A sensor grid approach for large scale acoustic monitoring

Félix Gontier, Mathieu Lagrange, Pierre Aumond, Arnaud Can, Catherine Lavandier

felix.gontier@reseau.eseo.fr

Abstract

The acoustic environment represents both societal and environmental concerns. Monitoring the acoustic environnment in urban but also rural or wilderness areas is therefore an important matter. Fostering on the recent development of low cost hardware acoustic sensors, we propose to consider a sensor grid approach to tackle this issue.

In this kind of approach, the crucial question is the nature of the data that is transmitted from the sensors to the processing and archival servers. To this end, we propose an efficient audio coding scheme based on third octave band spectral representation that allows 1) the estimation of most acousitc indicators and 2) the recognition of acoustic events at state of the art performance rate. The former is useful to provide quantitative information about the acoustic environment, while the latter is useful to gather more qualitative information and build perceptually motivated indicators using for example some emergence of a given sound source.

The coding scheme is also demonstrated to transmit spectrally encoded data that, reverted to the time domain using state of the art techniques, is not intelligible, thus protecting the privacy of citizens.

Keywords:
acoustic monitoring, soundscape assessment,

1. Introduction

The advent of low cost acoustic sensors together with the need to better monitor and comprehend the acoustic environment of urban and wilderness areas give rise to the deployment of experimental sensor networks such as the sonyc (wp.nyu.edu/sonyc) and cense (cense.ifsttar.fr) projects.

To do so, considering the so-called "sensor grid" approach [29, 43] has several advantages. The sensor grid approach put the focus on a) a distributed system of data acquisition with a large set of sensors, b) the production and storage of a large dataset, and c) its availability for intensive and open data analysis computation.

In our application setting, the requirements are the following. First, the number of sensor nodes shall be extended as needed without having to change the hardware and software architecture, *i.e.* the approach shall be scalable. Second, the nodes shall be energy efficient in order ot ease the deployment of the sensors grid to the desired topology, ideally autonomous. Third, the encoding scheme used to transmit the data from the sensor to the storage and processing servers shall be designed with care in order to ensure the privacy of the citizen. Lastly, the data stored shall be as rich as possible while remaining a low bitrate for efficient transmission and storage.

The data of interest is first acoustic indicators (LAeq, ...) and second the presence of sound sources of interest (bird calls, sirens, explosion, ...). The latter allow us to better assess soundscapes in terms of pleasantness and other perceptual indicators [28, 7]. This detection step can respectively be operated online (on the sensors) or offline (on the data servers).

The former is efficient in terms of data storage as only the detection events are transmitted and is thus currently considered in several approaches [18, 31, 32]. Though, it requires the availability of computing resources on the sensors in order to perform the detection step which as of today's efficiency of hardware do 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 posterior analyses.

The latter scheme has several benefits. First, the sensor is much simpler and can thus be autonomous in terms of energy, easing 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 analysed following newer classification schemes or using new indicators.

However, transmitting the raw audio through the network has several disavantages, in terms of required bandwidth, storage capabilities and privacy. Thus, for transmission from the sensors to the storage unit, the data can be encoded in a more efficient way than lossy audio encoding standards [**?**] as 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 ensure the privacy of the citizens.

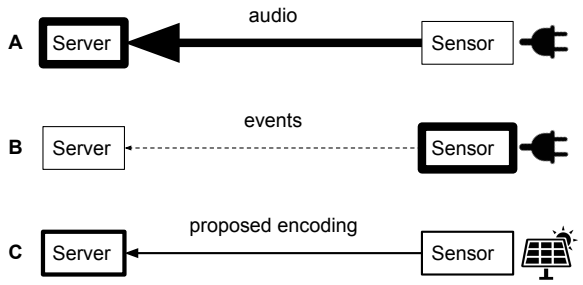

Figure 1: Alternative implementations of the sensor grid approach for the monitoring of the acoustic environnment. In A) the raw audio is transmitted, in B) only the detected events, and in C) a compressed spectral representation. Thickness of arrows indicate bandwidth and thickness of boxes indicates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

We thus propose in this paper a audio coding scheme specifically tailored for use in a sensor grid which has the following features. It is of low bitrate but still allows the computation of most of the standard acoustics indicators with high precision. As far as acoustic event detection is concerned, we report equivalent performance of several state of the art classification schemes from features computed using raw audio data and encoded data. Finally, according to preliminary perceptual evaluation, the proposed coding scheme very strongly degrades the intelligibility, thus ensuring citizen privacy.

