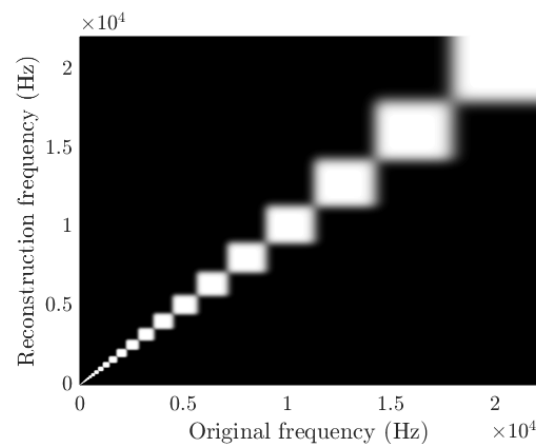
Figure 7. Third-octave bands analysis and approximate inverse transformation effects on energy location. Here a brighter gray indicates that part of the energy contained at the corresponding frequency of the original signal is transferred to the reconstructed signal at the indicated frequency. This process yields an important and heterogeneous loss in resolution, particularly at higher frequency points.

yields an overall higher size than their Mel equivalent, here estimated for 30 bands. It however compares with the 40-Mel bands representation which is the most used in sound event recognition literature. This higher bitrate is likely due to the distribution observed by the data prior to Huffman encoding.

A second set of parameters influences directly the time-frequency resolution of the analysis. By choosing a frame rate and number of bands, one can effectively control the size of the periodically transmitted data. We evaluate the bitrate for 10 to 40 Mel bands and a frame rate from 2 to 10 per second, with fixed $q = 8$. As can be expected, the bitrate for a given word size q can be approximated as a linear function of the representation dimensions for one second of analysis. Small variations are induced by data distributions on a per-frame level and their impact on the Huffman algorithm.

This experiment aims at an initial assessment of potential tradeoffs between bitrate and preservation of valuable information which is will be further discussed in Section ??4.2.

The measurement error is now computed for third-octave bands. Results are displayed in Figure 9 for white noise extracts and in Figure 10 for environmental sounds. Only full-octave bands are shown for the sake of visibility. When computing the slow analysis, we find that applying overlap ~~attenuated~~ attenuates the error. This is not observed for fast analysis where the error is similar or higher with the use of overlapping frames.

~~As can be expected when~~ When performing a short time analysis, strong errors appear in low frequency bands where ~~precision~~ the precision of the DFT is poor. The slow analysis yields better estimations as it provides a globally better resolution and thus also reduces the impact of spectral leakage effects in this region. When considering white noise the error is low on the whole frequency range. Its mean over the testing samples is close to zero and its standard deviation increases in low frequencies, confirming the above-mentioned considerations. This effectiveness does not translate well to environmental sound analysis, most likely ~~mainly~~ due to the sound level disparities absent in the former case but omnipresent in the latter. Spectral leakage induces a correlation between close frequency components. In lower frequencies, large differences between adjacent frequency bins can ~~thus~~ therefore cause important errors on the third-octave bands ~~computations~~ computation. This issue is however not specific to Fourier ~~transform based~~ transform-based schemes as it is also encountered with a time domain filtering approach.
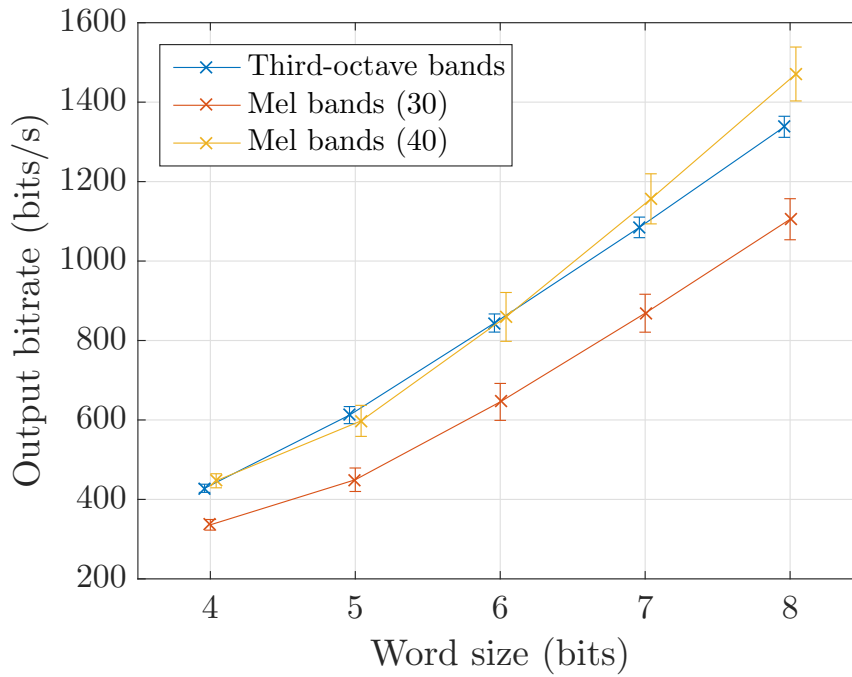
Figure 8. Coder output bitrate as a function of quantization for third-octave and Mel bands with 8 frames per second.

