JASA ARTICLE

Learning spectro-temporal representations of complex sounds with parameterized neural networks[a)]

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

[1]*Ecole des Hautes Etudes en Sciences Sociales, CNRS, Institut National de Recherche informatique et Automatique, Département d'Études Cognitives, Ecole Normale Supérieure-Paris Sciences et 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 et Lettres University, 29 Rue d'Ulm, 75005 Paris, France*

ABSTRACT:

Deep learning models have become potential candidates for auditory neuroscience research, thanks to their recent successes in 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, which computes specific spectro-temporal modulations based on Gabor filters [learnable spectro-temporal filters (STRFs)] and is fully interpretable. We evaluated this layer on speech activity detection, speaker verification, urban sound classification, and zebra finch call type classification. We found that models based on learnable STRFs are on par for all tasks with state-of-the-art and obtain the best performance for speech activity detection. As this layer remains a Gabor filter, it is fully interpretable. Thus, we used quantitative measures to describe distribution of the learned spectro-temporal modulations. 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 vocalization tasks closer to each other and bird vocalizations far away from human vocalizations and urban sounds tasks. © *2021 Acoustical Society of America*. https://doi.org/10.1121/10.0005482

(Received 18 February 2021; revised 4 June 2021; accepted 8 June 2021; published online 14 July 2021)

[Editor: Marie Roch] Pages: 353–366

I. INTRODUCTION

The main objective of auditory neuroscience is to build models that can predict both the brain neural responses to relevant sounds and the behaviors 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 of the importance of also matching 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.

DNN models typically take as input a spectral representation [although some new trends consist in sidestepping this representation and work directly from the raw waveform (Ravanelli and Bengio, 2018; Zeghidour *et al.*, 2018)]. 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, DNN 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 further 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 uncovering 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 an in depth explanation of 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

[a)]This paper is part of a special issue on Machine Learning in Acoustics.

[b)]Also at: NeuroPsychologie Interventionnelle, Ecole Normale Supérieure, 75005 Paris, France. Electronic mail: riadrachid3@gmail.com, ORCID: 0000-0002-7753-1219.

[c)]Also at: Facebook AI Research, Paris, France.

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), and 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 functional magnetic resonance imaging (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 (time-frequency analysis) of these modulations in the spectrogram have been proposed (Chang and Morgan, 2014; Chi *et al.*, 2005; Ezzat *et al.*, 2007; Schädler *et al.*, 2012), for instance, with a 2D discrete wavelet decomposition of the spectrogram. The idea is that on top of the spectrum, spectro-temporal wavelets (such as Morlet/Gabor wavelets) can be defined that drive both behavioral responses and brain signals. The problem of such analytic models is that they only propose a potentially very large representation space and 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 that explicitly represents spectro-temporal filters (STRFs), but whose parameters are differentiable and can therefore be tuned to each particular task. There are two advantages of this approach, as illustrated in Fig. 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 (Fig. 2). 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 the perceptual learning aspect of plasticity with a lot less training data than 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: Sec. 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 Sec. III, we describe the experimental setup with the different computational tasks, data, state-of-the-art systems, and evaluations; Sec. IV presents the performance results, the analysis of the learned distribution of spectro-temporal modulation for each setup, and the discussion. Section V presents our conclusions and potential future work. To encourage reproducible research, the developed learnable STRF layer, the learned STRF modulations, and the recipes to replicate results are available in an open-source package.[1]

II. MODELS AND METHODS

A. Overview

To learn spectro-temporal representations of sounds, we constructed a learnable front-end model of natural sounds that stays interpretable. The model is composed of an initial fixed frequency analysis of sounds. Then this time-frequency representation of the sound is convoluted with a parametrized layer that controls the parameters of a set of Gabor filters (Sec. II B). We used this layer as a replacement of the first stage of processing in the different neural network architectures to solve each individual task. As this layer adapted to each task under study, our ability to directly read out the parameters helped us quantify what was being learned by the models (Sec. II C).

B. Learnable STRFs

Here, $\Re$, $\Im$, $|\cdot|$, and $[\cdot,\cdot]$ represent the real part, imaginary part, modulus, and concatenation operators, respectively. $\{\cdot\}$ represents a set and $card(\cdot)$ 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 before spectral analysis (Ulyanov *et al.*, 2016). We computed Mel-filterbanks by filtering sounds with a set of 64 bandpass filters cascaded with a log compression, to mimic the cochlear frequency analysis. All the sounds are sampled at 16 kHZ; thus, the center frequency of the filters spanned $[0.0, 8000.0\,\text{Hz}]$. Frames are computed every 10 ms, with a Hamming window of 25 ms. The computations for the Mel-filterbanks of the audio can be performed on-the-fly directly on the graphics processing unit (GPU) thanks to Cheuk *et al.* (2020).

2. Definition of the learnable STRFs

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

The two-dimensional (2D) Gabor filter kernel g_k is a sine-wave w_k modulated by a 2D 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:

