Learning spectro-temporal representations of complex sounds with parameterized neural networks

Rachid Riad, Julien Karadayi, Anne-Catherine Bachoud-Lévi, Emmanuel Dupoux

Learning spectro-temporal representations of complex sounds with parameterized neural networks

Rachid Riad,[1,2] Julien Karadayi,[1] Anne-Catherine Bachoud-Lévi,[2] and Emmanuel Dupoux[1,3]

[1]*Ecole des Hautes Etudes en Sciences Sociales, Centre National de la Recherche Scientifique, INRIA, Ecole Normale Supérieure-Paris Sciences & Lettres University, 29 rue d'Ulm, 75005 Paris, France*

[2]*NeuroPsychologie Interventionnelle, Département d'Études Cognitives, Ecole Normale Supérieure, Institut national de la santé et de la recherche médicale, Institut Mondor de Recherche Biomédicale, Neuratris, Université Paris-Est Créteil, Paris Sciences & Lettres University, 29 rue d'Ulm, 75005 Paris, France*

[3]*Facebook AI Research, Paris, France*

(Dated: 15 March 2021)

Deep Learning models have become potential candidates for auditory neuroscience research, thanks to their recent successes on a variety of auditory tasks. Yet, these models often lack interpretability to fully understand the exact computations that have been performed. Here, we proposed a parametrized neural network layer, that computes specific spectro-temporal modulations based on Gabor kernels (Learnable STRFs) and that is fully interpretable. We evaluated predictive capabilities of this layer on Speech Activity Detection, Speaker Verification, Urban Sound Classification and Zebra Finch Call Type Classification. We found out that models based on Learnable STRFs are on par for all tasks with different toplines, and obtain the best performance for Speech Activity Detection. As this layer is fully interpretable, we used quantitative measures to describe the distribution of the learned spectro-temporal modulations. The filters adapted to each task and focused mostly on low temporal and spectral modulations. The analyses show that the filters learned on human speech have similar spectro-temporal parameters as the ones measured directly in the human auditory cortex. Finally, we observed that the tasks organized in a meaningful way: the human vocalizations tasks closer to each other and bird vocalizations far away from human vocalizations and urban sounds tasks.

I. INTRODUCTION

The main objective of auditory neuroscience is to build models that can both predict the brain neural responses to relevant sounds and the behaviours associated with these responses (Kell and McDermott, 2019; Pillow and Sahani, 2019). While most of the auditory neuroscience research has focused on the neural side, there is growing recognition for the importance to also match the performance of living organisms on a variety of behavioral tasks (Yarkoni and Westfall, 2017). In recent years, major progress has been achieved with Deep Neural Networks (DNNs) which, after training with supervised classification objectives on large datasets, proved able to perform near human performance on a variety of audio tasks such as automatic speech recognition (Amodei *et al.*, 2016), speaker verification (Snyder *et al.*, 2018) or audio scene classification (Salamon and Bello, 2017). These trained systems therefore become potential candidate models for auditory neuroscience (Koumura *et al.*, 2019), and have already started to be used to account for perceptual results (Saddler *et al.*, 2020) and brain data (Kell *et al.*, 2018) in humans.

DNNs models typically take as input a spectral representation (although some new trends consist in side-stepping this representation and work directly from the raw waveform). Working from a spectral representation has biological plausibility, since it matches approximately what we know about the first stage of auditory processing (Stevens *et al.*, 1937). However, DNNs models are less biologically motivated regarding the next steps. Most of them use rather generic connectivity patterns (fully connected, convolutional or recurrent networks), which while being very powerful in learning task-specific representations from an engineering

point of view, lack both interpretability and support in the auditory neuroscience. To push the understanding of both the artificial and real neural networks, there have been some attempts to decode the representation extracted from biological measurements or computed by deep learning models (Ondel *et al.*, 2019; Thoret *et al.*, 2020). Even though these methods allow to uncover the important aspects of the stimuli, they rely on simplifying hypotheses (linearity of the responses, independence across neurons (Meyer *et al.*, 2017; Shamma, 1996)) and they do not provide in depth explanation on how the DNNs made their decisions.

Fortunately, the stages beyond the extraction of the acoustic spectrum have been studied over the past few years with novel understanding of the representations and processing involved (McDermott, 2018). Slow spectral and temporal modulation built on top of the spectrum have been shown in psychophysical tests to be useful for several audio tasks solved by mammals: they contribute to speech intelligibility (Edraki *et al.*, 2019; Elhilali *et al.*, 2003; Elliott and Theunissen, 2009), they help to boost performance for speech processing in noisy environments (Chang and Morgan, 2014; Mesgarani *et al.*, 2006; Vuong *et al.*, 2020). In addition, the responses to such spectral and temporal modulations of natural sounds can be decoded from human fMRI (Santoro *et al.*, 2017) and have been measured directly with invasive techniques in ferrets (Depireux *et al.*, 2001), in birds (Woolley *et al.*, 2005), and also in the human brain (Hullett *et al.*, 2016).

Analytic models of these modulations have been proposed Chang and Morgan (2014); Chi *et al.* (2005); Ezzat *et al.* (2007); Schädler *et al.* (2012), on the basis of wavelet analysis. The idea is that on top of the spectrum, spectro-temporal wavelets or gabor patches can be defined which drive both behavioral responses and brain signals. The problem of such

analytic models is that they only propose a potentially very large representation space, but provide no method to select which gabor patch is relevant for which task. But analyses of brain signals show that the responses from the auditory cortex are not fixed, but vary depending on the task at hand (Francis *et al.*, 2018; Fritz *et al.*, 2003; Jääskeläinen *et al.*, 2007). Therefore, what is needed, is a model that can learn the characteristics of the spectro-temporal representations that are relevant to the task.

This is the goal of this work. We introduce a parametrized neural network which explicitly represent spectro-temporal filters, but which parameters are differentiable and can therefore be tuned to each particular task. There are two advantages of this approach as illustrated in Figure 1. First, as Analytic Models, and contrary to standard DNN models, this model is fully interpretable. The parameters of each filter can be directly read off the model and compared to physiological or neural data. Second, as DNNs, but contrary to Analytic Models, this model can be tuned to different tasks, accounting both for behavioral results and for the task-specificity of the brain representations. As a side issue, since the model is constrained and has few parameters, it has the potential to explain perceptual learning aspect of plasticity with a lot less training data than is typically used in generic DNNs. Therefore, the model makes direct and testable predictions about the auditory representation as a function of the task.

The paper is organized as follows: Section II presents the Methods with our parametrized neural network model, and the different ways to analyze the distribution of the learned spectro-temporal modulations; in Section III we described the experimental setup with the different computational tasks, data, toplines, and evaluations; Section IV presents the per-

formance results, the analysis of the learned distribution of spectro-temporal modulation for each setup and the discussions. Section V presents our conclusions and the potential future work.

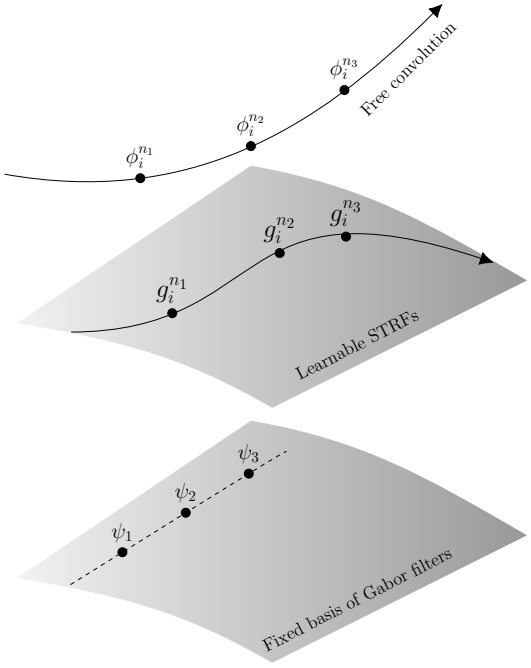

FIG. 1. Schematic illustration of the space of functions and different approaches to obtain spectro-temporal representations of sounds. (top) Free Learnable Convolution ($\phi_k^{n_i}$) (Młynarski and McDermott, 2018; Ondel *et al.*, 2019); (middle) Learnable STRFs (Learnable spectro-temporal filters) ($g_k^{n_i}$) (this study); (bottom) Fixed basis of Gabor filters (ψ_k) (Bellur and Elhilali, 2015; Chang and Morgan, 2014; Elie and Theunissen, 2016; Mesgarani *et al.*, 2006; Schädler *et al.*, 2012). The upper index n_i and lower index k_{th} represent the n_i-step during learning for the k_{th} filter.

