Privacy aware acoustic scene synthesis using deep spectral feature inversion

Félix Gontier, Mathieu Lagrange, Catherine Lavandier, Jean-François Petiot

PRIVACY AWARE ACOUSTIC SCENE SYNTHESIS USING DEEP SPECTRAL FEATURE INVERSION

Félix Gontier [1], *Mathieu Lagrange* [1], *Catherine Lavandier* [2], *Jean-François Petiot* [1]

[1] LS2N, UMR CNRS 6004, Ecole Centrale de Nantes, F-44321, France
[2] ETIS, UMR CNRS 8051, University of Paris Seine, University of Cergy-Pontoise, ENSEA, France

ABSTRACT

Gathering information about the acoustic environment of urban areas is now possible and studied in many major cities in the world. Part of the research is to find ways to inform the citizen about its sound environment while ensuring her privacy.

We study in this paper how this application can be cast into a feature inversion problem. We argue that considering deep learning techniques to solve this problem allows us to produce sound sketches that are representative and privacy aware. Experiments done considering the dcase2017 dataset shows that the proposed learning based approach achieves state of the art performance when compared to blind inversion approaches.

Index Terms— spectral feature inversion, privacy aware audio synthesis, deep neural network, environmental audio processing

1. INTRODUCTION

Together with sound quality concerns in urban environments, the advent of the Internet of Things has led to the implementation of large scale acoustic sensor networks for monitoring purposes in several projects [1, 2, 3, 4]. The aim of these sensor networks is to gather rich information about the sound environment and its content. Gathering, storing, and processing the data outputted from those networks are major technical challenges, but the production of indicators and materials that can be directly interpreted by the citizens raises important issues related to the quality and interpretability of the data, together with constraints about the privacy of the citizens.

On one hand, the sound pressure level clearly lacks richness, but one the other hand, presenting raw audio recordings is not possible due to privacy concerns. Defining perceptual indicators to go beyond the sound pressure level is one important avenue of current research [5]. In this paper, we want to study another direction where the citizen is allowed to listen to sound sketches that illustrate the acoustic content of the urban area where the sensor is located. Importantly, we need to guarantee that the processing pipeline is privacy aware, that is that no speech signals can be exposed. One solution studied in [6], is to perform speech source separation over audio recordings. An alternative approach is to assume that the output of the sensor are spectral features such as third octave sound levels, which are sufficient to compute most indicators used in monitoring applications and ensure privacy awareness of the data under specific time resolution settings [7].

In this paper, we propose to investigate the potential of a deep convolutional neural network approach operating in the spectral domain to recover fine spectral information from third octave spectra in acoustic scenes. In the general case where privacy is not ensured by the representation settings, the choice of training datasets may provide control on the type of generated content.

After introducing some background on the formal definition of the signal processing task considered in Section 2, the model is described in Section 3. The experimental protocol considered to evaluate the proposed approach[1] compared to two baselines is described in Section 4. Section 5 presents the results of the experiments, which are then discussed in Section 6.

2. BACKGROUND

Retrieving time domain audio signals from spectral features that embed rich information about the sound environment is beneficial for many applications. In environmental sound monitoring for example, it can be used for the manual annotation of sound scenes or the computation of more precise features for learning-based characterization.

Those applications all require the inversion of the processing steps applied to the original audio signal. In the case of third octave spectra this operation is an ill-posed problem, as the filterbank which summarizes spectral information on a logarithmic frequency scale is not invertible. Similar problems of spectral feature inversion have been studied in the speech processing community, through applications such as speech coding for communication or text-to-speech. Such

Work partially funded by ANR CENSE

[1] Code available at: https://github.com/felixgontier/paperPrivacyAwareSynthesisIcassp20

tasks consist in the estimation of speech waveforms primarily from Mel spectrograms, which are similar to third octave energies in construction though differently motivated, or Mel Frequency Cesptrum Coefficients (MFCC) [8]. Because Mel spectrograms are computed in the spectral domain, a commonly adopted approach is to first recover the linear scale spectrogram of the speech signal before time domain reconstruction. To do so, it is possible to compute an approximate inverse transformation as the pseudoinverse of the Mel filterbank [8]. The resulting inverse transform uses no *a priori* information on the properties of the studied signals. Such information can be used to enhance the quality of predictions. For example, in [9] the authors add non-negativity and sparsity constraints to the optimization of the inverse Mel filterbank given a clean speech spectrogram as the target representation. The sparsity property of speech spectra is further used in [10], where an inversion based on $\ell 1$ optimization is proposed. Other approaches use the invertibility of the discrete cosine transform used in the computation of MFCC to artificially increase the precision of Mel spectra before interpolating a spectrogram with linear frequency scaling [11].