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

https://doi.org/10.1121/10.0005482

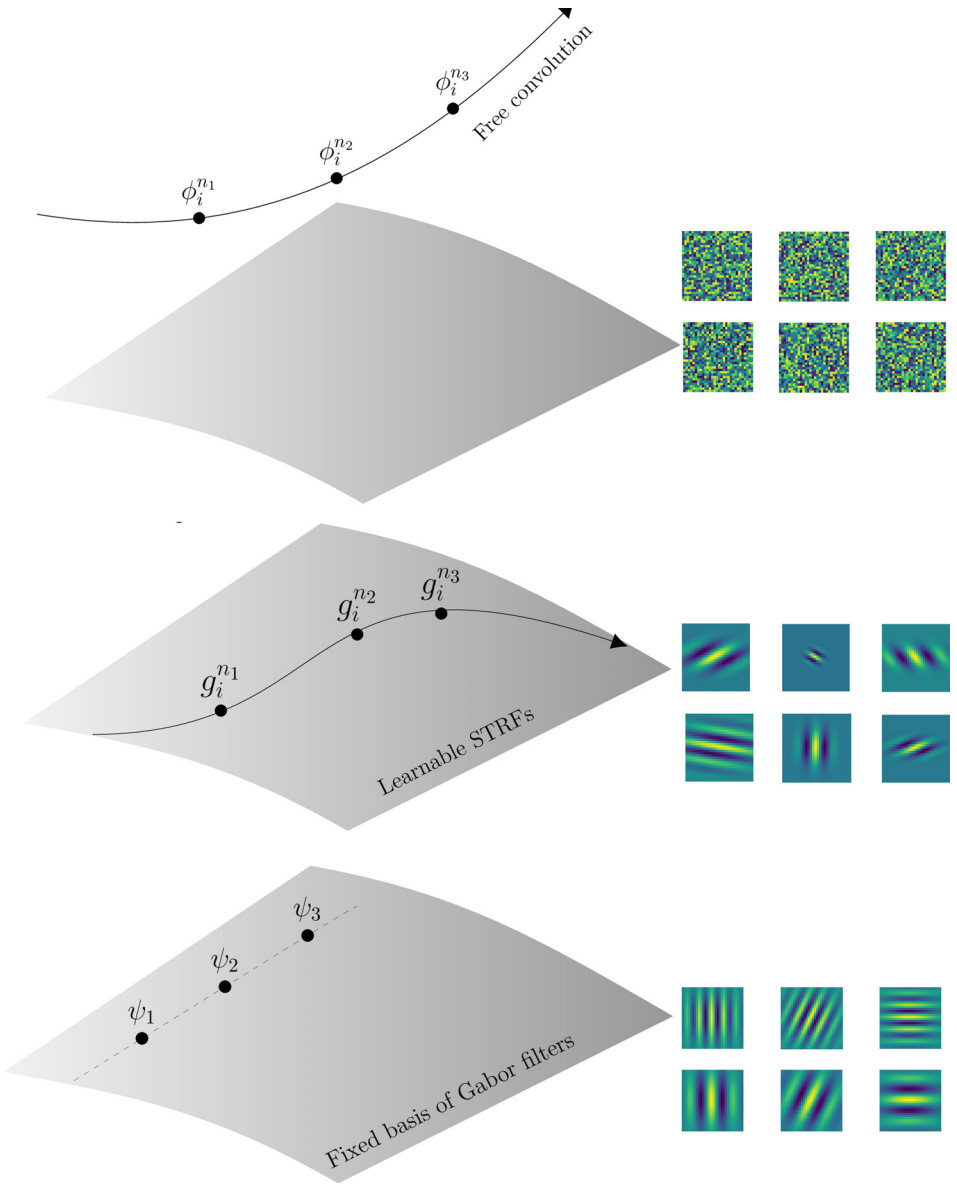

FIG. 1. (Color online) Schematic illustration of the different approaches to obtain spectro-temporal representations of sounds. The three dimensions represent all the functions that can be learned by a convolution. The parametric surface represents the space of Gabor functions. Top: Free learnable convolutions are unconstrained to move anywhere in the space of functions $(\phi_i^{n_k})$ (Młynarski and McDermott, 2018; Ondel *et al.*, 2019). The learned functions remain difficult to interpret. Middle: Learnable STRFs remain in the Gabor space of functions $(g_k^{n_i})$ (this study). The learned filters remain interpretable. Bottom: Fixed basis (ψ_i) are predefined by hand for each task. The 2D Gabor filterbanks are built with various scales and rotations but do not concentrate in specific modulations of interest (Bellur and Elhilali, 2015; Chang and Morgan, 2014; Elie and Theunissen, 2016; Mesgarani *et al.*, 2006; Schädler *et al.*, 2012). The upper index n_k and lower index i represent the n_k-step during learning for the filter indexed i.

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

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

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

We obtain a bank of N filters $\{g_k(t,f)\}_{k=0..N-1}$. This bank of filters is convolved 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. First, Gabor filters can be used as an initialization [free two-dimensional convolution Gabor initiation (free 2D conv. Gabor Init.)] of a 2D convolution neural network, and the 2D grid is tuned completely by backpropagation. In Chang and Morgan (2014), the authors compared the use of fixed Gabor features and the method free 2D conv. Gabor Init. (GCNN in their paper for automatic speech recognition). In our case, we also used Gabor filters for the learnable STRFs, but 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 filters are differentiable almost everywhere with respect to the parameters $(\sigma_{t_k}, \sigma_{f_k}, F_k, \gamma_k)$. 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.

Approaches such as Ezzat *et al.* (2007), Mesgarani *et al.* (2006), and Schädler *et al.* (2012) designed the Gabor filterbanks by hand for specific audio tasks. They designed each time a Gabor filterbank tailored by hand for each specific task under study (bottom panel of Fig. 1). Our learnable STRFs learned this Gabor filterbank (middle panel of Fig. 1).

On the other hand, Młynarski and McDermott (2018) and Ondel *et al.* (2019) did not use a prior and structure on

the convolution on top of the time-frequency representation. All the weights need to be learned, and the convolutions remain difficult to interpret (top panel of Fig. 1). This approach was evaluated with free two-dimensional convolution random initiation (free 2D conv. random Init.).

Images and spectrograms extracted from audio have very different properties. On one hand, images have important geometric variability due to perspective projections of 3D scenes under various viewpoints. On the other hand, classes of objects are invariant to image rotations (image rotation is an important data augmentation in computer vision).

These differences of modalities led us to add two main differences with our learnable STRF layer and the GaborNet (Alekseev and Bobe, 2019) introduced in computer vision. First, we introduced two different parameters (σ_t, σ_f) to give more freedom to the model in the temporal and spectral axes. Indeed, spectral and temporal axes play very different roles in the spectrogram. The second main difference is the shape of the receptive field. A 3×3 filter is limiting the computations of temporal modulations with an excerpt of 30 ms. Slow long-range modulations cannot be captured with a small receptive field, yet it was shown that different windows of integration play important parts in speech perception (Poeppel, 2003).