To mitigate this phenomenon, the choice of the window function shall be studied further. For a fast analysis with no overlap, the rectangular window seems a reasonable choice at first because of its energy conservation qualities. While it yields the lowest error in high frequency bands, its ~~high~~ important spectral leakage effect makes it unprecise ~~for~~ in lower frequency regions. Non-flat windows better account for this issue but require assuming that the signal is stationary in a given frame. The use of overlap then results in better estimates at the cost of an increased data size. For slow measurements, considering the rectangular window ~~impact~~ is less harmful to band analysis due to an increased frequency resolution.

The comparison of the error achieved by the proposed scheme to that of third-octave bands estimated by a reference time-filtering algorithm is now considered. It results in important similarities. For fast measurements and a rectangular window (Figure 10a), a student t-test ($p < 0.05$) shows that the errors are not significantly different between 125 Hz and 12 kHz bands. For the lower frequency bands, the two methods yield different error distributions but of similar importance. Using a Hann window function results in lower error in the low frequency range but adds a bias to the previously accurate band measurements. Again comparing the two computation methods shows a significant difference in the bands below 100 Hz only. A slow analysis results in statistically equivalent errors between the two methods apart from the very low frequencies: $p = 0.011$ with the rectangular window (Figure 10c), $p = 0.003$ at 20 Hz and $p = 0.021$ at 25 Hz for the Hann window and 66% overlap (Figure 10d).

The additional error caused by the encoding steps after obtaining the desired data representation is now studied. The only lossy operation applied is quantization, which is defined by an output word size q in bits. However, it corresponds to the word size *after* Δ-compression, meaning that the data is in fact quantized on 2^{q-1} values. Since these measurements are already expressed in dB, the error ε is computed as:
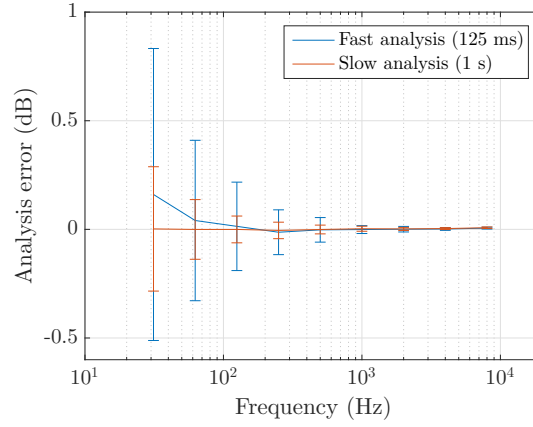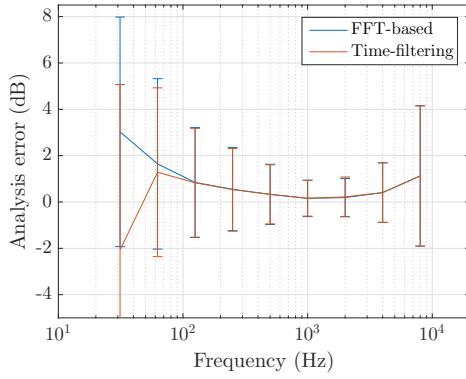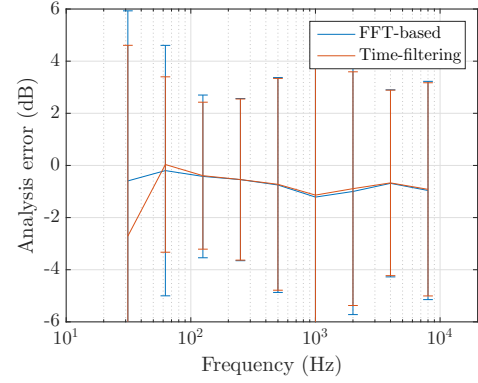
$$\varepsilon = |x_q - x|$$

Figure 9. Measurement error of third-octave bands over two-seconds white noise extracts, with rectangular window and no overlap. The analysis of short frames has an effect on energy estimation at low frequencies that is of lower importance as the frequency resolution increases.
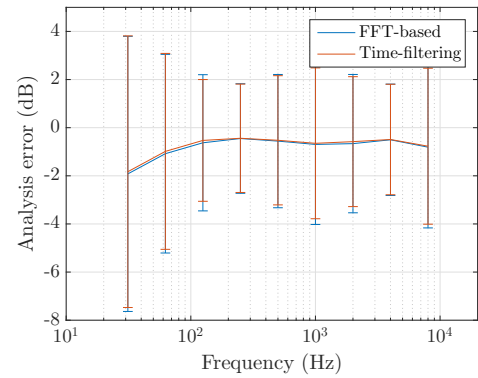