More recent approaches are based on deep learning techniques. In [12] a deep neural network is proposed that predicts linear scale spectrograms from MFCC, that outperforms non-parametric spectral inversion baselines on objective perceptual quality and distortion metrics. The authors of [13] further propose an encoder-decoder structure together with adversarial training to recover detail in speech spectrograms obtained from the Mel pseudoinverse. The problem of retrieving linear spectrograms from downsampled features also relates to applications in image processing, such as the inversion of both handcrafted and learned features [14] or super-resolution [15]. State of the art approaches in these areas of research primarily include convolutional neural networks, in supervised or unsupervised generative frameworks. Recently, some models exist that directly retrieve time domain audio using generative models. Notably, [16] propose a sample-based architecture conditioned on Mel spectrograms as a single-speaker vocoder. Though unprecedented performances are obtained, the generation process is computationally intensive.

Compared to speech signals, acoustic environments feature complex polyphonies with a wide range of sound sources yielding both harmonic and wide band structures in sound scenes. As such, designing a content-dependent inversion method is difficult. In this context, the use of deep learning techniques removes the need for model-based inversion process design. By training a model with sufficient capacity on a large dataset containing polyphonies with real-life complexity and diversity, it can learn to implicitly extract information about typical spectral patterns useful in the reconstruction of plausible spectrograms. Together with performance, using a learning-based inversion has the benefit of allowing us to control the type of content that it is trained to produce.

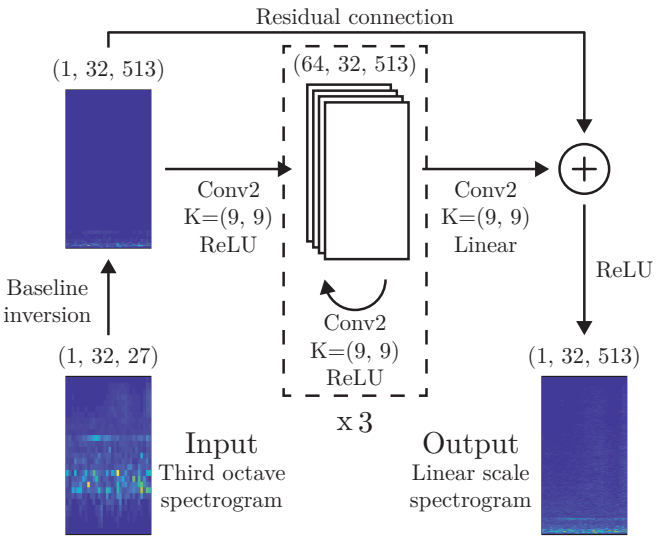

Fig. 1. Proposed fully convolutional neural network for third octave inversion.

3. PROPOSED APPROACH

The aim of the model is to predict a linearly scaled spectrogram from its energy summarized on third octave bands, thus on a logarithmically scaled frequency axis. Though, using third octave spectrograms directly as inputs of the model would require it to perform non-linear upsampling in the frequency dimension. Motivating an architecture based on convolutions with constant stride and kernel size would be difficult for such a task. Instead, the use of baseline linear spectrogram approximations obtained using the $\Phi^\dagger$ pseudoinverse is considered (see Section 4 for more details). The target representation is then a spectrogram of the same shape and domain.

The proposed architecture is a 5-layer fully convolutional neural network as illustrated in Figure 1. Each successive hidden layer is characterized by 9x9 (K) convolution kernels, 64 output channels and a rectified linear unit (ReLU) activation. Inputs are padded so that the size of internal representations does not change throughout the network. The output layer also uses 9x9 kernels, but outputs a single channel and is followed by a linear activation. The input linear spectrogram estimated using the $\Phi^\dagger$ pseudoinverse is added to the output of the network through a residual connection. As a result, the CNN architecture does not learn to model the overall envelope of the spectrum, but rather concentrates on adding fine scale variations to refine the baseline approximation. A final ReLU activation is applied to ensure that the model's predictions are valid (*i.e.* non-negative). The considered architecture contains a total of about 1M parameters.