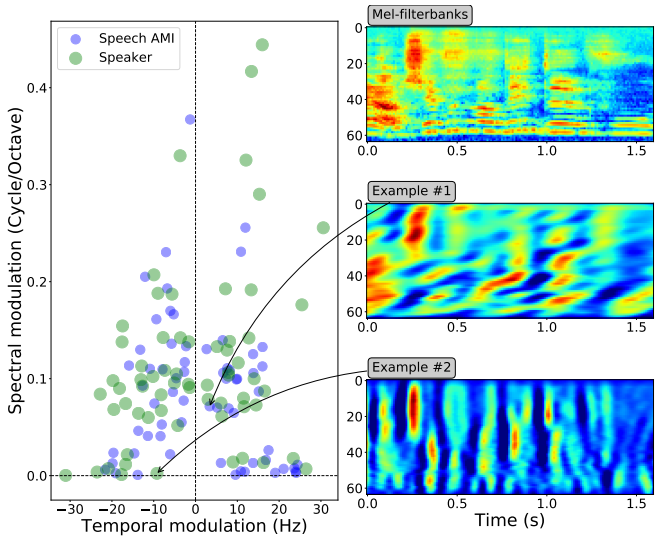

FIG. 2. Temporal and Spectral Modulations populations learned to tackle Speech Activity Detection (Speech AMI) on the AMI dataset, and Speaker Verification (Speaker) (left). Mel-filterbanks representation of a sentence pronounced by a Female Speaker (top right). Outputs examples computed by the convolution of specific learned STRF kernels with the input Mel-filterbanks (middle and bottom right).

II. MODELS AND METHODS

A. Learnable spectro-temporal filters

Here, $\Re$, $\Im$, $|.|$, $[.,.]$ represent the real part, imaginary part, modulus and the concatenation operators respectively. $\{.\}$ represents a set and $|.|$ the cardinal of a specific set.

1. First stage of processing

The first audio processing step is the transformation of the audio signal from the time domain into the frequency domain $\mathbf{Y}(t, f)$. Each excerpt of sound given to the network is normalized per instance with a 1 dimensional Instance Normalization layer Ulyanov *et al.*

(2016). Then the sound is decomposed into a Filter banks spaced based on the Mel scale after a log compression (Mel-filterbanks) similarly as the resolution of the human cochlea. There are 64 filters with center frequencies in the range $[0.0, 8000.0Hz]$. The computations for the Mel-filterbanks of the audio can be performed *on-the-fly* directly on GPU thanks to Cheuk *et al.* (2020).

2. Definition of the Learnable spectro-temporal filters

The second step of front-end processing is a set of convolution between the time-frequency representation of the audio and a set of Gabor Filters (Gabor, 1946).

The 2-D Gabor filter kernel g_k is a sine-wave w_k modulated by a 2-D Gaussian envelope s_k. Each Gabor filter g_k is expressed based on the set of parameters $(\sigma_t, \sigma_f, \gamma_k, F_k)$ in polar coordinates. We used the following formulation in this work:

$$g_k(t,f) = s_k(t,f) \cdot w_k(t,f) \tag{1a}$$

$$s_k(t,f) = \frac{1}{2\pi\sigma_{t_k}\sigma_{f_k}} e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_{t_k}^2}+\frac{f^2}{\sigma_{f_k}^2}\right)} \tag{1b}$$

$$w_k(t,f) = e^{j\left(2\pi\left(F_k R_{\gamma_k}\right)\right)} \tag{1c}$$

$$R_{\gamma_k} = t\cos(\gamma_k) + f\sin(\gamma_k) \tag{1d}$$

We obtain a bank of N filters $\{g_k(t,f)\}_{k=0..N-1}$. These bank of filters is convoluted with the time-frequency representation $\mathbf{Y}$ to obtain the 3D representation $\mathbf{Z}$.

$$\mathbf{Z}(t,f,k) = \sum_{u,v} \mathbf{Y}(u,v)\, g_k(t-u, f-v) \in \mathbb{C} \tag{2}$$

These filters and their parameters can be used in 2D Convolution neural Networks (Alekseev and Bobe, 2019) in different ways: (1) as an *Initialisation* (Free 2D conv. Gabor

Init.) of a 2D convolution neural network and the 2D grid is tuned completely by back-propagation (as in (Chang and Morgan, 2014; Ezzat *et al.*, 2007; Schädler *et al.*, 2012)), (2) or finally as *Learnable spectro-temporal Filters* (Learnable STRFs), so the gradient descent is only performed on the set of parameters $(\sigma_{t_k}, \sigma_{f_k}, F_k, \gamma_k)$. Indeed, all the operators to derive the Gabor Filter based on $(\sigma_{t_k}, \sigma_{f_k}, F_k, \gamma_k)$ are differentiable almost everywhere: $x \to e^x, x \to \cos(x), x \to \sin(x), x \to x^2, x \to \frac{1}{x}, x \to \text{constant} \cdot x$. The 2D grid instantiated by the Gabor Filter used the parameters $(\sigma_{t_k}, \sigma_{f_k}, F_k, \gamma_k)$ in each cell, therefore, the gradients are summed over the 2D grid for each parameter. Each Learnable STRF filters were given 9 Mel-frequency spectral filters and 1.1s of context, thus yielding a size of 9x111 for each filter.

Finally, the output representation $\mathbf{Z}$ is in the complex domain $\mathbb{C}$. To be used by classic neural network architectures, we concatenated $[\Re(\mathbf{Z}), \Im(\mathbf{Z})]$ to obtain the representation to be fed to the rest of each network. We denoted by *Learnable STRFs* this specific front-end in all result tables, and by *Learned STRFs* once we examined these representations.

3. Descriptive quantifiers of the distribution

We used quantitative measures to describe the structure of the distribution of the learned spectro-temporal Convolutions. This is in the same spirit as Singh and Theunissen (2003) for the 2-D Modulation Power Spectrum of sound ensembles, and Modulation Power Spectrum of the spectral-temporal receptive field of auditory neurons in ferrets Depireux *et al.* (2001). We converted the learned parameters in Cartesian coordinates (Schädler *et al.*, 2012) with the Temporal Modulation ω_k and Spectral Modulation Ω_k: $(\sigma_t, \sigma_f, \omega_k, \Omega_k)$, where $\omega_k =$

$F_k \cos(\gamma_k)$ and $\Omega_k = F_k \sin(\gamma_k)$. We took the same convention as Chi *et al.* (2005); Singh and Theunissen (2003) for the up-sweep and down-sweep modulations, and represented only half the plan due to the symmetry.

We adapted the measures of Separability, Asymmetry, Low-pass coefficient and Starriness coefficients with the interpretable parameters obtained for each Supervised Learning tasks. As the Learned spectro-temporal receptive fields (Learned STRFs) self-organized to solve each task, we examined each of this parameter for each task. Each α is estimated with the bootstrap re-sampling method (Efron and Tibshirani, 1994) on the Learned STRFs (100 bootstraps).

a. Asymmetry. The distribution of the learned STRFs can show asymmetry preferences. The distribution is considered asymmetric if there are preferences for either down-sweeps or up-sweeps Learned STRFs.

$$\alpha_{\text{asymmetry}} = \frac{|\{g_k \text{ s.t. } \omega_k > 0\}|}{|\{g_k\}|} = \frac{|\{g_k \text{ s.t. } \omega_k > 0\}|}{N} \tag{3}$$

If $\alpha_{\text{asymmetry}} \approx 0$, the distribution of STRFs filters $\{g_k(t,f)\}_{k=0..N-1}$ is considered symmetric. If $\alpha_{\text{asymmetry}} > 0$, there are more up-sweeps than down-sweeps.