(a) Fast analysis, rectangular window, no overlap

(b) Fast analysis, Hann window, 66% overlap
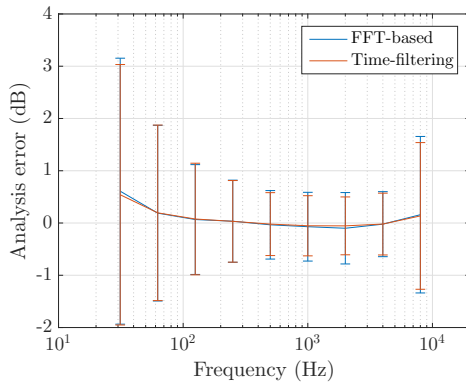
(c) Slow analysis, rectangular window, 50% overlap

(d) Slow analysis, Hann window, 66% overlap

Figure 10. Measurement error of third-octave bands over two-seconds environmental sounds recordings (US8k dataset). For most settings the FFT-based method yields similar errors compared to the time-domain filtering reference.

where x_q and x are the quantized and clean representations repectively. x_q is given by:

$$x_q(n) = \frac{\Delta_x}{2^{q-1}-1} \text{round}\left(\frac{(2^{q-1}-1)x(n)}{\Delta_x} \right), x \in [0, \Delta_x]$$

Figure 11 shows an estimation of ε as a function of q for third-octave and Mel bands. In both cases, the mean and standard deviation seem to decrease by a factor of 2 as q increases. To better understand this phenomenon, let us model x as a uniform distribution such as $x \sim U\{0, \Delta_x\}$. While this is not exact it matches our objective when using a logarithm to flatten given PDFs. The error ε will follow a uniform distribution $U \sim \{0, \frac{\Delta_x}{2 \times (2^{q-1}-1)}\}$. Therefore, the mean and standard deviation are:

$$\begin{cases} \mu_\epsilon = \frac{\Delta_x}{4 \times (2^{q-1}-1)} \\ \sigma_\epsilon = \frac{1}{12} \frac{\Delta_x}{2 \times (2^{q-1}-1)} \end{cases}$$

justifying the decaying ratio as q increases. In practice, we assume that $\varepsilon = f(\Delta_x, \frac{1}{2^{q-1}-1})$ with the heterogeneity of the data's PDF inducing small variations.
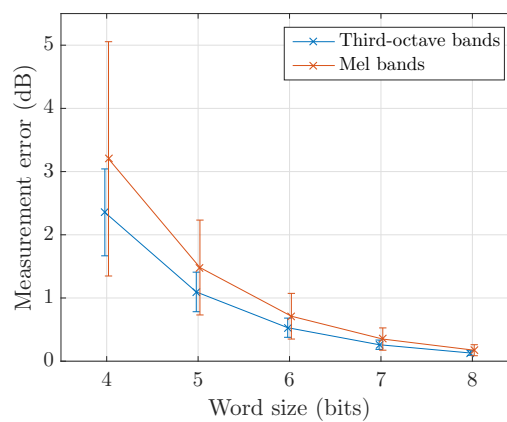
Figure 11. Measurement error induced by encoding for different quantization resolutions.

As Δ_x is sampled for each analysis frame in our implementation, the error has the same statistics across all bands. It is to be added to the measurement error discussed in the same section before comparing to the values specified in the IEC 61672-1:2013 [57] standard on class 1 and 2 sound level meters.

4.2. Event recognition

The same set of parameters as in bitrate evaluation is used to study the impact of their tuning ~~to the~~ on sound event recognition performance. ~~We thus aim~~ This experiment aims at finding ways to further reduce the encoded data size without strongly affecting ~~recognition~~ detection performance. Results of the four presented classification methods 3.2 are provided for the sake of completeness.

First, we observe the impact of quantization on classification accuracy. The models are trained on the most complete implemented representation to replicate the results of [47], ie. 40 Mel bands, 23.2 ms, 50% overlap analysis and no time averaging (85 frames per second). Figure 12 shows that ~~this process yields equivalent accuracy for higher resolutions~~ equivalent accuracies are obtained for all considered values of the quantization parameter q down to 5 bits.