4. EXPERIMENTAL PROTOCOL

We use the DCASE2017 task 1 dataset [17] in this study. It is composed of a development set and an evaluation set containing 4680 and 1620 sound scenes of 10s each respectively. The development set is further split into train and validation subsets of 3510 and 1170 extracts respectively, according to split 1 in the provided cross-validation setup. Though this dataset is intended for acoustic scene classification tasks, its wide range of ambiances covers most of the polyphonies that could be recorded from large scale sensor networks in urban environments. Furthermore, it is appropriate for the privacy aware constraint of the considered application as it does not contain intelligible speech.

All extracts are resampled to 16kHz and converted to mono. Target spectrograms are extracted using a short-term Fourier transform (STFT) on windows of 1024 samples (64ms) with an overlap of 512 samples (32ms) and Hann windowing. These spectra are then partitioned into texture frames of 32 frames, or about 1s, corresponding to individual examples. The input third octave spectra are then obtained by applying a fixed filterbank in the frequency domain on the squared magnitude spectrum. At the sampling rate of 16kHz, this yields 24 logarithmically spaced bands in the range 20Hz-8kHz.

To evaluate the performance of the proposed method, two baselines are considered for the feature inversion task. First, an approximation of the linear spectrogram is obtained using the Moore-Penrose pseudoinverse $\Phi^\dagger$ of the third octave filterbank matrix Φ [8]. As Φ does not have full-rank, $\Phi^\dagger$ must be computed using either a least squares solver or the singular value decomposition (SVD) of the forward transformation matrix. Note that in both cases, the computation $\Phi^\dagger$ is not subject to a non-negativity constraint, which may lead to negative magnitude estimates in linear spectrograms. To palliate this issue a threshold $\hat{X} = max(0, \hat{X})$ is applied to spectrograms computed with this baseline.

The second baseline considered is the estimation of the linear spectrogram using a non-linear least squares (NNLS) method, defined as:

$$\hat{Y} = \arg\min_{Y} |\Phi Y - X|^2 \tag{1}$$

where $\hat{Y}$ is the linear spectrogram estimation and X is the third octave spectrogram.

The proposed model described in Section 3 is implemented in the *Pytorch* framework and trained on minibatches of 32 examples for 20 epochs. The loss function used is the mean squared error:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{n=1}^{N_b} \sum_{t=1,f=1}^{T,F} (\hat{y}_n(t,f) - y_n(t,f))^2 \tag{2}$$

where y is the ground truth linear spectrogram, $\hat{y}$ is the output of the model, N_b is the batch size, and T and F are the time and frequency dimensions of the output respectively. Optimization is performed using the Adam algorithm with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and learning rate $\lambda = 0.001$.

The reconstruction of audio signals in the time domain from estimated magnitude spectrograms using an overlap-add method requires phase information. Here we consider either using the original phase or an estimate obtained using the Griffin-Lim algorithm [18] with 100 iterations.

Two metrics are proposed for the evaluation of the spectral feature inversion techniques on the evaluation set. The first is the spectral root mean squared error (RMSE) defined as the square root of the loss function used to train the model as described in eq. 2. As a time-domain metric, the signal to reconstruction ratio (SRR) is considered. The SRR is computed for each 10s extract of the evaluation set as:

$$SRR = 10\, log_{10} \left(\frac{\sum_t |x(t)|^2}{\sum_t |x(t) - \hat{x}(t)|^2} \right) \tag{3}$$

where x is the target time-domain signal and $\hat{x}$ its estimation.

While both metrics give quantitative estimates of the performance of the proposed methods, they do not necessarily reflect the perceived quality of reconstructed sound scenes. A listening test would be the most appropriate method to do so. Though, conducting one is time-consuming and covering the entire range of ambiances represented in the evaluation set is difficult. As an alternative, some objective metrics have been proposed that are based on perceptual models to assess the quality of audio signals in estimation tasks. In addition to the spectral MSE and time-domain SRR, the objective difference grade (ODG) obtained as part of the perceptual evaluation of audio quality (PEAQ) algorithm proposed in the ITU-R BS.1387 norm [19] is considered.