b. Low Pass Coefficient and Starriness. It has been observed in Singh and Theunissen (2003), that most energy in Modulation Power Spectrum was concentrated in low spectral and temporal modulations for natural sounds. In addition, the higher spectral and temporal

modulations were not distributed uniformly but were mostly along the axes. We derived two coefficients to quantify these phenomena with the Learned STRFs:

$$\alpha_{\text{low}} = \frac{|\{g_k \text{ s.t. } |\omega_k| < \Delta_t, \Omega_k < \Delta_f\}|}{N} = \frac{N_{\text{low}}}{N} \tag{4}$$

For the temporal modulation low limit we opt as Singh and Theunissen (2003) for $\Delta_t = 16Hz$. The spectral modulation low limit is set to $\Delta_f = 0.08\ Cycle/Octave$. These parameters were chosen deliberately low as most learned modulations were mostly concentrated in low temporal and spectral modulations. The parameter α_{star} to measure "stariness" of the distribution is the relative measure of distribution excluding regions with high joint temporal and spectral modulations and the regions with low joint temporal and spectral modulations.

$$\alpha_{\text{star}} = \frac{N_{\Delta_t} + N_{\Delta_f} - 2 \times N_{\text{low}}}{N - N_{\text{low}}} \tag{5}$$

The quantities $N_{\Delta_t} = |\{g_k \text{ s.t. } |\omega_k| < \Delta_t\}|$ and $N_{\Delta_f} = |\{g_k \text{ s.t. } \Omega_k < \Delta_f\}|$ are the regions near the axes.

c. Separability. To obtain a separability measure from the learned STRFs, we approximated the 2D-distribution $\mathcal{P}(\omega, \Omega)$ of the filters with Kernel Density Estimation with Gaussian Filters. Then we evaluate if the normalized 2D-distribution $\mathcal{P}$ can be factorized into a product of two independent functions $\mathcal{P}(\omega, \Omega) = G(\omega) \cdot F(\Omega)$. To quantify the separability, we calculated as Singh and Theunissen (2003) the singular value decomposition of the $\mathcal{P}(\omega, \Omega)$ obtained from each task:

$$\mathcal{P}(\omega, \Omega) = \sum_{i=1}^{n} \lambda_i g_i(\omega) \cdot h_i(\Omega), \lambda_1 > \lambda_2 > \cdots > \lambda_n \tag{6}$$

Then, we computed the ratio of first singular value relative to the sum of all singular values.

$$\alpha_{\text{sep}} = \frac{\lambda_1}{\sum_{i=1}^{n} \lambda_i} \tag{7}$$

If $\alpha_{\text{sep}} \approx 1$, the distribution of the learned STRFs can be considered separable.

4. Measuring distance between tasks based on the learned STRF filters and optimal transport

The α measures provide some descriptors allowing some comparison between the learned distributions. However, they only look at one view and aspect of the learned distributions at a time. There is no clear way to measure the distances between each task based on the α. Besides, these α measures do not take into account the learned Gaussian envelope parameters $(\sigma_{t_k}, \sigma_{f_k})$. Here, the goal is to obtain a quantitative metric able to compare the distributions obtained from each task. Usually, researchers fall back to the Mahalanobis distance or an approximation of the KL-divergence to compare observations of two sets of points. Yet, these metrics either make modelling assumptions about the data (approximation of the underlying density functions that generated the data), or it is impossible to compare set of points with different cardinals.

The non-parametric, natural and most powerful way to compare distributions is to use optimal transport distances (Peyré *et al.*, 2019). We compared the different tasks by comparing the learned STRFs using the regularized version Sinkhorn distance (Cuturi, 2013;

Flamary and Courty, 2017). Especially, this regularized version of the optimal transport distance allow fast computation of distances and multiple assignments between points.

We made the choice to compare two individual learned STRFs with the Euclidean distance ||.|| We normalized along each axis/parameter to not privilege for a specific parameter variability. Based on each task we tackled in this work, we obtained a distribution of normalized learned parameters $\{\sigma_{t_k}, \sigma_{f_k}, \omega_k, \Omega_k\}_{task}$ with the size n_{task} being the total number of filters used for this task. Therefore, equipped with the Euclidean distance to compare the individual filters, we can obtain the cost matrix between two tasks $M_{(task_a, task_b)} \in \mathbb{R}^{n_{task_a} \times n_{task_b}}$. We did not privilege any learned STRFs to build the distribution, therefore we attributed equal weight to each individual filter $w_{task} = (1/n_{task})1_{n_{task}}$. This allows to compare the different task if we have several models due to cross validation (*Urban* and *Bird*) or less filters for a specific task (*Bird*). If we denote, by $< ., . >_F$ the norm of Froebenius between matrices, the regularized distance d_λ between two tasks is defined as:

$$
\begin{aligned}
d_\lambda &= \min_P < P, M >_F - \lambda \cdot h(P) \\
\text{s.t. } & P1_{n_{task_a}} = w_{task_a} \\
& P^T 1_{n_{task_b}} = w_{task_b} \\
& P \in \mathbb{R}_+^{n_{task_a} \times n_{task_b}} \\
& h(P) = -\sum_{i,j} P_{i,j} \log(P_{i,j}) \\
& \lambda = 10^{-3}
\end{aligned}
\tag{8}
$$

Therefore, we were able to obtain a proxy on how close/far are two different tasks $task_a$ and $task_b$ to each other based on the Sinkhorn-distance $d_\lambda(\{\sigma_{t_k}, \sigma_{f_k}, \omega_k, \Omega_k\}_{task_a}, \{\sigma_{t_k}, \sigma_{f_k}, \omega_k, \Omega_k\}_{task_b})$. Based on the distances between all tasks, we built a hierarchical cluster tree and represent these distances with a dendogram (See Figure 5).

III. EXPERIMENTAL SETUP

We compared the Learnable STRFs layer with strong toplines that have been lately introduced to solve each task as well more classic baselines for each task. We tried to keep the systems with Learnable STRFs as close as possible from topline systems.

A. Speech Activity Detection

The goal of Speech Activity Detection is to segment a given stream audio into portions of Speech or Non-Speech. We choose this task, as it allows us to examine what are the exact spectro-temporal modulations that makes standout speech in a audio stream with silences and background noises (Mesgarani *et al.*, 2006).

We conduct experiments with 2 challenging datasets with different characteristics:

The AMI database (McCowan *et al.*, 2005) a meeting dataset in English recorded with multiple microphones in 3 different rooms. There are 180 different speakers in the datasets. Here, we focus on the *AMI.SpeakerDiarization.MixHeadset* protocol as we are working only single channel feature analysis. We denoted by *Speech AMI* the experiments and the distribution of learned STRFs on this dataset and this task. The CHiME5 database (Barker *et al.*, 2018) is a dataset recorded at home during parties. Here, we focus also on single

channel feature analysis with the *CHiME5.SpeakerDiarization.U01* protocol. We denoted by *Speech CHIME5* the experiments and the distribution of learned STRFs on this dataset and this task.

We compared different input front-end to tackle this task. We evaluated the Learnable STRFs (64 filters) with a contraction layer (CL) as well the Free 2D convolution with a contraction layer (CL). The contraction layer is a convolution layer taking the outputs at each time step of the Learnable STRFs to reduce the number of dimensions of $\mathbf{Z}$. We compared these techniques with classic signal processing baselines used in speech processing: Mel-filterbanks (64 filters) and MFCC (19 coefficients, with their deltas and their delta-deltas). We also compared with the more recent parametrized neural network SincNet. SincNet is composed of parametrized sinc functions, which implement 80 band-pass filters (to replace directly more classic input spectral representations), and 3 temporal convolution/pooling layers. All the input front-end are then fed to a stack of two layers of BiLSTM layers of dimension 128 and two forward layers of dimension 32 before a final decision layer. The learning rate is controlled by a cyclical scheduler, each cycle lasting for 21 epochs. Data augmentation is performed directly on the waveform using additive noise based on the MUSAN database (Snyder *et al.*, 2015) with a random target signal-to-noise ratio ranging from 5 to 20 dB. To evaluate Speech Activity detection, we used the Detection Error Rate (DetER):

$$\text{Detection Error Rate} = \frac{T_{\text{false alarm}} + T_{\text{missed detection}}}{T_{\text{total speech}}}$$

We also reported the Missed detection Rate(%) and False Alarm Rate (%). We used the implementation of the metrics from `pyannote.metrics` (Bredin, 2017) and all experiments were run with `pyannote.audio` Bredin *et al.* (2020).