Therefore, each learnable STRF filter takes as input nine Mel-frequency bands and 1.1 s of context, thus yielding a size of 9×111 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 this specific front-end by "learnable STRFs" in Tables I–V and denoted it by "learned STRFs" once we examined these representations.

C. Descriptive quantifiers of the distribution

We used quantitative measures to describe the structure of the distribution of the learned STRFs. These measures are used in auditory neuroscience to study the characteristics of the transfer function of biological neurons in several species: zebra finch (Depireux *et al.*, 2001; Theunissen *et al.*, 2000), monkeys (Massoudi *et al.*, 2015), and humans (Schönwiesner and Zatorre, 2009).

These measures (Singh and Theunissen, 2003) can also be extracted directly from the sound ensembles and compared to the ones extracted in the brain.

To extract such measures from our models, we read out directly the parameters of the learned STRFs and converted them 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) and 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 task. As the learned STRFs self-organized to solve each task, we examined each of these parameters for each task. Each α is estimated with the bootstrap re-sampling method (Efron and Tibshirani, 1994) on the learned STRFs (100 bootstraps).

1. Asymmetry

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

$$\alpha_{\text{asymmetry}} = \frac{card(\{g_k \text{s.t.} \omega_k > 0\})}{card(\{g_k\})} = \frac{card(\{g_k \text{s.t.} \omega_k > 0\})}{N}. \quad (3)$$

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. For instance, zebra finches exploit these degrees of freedom during their calls. The distribution of down-sweeps of zebra finch calls differs between male and female (Theunissen *et al.*, 2000).

2. Low-pass coefficient and starriness

It has been observed in Singh and Theunissen (2003) that most energy in the 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:

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

For the temporal modulation low limit, we opt, as did Singh and Theunissen (2003), for $\Delta_t = 16$ Hz. The spectral modulation low limit is set to $\Delta_f = 0.08$ cycle/octave. These parameters were chosen deliberately low as in Singh and Theunissen (2003) to observe differences between tasks. The parameter α_{star} measures the "starriness" of the distribution. This measure examines portions of the distribution that do not have high joint modulation and is an indicator of the importance of low modulation in either time or frequency, but not both,

$$\alpha_{\text{star}} = \frac{N_{\Delta_t} + N_{\Delta_f} - 2 \times N_{\text{low}}}{N - N_{\text{low}}}. \quad (5)$$

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

https://doi.org/10.1121/10.0005482

3. 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

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

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

$$\alpha_{\text{sep}} = \frac{\lambda_1}{\sum_{i=1}^{n}\lambda_i}. \quad (7)$$

If $\alpha_{\text{sep}} \approx 1$, the distribution of the learned STRFs can be considered separable. Indeed, the magnitudes of the singular values λ_i of the decomposition provide information about the stretching/shrinking in the corresponding directions.[2] Complete separability suggests that there are fully independent temporal and spectral processing stages in the brain (Depireux *et al.*, 2001; Flinker *et al.*, 2019).

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 on the Mahalanobis distance or an approximation of the Kullback–Leibler (KL) divergence to compare observations of two sets of points, yet either these metrics make modeling assumptions about the data (approximation of the underlying density functions that generated the data), or it is impossible to compare sets of points with different cardinals.

The non-parametric, natural, and most powerful way to compare distributions is to use optimal transport distances (Peyré and Cuturi, 2019). Instead of computing distance between two individual items at a time, optimal transport is concerned with the problem of moving simultaneously several items (i.e., a distribution) from one configuration onto another. 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 allows fast computation of distances and multiple assignments between points. Optimal transport distances require a metric space to find the transport between the sets of 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 us to compare the different tasks if we have several models due to cross-validation (urban and bird) or fewer filters for a specific task (bird). If we denote, by $\langle .,. \rangle_F$, the norm of Froebenius between matrices, the regularized distance d_λ between two tasks is defined as

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

Therefore, we were able to obtain a proxy on how close two different tasks $task_a$ and $task_b$ are 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 dendrogram (see Fig. 5).

III. EXPERIMENTAL SETUP

We compared the learnable STRF layer with state-of-the-art systems that have recently been introduced to solve each task as well more classic baselines for each task. We tried to keep the experimental methods as close as possible to the methods used in the reported experiments of state-of-the-art 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 make standout speech in an audio stream with silences and background noises (Mesgarani *et al.*, 2006).

We conduct experiments with two challenging datasets with different characteristics.

The AMI database (McCowan *et al.*, 2005) is a meeting dataset in English recorded with multiple microphones in three different rooms. There are 180 different speakers in the dataset. Here, we focus on the *AMI.SpeakerDiarization.MixHeadset* protocol, as we are working only on 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-ends to tackle this task. We evaluated the learnable STRFs (64 filters) with a contraction layer (CL) as well the free 2D convolution with a CL. The CL is a convolution layer taking the outputs at each time step of the learnable STRFs to reduce the number of dimensions of output tensor of the learnable STRF layer $\mathbf{Z}$ [see Eq. (2)]. We compared these techniques with classic signal processing baselines used in speech processing: Mel-filterbanks with 64 filters and Mel-frequency cepstral coefficients (MFCC) with 19 coefficients, with their deltas and their delta-deltas. We also compared them with the more recent parametrized neural network SincNet. SincNet is composed of parametrized sinc functions, which implement 80 bandpass filters (to replace directly more classic input spectral representations), and three temporal convolution/pooling layers. All the input front-ends 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),

$$\text{DetER} = \frac{T_{\text{false alarm}} + T_{\text{missed detection}}}{T_{\text{total speech}}}.$$

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 the Appendix).