The effect of changing the time-frequency resolution of the representation is presented in Table ~~??.~~ 1. In this table, equivalence of performance with respect to the best performing setting (in red) is depicted with a bold font evaluated using the following procedure. The null hypothesis that the subtraction of the two compared distribution (the distribution of the setting we consider minus the distribution of the best performing one) comes from a normal distribution with mean equal to zero and unknown variance is evaluated using the paired-sample t-test at the 0.05 significance level. If the
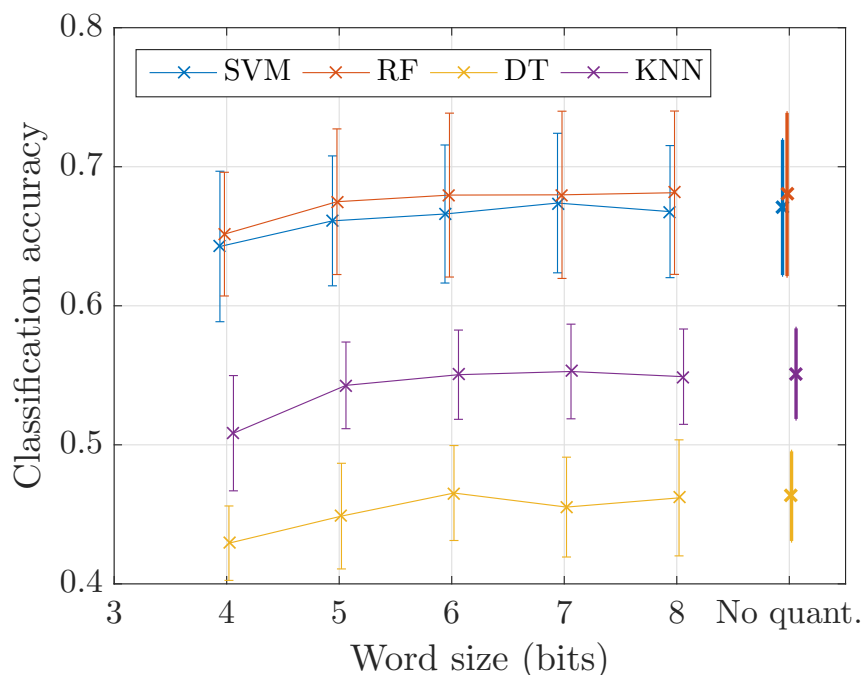
Figure 12. Classification accuracy as a function of word size before encoding. The baseline plotted on the right is computed without quantizing the representation.

null hypothesis is not rejected, the setting is considered as equivalent in terms of performance to the best performing one.

Similarly to baseline results, the random forest and SVM classifiers are the ~~higher~~ best performing systems at 0.69 ± 0.06 and 0.68 ± 0.04 respectively. However, we find that given our setup, reducing the number of analysis bands up to a given level ~~most~~ often does not induce a ~~strong~~ severe drop of performance. The ~~performance~~ accuracy of the decision tree classifier is a good example of this behavior, with best performances for 10 Mel bands only and a 4% lower accuracy for 40. ~~Even if the difference is small, all other models perform best with 30 or~~ Other models trained with 20 to 40 Mel bands as a representation show no significant performance difference. Time decimation can also be considered, as the original representation can be averaged and consistently yield a good classification accuracy. The loss of information can be considered negligible for ~~FPS~~ frame-per-second values as low as 20 Hz or 10 Hz depending on the method. Even below, the classification accuracy only drops by one ~~or~~ to a few percents. This means that we can effectively divide the data size by a factor of at most 10 without strongly affecting sound event recognition performance. It also provides us with a preliminary confirmation of the possibility for "fast"~~-sampled~~, 8 Hz third-octave bands cepstra to match MFCC performances.

4.3. Third-octave bands as base descriptors

We now compare the efficiency of third-octave bands at ~~characterizing~~ differenciating sound sources in urban acoustic environments to that of Mel spectrograms. Following the previous discussions, the classification task is run on corresponding cepstrum coefficients, computed with the following parameters: 125 ms analysis frames with rectangular windowing and no overlap are used to perform the STFT, and 31 third-octave bands (20 Hz - 20 kHz) are considered. Figure 13 displays similar results to the 40 bands, 8 fps Mel spectrograms seen in Table ~~??~~ 1 for all classification schemes.~~In this table, equivalence of performance with respect to the best performing setting (in red)~~

Table 1. Classification accuracy in percentage for different ~~classifiers~~models: SVMs (a), Random Forests (b), Decision Trees (c), and Nearest Neigbors (d) with respect to varying representation resolutions. Numbers in red indicate best performance and numbers in bold indicate that no ~~statisticaly~~ statistically significant difference is found when comparing to the best performing setting ($p < .05$).

(a)

SVM		Frames per second						
		2	4 (4.1)	6 (6.1)	8 (7.7)	10 (9.5)	20 (21)	85
Mel bands	10	55±3	60±3	61±4	62±3	62±4	**63±6**	**65±6**
	20	58±4	62±4	63±4	64±4	63±4	**65±5**	**67±6**
	30	60±3	64±4	64±4	**65±4**	**65±3**	**67±4**	**68±4**
	40	60±3	63±4	**64±4**	64±4	**64±4**	**66±4**	**68±5**

(b)

RF-500		Frames per second						
		2	4 (4.1)	6 (6.1)	8 (7.7)	10 (9.5)	20 (21)	85
Mel bands	10	60±3	62±3	62±3	63±3	63±3	**65±4**	**67±5**
	20	61±4	63±3	64±3	64±3	64±4	**66±6**	**69±6**
	30	62±3	63±3	64±3	64±3	64±4	**67±5**	**69±6**
	40	62±4	63±4	63±4	64±4	63±4	**67±6**	**68±6**

(c)