We ran an additional analysis for the Speech Activity Detection Task to compare the use of $\Re(\mathbf{Z})$, $\Im(\mathbf{Z})$, $|\mathbf{Z}|$ and $[\Re(\mathbf{Z}), \Im(\mathbf{Z})]$ (See Table V in Appendix A).

B. Speaker Verification

The goal of the Speaker Verification task in speech processing is to accept or reject the hypothesis that a given speaker pronounced a given sentence. To do so, we learned an embedding function of any speech sequence of variable length. We examined this task, as it is believed that spectro-temporal modulations encode specifically the speaker information (Elliott and Theunissen, 2009; Lei *et al.*, 2012).

We followed the same procedure as Coria *et al.* (2020) to conduct experiments with the two version of the VoxCeleb databases: VoxCeleb2 (Chung *et al.*, 2018) is used for training, and VoxCeleb1 (Nagrani *et al.*, 2017) is split into two parts for a development and test sets. We compared two different input front-end for the speaker verification task. We compared the Learnable STRFs (64 filters) with a contraction layer (CL) and the SincNet front-end, as described in the Speech Activity Detection setup. Each model is trained with the Additive Angular Margin Loss ($\alpha = 10, m = 0.05$) with stochastic gradient descent with a learning rate of 0.01. We compared the different Speaker Verification approaches with the Equal Error Rate (EER). We also measured the performance of each approach when using the the S-normalization. We also reported the baseline performance of the I-vector system trained

on VoxCeleb1 combined with Probabilistic Linear Discriminant Analysis (PLDA) (Dehak *et al.*, 2010). We denoted by *Speaker* the experiments and the distribution of learned STRFs on this dataset and this task.

C. Urban Sound Classification

The problem of Urban Sound Classification is to classify short excerpt of audio sounds into broad categories (Ex: Car horns, Air Conditioner, Drilling). We investigated the use of the Learnable STRFs for Urban sound classification, especially to test the use of spectro-temporal modulations for other type of sounds than animal (human or bird) vocalizations (Młynarski and McDermott, 2018).

We followed the same evaluation procedure as Salamon and Bello (2017) to evaluate the experiments with the UrbanSound8K database (Salamon *et al.*, 2014). The dataset is composed of 8732 excerpts of urban sounds from 10 categories (air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music), and split in 10 separate folds. To compare with previous approaches, each model is evaluated by cross-validation on the 10 folds. We reported the Mean, Min and Max of the Accuracy across the 10 folds. We used the code-base from Arnault *et al.* (2020) for the training and evaluations of the 2 approaches. The topline approach is the use of a Mel-filterbanks with the CNN10 architecture from Kong *et al.* (2020). For the Learnable STRFs approach, the first convolution layer of the CNN10 architecture (Free 2D convolution with size 3x3 with 64 filters) is replaced by the Learnable STRFs layer (64 filters) on top of the Mel-filterbanks described in section II. The models are trained with the RAdam optimizer (Liu *et al.*, 2019)

with LookAhead (Zhang *et al.*, 2019). We also reported the results from Salamon and Bello (2017) as baseline. We denoted by *Urban* the experiments and the distribution of learned STRFs on this dataset and this task.

D. Zebra Finch Call Type classification

Finally, we examined the Zebra Finch Call Type classification task (Elie and Theunissen, 2016). The goal of this task is to classify short-excerpt of sounds into Call Type categories for the Zebra Finch bird. Indeed, it has been found by Elie and Theunissen (2016) that several properties the acoustic space allow to separate to some extent the Call Types in the repertoire of Zebra Finches. We tried to stay as close as possible from the experimental protocol of Elie and Theunissen (2016). The dataset is composed of 3433 excerpts of Zebra Finches' calls from 11 categories ('Wsst or aggressive call','Begging calls', 'Distance call', 'Distress call','Long Tonal call', 'Nest call', 'Song','Tet call', 'Thuk call','Tuck call', 'Whine call') produced by Adults and Chicks. The calls were segmented to keep only the 3 first seconds of each excerpt and if the file is too short, the sound was zero-padded.

Each set of features and model was evaluated with a random cross-validation procedure, that took into account the nested format the database. 80% of the birds were kept for training and 20% for testing. 50 different permutations of excluded birds were obtained to generate 50 training and validation data sets. To compare the approaches, we computed the Mean, Min, Max of the Accuracy over the permutations.

We ran 4 different baselines for this task based on 2 different input features and 2 types of classifiers. We extracted the features introduced by Elie and Theunissen (2016): Predefined

Acoustical Features (PAF) and the Modulation Power Spectrum (MPS). The PAF features are composed of 23 parameters extracted from Spectral envelope, Temporal Envelope and the Fundamental Frequency. (Mean, Min, Max, Std of the F0; Mean of F1; Mean of F2; Mean of F3; Saliency; RMS energy; Max of the Amplitude; Mean, Std, Skewness, Kurtosis, Entropy, first, second and third quartiles of the frequency power spectrum; Mean, Std, Skewness, Kurtosis, Entropy of the temporal envelope) The MPS representation is the amplitude spectrum of the 2D Fourier Transform applied on the spectrum representation of the sound waveform. The MPS extracts the spectro-temporal modulations in a fine-grained fashion and sum the contribution along the frequency axis. We tested both these input features with Linear Discriminant Analysis (LDA) and Random Forest (RF) classifiers as in Elie and Theunissen (2016).

Finally, we evaluated the potential of the Learnable STRFs (24 filters) for this task. We combined the Learnable STRFs with a simple linear layer to output directly the decision layer. The models were trained with the Adam optimizer (Kingma and Ba, 2014). We denote by *Bird* the experiments and the distribution of learned STRFs on this dataset and this task.

IV. RESULTS AND DISCUSSIONS

First, we analyzed the quantitative performances to perform the tasks for the Learnable STRFs for the different audio benchmarks. Then, we examined and compared, qualitatively and quantitatively, the statistics of the Learned STRFs representations.

A. Quantitative performance on audio benchmarks

Overall, the performances of the Learnable STRFs is on par for all tasks with the different baselines. There is no skip connection between the Mel-filterbanks and the rest of each neural network that has been considered. This means that these Learned STRFs are in someway useful to perform each task, as this layer act as a filter. A degradation of performance means that it might be not fully sufficient to use spectro-temporal modulations to perform this specific task.

The objective results for the *Speech Activity Detection* task are shown for all models in Table I. Overall, learnable front-end approaches with injected prior improved over the classic signal processing baselines, Mel-filterbanks and MFCCs. Yet, the approaches with Free 2D conv. were not capable to improve over the classic signal processing baselines and had the worst performance, even for the convolution initialized with Gabor filters. The best-performing models for this task, were the ones trained with the Learnable STRFs, and outperformed all the baselines. It is improving over the State-of-the-Art model with SincNet on the AMI dataset, and matched the performance on the CHIME5 dataset. Therefore, adding prior for spectro-temporal modulations was beneficial for Speech Activity Detection. The closest work to our knowledge around speech activity detection is Vuong *et al.* (2020), where they derived a layer that learn the spectro temporal modulation especially for Voice Type Discrimination in an industrial environment. The main difference with our work, is that they relied on the expression of the discrete implementation of the Hilbert transform. They also reported that parametrized Neural Network were better than free convolutions.

They also used a long receptive field along the time axis, and a small receptive field along the frequency axis.

TABLE I. Speech Activity Detection results for the different approaches described. Init. stands for Initialisation. CL stands for contraction layer, it is a convolution (conv.) layer reducing the size of the tensor dimension after the convolution (Free 2D conv. or Learnable STRFs) on the Mel-filterbanks. The Free 2D conv. had the same grid size as the Learnable STRFs (9x111). Each input front-end is then fed to a 2-layer BiLSTM and 2 feed-forward layers. The best scores for each metric overall are in **bold**. MD stands for Missed detection rate. FA stands for False Alarm rate. DetER stands for Detection Error Rate. For all metrics, lower is better.