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 versions 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 development and test sets. We compared two different input front-ends for the speaker verification task. We compared the learnable STRFs (64 filters) with a 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 S-normalization. We also reported the baseline performance of the I-vector system trained on VoxCeleb1 combined with probabilistic linear discriminant analysis (PLDA). 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 excerpts of audio sounds into broad categories (e.g., car horns, air conditioners, drilling). We investigated the use of the learnable STRFs for urban sound classification, especially to test the use of spectro-temporal modulations for types of sounds other 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 ten categories (air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music) and split into ten separate folds. To compare with previous approaches, each model is evaluated by cross-validation on the ten folds. We reported the mean, minimum, and maximum of the accuracy across the ten folds. We used the code-base from Arnault *et al.* (2020) for the training and evaluations of the two approaches. The state-of-the-art approach is the use of 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 3×3 with 64 filters) is replaced by the learnable STRFs layer (64 filters) on top of the Mel-filterbanks described in Sec. 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

https://doi.org/10.1121/10.0005482

task is to classify short excerpts of sounds into call type categories for the zebra finch bird. Indeed, it has been found by Elie and Theunissen (2016) that several properties of the acoustic space allow the separation, to some extent, of the call types in the repertoire of zebra finches. We tried to stay as close as possible to 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 first 3 s of each excerpt, and if the file was 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 of the database. Eighty percent of the birds were kept for training and 20% for testing. Fifty different permutations of excluded birds were obtained to generate 50 training and validation data sets. To compare the approaches, we computed the mean, minimum, and maximum of the accuracy over the permutations.

We ran four different baselines for this task based on two different input features and two types of classifiers. We extracted the features introduced by Elie and Theunissen (2016): predefined acoustical features (PAF) and the modulation power spectrum (MPS). The PAFs are composed of 23 parameters extracted from the spectral envelope, temporal envelope, and fundamental frequency [mean, minimum, maximum, standard deviation (SD) of the F0; mean of F1; mean of F2; mean of F3; saliency; root mean square (RMS) energy; maximum of the amplitude; mean, SD, skewness, kurtosis, entropy, first, second, and third quartiles of the frequency power spectrum; mean, SD, 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 sums 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 directly output 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 DISCUSSION

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 STRF representations.

A. Quantitative performance on audio benchmarks

Overall, the performances of the learnable STRFs are 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 some way useful to perform each task, as this layer acts 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 convolution were not capable of improving 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 they outperformed all the baselines. They improved 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 learned the spectro-temporal modulation especially for voice type discrimination in an industrial environment. The main difference from our work is that they relied on the expression of the discrete implementation of the Hilbert transform. They also reported that parametrized neural networks 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.

The results for the speaker verification task are reported in Table II. We found that the SincNet that was designed initially for speaker recognition (Ravanelli and Bengio, 2018) got better results than the learnable STRFs + CL. The S-normalization improved both systems. This result differs from previous findings reported by Lei *et al.* (2012) that the spectro-temporal modulations were useful for speaker recognition. One difference, which could explain this discrepancy, is the use of Bayesian models after the different features [hidden Markov model-Gaussian mixture model (HMM-GMM)]. Indeed, the X-vector (Snyder *et al.*, 2015) was designed based on the latest advances of deep learning research to tackle the speaker recognition task and was 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 hypotheses is that the learning of the global harmonic structure is more difficult with the output learnable STRF layer than directly with the Mel-filterbanks. The Gabor filters applied on log Mel-spectrograms are capable of capturing local harmonic

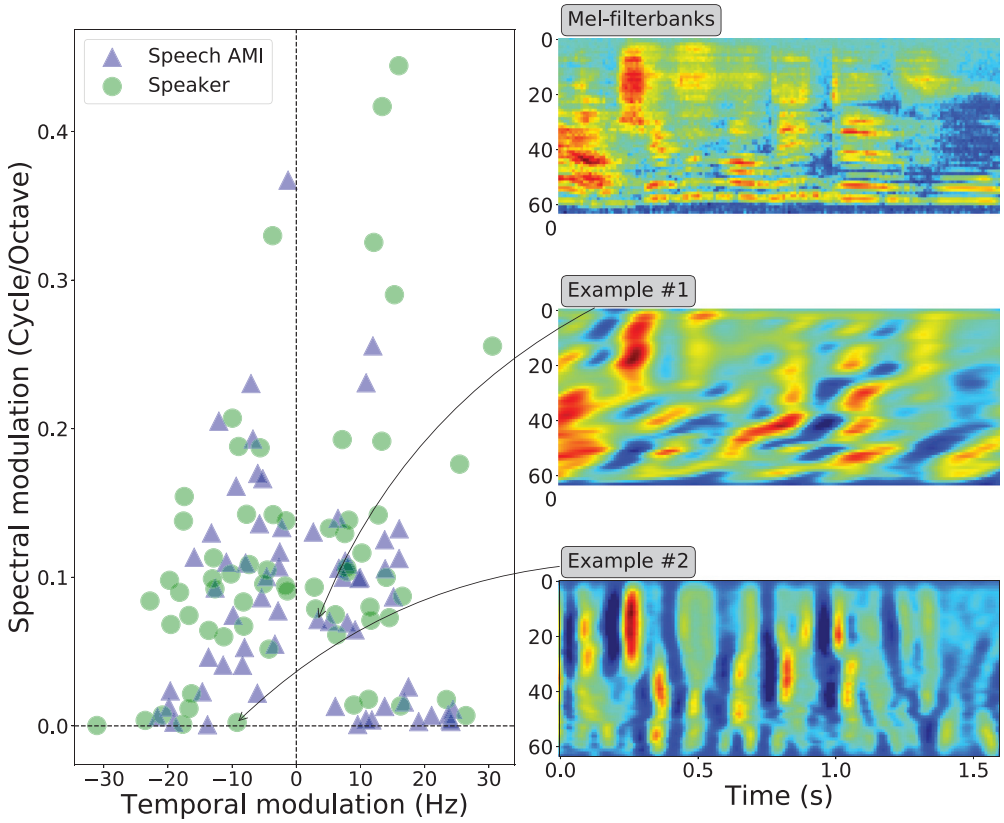

FIG. 2. (Color online) Left: Temporal and spectral modulation populations learned to tackle speech activity detection (speech AMI) on the AMI dataset and speaker verification (speaker). Top right: Mel-filterbanks representation of a sentence pronounced by a female speaker. Middle and bottom right: Output examples computed by the convolution of specific learned STRF kernels with the input Mel-filterbanks displayed in the first row.

dependencies, yet, as the receptive field along the frequency axis remained small (nine frequency bands), the Gabor filters cannot capture the global harmonic structure. The global harmonic structure is left to be learned by the rest of the models after the learnable STRFs. In contrast, specific global harmonic convolutions were introduced in Lostanlen (2017) and capture long dependencies across multiple octaves.