DT		Frames per second						
		2	4 (4.1)	6 (6.1)	8 (7.7)	10 (9.5)	20 (21)	85
Mel bands	10	42±3	**46±4**	46±2	44±3	45±3	**46±5**	**49±3**
	20	43±5	43±3	43±2	44±3	45±3	45±5	**47±5**
	30	42±4	43±2	44±3	45±5	43±3	43±4	**45±5**
	40	42±6	43±3	43±3	42±3	44±3	46±3	**46±4**

(d)

KNN-5		Frames per second						
		2	4 (4.1)	6 (6.1)	8 (7.7)	10 (9.5)	20 (21)	85
Mel bands	10	43±2	51±4	53±4	53±5	53±4	54±4	**56±3**
	20	44±3	52±4	53±3	54±4	**54±4**	55±4	**58±4**
	30	45±4	54±5	**55±5**	**55±4**	**55±4**	56±4	56±4
	40	46±3	53±5	**55±5**	**55±4**	**55±5**	57±4	57±3

~~is depicted with a bold font evaluated using the following procedure. The null hypothesis that the subtraction of the two compared distribution (the distribution of the setting we consider minus the distribution of the best performing one) comes from a normal distribution with mean equal to zero and unknown variance is evaluated using the paired-sample t-test at the 0.05 significance level. If the null hypothesis is not rejected, the setting is considered as equivalent in terms of performance to the best performing one.~~

To further analyze both representations advantages, the confusion matrix is a reliable tool. It aims at providing information regarding class-by-class accuracy and misclassification rates. We thus compute these metrics for the SVM classifier, with predictions accumulated over the ten folds in test configuration. The analogous designs of the two descriptors is highlighted by close one-versus-one differentiation performances, with a slightly lower accuracy for third-octave bands on average. Both representations yield best results for the *Gun shot* class with 88.2% for third-octave cepstra and 86.6% for Mel cepstra. Their poorest accuracy is on the *Air conditioning* class with 32.0% and 38.1% respectively. However, a noticeable difference between them is that on the log-frequency scale the bandwidth of Mel filters narrows as frequency increases, while third-octave are evenly distributed. The effect of this can be seen between classes *Drilling* and *Jackhammer*. These sounds involve important low-frequency information which can generally differentiate them, leading third-octave descriptors to perform better. Conversely, using Mel-based cepstra improves globally *Air conditioning* recognition as most of its defining components are situated in higher frequencies.
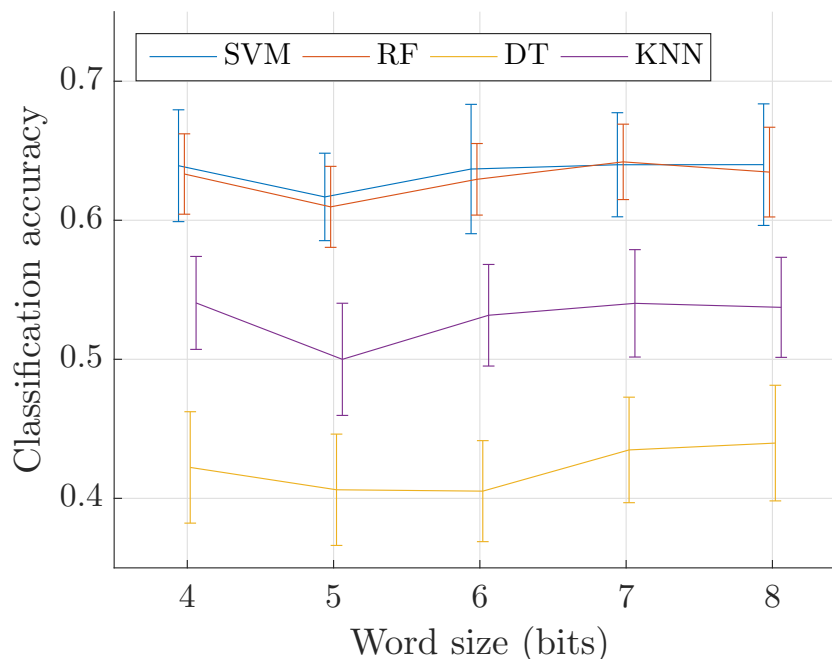
Figure 13. Classification accuracy with third-octave bands and varying word size.

Both descriptor are thus found to have similar representational capabilities despite minor differences due to their respective designs.

4.4. Inintelligibility

The perceptual test [5] then evaluates the impact of analysis frame rate on intelligibility using the average intelligibility score (AIS) given by the subjects and the intelligibility ratio (IR) computed as the number of correctly transcribed words versus the number of words in the spoken utterance. Given the test conditions presented in Section ??3.3 and the number of subjects, the mean and standard deviation are computed on an average of 54 results per frame rate parameter: 12 subjects evaluating each analysis resolution 4.5 times on average.

The results of the perceptual test, visible in Figure 14, confirm the qualitative evaluation of speech properties discussed in Section ??3.3. Phonemes separation is indeed mandatory in order to have a good intelligibility of speech. This implies an analysis frame rate significantly larger than the phonemes rate the considered language. A limit at about 10 Hz corresponding to a frame duration of 100 ms is observed, as the studied 8 Hz setting is found to be almost completely inintelligible as very few words are understood correctly at this rate. We assume that the correctly identified words are a result of coincidentally adequate frame timings, *i.e.* analysis frames matching the precise location of phoneme utterances for a short duration. Similarly, framing effects induced by the proposed processing scheme yielded errors for higher frame rates transcriptions and a globally lower perceptual intelligibility.