Database	AMI			CHIME5		
Metric	DetER	MD	FA	DetER	MD	FA
Input front-end						
Mel-filterbanks	7.7	2.6	5.1	24.1	2.8	21.3
MFCC	6.3	2.7	3.5	19.6	1.6	18.0
SincNet (Ravanelli and Bengio, 2018)	6.0	**2.4**	3.6	**19.2**	1.7	17.6
Free 2D conv. Random Init. + CL	8.0	3.0	5.0	26.5	0.6	25.9
Free 2D conv. Gabor Init. + CL	7.9	2.5	5.3	26.4	**0.2**	26.1
Learnable STRFs + CL	**5.8**	**2.4**	**3.4**	**19.2**	3.1	**16.1**

TABLE II. Speaker Verification results for the different approaches described. CL stands for contraction layer, it is a convolution layer reducing the size of the tensor dimension after the convolution (Learnable STRFs). The X-vector Snyder *et al.* (2018) is used after each input front-end. We evaluated the performance of the Speaker Verification with and without S-normalization Coria *et al.* (2020). We also reported the baseline performance of the I-vector combined with Probabilistic Linear Discriminant Analysis (PLDA) (Dehak *et al.*, 2010). The best scores for each metric overall are in **bold**. For the EER, lower is better.

Metric	EER	EER w/ S-norm
Baseline		
I-vectors+PLDA (Dehak *et al.*, 2010)[a]	8.8	–
Input front-end		
SincNet (Coria *et al.*, 2020)	**3.9**	**3.5**
Learnable STRFs + CL	6.4	6.1

[a] This result is directly extracted from their paper and was not replicated for this study.

The results for the Speaker Verification task are reported in Table II. We found out that the SincNet that was designed initially for Speaker Recognition (Ravanelli and Bengio, 2018) is getting better results than the Learnable STRFs + CL. The S-normalization improved both systems. This result differ with previous results reported by Lei *et al.* (2012) that the spectro-temporal modulations were useful for speaker recognition. One difference, that could explain this discrepancy, is the use of Bayesian models after the different features

(HMM-GMM). Indeed, the X-vector (Snyder *et al.*, 2015) was designed based on the latest progresses of Deep Learning research to tackle the Speaker recognition task and were validated initially on spectral representations of the audio. Our results suggest that spectro-temporal modulations are not fully sufficient to distinguish speakers. Harmonic structure was found useful for the Speaker Verification and Recognition task Imperl *et al.* (1997). One of the hypothesis is that the learning of the harmonic structure is more difficult with the outputs Learnable STRFs layer than directly with the Mel-filterbanks.

TABLE III. Urban Sound Classification results for the different approaches described. The best score for the mean Accuracy over the 10 folds overall is in **bold**. the CNN10 architecture from Kong *et al.* (2020) is used after each input front-end. Higher is better.

Accuracy	Mean [Min - Max]
Baseline	
SB-CNN (Salamon and Bello, 2017) [a]	79 % [71%-85%]
Input front-end	
Free 2D conv. 3by3 (Kong *et al.*, 2020)	**84%** [76%-93%]
Learnable STRFs	82% [74%-90%]

[a] This result is directly extracted from their paper and was not replicated for this study.

The performances for the Urban Sound Classification task are reported in Table III. The accuracy of the Learnable STRFs is above the baseline approach from Salamon and Bello (2017) and is on par (slightly below) with the CNN10 architecture using Mel-filterbanks

Kong *et al.* (2020). It was studied before in Espi *et al.* (2015), that the use of different sizes of the Spectral representation was increasing the performance of deep learning models for acoustic Event Detection. This suggests that the varying sizes of the focus on the Mel-filterbanks representations boost the performances, both in time and frequency. In our case, the model learned to focus through the fitting of the (σ_f, σ_t) parameters.

TABLE IV. Zebra Finch Call Type classification results for the different approaches. The best scores for each metric overall are in **bold**. PAF stands for Predefined Acoustical Features and MPS for the Modulation Power Spectrum. LDA stands for Linear Discriminant Analysis. RF stands for Random Forest. The Learnable STRFs input front-end is combined with a simple linear model to output directly the decisions. Higher is better.

Accuracy	Mean	[Min - Max]
Chance level	17%	[6%-23%]
Features + Model		
PAF Elie and Theunissen (2016) + LDA	57%	[43%-71%]
PAF Elie and Theunissen (2016) + RF	59%	[47%-68%]
MPS Elie and Theunissen (2016) + LDA	41%	[23%-53%]
MPS Elie and Theunissen (2016) + RF	**69%**	[49%-84%]
Input front-end		
Learnable STRFs	43%	[23 %-73%]

Finally, the results for the Zebra Finch Call Type classification are in Table IV. On one hand, the PAF features depended slightly on the model used after for classification (LDA 57% to RF 58%) while the MPS had the worst performance overall with linear model such

as LDA while the MPS had the best performance overall when combined with the RF (going from 41 % to 69%). The Learnable STRFs models was decoded with a simple linear layer, so the closest baseline is the combination of the MPS with the LDA. The Learnable STRFs perform below the PAF features and the MPS with RF. The performance with combination of features in the MPS with RF suggest that the model with Learnable STRFs could benefit greatly from Adaptive Neural Trees Tanno *et al.* (2019) to perform the task. In addition, this encourages the use co-occurrences or anti-occurrences of the spectro-temporal patterns in models as in Młynarski and McDermott (2018, 2019), since the routing in RF imply to measure joint patterns in the feature space of the MPS.

B. Description of the learned filters

First, we observed that the Learned STRFs organized differently for each task, both the modulations (ω, Ω) (See Figure 3) and the size of the Gaussian envelopes through (σ_t, σ_f) (See Figure 6 in Appendix). Within the space allowed by the Nyquist theorem and the size of the convolutions, all the learned STRFs still concentrated in low spectral and temporal modulations (See Figure 3). We also observed as Singh and Theunissen (2003) that higher spectral modulations were found at low temporal modulations (and vice-versa). We found out that the Gaussian envelopes of the Learned STRFs can be characterized more by a continuum of values, not in a set of specific values. The Gaussian envelopes are more concentrated in the low values, and exhibit preferences depending on the task for temporal or spectral shapes. Finally, the distributions of the Learned STRFs modulations and Gaussian envelopes of *Speech* tasks on AMI and CHIME5 datasets, and the Speaker task look more

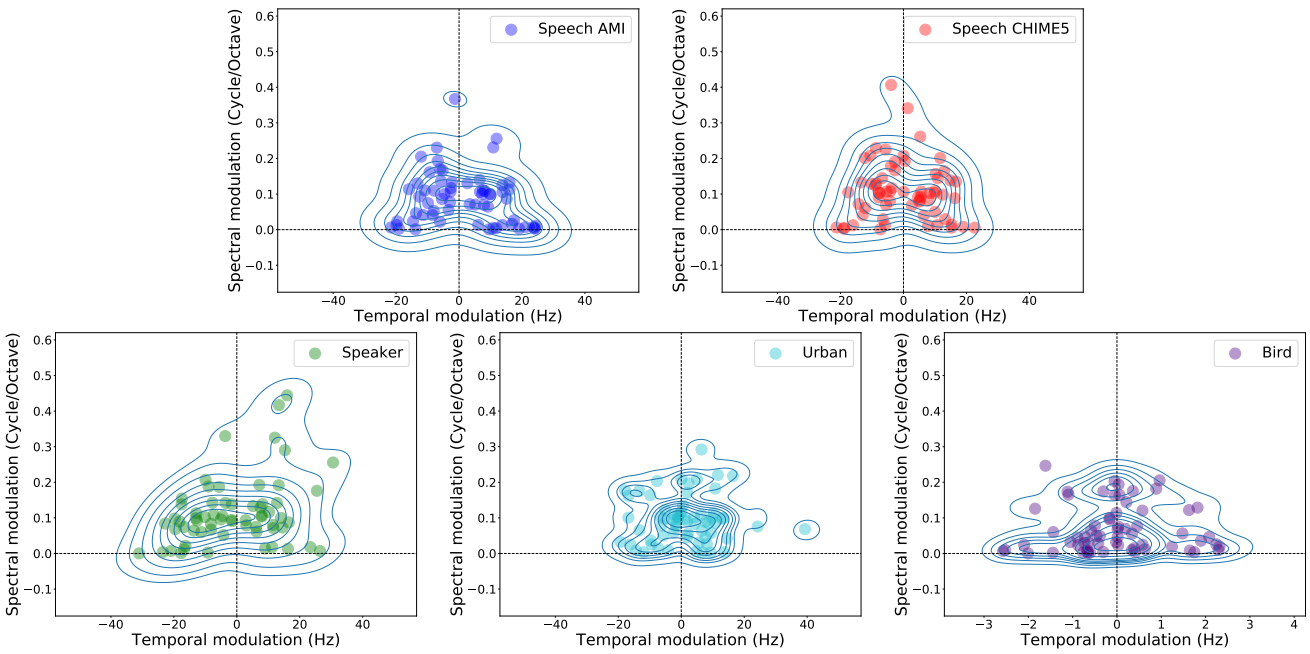

FIG. 3. Temporal and Spectral Modulation of the Learned STRFs to tackle Speech Activity Detection on the AMI dataset (Speech AMI) and on the CHIME5 (Speech CHIME5), Speaker Verification on VoxCeleb (Speaker), Urban Sound Classification on Urban8k (Urban), Zebra Finch Call Type Classification (Bird). We displayed only a subset of the Learned STFRs of the Bird and Urban tasks for clarity. We also plotted the bi-variate distributions using kernel density estimation for each task.

similar than the *Bird* and *Urban* ones. We quantified these observations with the description of the parameters described in Sub-section II A 3 and the measure of distance between tasks with optimal transport in Sub-section II A 4.