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 found previously by Espi *et al.* (2015), that the use of different sizes of the spectral representation increased the performance of deep learning models for acoustic event detection. This suggests that the varying sizes of the focus on the Mel-filterbank 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 I. Speech activity detection results for the different approaches described. The CL is a convolution layer reducing the size of the tensor dimension after the convolution (free 2D convolution or learnable STRFs) on the Mel-filterbanks. The free 2D convolution had the same grid size as the learnable STRFs (9×111). Each input front-end is then fed to a two-layer BiLSTM and two feed-forward layers. The best scores for each metric overall are in bold. MD, missed detection rate; FA, false alarm rate; DetER, detection error rate. For all metrics, lower is better.

	AMI database			CHIME5 database		
Input front-end	DetER	MD	FA	DetER	MD	FA
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. The CL 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 PLDA (?). The best scores for each metric overall are in bold. For the EER, lower is better.

Metric	EER	EER with S-normalization
Baseline		
I-vectors + PLDA (?)[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 Coria *et al.* (2020) and was not replicated for this study.

https://doi.org/10.1121/10.0005482

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	(Minimum–maximum)
Baseline		
SB-CNN (Salamon and Bello, 2017)[a]	79	(71–85)
Input front-end		
Free 2D conv. 3 × 3 (Kong *et al.*, 2020)	**84**	(76–93)
Learnable STRFs	82	(74–90)

[a]This result is directly extracted from Salamon and Bello (2017) and was not replicated for this study.

TABLE IV. Zebra finch call type classification results for the different approaches. The best scores for each metric overall are in **bold**. PAF, predefined acoustical features; MPS, modulation power spectrum; LDA, linear discriminant analysis. RF, random forest. The learnable STRF input front-end is combined with a simple linear model to output directly the decisions. Higher is better.

	Accuracy (%)	
	Mean	(Minimum–maximum)
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 shown 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 a linear model, such as LDA, and the MPS had the best performance overall when combined with the RF (going from 41% to 69%). The learnable STRF models were 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 a combination of features in the MPS with RF suggests 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 of co-occurrences or anti-occurrences of the spectro-temporal patterns in models as in Młynarski and McDermott (2018, 2019), since the routing in RF implies measurement of 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 Fig. 3) and the size of the Gaussian envelopes through (σ_t, σ_f) (see Fig. 6 in the Appendix). Within the space allowed by the Nyquist theorem and the size of the convolutions, all the learned STRFs concentrated in low spectral and temporal modulations (see Fig. 3). We also observed, as did Singh and Theunissen (2003), that higher spectral modulations were found at low temporal modulations (and vice versa). We found that the Gaussian envelopes of the learned

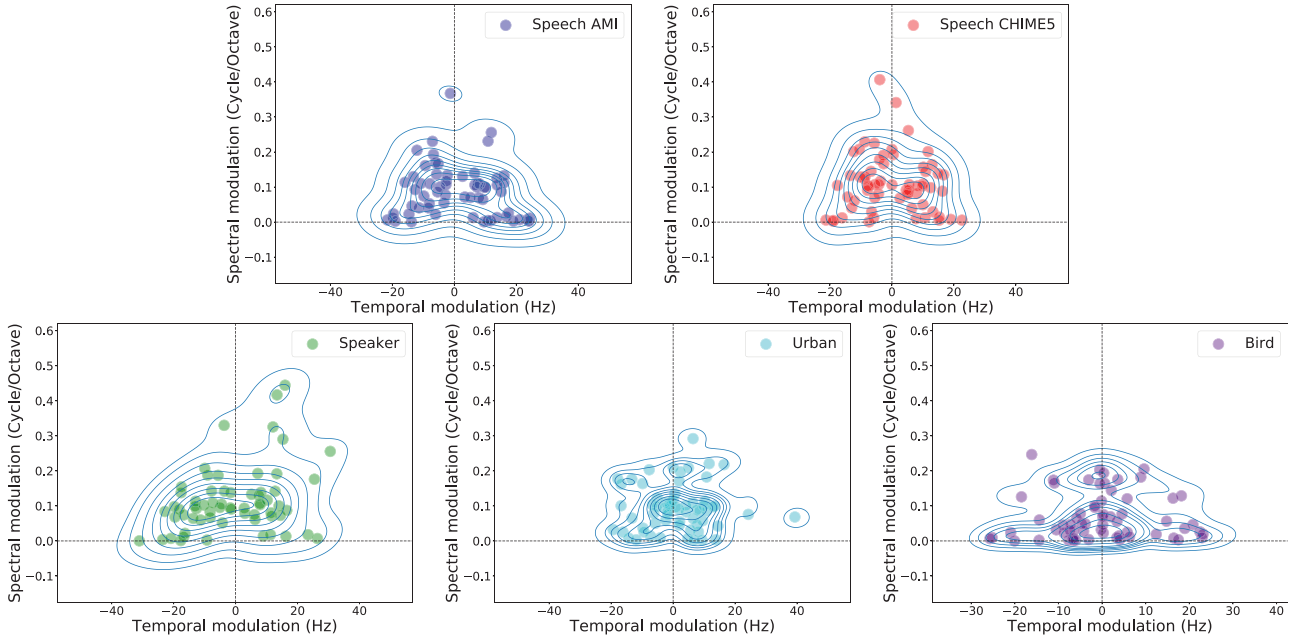

FIG. 3. (Color online) Temporal and spectral modulation of the learned STRFs to tackle speech activity detection on the AMI dataset (speech AMI) and on the CHIME5 (speech CHIME5), speaker verification on VoxCeleb (speaker), urban sound classification on Urban8k (urban), and zebra finch call type classification (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.