---

[5] Decoded speech utterances for various conditions can be listened to at https://felixgontier.github.io/cense-coder/
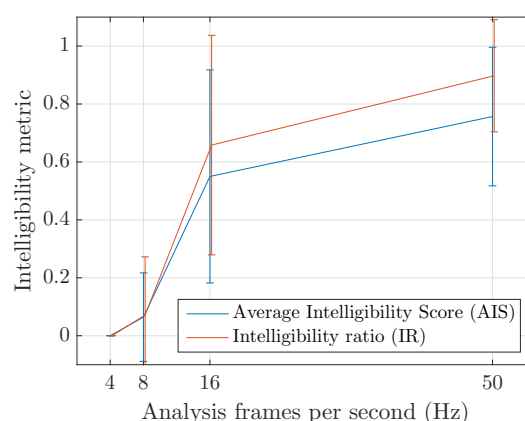
Figure 14. Mean and standard deviation of the Normalized Average intellibility score (AIS) given by the subjects and the inteligibility ratio (IR) as a function of the frame rate. Inintelligibility appears at around 10 Hz, corresponding to the average duration of a phoneme.

5. Conclusion

An audio encoding scheme has been described and motivated for use in a sensor grid approach to tackle the problem of large scale acoustic monitoring. It has been designed to transfer data from acoustic sensors to processing and storage servers in an efficient and effective manner. It relies on a lossy audio coding scheme designed to preserve the privacy of the citizens while retaining the needed amount of acoustic precision to compute standard acoustic descriptors and to allow the recognition of events of interest at a ~~state of the art~~ state-of-the-art level of recognition accuracy.

It is based on a third octave spectral representation computed at an adjustable frame rate, encoded using an Huffman encoding scheme after logarithmic compression and differential encoding.

The proposed approach has the following benefits:

- precision: the impact of the quantization step for 8 bits codewords is found to be very small (< ~~0.5~~0.2dB) and thus:
  - does not impact the quality of the acoustics indicators ~~.~~(Sections 3.1, 4.1).
  - enables the sound recognition with the same quality than directly from the spectrogram (Sections 3.2, 4.2 and 4.3).
- efficiency: using the high precision parameters, the bitrate is about 1.4 kbps and 0.4 kbps for fast and slow modes respectively (Section 4.1).
- privacy: at all operating modes, the reverted audio signal is found to be inintelligible according to a perceptual test done using clean speech recorded at very low background level. Additionally, the sensors being typically hanged at 4 meters high in a usually very noisy environment, we believe that the intelligibility risk of the proposed approach is negligible (Sections 3.3, 4.4).

The encoding scheme is designed to be easy to implement and to lead to an efficient data format for storage and sharing across the research community. To encourage its use for other research projects, an open source Matlab and Python implementations of the coder are available.

~~Future~~ This scheme will be implemented with a large-scale sensor grid as part of the CENSE project, where sensors are designed to be low-cost and energy efficient. The aim of the project is to create noise maps by linking standard acoustic measurements with novel perceptually motivated metrics. The presented coding scheme will enable a thorough study of the correlation between sound sources and context with pleasantness as perceived by the citizens to better understand the content of urban acoustic scenes and its perception by city dwellers.

To do so, future work will focus on the use of such data encoder in a sensor grid approach as part of the ~~Cense~~ CENSE project where its effectiveness will be assessed in field conditions. ~~The~~ As

far as the proposed encoding scheme is concerned, the usefulness of more sophisticated encoding paradigms should also be studied in order to further improve bitrate efficiency~~should also be studied~~, as well as the effect of multiple channel coding and its interest in sensor grids applications.

6. Acknowlegments

References

1. Mydlarz, C.; Salamon, J.; Bello, J. The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics* **2017**, *117*, 207–218.

2. Picaut, J.; Can, A.; Ardouin, J.; Crépeaux, P.; Dhorne, T.; Écotière, D.; Lagrange, M. Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling. The Journal of the Acoustical Society of America. Boston, 2017.

3. Lim, H.; Teo, Y.; Mukherjee, P.; Lam, V.; Wong, W.; See, S. Sensor grid: integration of wireless sensor networks and the grid. Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on. IEEE, 2005, pp. 91–99.

4. Tham, C.K.; Buyya, R. SensorGrid: Integrating sensor networks and grid computing. *CSI communications* **2005**, *29*, 24–29.