In order to promote reproducible research, the coder as well as the experiments needed to generate the figures will be made available to the community. The remaining of the paper is organized as follows: Section 2 reviews the types of data that needs to be estimated using the transmitted data and Section 3 describes the proposed encoding scheme. The experimental protocol designed to validate the proposed approach is described in Section 4 and results are discussed in Section 5.

2. Background

The present work is constrained to allow both acoustic monitoring and sound event recognition in urban

soundscapes. Here we briefly review state-of-the-art methods in both subjects in order to better motivate our choices.

2.1. Acoustic monitoring

The considered sensor network primarily aims at monitoring urban soundscapes, that is, continously assessing their content and impact on the population. In the literature, this is typically enabled by measuring energetic acoustic indicators such as the equivalent sound pressure level L_{eq} in dB SPL or its A-weighted equivalent L_{Aeq} in dBA. However, while these indicators have proven to correlate well with perceptual evaluations for negative impact sound environments [37], they are not sufficient to fully describe soundscapes [38]. Many other variables can be derived to better account for previously implicit properties [11] including percentile values or time evolution. Studies have been conducted to select relevant subsets of descriptors in sound environment characterization [12, 9, 33].

All the mentionned indicators are measured at periods ranging from 0.125 ms or 1 s (resp. fast and slow) to longer periods of several minutes. Therefore, sound frequency information remains unused despite being closely related with subjective evaluations [24]. A simple solution is the calculation of energetic indicators for the 31 third-octave bands within the human audition range 20 Hz - 20 kHz.

This measurement appears well-suited for this work's purpose: in addition to being an efficient descriptor [44], it allows for the computation of most cited indicators while representing reasonably small, fixed amounts of data to be transmitted.

2.2. Event detection

Eventhough the above discussed indicators provide more richness than L_{eq}, recent studies show that abstract statistics are still not sufficient to fully model the human perception of soundscapes. The recognition of sources of interest is thus of importance. As far as the encoding scheme is concerned, it is thus important that the encoded data allows the computation of the above cited indicators but also the recognition of sources of interest with state of the art methods.

The recognition of sources of interest from audio streams has been the subject of extensive research in the past on speech [5], music [45], and lately more complex scenes in which the current work falls. Studied classification methods are diverse, ranging from time-dependant modeling with HMMs [34] to "bag-of-frames" approaches [6, 20]. Common architectures include learning-based classifiers such as support vector machines (SVM) [27], gaussian mixture models (GMM) [36] or neural networks [41, 35]. However, the selection of relevant features is still an open debate. The most used are certainly spectral [26] or cepstral [16] representations of the signal. Among them, mel spectrograms and their cepstral-domain derivation, the mel-frequency cepstrum coefficients (MFCC), are the most recurrent. These representations effectively try to model the human cochlear response to sounds by grouping frequency components around critical bands in a logarithmic scale, thus bearing important physical significance. They may also be exploited together with features computed in other domains to better model signal properties. For instance, [10] adds spectral features related to harmonicity and salient frequencies, and [14] uses a matching pursuit (MP) algorithm to deduce time-domaid features. Another promising solution is feature engineering via unsupervised learning, which [40] implements with a k-means clustering technique. Alternative data representations such as the scattering transform [8] show good results in environmental sound classification tasks [39].
A more detailed review of used methods is available in [13].

3. Encoding scheme

The technical constraints of size inherent to the studied setup make it impossible to transmit raw audio recordings. Following the considerations briefly exposed in the previous section, we conclude that third-octave bands are a reasonable choice. As mentionned, they provide advantages in both acoustic monitoring quality and reduced data volume per measurement period. The physical content is also close to that of mel spectrograms, being just another filterbank transform with logarithmic frequency scaling.

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 [17]. In practice, however, this technique presents multiple issues. [4] presents an alternative analysis method based on direct frequency weighting. The weights matrix design procedure complies with both ANSI S1.1-1986 [1] and IEC 61260-1:2014 [3] standards. It also respects 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 [15] of the time-domain filtering, as shown in Figure 2. The major difference is that gains at cutoff frequencies are fixed at the optimal -3dB by design.