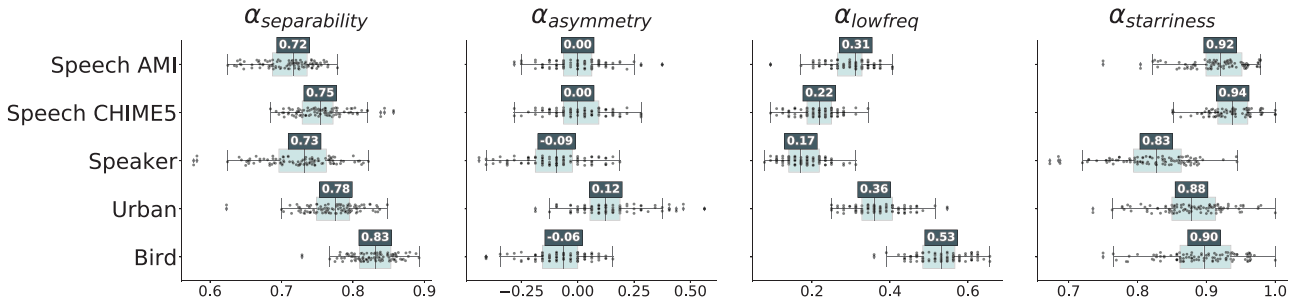

FIG. 4. (Color online) Separability, asymmetry, low-pass, starriness coefficients. Four quantifiers measured different aspects of learned distribution for the different tasks under study: speech activity detection on the CHIME5 dataset (Speech CHIME5) and on the AMI (Speech AMI), urban sound classification on Urban8k (Urban), speaker verification on VoxCeleb (Speaker), and zebra finch call type classification (Bird). We display the median value for each α and task above each box-plot.

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 STRF modulations and Gaussian envelopes of speech tasks on AMI and CHIME5 datasets and the speaker task look more similar than the bird and urban ones. We quantified the learned parameters with the distributional parameters that measured modulation (Sec. II C) and the optimal transport distance between distributions (Sec. II C 4).

First, the separability index $\alpha_{separability}$ showed that most learned STRFs are quite separable and that the tasks related to human vocalizations (speech and speaker) were less separable than the other ones (Fig. 4). The results for the separability for speech tasks are consistent with speech perception in humans, as found very recently by Flinker *et al.* (2019). There seems to be independence for the processing of spectral and temporal modulations. 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 the starriness for the ensembles of sounds of speech corpora, zebra finch 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 performance for speech recognition tasks.

In addition, the learned STRFs for the speech did not show preferences for up- or down-sweep 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 deciphering of the different factors and contributions that influenced the acoustic properties of vocalizations. Finally, the bird task focused more on the low frequency modulations ($\alpha_{lowfreq} = 0.53$) 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 ($\alpha_{lowfreq} = 0.19$). Especially, Elliott and Theunissen (2009) also found that the removal 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; Schönwiesner and Zatorre, 2009). Hullett *et al.* (2016), Massoudi *et al.* (2015), and Schönwiesner and Zatorre (2009) 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 ± 11 Hz, 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 Sec. II C 4 for the full description). We obtained the clustering tree in Fig. 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 that another 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 vocalization tasks (speech and speaker). In future work, this method could be used to discover automatically organization trees based only on acoustic properties of spectro-temporal modulations and test these predictions against what is known about the phylogeny, acoustic environment, and con-species living nearby (McCracken and Sheldon, 1997).

V. CONCLUSIONS 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 an engineering-oriented neural

https://doi.org/10.1121/10.0005482

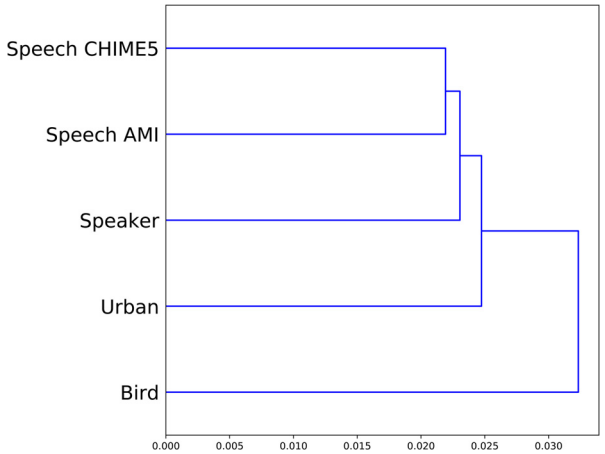

FIG. 5. (Color online) Hierarchical clustering of the tasks: speech activity detection on the CHIME5 dataset (Speech CHIME5) and on the AMI (Speech AMI), urban sound classification on Urban8k (Urban), speaker verification on VoxCeleb (Speaker), and zebra finch call type classification (Bird). The distance 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 $\lambda = 10^{-3}$).

network for speaker verification, urban sound classification, and 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 by 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 "starriness" and "separability," corresponding to the fact that filters tend to remain close to either the temporal or spectral axis, with low occupation of joint spectral and temporal responses. This is consistent with Singh and Theunissen (2003).

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 had 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 by Bellur and Elhilali (2015) has investigated the online adaptation of modulations in analytical models and witnessed several improvements in terms of engineering performance, suggesting that this is also an interesting avenue in terms of behavioral modeling. Second, the analyses from Hullett *et al.* (2016) showed that neurons not only have spectro-temporal selectivity, but they 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 map objective in addition to the task-specific loss