~~Lavandier, C.; Defréville, B. The contribution of sound source characteristics in the assessment of urban soundscapes. Acta Acustica united with Acustica2006, 92, 912–921.~~

~~Aumond, P.; Can, A.; De Coensel, B.; Botteldooren, D.; Ribeiro, C.; Lavandier, C. Modeling Soundscape Pleasantness Using perceptual Assessments and Acoustic Measurements Along Paths in Urban Context. Acta Acustica united with Acustica2017, 103, 430–443.~~

~~Defréville, B.; Pachet, F.; Rosin, C.; Roy, P. Automatic recognition of urban sound sources. Audio Engineering Society Convention 120. Audio Engineering Society, 2006.~~

~~Mydlarz, C.; Shamoon, C.; Baglione, M.; Pimpinella, M. The design and calibration of low cost urban acoustic sensing devices. EuroNoise 2015. Euronoise, 2015.~~

~~Pan, D. A tutorial on MPEG/audio compression. IEEE multimedia1995, 2, 60–74.~~

5. Rey Gozalo, G.; Trujillo Carmona, J.; Barrigon Morillas, J.; Gomez Escobar, V. Relationship between objective acoustic indices and subjective assessments for the quality of soundscapes. *Applied Acoustics* **2015**, *97*, 1–10.

6. Rychtarikova, M.; Vermeir, G. Soundscape categorization on the basis of objective acoustical parameters. *Applied Acoustics* **2013**, *74*, 240–247.

7. Can, A.; Aumond, P.; Michel, S.; De Coensel, B.; Ribeiro, C.; Botteldooren, D.; Lavandier, C. Comparison of noise indicators in an urban context. 45th International Congress and Exposition on Noise Control Engineering, 2016, pp. 5678–5686.

8. Can, A.; Gauvreau, B. Describing and classifying urban sound environments with a relevant set of physical indicators. *J. Ac. Soc. Am.* **2015**, *137*, 208–218.

9. Brocolini, L.; Lavandier, C.; Quoy, M.; Ribeiro, C. Measurements of acoustic environments for urban soundscapes: Choice of homogeneous periods, optimization of durations, and selection of indicators. *J. Ac. Soc. Am.* **2013**, *134*, 813–821.

10. Nilsson, M.; Botteldooren, D.; De Coensel, B. Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas. 19th International Congress on Acoustics, 2007.

~~Ishiyama, T.; Hashimoto, T~~

11. Lavandier, C.; Defréville, B. The ~~impact of sound quality on annoyance caused by road traffc noise: an influence of frequency spectra on annoyance~~contribution of sound source characteristics

in the assessment of urban soundscapes. ~~*JSAE Review*~~*Acta Acustica united with Acustica* ~~2000~~2006, ~~21~~92, ~~225–230~~912–921.

12. Aumond, P.; Can, A.; De Coensel, B.; Botteldooren, D.; Ribeiro, C.; Lavandier, C. Modeling Soundscape Pleasantness Using perceptual Assessments and Acoustic Measurements Along Paths in Urban Context. *Acta Acustica united with Acustica* **2017**, *103*, 430–443.

13. Alsina-Pagès, R.; Hernandez-Jayo, U.; Alías, F.; Angulo, I. Design of a Mobile Low-Cost Sensor Network Using Urban Buses for Real-Time Ubiquitous Noise Monitoring. *Sensors* **2016**, *17*, 57.

14. Alías, F.; Socoró, J. Description of Anomalous Noise Events for Reliable Dynamic Traffic Noise Mapping in Real-Life Urban and Suburban Soundscapes. *Applied Sciences* **2017**, *7*.

15. Gloaguen, J.R.; Can, A.; Lagrange, M.; Petiot, J.F. Estimating traffic noise levels using acoustic monitoring: a preliminary study. DCASE 2016, Detection and Classification of Acoustic Scenes and Events, 2016, p. 4p.

16. Aumond, P.; Can, A.; De Coensel, B.; Botteldooren, D.; Ribeiro, C.; Lavandier, C. Modeling Soundscape Pleasantness Using perceptual Assessments and Acoustic Measurements Along Paths in Urban Context. *Acta Acustica united with Acustica* **2017**, *103*, 430–443.

17. Defréville, B.; Pachet, F.; Rosin, C.; Roy, P. Automatic recognition of urban sound sources. Audio Engineering Society Convention 120. Audio Engineering Society, 2006.

18. Mydlarz, C.; Shamoon, C.; Baglione, M.; Pimpinella, M. The design and calibration of low cost urban acoustic sensing devices. EuroNoise 2015. Euronoise, 2015.

19. Pan, D. A tutorial on MPEG/audio compression. *IEEE multimedia* **1995**, *2*, 60–74.

20. Ishiyama, T.; Hashimoto, T. The impact of sound quality on annoyance caused by road traffc noise: an influence of frequency spectra on annoyance. *JSAE Review* **2000**, *21*, 225–230.

21. Anusuya, M.; Katty, S. Speech recognition by machine, a review. *International Journal of Computer Science and Information Security* **2009**, *6*, 181–205.

22. Tzanztakis, G.; Essl, G.; Cook, P. Automatic musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* **2002**, *10*, 293–302.

23. Ntalampiras, S. Universal background modeling for acoustic surveillance of urban traffic. *Digital Signal Processing* **2014**, *31*, 69–78.