First, the separability index $\alpha_{separability}$ showed that most Learned STRFs are quite separable, and that the task related to human vocalizations (*Speech* and *Speaker*) were less separable than the other ones. We also found that all modulations have quite high $\alpha_{starriness}$ indexes. Similar results were found in Singh and Theunissen (2003) for the separability and

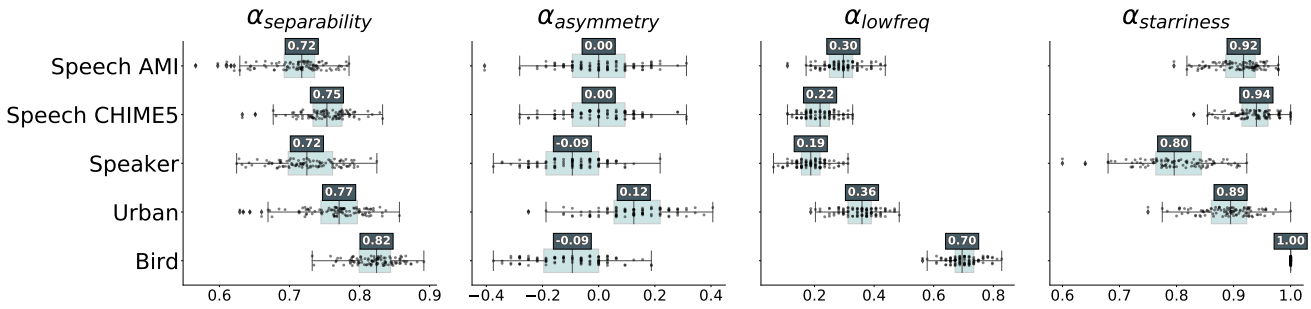

FIG. 4. Separability, asymmetry, low-pass, starriness coefficients. Four quantifiers that measured different aspects learned distribution for the different tasks under study: Speech Activity Detection on the CHIME5 dataset (Speech CHIME5) and on the AMI (Speech AMI), Urban Sound Classification on Urban8k (Urban), Speaker Verification on VoxCeleb (Speaker) Zebra Finch Call Type Classification (Bird). We displayed the median value for each α and task above each box-plot.

the starriness for the ensembles of sounds of Speech corpora, Zebra Finches vocalizations, and environmental sounds. Schädler *et al.* (2012) evaluated the use of high joint spectral and temporal modulation, and also found that they were degrading the performances for Speech Recognition tasks.

Besides, the Learned STRFs for the *Speech* did not show preferences for up or down sweeps modulations ($\alpha_{asymmetry} \approx 0.0$), while the *Speaker* and *Bird* tasks exhibit slight preferences for down-sweeps and the *Urban* for up-sweeps. The result for the *Bird* task differed from Singh and Theunissen (2003). This could be explained by the fact that Singh and Theunissen (2003) used a quantification of these parameters with an ensemble of sounds. The information about the specific characteristic of an individual Zebra Finch is mixed with the information of the Call Type. This suggests that a fully interpretable supervised approach might allow to decipher the different factors and contributions that influenced the acoustic properties of vocalizations. Finally, the *Bird* task focused more on the low frequency

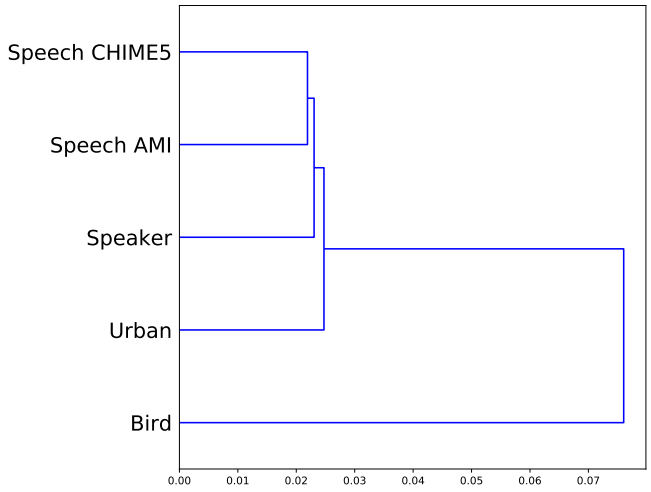

FIG. 5. Hierarchical clustering of the tasks: Speech Activity Detection on the CHIME5 dataset (Speech CHIME5) and on the AMI (Speech AMI), Urban Sound Classification on Urban8k (Urban), Speaker Verification on VoxCeleb (Speaker) Zebra Finch Call Type Classification (Bird). The distance between tasks is computed between the learned STRF filters of each task with the Sinkhorn distance (we used the euclidean distance between each filter and the regularization parameter of the Sinkhorn distance is $\lambda = 10^{-3}$).

modulations ($\alpha_{lowfreq} \approx 0.70$) than the other tasks ($\alpha_{lowfreq} \leq 0.35$). We also observed that the Learned STRFs of the *Speaker* task moved away from the low spectral modulations and yielded the lowest low-pass coefficient ($\alpha_{lowfreq} \approx 0.19$). Especially, Elliott and Theunissen (2009) also found out that the removing of spectral modulations between 3 and 7 cycles/kHz significantly increases the gender mis-identifications of female speakers. In addition, the results for the *Speech* on the AMI and CHIME5 datasets and the *Speaker* are very similar to the ones found directly in the auditory cortex neurons in awake monkeys (Massoudi *et al.*, 2015) and in awake humans (Hullett *et al.*, 2016). Hullett *et al.* (2016); Massoudi *et al.* (2015) measured responses of natural sounds directly in the superior temporal gyrus

and found specific spectral modulation selectivity for 0.4 ± 0.55 Cycle/Octave and specific temporal modulation 16 ± 11Hz; and most of the modulations were concentrated along the axes with high separability.

Finally, we examined the structure obtained from the hierarchical clustering based on the distances between tasks, see Sub-section II A 4 for the full description. We obtained the clustering tree in Figure 5. We observed that the learned STRFs of the different tasks organized in a meaningful disposition. The Learned STRFs for *Speech* on the CHIME5 and AMI are the closest to each other. Then, we found out that other Human vocalization task, *Speaker*, is closer to the *Speech* ones. On the other hand, the *Bird* task organized far away from both the *Urban* and the Human vocalizations tasks (*Speech* and *Speaker*). In future work, this method could be used to discover automatically phylogeny trees based only on acoustic properties of spectro-temporal modulations and test these predictions against a molecular-based phylogeny (McCracken and Sheldon, 1997).