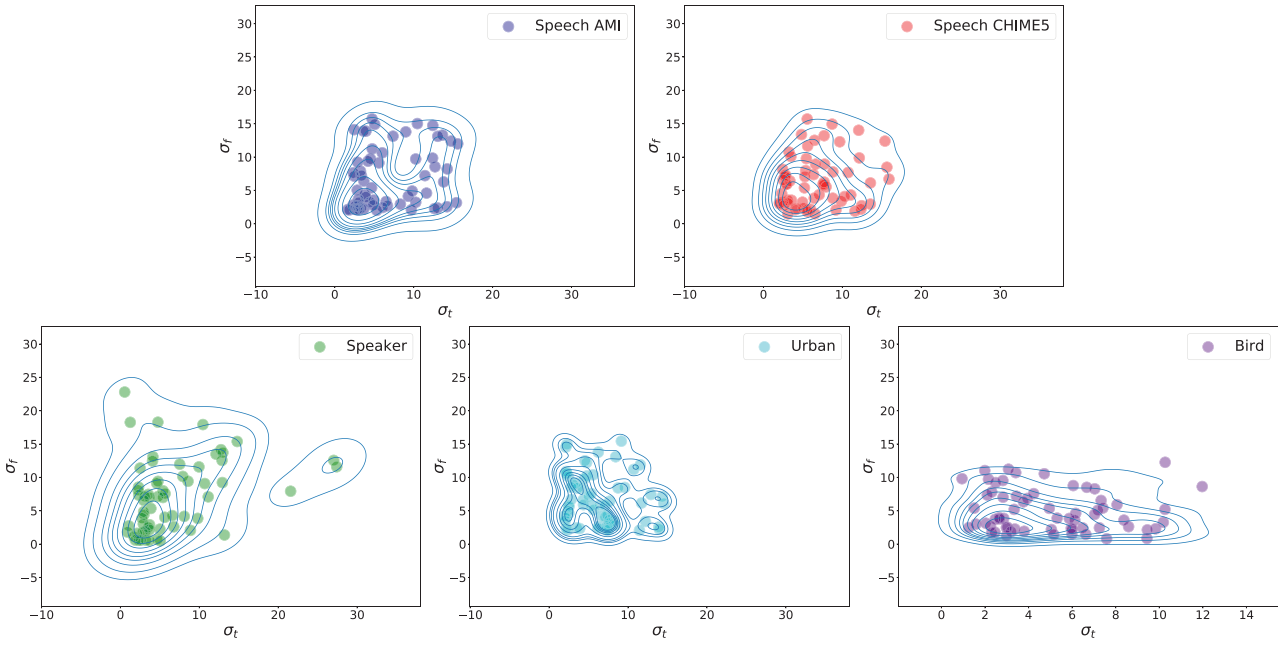

FIG. 6. (Color online) Gaussian envelopes (σ_t, σ_f) of the learned STRFs to tackle speech activity detection on the AMI dataset (Speech AMI) and on the CHIME5 (Speech CHIME5), speaker verification on VoxCeleb (Speaker), urban sound classification on Urban8k (Urban), and zebra finch call type classification (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.

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 (Grant Nos. ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, and ANR-19-P3IA-0001 PRAIRIE 3IA Institute). A.C.B.L. was funded through Neuratris, and E.D. in his Ecole des Hautes Etudes en Sciences Sociales (EHESS) role by Facebook AI Research (Research Gift) and CIFAR (Learning in Minds and Brains). The university [EHESS, CNRS, INRIA, Ecole Normale Supérieure (ENS)-Paris Sciences et Lettres] obtained the datasets reported in this paper, and the experiments were run on its computer resources.

APPENDIX: IMPORTANCE OF THE REPRESENTATION OF THE LEARNABLE STRFS

We performed an additional analysis of speech activity detection of the choice of representations **Z** from the learnable STRFs used in the subsequent neural network (Fig. 6). The performance of the real part and 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 performances obtained for each part were in the same range on the AMI dataset but were lower 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. The CL 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 two-layer BiLSTM and two feed-forward layers. The best scores for each metric overall are in bold. MD, missed detection rate. FA, false alarm rate. DetER, detection error rate. For all metrics, lower is better.

Input front-end (learnable STRFs + CL)	AMI database			CHIME5 database		
	DetER	MD	FA	DetER	MD	FA
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 (Last viewed 7/7/2021).
[2]For more information on separability, see https://bartwronski.com/2020/02/03/separate-your-filters-svd-and-low-rank-approximation-of-image-filters/ (Last viewed 7/7/2021).

Alekseev, A., and Bobe, A. (**2019**). "Gabornet: Gabor filters with learnable parameters in deep convolutional neural network," in *Proceedings of the 2019 International Conference on Engineering and Telecommunication (EnT)*, November 20–21, Dolgoprudny, Russia, pp. 1–4.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L. V., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. (**2016**). "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of the 33rd International Conference on Machine Learning*, New York, June 20–22, pp. 173–182.

Arnault, A., Hanssens, B., and Riche, N. (**2020**). "Urban sound classification: Striving towards a fair comparison," 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)*, September 2–6, Hyderabad, India.

Bellur, A., and Elhilali, M. (**2015**). "Detection of speech tokens in noise using adaptive spectrotemporal receptive fields," in *Proceedings of the 2015 49th Annual Conference on Information Sciences and Systems (CISS)*, March 18–20, Baltimore, MD, pp. 1–6.

Bredin, H. (**2017**). "pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, August 20–24, Stockholm, Sweden.

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 *Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4–8, Barcelona, Spain, pp. 7124–7128.

Chang, S.-Y., and Morgan, N. (**2014**). "Robust CNN-based speech recognition with Gabor filter kernels," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, September 8–12, San Francisco, CA.

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 **8**, 161981–162003.

Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**(2), 887–906.

Chung, J. S., Nagrani, A., and Zisserman, A. (**2018**). "Voxceleb2: Deep speaker recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, September 2–6, Hyderabad, India, pp. 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, Switzerland), pp. 137–148.