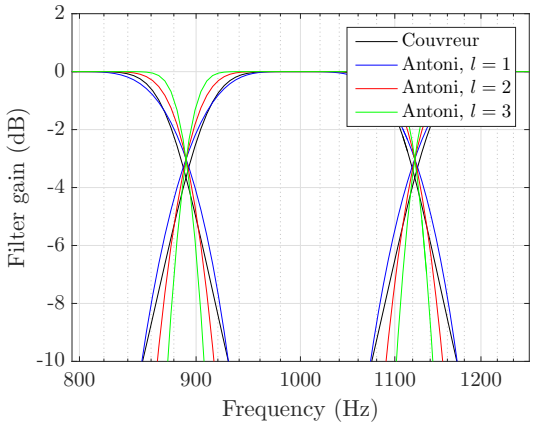

Figure 2: Comparison of Couvreur's and Antoni's implementations of third-octave filters. Frequency-weighting allows for arbitrary transfer functions and thus more accurate gains as standards impose.

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 125 ms frames to provide "fast" acoustic measurements. The data is zero-padded to the next power of two to allow best fft performances. No overlap is used as it doubles both computational and memory costs and is not required for our application. A rectangular window is thus applied to ensure energy conservation in a given analysis frame. The phase is unimportant to third-octave analysis, thus it is discarded. Similarly, negative frequency components contain the same information as positive ones because the base signal is real. We then compute third-octave bands from the squared spectral magnitude by matrix multiplication with the frequency weights proposed in [4] for $l = 2^1$. We include analysis for bands i from -17 to 13, ie. center frequencies f_i ranging from 20 Hz to 20 kHz with $f_0 = 1\ kHz$. This means the main representation is composed of 31 values sampled every 125 ms.

Alternatively, we also implement the mel filterbank for comparison purposes, as the mel filterbank is the root representation of many state of the art classification algorithms. Mel spectrograms can thus be used as a baseline method to evaluate the relative performance of third-octave bands in sound event recognition.

As the Mel representation is not bounded by constraints such as conservation of energy, specific framing parameters or fixed resolution, it also offers a more flexible tool to study intelligibility alteration and coder performance, both of which we take interest in.

Mel spectrograms are computed with the *rastamat* library [19]. 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.

3.1. Data Encoding

A Huffman coding scheme [23] is then used to further reduce data dimensionality. As in most entropy coding algorithms, the efficiency of this technique depends on two important factors. A reduced amount of symbols in the dictionnary yields smaller code size on lower probability symbols. Lower data entropy, ie. lower average information content of the signal distribution, also increases performance. The first is generally obtained by applying a quantization process to the signal, whereas the second is directly linked with the probability density function (PDF) of these symbols. 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 and n the numerical base in which information is represented.

As such, 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 3a, immediately applying a linear quantization clearly results in most of the information being lost. We therefore spread the PDF by taking the logarithm of the representation. A linear quantization is then applied with $2^{q-1}-1$ output values to obtain Figure 3b. Finally, we use a Δ-encoding algorithm along the time dimension to reduce redundancies. This also effectively concentrates higher probabilities on symbols around zero (Figure 3c). This yields a higher amount of symbols at $2^q - 1$ but a nevertheless smaller entropy.

In the example shown here, the former is $H_{log} = 6.24\ sh$ and the latter $H_\Delta = 3.54\ sh$. Huffman encoding is then computed with either a frame-specific symbol-code dictionnary or a constant one generated from an entire dataset. A comparison of both methods is shown in Figure 4. For most short texture frame durations, we found the local Huffman algorithm to be

[1] 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.

much faster. In fact, encoding complexity is function of the amount of dictionnary elements. If it is specific to a given data packet, then it does not need to contain every possible value. This gain seem to outweight the additional complexity induced by the tree generation algorithm. Alternatively, short texture frames or single analysis windows can be encoded using one general dictionnary to improve bitrate. This choice depends on which of the size and complexity parameters must be reduced most.

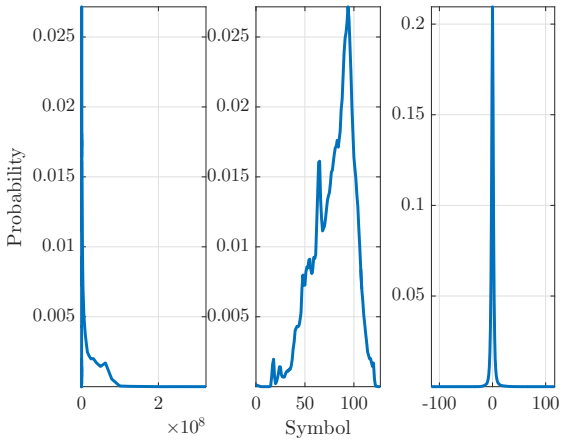

Figure 3: Example of estimated probability density functions of the data throughout the encoding step. (left) Unchanged output of the representation step, concentrated towards very low values. (middle) PDF "flattening" effect induced by logarithm application. Here values are mapped to the range $[0, 2^7 - 1]$ and rounded to perform quantization. (right) Output of the Δ compression, with desirable probabilities as the input to a Huffman algorithm.