5. RESULTS

First, the performance metrics are computed on the whole evaluation dataset and shown in Table 1. Compared to both baselines the proposed model performs better in terms of spectral RMSE. This is expected as the model is trained to directly minimize this loss, while the baselines are both blind approaches. The SRR with oracle phase reconstruction is correlated to the spectral RMSE. The proposed method shows similar performance to NNLS estimation, and both are beneficial compared to the use of the pseudoinverse only. Using Griffin-Lim phase reconstruction results in an important decrease in SRR values for the three methods. Though, the SRR obtained by applying phase retrieval to the reference magnitude spectrograms yields similar results. This indicates that the SRR is sensitive to the phase shift between the estimated and reference time-domain signals introduced by this algorithm, which reduces its interpretability in this case. The improvements in spectral RMSE and oracle phase SRR are not reflected by the ODG: phase retrieval through the Griffin-Lim algorithm yields similar results for the three methods,

Table 1. Overall performance metrics of the proposed CNN compared to the pseudoinverse and NNLS baselines.

	Spectral RMSE	SRR (dB)		ODG	
	-	Griffin-Lim	Oracle	Griffin-Lim	Oracle
Pseudoinverse $\Phi^{\dagger}$	0.047 ±0.061	-2.96 ±0.28	13.1 ±3.8	-3.49 ±0.42	-2.83 ±0.44
NNLS	0.041 ±0.054	-2.96 ±0.27	14.4 ±5.1	-3.49 ±0.42	-2.84 ±0.44
Proposed	0.034 ±0.044	-2.78 ±0.33	14.2 ±5.3	-3.45 ±0.50	-3.00 ±0.48

Table 2. Class-wise differences between metrics for the proposed approach and the pseudoinverse baseline, computed on $n = 108$ extracts for each class.

	Spectral RMSE	SRR (dB)
	-	Oracle
beach	-0.003±0.002	0.42±0.27
bus	-0.040±0.035	3.63±2.16
cafe/restaurant	-0.005±0.004	0.49±0.49
car	-0.047±0.034	4.44±2.24
city center	-0.012±0.011	1.18±0.71
forest path	-0.001±0.002	0.11±1.63
grocery store	-0.005±0.004	0.76±0.58
home	-0.001±0.001	-0.52±1.23
library	-0.001±0.001	0.67±1.34
metro station	-0.002±0.002	0.28±1.11
office	0.000±0.000	-1.46±1.15
park	-0.001±0.001	0.18±0.70
residential area	-0.004±0.007	0.79±1.38
train	-0.054±0.029	4.14±1.31
tram	-0.025±0.024	2.28±1.46

and waveform reconstructions using the proposed CNN and the original phase perform slightly worse than both baselines.

All metrics in Table 1 display high variances compared to their respective mean values. This may be due to the variety of sound environments represented in the evaluation dataset, as the model may perform differently depending on the type of spectral content or complexity of the scene. Thus, the 15 ambiances labeled in the dataset are considered separately. As the evaluation set is balanced, each class is represented by $n = 108$ sound scenes. Table 2 shows the ambiance-wise differences between metrics computed for the proposed CNN and the pseudoinverse baseline, as a way to assess the gain associated with the CNN and residual connection. For both the spectral RMSE and the SRR, the proposed model improves the *bus*, *car*, *train* and *tram* classes. These ambiances are characterized by predominant low-frequency content. An interpretation of this result is that the model is trained to minimize the error between the estimated and reference linear scale spectrograms. Due to the typical exponential decay of magnitude as a function of increasing frequency in environmental sound scenes, there is an imbalance between values of the loss function in lower and higher frequencies. Thus, the model primarily concentrates on learning to reproduce low-frequency content. Conversely, a decrease in quality is found for the *home* and *office* ambiances, both of which are quieter indoor environments. Without normalization the magnitudes and thus errors are lower, and the model may not learn as well to represent the polyphonies found in these ambiances. An informal listening test confirms these results for the CNN outputs, with improvements in low-frequency content (traffic sources) but poor resolution in high frequencies characterized by a lack of clearly defined pitches for sources such as birdsong or sirens. Sound scene examples can be listened to here[2].