24. Aucouturier, J.; Defreville, B.; Pachet, F. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *J. Ac. Soc. Am.* **2007**, *122*, 881–891.

25. Foggia, P.; Petkov, N.; Saggese, A.; Strisciuglio, N.; Vento, M. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters* **2015**, *65*, 22–28.

26. Kumar, A.; Raj, B. Features and Kernels for Audio Event Recognition. https://arxiv.org/abs/1607.05765, 2016.

27. Radhakrishnan, R.; Divakaran, A.; Smaragdis, P. Audio analysis for surveillance applications. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.

28. Salamon, J.; Bello, J. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* **2017**, *24*, 279–283.

29. Piczak, K. Environmental sound classification with convolutional neural networks. IEEE 25th International Workshop on Machine Learning for Signal Processing, 2015.

30. Khunarsal, P.; Lursinsap, C.; Raicharoen, T. Very short time environmental sound classification based on spectrogram pattern matching. *Information Sciences* **2013**, *243*, 57–74.

31. Couvreur, L.; Laniray, M. Automatic noise recognition in urban environments based on artificial neural networks and hidden markov models. The 33rd International Congress and Exposition on Noise Control Engineering, 2004.

32. Cai, R.; Lu, L.; Hanjalic, A.; Zhang, H.; Cai, L. A flexible framewok for key audio effects detection and auditory context inference. *IEEE Transactions on Audio, Speech, and Language Processing* **2006**, *14*, 1026–1039.

33. Chu, S.; Narayanan, S.; Kuo, C. Environmental sound recognition with time-frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing* **2009**, *17*, 1142–1158.

~~Salamon, J.; Bello, J. Unsupervised feature learning for urman sound classification. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, 2015.~~

34. Baugé, C.; Lagrange, M.; Andén, J.; Mallat, S. Representing environmental sounds using the separable scattering transform. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

35. Salamon, J.; Bello, J. Feature learning with deep scattering for urban sound analysis. 23rd European Signal Processing Conference, 2015.

36. Salamon, J.; Bello, J. Unsupervised feature learning for urban sound classification. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, 2015.

37. Ntalampiras, S.; Potamitis, I.; Fakotakis, N. Acoustic Detection of Human Activities in Natural Environments. *J. Audio Eng. Soc* **2012**, *60*, 686–695.

38. Chachada, S.; Kuo, C. Environmental sound recognition: A survey. 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2013.

39. Torija, A.; Ruiz, D.; Ramos-Ridao, A. Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes. *J. Ac. Soc. Am.* **2013**, *134*, 791–802.

40. Davis, S. Octave and fractional-octave band digital filtering based on the proposed ANSI standard. 1986 IEEE International Conference on Acoustics, Speech and Signal Processing, 1986.

41. Antoni, J. Orthogonal-like fractional-octave-band filters. *J. Ac. Soc. Am.* **2010**, *127*, 884–895.

42. ANSI S1.1-1986, (ASA 65-1986)—Specifications for Octave-Band and Fractional-Octave-Band Analog and Digital Filters, 1993.

43. IEC 61260-1:2014 —Electroacoustics - Octave-band and fractional-octave-band filters - Part 1: Specifications, 2014.

44. Couvreur, C. Implementation of a one-third-octave filter bank in Matlab. http://citeseer.ist.psu.edu/24150.html.

45. Huffman, D. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE* **1952**, *40*, 1098–1101.

46. Berzborn, M.; Bomhardt, R.; Klein, J.; Richter, J.G.; Vorländer, M. The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing. 43th Annual German Congress on Acoustics, Kiel (Germany), 6 Mar 2017 - 9 Mar 2017, 2017.

47. Salamon, J.; Jacoby, C.; Bello, J. A Dataset and Taxonomy for Urban Sound Research. 22nd ACM international conference on Multimedia, 2014.

48. Ntalampiras, S. A Novel Holistic Modeling Acoustic Approach for Generalized Sound Recognition. *IEEE Signal Processing Letters* **2013**, *20*, 185–188.

49. Ellis, D. PLP and RASTA (and MFCC, and inversion) in Matlab. http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/, 2005.

50. Baldwin, J.; French, P. *Forensic phonetics*; Pinter Publishers, 1990.

51. Boë, L.J. Forensic voice identification in France. *Speech Communication* **2000**, *31*, 205–224.

52. Kuwabara, H. Acoustic properties of phonemes in continuous speech for different speaking rate. Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on. IEEE, 1996, Vol. 4, pp. 2435–2438.

53. Rosen, S. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **1992**, *336*, 367–373.

54. ITU-T P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, 2003.

55. Ntalampiras, S.; Ganchev, T.; Potamitis, I. A Novel Holistic Modeling Acoustic Approach for Generalized Sound Recognition. 1st international conference on PErvasive Technologies Related to Assistive Environments, 2008.

56. Griffin, D.; Lim, J. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1984**, *32*, 236–243.

57. IEC 61672-1:2013 —Electroacoustics - Sound level meters - Part 1: Specifications, 2013.