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

4. Validation protocol

A set of metrics is computed to assess the efficiency of the proposed coder scheme and determine the impact of the algorithm's parameters, which are as follow: the desired word size prior to Huffman encoding, set by quantization for both signal representation choices, and only for Mel bands the number of coefficients between 10 and 40. We also study the effect of time averaging analysis frames: to force a fixed amount of frames per second lesser than the average speech rate in phonemes per second is a possible way to alter intelligibility at the cost of temporal information resolution.

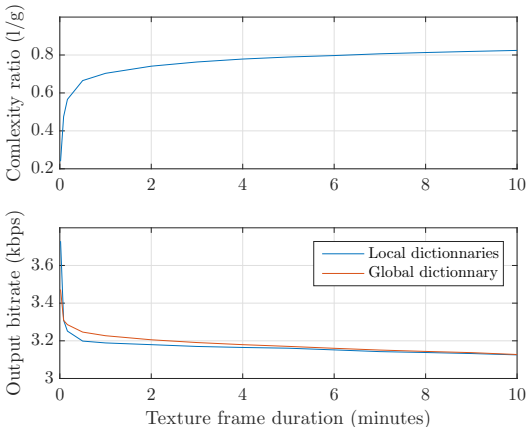

Figure 4: Performance comparison of locally and globally generated Huffman dictionnaries. (top) The mean execution time ratio favors the use of frame-specific dictionnaries for tested texture frame lengths. (bottom) The mean output bitrate is close for both algorithms, showing that the necessity of sending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

4.1. Efficiency

The coder's output bitrate as well as additional measurement error is computed for both data representations : third-octave and Mel. It is expressed in dB and compared to the IEC 61672-1:2013 standard [2] on sound level meter tolerance. This is to verify that the quantization step yields a maximum absolute error inferior to the precision of the target recording device. To provide relevant statistics, these quantities are estimated for a total of 53 10-minutes texture frames. They are obtained from the UrbanSound8k dataset presented in the next section.

4.2. Event recognition

In order to ensure that the proposed scheme allows the recognition of events using state of the art methods, we consider the UrbanSound8k dataset[42]. It features about 9 hours of urban environmental sounds recordings separated in 8732 wave files ranging from 1 to 4 seconds each. The recordings are labeled in 10 classes (air conditioner, car horn, children playing, dog bark, drilling, enginge idling, gun shot, jackhammer, siren, street music), and distributed in 10 independant 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 files are resampled 44.1 kHz, normalized and reduced to a single channel.

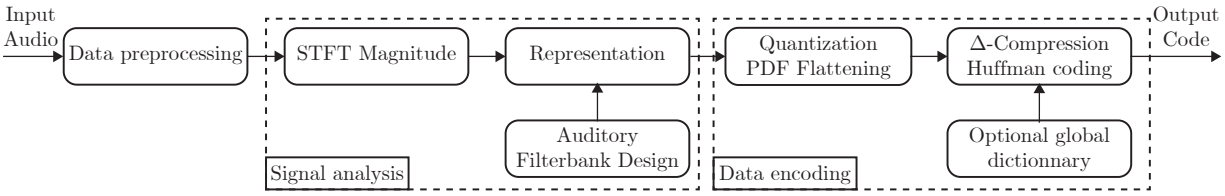

Figure 5: Overview of the coder process.

We evaluate the global loss of information by implementing the four classification models proposed in [42]: 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 found with a grid search. We apply a discrete cosine transform to the critical band signal representation to obtain cepstra (known as Mel Frequency Cepstrum Coefficients or MFCC in the first case) of which we conserve the 25 first coefficients, and summarize along time with the mean, variance, skewness, kurtosis, minimum, maximum, median, derivative mean and variance, second order derivative mean and variance. The feature vector is thus comprised of 275 values to reproduce available results and compare with our own. Models are trained for each setup using a 10-fold cross validation method provided by the authors of [42], that is, every combination of testing one fold on models trained with the other nine.