6. DISCUSSION

The potential of using a fully convolutional neural network with a residual connection to refine baseline spectral feature inversion predictions has been studied. Overall, the proposed approach performs comparably to a baseline that uses extract-wise optimization to recover the linear spectrogram from third octave energies. The model outperforms baselines on ambiances with dominant low frequency activity, but is worse for quiet environments. Some design considerations can be outlined from these results: spectral whitening, pre-emphasis filtering, example-wise normalization [20] or scale-invariant metrics could be beneficial the model to produce sound environments with high variations in source types, polyphonies and sound levels.

Because of the amount of information missing in high frequencies, generative models could improve the quality of generated extracts compared to deterministic approaches. Furthermore, the loss of phase information yields a significant decrease in objective prediction quality. Sample-based approaches are a potential solution to this problem, and will be studied in future work.

Still, learning-based models may allow us to control the content that is generated, both implicitly through the training data and explicitly through the loss function. This is a key advantage in the context of privacy aware environmental scene generation compared to blind approaches. In addition to the previous considerations, future work will thus aim at determining the impact of the content in training data on the intelligibility of reconstructed speech segments, as well as possible custom loss functions penalized by speech content quality.

[2] http://soundthings.org/research/paperPrivacyAwareSynthesisIcassp20

7. REFERENCES

[1] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," *Computational Analysis of Sound Scenes and Events*, pp. 373–397, 2018.

[2] J. Picaut, A. Can, J. Ardouin, P. Crépeaux, T. Dhorne, D. Écotière, M. Lagrange, C. Lavandier, V. Mallet, C. Mietlicki, and M. Paboeuf, "Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling," *J. Ac. Soc. Am.*, vol. 141, pp. 3808, 2017.

[3] P. Bellucci, L. Peruzzi, and G. Zambon, "LIFE DYNAMAP project: The case study of Rome," *Applied Acoustics*, vol. 117, pp. 193–206, 2017.

[4] J. Abeßer, M. Gotze, S. Kuhnlenz, R. Grafe, C. Kuhn, T. Clauß, and H. Lukashevich, "A distributed sensor network for monitoring noise level and noise sources in urban environments," in *2018 IEEE International Joint Conference on Future Internet of Things and Cloud (FiCloud)*, 2018.

[5] 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 of Noise Control Engineering*, 2016.

[6] A. Cohen-Hadria, M. Cartwright, B. McFee, and J.P. Bello, "Voice anonymization in urban sound recordings," in *2019 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[7] F. Gontier, M. Lagrange, P. Aumond, A. Can, and C. Lavandier, "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach," *Sensors*, vol. 17, 2017.

[8] L. Boucheron, P. De Leon, and S. Sandoval, "Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 610–19, 2012.

[9] G. Min, X. Zhang, J. Yang, X. Zou, and Z. Pan, "Speech reconstruction from MFCC based on nonnegative and sparse priors," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E98.A, pp. 1540–3, 2015.

[10] G. Min, X. Zhang, J. Yang, and X. Zou, "Speech reconstruction from mel-frequency cepstral coefficients via l1-norm minimization," in *2015 IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2015.

[11] T. Ramabadran, J. Meunier, M. Jasiuk, and B. Kushner, "Enhancing distributed speech recognition with back-end speech reconstruction," in *Eurospeech 2001*, 2001.

[12] W. Jiang, P. Liu, and F. Wen, "Speech magnitude spectrum reconstruction from MFCCs using deep neural network," *Chinese Journal of Electronics*, vol. 27, pp. 393–8, 2018.

[13] P. Neekhara, C. Donahue, M. Puckette, S. Dubnov, and J. McAuley, "Expediting TTS synthesis with adversarial vocoding," in *Interspeech 2019*, 2019, pp. 186–190.

[14] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *2016 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. Saurous, Y. Agiomvrgiannis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[17] A. Mesaros, T. Heitolla, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: tasks, datasets and baseline system," in *Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 workshop*, 2017.

[18] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–43, 1984.

[19] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[20] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J.-P. Bello, "Per-channel energy normalization: why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2018.