Cuturi, M. (**2013**). "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proceedings of Advances in Neural Information Processing*

Systems 26 (NIPS 2013), December 5–10, Lake Tahoe, NV, pp. 2292–2300.
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," J. Neurophysiol. **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 *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019)*, September 15–19, Graz, Austria, pp. 1378–1382.
Efron, B., and Tibshirani, R. J. (**1994**). *An Introduction to the Bootstrap* (CRC, Boca Raton, FL).
Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**(2), 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," Anim. Cogn. **19**(2), 285–315.
Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol. **5**(3), e1000302.
Espi, M., Fujimoto, M., Kinoshita, K., and Nakatani, T. (**2015**). "Exploiting spectro-temporal locality in deep learning based acoustic event detection," EURASIP J. Audio Speech Music Process. **2015**(1), 1–12.
Ezzat, T., Bouvrie, J., and Poggio, T. (**2007**). "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proceedings of the Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, August 27–31, Antwerp, Belgium.
Flamary, R., and Courty, N. (**2017**). "POT: Python optimal transport," https://pythonot.github.io/ (Last viewed 7/7/2021).
Flinker, A., Doyle, W., Mehta, A., Devinsky, O., and Poeppel, D. (**2019**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Hum. Behav. **3**(4), 393–405.
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," Sci. Rep. **8**(1), 16375.
Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (**2003**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Neurosci. **6**(11), 1216–1223.
Gabor, D. (**1946**). "Theory of communication. part 1: The analysis of information," J. Inst. Electr. Eng. Part III Radio Commun. Eng. **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," J. Neurosci. **36**(6), 2014–2026.
Imperl, B., Kačič, Z., and Horvat, B. (**1997**). "A study of harmonic features for the speaker recognition," Speech Commun. **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 Neurosci. **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," Curr. Opin. Neurobiol. **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: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 Trans. Audio Speech Lang. Process. **28**, 2880–2894.
Koumura, T., Terashima, H., and Furukawa, S. (**2019**). "Cascaded tuning to amplitude modulation for natural sound recognition," J. Neurosci. **39**(28), 5517–5533.
Lei, H., Meyer, B. T., and Mirghafori, N. (**2012**). "Spectro-temporal Gabor features for speaker recognition," in *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 25–30, Kyoto, Japan, 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 *Proceedings of the International Conference on Learning Representations*, May 6–9, New Orleans, LA.
Lostanlen, V. (**2017**). "Convolutional operators in the time-frequency domain," Ph.D. thesis, Université Paris Sciences et Lettres, Paris, France.
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., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (**2005**). "The AMI meeting corpus," in *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, August 30–September 2, Wageningen, Netherlands, pp. 137–140.
McCracken, K. G., and Sheldon, F. H. (**1997**). "Avian vocalizations and phylogenetic signal," Proc. Nat. Acad. Sci. U.S.A. **94**(8), 3833–3836.
McDermott, J. H. (**2018**). "Audition," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, Vol. 2 (Wiley, New York), pp. 1–57.
Mesgarani, N., Slaney, M., and Shamma, S. A. (**2006**). "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," IEEE Trans. Audio Speech Lang. Process. **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," Front. Syst. Neurosci. **10**, 109.
Młynarski, W., and McDermott, J. H. (**2018**). "Learning midlevel auditory codes from natural sound statistics," Neural Comput. **30**(3), 631–669.
Młynarski, W., and McDermott, J. H. (**2019**). "Ecological origins of perceptual grouping principles in the auditory system," Proc. Natl. Acad. Sci. U.S.A. **116**(50), 25355–25364.
Nagrani, A., Chung, J. S., and Zisserman, A. (**2017**). "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, pp. 2616–2620.
Ondel, L., Li, R., Sell, G., and Hermansky, H. (**2019**). "Deriving spectro-temporal properties of hearing from speech data," in *Proceedings of ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 12–17, Brighton, UK, pp. 411–415.
Peyré, G., and Cuturi, M. (**2019**). "Computational optimal transport: With applications to data science," Found. Trends Mach. Learn. **11**(5), 355–607.
Pillow, J., and Sahani, M. (**2019**). "Editorial Overview: Machine Learning, Big Data, and Neuroscience," Curr. Opin. Neurobiol. **55**, iii–iv.
Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time,'" Speech Commun. **41**(1), 245–255.
Ravanelli, M., and Bengio, Y. (**2018**). "Speaker recognition from raw waveform with sincnet," in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, December 18–21, Athens, Greece, 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 2020.11.19.389999.
Salamon, J., and Bello, J. P. (**2017**). "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Process. Lett. **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*, November 3–7, Orlando, FL, 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," Proc. Natl. Acad. Sci. U.S.A. **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," J. Acoust. Soc. Am. **131**(5), 4134–4151.
Schönwiesner, M., and Zatorre, R. (**2009**). "Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI," Proc. Natl. Acad. Sci. U.S.A. **106**(34), 14611–14616.

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

Singh, N. C., and Theunissen, F. E. (**2003**). "Modulation spectra of natural sounds and ethological theories of auditory processing," J. Acoust. Soc. Am. **114**(6), 3394–3411.

Snyder, D., Chen, G., and Povey, D. (**2015**). "Musan: A music, speech, and noise corpus," 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)*, April 15–20, Calgary, Canada, pp. 5329–5333.

Stevens, S. S., Volkmann, J., and Newman, E. B. (**1937**). "A scale for the measurement of the psychological magnitude pitch," J. Acoust. Soc. Am. **8**(3), 185–190.

Tanno, R., Arulkumaran, K., Alexander, D., Criminisi, A., and Nori, A. (**2019**). "Adaptive neural trees," in *Proceedings of the 36th International Conference on Machine Learning*, June 9–15, Long Beach, CA, pp. 6166–6175.

Theunissen, F., Sen, K., and Doupe, A. (**2000**). "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," J. Neurosci. **20**(6), 2315–2331.

Thoret, E., Andrillon, T., Léger, D., and Pressnitzer, D. (**2020**). "Probing machine-learning classifiers using noise, bubbles, and reverse correlation," bioRxiv 2020.06.22.165688.

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

Vuong, T., Xia, Y., and Stern, R. M. (**2020**). "Learnable spectro-temporal receptive fields for robust voice type discrimination," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020)*, October 25–29, Shanghai, China, pp. 1957–1961.

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," Nat. Neurosci. **8**(10), 1371–1379.

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

Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., and Dupoux, E. (**2018**). "End-to-end speech recognition from the raw waveform," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, September 2–6, Hyderabad, India, pp. 781–785.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. (**2019**). "Lookahead optimizer: k steps forward, 1 step back," in *Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, December 8–14, Vancouver, Canada, Vol. 32, pp. 9597–9608.