4.3. Inintelligibility

Intelligibility in decoded and reconstructed audio is also oan concerns of this study. It is indeed important that the produced scheme maintains a high level of recognition of acoustic events but, as the dat is transmitted over the network and potentially stored, the level of intelligibility of the decoded stream should be as low as possible.

We thus provide the computation of two objective metrics: the Coherence Speech Intelligibility Index[25] (CSII) and frequency-weighted segmental SNR[22] (fwSNRseg) by comparing original unaltered samples and recovered audio. These two indicators have been shown to correlate well with perceptual tests[30], although results are only available for lesser degradations and distortions than the ones we consider in this study.

These metrics require a recovery of the time-domain signal. To do so, a linearly-scaled spectrogram from the band-focused representations is estimated. This estimation can be achieved by multiplying 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 6. Signal phase is approximated by either white noise spectrogram scaling or a Griffin & Lim algorithm [21], and the signal is retrieved using concatenation or overlap-add. When using overlap to compute third-octave bands, one can also avoid framing effects produced by rectangular windowing by convoluting the signal with another windowing function prior to inverse-STFT computation.

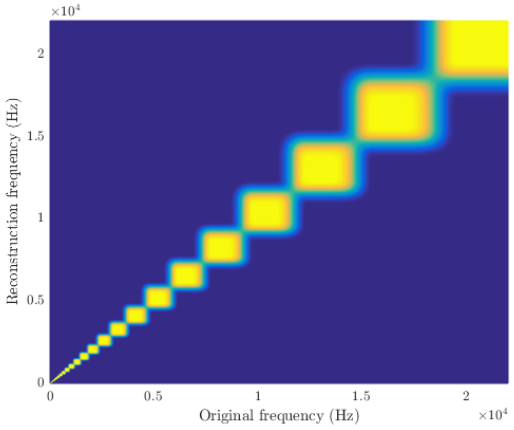

Figure 6: Third-octave bands analysis and approximate inverse transformation effects on energy location. This process yields an important and heterogeneous loss in resolution, particularly at higher frequency points.

5. Results

5.1. 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 7 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, we averaged mel frames 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 yielded 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 litterature. This occurence is likely due to the distribution observed by the data prior to Huffman encoding.

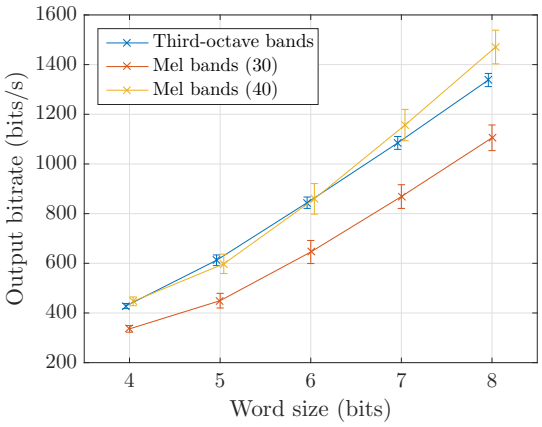

Figure 7: Coder output bitrate as a function of quantization for third-octave and mel bands with 8 frames per second.

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 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 $q = 8$. Results are exposed in Figure 8. As expected, the bitrate for a given word size q can be modeled 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 effect on the Huffman algorithm.

This experiment aims at assessing potential tradeoffs between bitrate and information. As such, it will be further discussed in the *Event recognition* subsection.

We also examine the additionnal error caused by the encoding steps after obtaining the desired data representation. The only lossy operation applied is quan-

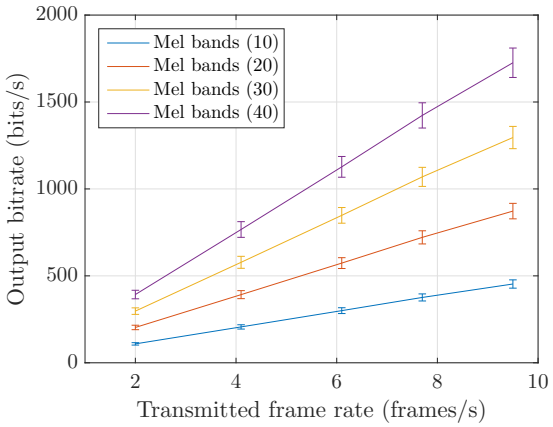

Figure 8: Impact of representation resolution on the encoded data bitrate.

tization, 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:

$$\varepsilon = |x_q - x|$$

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 9 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 explain 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.