V. CONCLUSION AND FUTURE WORK

In summary, we examined the use of a parametrized neural network front-end to learn spectro-temporal modulations optimal for different behavioral tasks. This front-end, the Learnable STRFs, yielded performances close to published state-of-the-art using engineering-oriented neural network for Speaker Verification, Urban Sound Classification, Zebra Finch Call Type Classification, and obtained the best results on two datasets for Speech Activity Detection. As our front-end is fully interpretable, we found markedly different spectro-temporal modulations as a function of the task, showing that each task relies on a specific

set of modulations. These task-specific modulations were globally congruent with previous work based on three approaches: spectro-temporal analysis of different audio signals (Elliott and Theunissen, 2009), analysis of trained neural networks (Schädler *et al.*, 2012), and analysis of the auditory cortex (Hullett *et al.*, 2016; Santoro *et al.*, 2017). In particular, for the Speech Activity Detection task, we observed the same modulation distributions as the ones found directly the human auditory cortex listening while listening to naturalistic speech Hullett *et al.* (2016). The modulations also displayed generic characteristics across tasks, namely, a predominance of low frequency spectral and temporal modulations and a high degree of 'stariness' and 'separability', corresponding to the fact that filters tend to remain close to either the temporal or spectral axis, with low occupation of joint spectro and temporal responses. This is consistent with Singh and Theunissen (2003). In order to encourage reproducible research, the developed Learnable STRFs layer, the learned STRFs modulations, and the recipes to replicate results are available in a open-source package [1].

Several avenues of extensions are possible for this work, based on what is known in auditory neuroscience. First, this work only modelled the final outcome of plasticity after each task has been fully learned, starting from a random initialization. Yet, the same model could be used to address a range of issues relevant to changes occurring during task learning (top-down plasticity) or due to modification of the distribution of audio input (bottom-up plasticity). Recent work Bellur and Elhilali (2015) have investigated the adaptation of modulations in analytical models and witnessed several improvements in terms of engineering performances suggesting that this is also an interesting avenue in terms of behavioral modeling. Second, the analyses from Hullett *et al.* (2016) showed that not only neurons have

spectro-temporal selectivity but are also topographically distributed along the posterior-to-anterior axis in the superior temporal gyrus. In future work, it would be interesting to reproduce such topography by using an auxiliary self-organizing maps objective in addition to the task-specific loss function for the STRFs. Finally, despite their wide use in auditory neuroscience, the spectro-temporal modulations do not provide a complete picture of computations in the auditory cortex (Williamson *et al.*, 2016). A potential extension of our work would be to add an extra layer able to express co-occurrences and anti-occurrences of pairs of spectro-temporal receptive fields as in Młynarski and McDermott (2018, 2019). Such an extra layer would provide a learnable and interpretable extension to spectro-temporal representations.

To conclude, we emphasize that neuroscience-inspired parametrized neural networks can provide models that are both efficient in terms of behavioral tasks and interpretable in terms of auditory signal processing.

ACKNOWLEDGMENTS

This work is funded in part from the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute). ACBL was funded through Neuratris, and ED in his EHESS role by Facebook AI Research (Research Gift) and CIFAR (Learning in Minds and Brains). The university (EHESS, CNRS, INRIA, ENS Paris-Sciences Lettres) obtained the datasets reported in this paper, and the experiment were run on it's computer resources.

APPENDIX A:

1. Importance of the representation of the Learnable STRFs

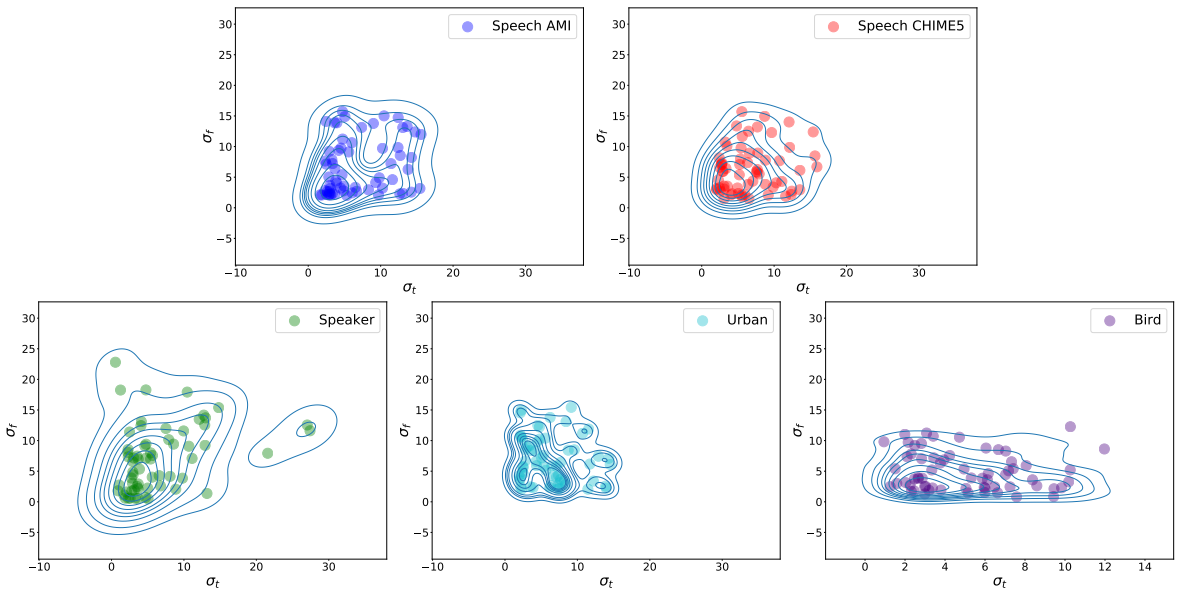

FIG. 6. Gaussian envelopes (σ_t, σ_f) of the Learned STRFs to tackle Speech Activity Detection on the AMI dataset (Speech AMI) and on the CHIME5 (Speech CHIME5), Speaker Verification on VoxCeleb (Speaker), Urban Sound Classification on Urban8k (Urban), Zebra Finch Call Type Classification (Bird). We displayed only a subset of the Learned STFRs of the Bird and Urban tasks for clarity. We also plotted the bi-variate distributions using kernel density estimation for each task.

We performed an additional analysis of Speech Activity Detection of the choice of representations $\mathbf{Z}$ from the Learnable STRFs used in the subsequent neural network. The performance of the real part, the imaginary part and absolute values of the filter output are compared. The results are presented in Table V. In comparison with the concatenation of the real and imaginary parts, the performance obtained for each part were in the same

range on the AMI dataset, but were below on the CHIME5 dataset. As in Schädler *et al.* (2012), this indicates that phase information contained in the real and imaginary parts is important for the Learnable STRFs.

TABLE V. Speech Activity Detection results for the different uses of the **Z** for the Learnable STRFs. CL stands for contraction layer, it is a convolution layer reducing the size of the tensor dimension after the convolution (Learnable STRFs). Each input front-end is then fed to a 2-layer BiLSTM and 2 feed-forward layers. The best scores for each metric overall are in **bold**. MD stands for Missed detection rate. FA stands for False Alarm rate. DetER stands for Detection Error Rate. For all metrics, lower is better.

Database	AMI			CHIME5		
Metric	DetER	MD	FA	DetER	MD	FA
Input front-end (Learnable STRFs + CL)						
Real part $\Re(\mathbf{Z})$	5.9	2.4	3.5	20.1	2.6	17.5
Imaginary part $\Im(\mathbf{Z})$	5.9	**2.2**	3.7	22.1	**1.0**	21.1
Magnitude $\lvert\mathbf{Z}\rvert$	5.9	**2.2**	3.7	19.8	3.1	16.7
Concatenation $[\Re(\mathbf{Z}), \Im(\mathbf{Z})]$	**5.8**	2.4	**3.4**	**19.2**	3.1	**16.1**

[1]https://github.com/bootphon/learnable-strf

Alekseev, A., and Bobe, A. (**2019**). "Gabornet: Gabor filters with learnable parameters in deep convolutional neural network," pp. 1–4, doi: 10.1109/EnT47717.2019.9030571.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G. *et al.* (**2016**). "Deep speech 2: End-to-end speech recognition in english and mandarin," in *ICML*, pp. 173–182.

Arnault, A., Hanssens, B., and Riche, N. (**2020**). "Urban sound classification: striving towards a fair comparison," arXiv preprint arXiv:2010.11805 .

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (**2018**). "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India.

Bellur, A., and Elhilali, M. (**2015**). "Detection of speech tokens in noise using adaptive spectrotemporal receptive fields," in *2015 49th Annual Conference on Information Sciences and Systems (CISS)*, IEEE, pp. 1–6.

Bredin, H. (**2017**). "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, http://pyannote.github.io/pyannote-metrics.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (**2020**). "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7124–7128.

Chang, S.-Y., and Morgan, N. (**2014**). "Robust cnn-based speech recognition with gabor filter kernels," in *Fifteenth annual conference of the international speech communication association*.