As Δ_x is sampled for each analysis frame in our implementation, the error has the same statistics across all bands. We compare the observed values against those specified in the IEC 61672-1:2013 [2] for $f = 1\ kHz$

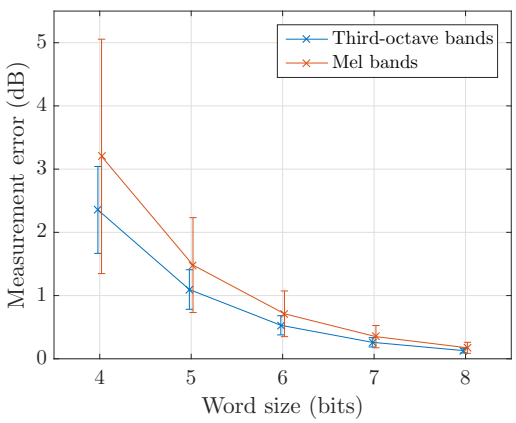

Figure 9: Measurement error induced by encoding for different quantization resolutions.

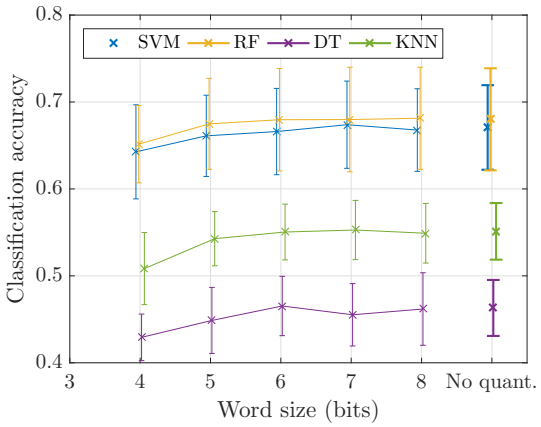

Figure 10: Classification accuracy as a function of word size before encoding. The baseline on the right is computed without quantizing the representation.

as it is where the most accurate measurements are expected. For class 1 and 2 sound level meters, the tolerance is set at about $\pm 1\ dB$ and $\pm 1.5\ dB$ respectively. Results for 6, 7 or 8-bit words lie within the shorter range and are therefore compliant with those standard's measurement tolerance.

5.2. Event recognition

The same set of parameters as in bitrate evaluation is used to study the impact of their tuning to the sound event recognition performance. We thus aim at finding ways to further reduce the encoded data size without strongly affecting recognition performance. Results of the four presented classification methods 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, ie. 40 mel bands and no time averaging (85 frames per second). Figure 10 shows that this process yields equivalent accuracy for higher resolutions.

The effect of changing representation time-frequency resolution is presented in Table 1. Similarly to baseline results, the random forest and SVM classifiers are the higher performing systems at 0.69 ± 0.06 and 0.68 ± 0.04 respectively. However, we find that given our setup, diminishing the number of analysis bands to some extent most often does not induce a strong drop of performance. The performance 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 40 mel bands as a representation. Time decimation is also possible, as the original representation can be averaged and consistently yield a good classification accuracy. The loss of information can be considered negligible for FPS values as low as 20 or 10 depending on the method. Even below, classification accuracy only drops by one or a few percents. This means that we can effectively divide the data size by a factor of at most 10 without affecting sound event recognition performance. It also provides us with a preliminary confirmation of the possibility for "fast"-sampled third-octave bands cepstra to match MFCC performances.

5.3. Third-octave bands as base descriptors

We now compare the efficiency of third-octave bands at characterizing urban soundscapes to that of mel spectrograms. Following the previous discussions, the classification task is run on corresponding cepstra with a fixed 8 frames per second and 31 bands. Figure 11 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) is evaluated using the following procedure. We evaluate the null hypothesis that the substraction 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, 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

Table 1: Classification accuracy in percentage for different classifiers: 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 statistical equivalent results compared to teh best performing setting.

(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±0.05**	**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**

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 cumulated over the ten folds in test configuration. The analogous natures 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

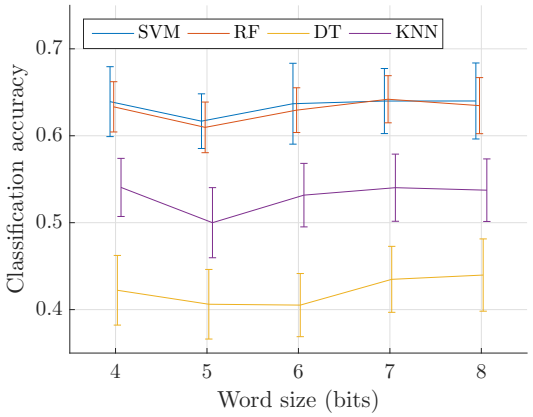

Figure 11: Classification accuracy with third-octave bands and varying word size.

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.

Both descriptor are thus found to have similar representational capabilities despite minor construction differences.

5.4. Inintelligibility

Finally, objective intelligibility metrics are estimated for the studied parameters and compared to an informal perceptual evaluation [2]. Figures 12 and 13 shows both the CSII and fwSNRseg indicators for varying representation dimensions. Figure 14 presents the results of the preliminar perceptual test for the same setup.

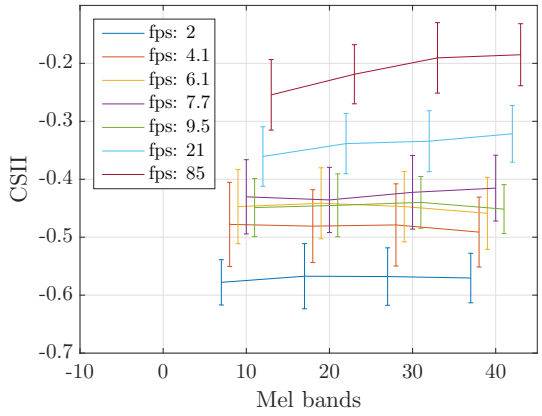

Figure 12: The Coherence SII objective indicator computed for intelligibility assessment. Negative CSII values indicate a signal-to-noise ratio too low for the recommended normalization, thus severe disparities between the original and reconstructed signal.

6. Conclusion

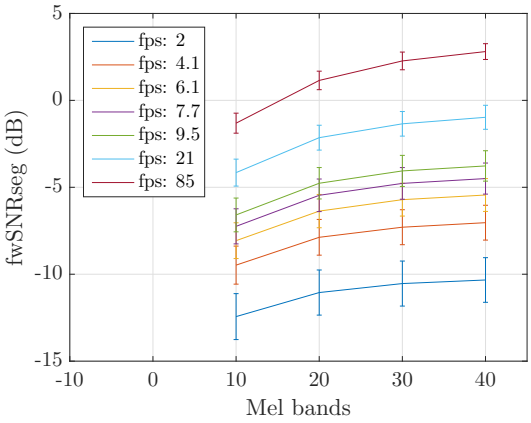

Figure 13: Frequency-weighted segmental SNR for the same parameters.

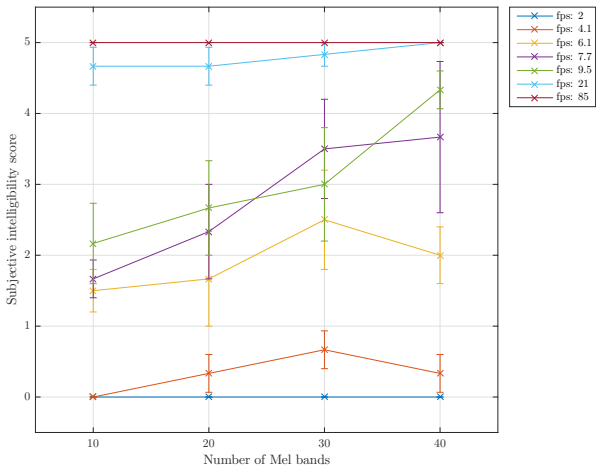

Figure 14: Results of the perceptual intelligibility test.

[2] Decoded speech utterances for various conditions can be listened to at `http://`

References

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

[2] IEC 61672-1:2013 —Electroacoustics - Sound level meters - Part 1: Specifications, 2013.

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

[4] J. Antoni. Orthogonal-like fractional-octave-band filters. *J. Ac. Soc. Am.*, 127(2):884895, 2010.

[5] M. Anusuya and S. Katty. Speech recognition by machine, a review. *International Journal of Computer Science and Information Security*, 6(3):181–205, 2009.

[6] J. Aucouturier, B. Defreville, and F. Pachet. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *J. Ac. Soc. Am.*, 122(2):881–891, 2007.