Cheuk, K. W., Anderson, H., Agres, K., and Herremans, D. (**2020**). "nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks," IEEE Access **In press**.

Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," The Journal of the Acoustical Society of America **118**(2), 887–906.

Chung, J. S., Nagrani, A., and Zisserman, A. (**2018**). "Voxceleb2: Deep speaker recognition," Proc. Interspeech 2018 1086–1090.

Coria, J. M., Bredin, H., Ghannay, S., and Rosset, S. (**2020**). "A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification," in *Statistical Language and Speech Processing*, edited by L. Espinosa-Anke, C. Martín-Vide, and I. Spasić, Springer International Publishing, Cham, pp. 137–148.

Cuturi, M. (**2013**). "Sinkhorn distances: Lightspeed computation of optimal transport," Advances in neural information processing systems **26**, 2292–2300.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (**2010**). "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," Journal of neurophysiology **85**(3), 1220–1234.

Edraki, A., Chan, W.-Y., Jensen, J., and Fogerty, D. (**2019**). "Improvement and assessment of spectro-temporal modulation analysis for speech intelligibility estimation," in *Interspeech 2019Annual Conference of the International Speech Communication Association*, ISCA, pp. 1378–1382.

Efron, B., and Tibshirani, R. J. (**1994**). *An introduction to the bootstrap* (CRC press).

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," Speech communication **41**(2-3), 331–348.

Elie, J. E., and Theunissen, F. E. (**2016**). "The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals," Animal cognition **19**(2), 285–315.

Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS computational biology **5**(3).

Espi, M., Fujimoto, M., Kinoshita, K., and Nakatani, T. (**2015**). "Exploiting spectro-temporal locality in deep learning based acoustic event detection," EURASIP Journal on Audio, Speech, and Music Processing **2015**(1), 1–12.

Ezzat, T., Bouvrie, J., and Poggio, T. (**2007**). "Spectro-temporal analysis of speech using 2-d gabor filters," in *Eighth Annual Conference of the International Speech Communication Association*.

Flamary, R., and Courty, N. (**2017**). "Pot python optimal transport library" https://pythonot.github.io/.

Francis, N. A., Elgueda, D., Englitz, B., Fritz, J. B., and Shamma, S. A. (**2018**). "Laminar profile of task-related plasticity in ferret primary auditory cortex," Scientific reports **8**(1),

1–10.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (**2003**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nature neuroscience **6**(11), 1216–1223.

Gabor, D. (**1946**). "Theory of communication. part 1: The analysis of information," Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering **93**(26), 429–441.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. (**2016**). "Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli," Journal of Neuroscience **36**(6), 2014–2026.

Imperl, B., Kačič, Z., and Horvat, B. (**1997**). "A study of harmonic features for the speaker recognition," Speech communication **22**(4), 385–402.

Jääskeläinen, I. P., Ahveninen, J., Belliveau, J. W., Raij, T., and Sams, M. (**2007**). "Short-term plasticity in auditory cognition," Trends in neurosciences **30**(12), 653–661.

Kell, A. J., and McDermott, J. H. (**2019**). "Deep neural network models of sensory systems: windows onto the role of task constraints," Current opinion in neurobiology **55**, 121–132.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (**2018**). "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," Neuron **98**(3), 630–644.

Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 .

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (**2020**). "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2880–2894.

Koumura, T., Terashima, H., and Furukawa, S. (**2019**). "Cascaded tuning to amplitude modulation for natural sound recognition," Journal of Neuroscience **39**(28), 5517–5533.

Lei, H., Meyer, B. T., and Mirghafori, N. (**2012**). "Spectro-temporal gabor features for speaker recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4241–4244.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. (**2019**). "On the variance of the adaptive learning rate and beyond," in *International Conference on Learning Representations*.

Massoudi, R., Van Wanrooij, M. M., Versnel, H., and Van Opstal, A. J. (**2015**). "Spectrotemporal response properties of core auditory cortex neurons in awake monkey," PLoS One **10**(2), e0116118.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V. *et al.* (**2005**). "The ami meeting corpus," in *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, Noldus Information Technology, pp. 137–140.

McCracken, K. G., and Sheldon, F. H. (**1997**). "Avian vocalizations and phylogenetic signal," Proceedings of the National Academy of Sciences **94**(8), 3833–3836.

McDermott, J. H. (**2018**). "Audition," Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience **2**, 1–57.

Mesgarani, N., Slaney, M., and Shamma, S. A. (**2006**). "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations," IEEE Transactions on Audio, Speech, and Language Processing **14**(3), 920–930.

Meyer, A. F., Williamson, R. S., Linden, J. F., and Sahani, M. (**2017**). "Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation," Frontiers in systems neuroscience **10**, 109.

Młynarski, W., and McDermott, J. H. (**2018**). "Learning midlevel auditory codes from natural sound statistics," Neural computation **30**(3), 631–669.

Młynarski, W., and McDermott, J. H. (**2019**). "Ecological origins of perceptual grouping principles in the auditory system," Proceedings of the National Academy of Sciences **116**(50), 25355–25364.

Nagrani, A., Chung, J. S., and Zisserman, A. (**2017**). "Voxceleb: a large-scale speaker identification dataset," Telephony **3**, 33–039.

Ondel, L., Li, R., Sell, G., and Hermansky, H. (**2019**). "Deriving spectro-temporal properties of hearing from speech data," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 411–415.

Peyré, G., Cuturi, M. *et al.* (**2019**). "Computational optimal transport: With applications to data science," Foundations and Trends® in Machine Learning **11**(5-6), 355–607.

Pillow, J., and Sahani, M. (**2019**). "Editorial overview: Machine learning, big data, and neuroscience" .

Ravanelli, M., and Bengio, Y. (**2018**). "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp. 1021–1028.

Saddler, M. R., Gonzalez, R., and McDermott, J. H. (**2020**). "Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception," bioRxiv .

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

Salamon, J., Jacoby, C., and Bello, J. P. (**2014**). "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., and Formisano, E. (**2017**). "Reconstructing the spectrotemporal modulations of real-life sounds from fmri response patterns," Proceedings of the National Academy of Sciences **114**(18), 4799–4804.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (**2012**). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," The Journal of the Acoustical Society of America **131**(5), 4134–4151.

Shamma, S. A. (**1996**). "Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method," Network: Computation in Neural Systems **7**(3), 439–476.

Singh, N. C., and Theunissen, F. E. (**2003**). "Modulation spectra of natural sounds and ethological theories of auditory processing," The Journal of the Acoustical Society of America **114**(6), 3394–3411.

Snyder, D., Chen, G., and Povey, D. (**2015**). "Musan: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484 .

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (**2018**). "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5329–5333.

Stevens, S. S., Volkmann, J., and Newman, E. B. (**1937**). "A scale for the measurement of the psychological magnitude pitch," The Journal of the Acoustical Society of America **8**(3), 185–190.

Tanno, R., Arulkumaran, K., Alexander, D., Criminisi, A., and Nori, A. (**2019**). "Adaptive neural trees," in *International Conference on Machine Learning*, PMLR, pp. 6166–6175.

Thoret, E., Andrillon, T., Léger, D., and Pressnitzer, D. (**2020**). "Probing machine-learning classifiers using noise, bubbles, and reverse correlation," BioRxiv .

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (**2016**). "Instance normalization: The missing ingredient for fast stylization," arXiv preprint arXiv:1607.08022 .

Vuong, T., Xia, Y., and Stern, R. M. (**2020**). "Learnable Spectro-Temporal Receptive Fields for Robust Voice Type Discrimination," in *Proc. Interspeech 2020*, pp. 1957–1961, doi: 10.21437/Interspeech.2020-1878.

Williamson, R. S., Ahrens, M. B., Linden, J. F., and Sahani, M. (**2016**). "Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds," Neuron **91**(2), 467–481.

Woolley, S. M., Fremouw, T. E., Hsu, A., and Theunissen, F. E. (**2005**). "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds,"

Nature neuroscience **8**(10), 1371–1379.

Yarkoni, T., and Westfall, J. (**2017**). "Choosing prediction over explanation in psychology: Lessons from machine learning," Perspectives on Psychological Science **12**(6), 1100–1122.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. (**2019**). "Lookahead optimizer: k steps forward, 1 step back," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Vol. 32, pp. 9597–9608.