[7] Pierre Aumond, Arnaud Can, Bert De Coensel, Dick Botteldooren, Carlos Ribeiro, and Catherine Lavandier. Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. *Acta Acustica united with Acustica*, 103(3):430–443, 2017.

[8] C. Baug, M. Lagrange, J. Andén, and S. Mallat. Representing environmental sounds using the separable scattering transform. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[9] L. Brocolini, C. Lavandier, M. Quoy, and C. Ribeiro. Measurements of acoustic environments for urban soundscapes: Choice of homogeneous periods, optimization of durations, and selection of indicators. *J. Ac. Soc. Am.*, 134(1):813–821, 2013.

[10] R. Cai, L. Lu, A. Hanjalic, H. Zhang, and L. Cai. A flexible framewok for key audio effects detection and auditory context inference. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):1026–1039, 2006.

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

[12] A. Can and B. Gauvreau. Describing and classifying urban sound environments with a relevant set of physical indicators. *J. Ac. Soc. Am.*, 137(1):208–218, 2015.

[13] S. Chachada and C. Kuo. Environmental sound recognition: A survey. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013.

[14] S. Chu, S. Narayanan, and C. Kuo. Environmental sound recognition with time-frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009.

[15] C. Couvreur. Implementation of a one-third-octave filter bank in matlab. `http://citeseer.ist.psu.edu/24150.html`.

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

[17] S. Davis. Octave and fractional-octave band digital filtering based on the proposed ansi standard. In *1986 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1986.

[18] Boris Defréville, François Pachet, Christophe Rosin, and Pierre Roy. Automatic recognition of urban sound sources. In *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.

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

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

[21] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[22] Y. Hu and P. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229238, 2008.

[23] D. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[24] T. Ishiyama and T. Hashimoto. The impact of sound quality on annoyance caused by road traffc noise: an influence of frequency spectra on annoyance. *JSAE Review*, 21(2):225–230, 2000.

[25] J. Kates and K. Arehart. Coherence and the speech intelligibility index. *J. Ac. Soc. Am.*, 115(5):22242237, 2005.

[26] P. Khunarsal, C. Lursinsap, and T. Raicharoen. Very short time environmental sound classification based on spectrogram pattern matching. *Information Sciences*, 243:57–74, 2013.

[27] A. Kumar and B. Raj. Features and kernels for audio event recognition. `https://arxiv.org/abs/1607.05765`, 2016.

[28] Catherine Lavandier and Boris Defréville. The contribution of sound source characteristics in the assessment of urban soundscapes. *Acta Acustica united with Acustica*, 92(6):912–921, 2006.

[29] Hock Beng Lim, Yong Meng Teo, Protik Mukherjee, Vinh The Lam, Weng Fai Wong, and Simon See. Sensor grid: integration of wireless sensor networks and the grid. In *Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on*, pages 91–99. IEEE, 2005.

[30] J. Ma, Y. Hu, and P. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Ac. Soc. Am.*, 125(5):33873405, 2009.

[31] Charlie Mydlarz, Justin Salamon, and Juan Pablo Bello. The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics*, 117:207–218, 2017.

[32] Charlie Mydlarz, Charles Shamoon, Melody Baglione, and Michael Pimpinella. The design and calibration of low cost urban acoustic sensing devices. Euronoise, 2015.

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

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

[35] K. Piczak. Environmental sound classification with convolutional neural networks. In *IEEE 25th International Workshop on Machine Learning for Signal Processing*, 2015.

[36] R. Radhakrishnan, A. Divakaran, and P. Smaragdis. Audio analysis for surveillance applications. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

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

[38] M. Rychtarikova and G. Vermeir. Soundscape categorization on the basis of objective acoustical parameters. *Applied Acoustics*, 74(2):240–247, 2013.

[39] J. Salamon and J. Bello. Feature learning with deep scattering for urban sound analysis. In *23rd European Signal Processing Conference*, 2015.

[40] J. Salamon and J. Bello. Unsupervised feature learning for urman sound classification. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[41] J. Salamon and J. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[42] J. Salamon, C. Jacoby, and J. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM international conference on Multimedia*, 2014.

[43] Chen-Khong Tham and Rajkumar Buyya. Sensorgrid: Integrating sensor networks and grid computing. *CSI communications*, 29(1):24–29, 2005.

[44] A. Torija, D. Ruiz, and A. Ramos-Ridao. Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes. *J. Ac. Soc. Am.*, 134(1):791–802, 2013.

[45] G. Tzanztakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.