Fitting Auditory Filterbanks with Multiresolution Neural Networks

Vincent Lostanlen, Daniel Haider, Han Han, Mathieu Lagrange, Peter Balazs, Martin Ehler

FITTING AUDITORY FILTERBANKS WITH MULTIRESOLUTION NEURAL NETWORKS

Vincent Lostanlen[1], Daniel Haider[2,3], Han Han[1], Mathieu Lagrange[1], Peter Balazs[2], and Martin Ehler[3]

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France.
[2] Acoustics Research Institute, Austrian Academy of Sciences, A-1040 Vienna, Austria.
[3] University of Vienna, Department of Mathematics, A-1090 Vienna, Austria.

ABSTRACT

Waveform-based deep learning faces a dilemma between nonparametric and parametric approaches. On one hand, convolutional neural networks (convnets) may approximate any linear time-invariant system; yet, in practice, their frequency responses become more irregular as their receptive fields grow. On the other hand, a parametric model such as LEAF is guaranteed to yield Gabor filters, hence an optimal time–frequency localization; yet, this strong inductive bias comes at the detriment of representational capacity. In this paper, we aim to overcome this dilemma by introducing a neural audio model, named multiresolution neural network (MuReNN). The key idea behind MuReNN is to train separate convolutional operators over the octave subbands of a discrete wavelet transform (DWT). Since the scale of DWT atoms grows exponentially between octaves, the receptive fields of the subsequent learnable convolutions in MuReNN are dilated accordingly. For a given real-world dataset, we fit the magnitude response of MuReNN to that of a well-established auditory filterbank: Gammatone for speech, CQT for music, and third-octave for urban sounds, respectively. This is a form of knowledge distillation (KD), in which the filterbank "teacher" is engineered by domain knowledge while the neural network "student" is optimized from data. We compare MuReNN to the state of the art in terms of goodness of fit after KD on a hold-out set and in terms of Heisenberg time–frequency localization. Compared to convnets and Gabor convolutions, we find that MuReNN reaches state-of-the-art performance on all three optimization problems.

Index Terms— Convolutional neural network, digital filters, filterbanks, multiresolution analysis, psychoacoustics.

1. INTRODUCTION

Auditory filterbanks are time-invariant systems whose design takes inspiration from domain-specific knowledge in hearing science [1]. For example, the critical bands of the human cochlea inspires frequency scales such as mel, bark, and ERB [2]. The phenomenon of temporal masking calls for asymmetric impulse responses, motivating the design of Gammatone filters [3]. Lastly, the constant-Q transform (CQT), in which the number of filters per octave is fixed, reflects the principle of octave equivalence in music [4].

In recent years, the growing interest for deep learning in signal processing has proposed to learn filterbanks from data rather than design them a priori [5]. Such a replacement of feature engineering to feature learning is motivated by the diverse application scope of audio content analysis: i.e., conservation biology [6], urban science [7], industry [8], and healthcare [9]. Since these applications differ greatly in terms of acoustical content, the domain knowledge which prevails in speech and music processing is likely to yield suboptimal performance. Instead, gradient-based optimization has the potential to reflect the spectrotemporal characteristics of the data at hand.

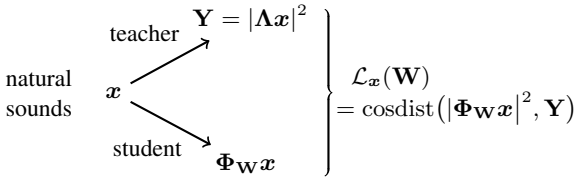

Figure 1: Graphical outline of the proposed method. We train a neural network "student" $\mathbf{\Phi_W}$ to regress the squared magnitudes $\mathbf{Y}$ of an auditory filterbank "teacher" $\mathbf{\Lambda}$ in terms of spectrogram-based cosine distance $\mathcal{L}_{\boldsymbol{x}}$, on average over a dataset of natural sounds $\boldsymbol{x}$.

Enabling this potential is particularly important in applications where psychoacoustic knowledge is lacking; e.g., animals outside of the mammalian taxon [10, 11]. Beyond its perspectives in applied science, the study of learnable filterbanks has value for fundamental research on machine listening with AI. This is because it represents the last stage of progress towards general-purpose "end-to-end" learning, from the raw audio waveform to the latent space of interest.

Yet, success stories in waveform-based deep learning for audio classification have been, up to date, surprisingly few—and even fewer beyond the realm of speech and music [12]. The core hypothesis of our paper is that this shortcoming is due to an inadequate choice of neural network architecture. Specifically, we identify a dilemma between nonparametric and parametric approaches, where the former are represented by convolutional neural networks (convnets) and the latter by architectures used in SincNet [13] or LEAF [14]. In theory, convnets may approximate any finite impulse response (FIR), given a receptive field that is wide enough; but in practice, gradient-based optimization on nonconvex objectives yields suboptimal solutions [12]. On the other hand, the parametric approaches enforce good time–frequency localization, yet at the cost of imposing a rigid shape for the learned filters: cardinal sine (inverse-square envelope) for SincNet and Gabor (Gaussian envelope) for LEAF.

Our goal is to overcome this dilemma by developing a neural audio model which is capable of learning temporal envelopes from data while guaranteeing near-optimal time–frequency localization. In doing so, we aim to bypass the explicit incorporation of psychoacoustic knowledge as much as possible. This is unlike state-of-the-art convnets for filterbank learning such as SincNet or LEAF, whose parametric kernels are initialized according to a mel-frequency scale. Arguably, such careful initialization procedures defeat the purpose of deep learning; i.e., to spare the human effort of feature engineering.

Companion website: https://github.com/lostanlen/lostanlen2023waspaa

Furthermore, it contrasts with other domains of deep learning (e.g., image processing) in which all convnet layers are simply initialized with i.i.d. Gaussian weights [15].

Prior work on this problem has focused on advancing the state of the art on a given task, sometimes to no avail [16]. In this article, we take a step back and formulate a different question: before we try to outperform an auditory filterbank, can we replicate its responses with a neural audio model? To answer this question, we compare different "student" models in terms of their ability to learn from a black-box function or "teacher" by knowledge distillation (KD).

Given an auditory filterbank $\mathbf{\Lambda}$ and a discrete-time signal $\boldsymbol{x}$ of length T, let us denote the squared magnitude of the filter response at frequency bin f by $\boldsymbol{Y}[f,t] = |\mathbf{\Lambda}\boldsymbol{x}|^2[f, 2^J t]$, where 2^J is the chosen hop size or "stride". Then, given a model $\Phi_{\mathbf{W}}$ with weights $\mathbf{W}$, we evaluate the dissimilarity between teacher $\mathbf{\Lambda}$ and student $\mathbf{\Phi}_{\mathbf{W}}$ as their (squared) spectrogram-based cosine similarity $\mathcal{L}_{\boldsymbol{x}}(\mathbf{W})$. The distance of student and teacher in this similarity measure can be computed via the L^2 distance after normalizing across frequency bins f, independently for each time t. Let $|\widetilde{\mathbf{\Phi}}_{\mathbf{W}}\boldsymbol{x}|^2$ and $\widetilde{\mathbf{Y}}$ denote these normalized versions of student and teacher, then

$$\begin{aligned}\mathcal{L}_{\boldsymbol{x}}(\mathbf{W}) &= \mathrm{cosdist}\big(|\mathbf{\Phi}_{\mathbf{W}}|^2, \mathbf{Y}\big) \\ &= \frac{1}{2}\sum_{t=1}^{T/2^J}\sum_{f=1}^{F}\big||\widetilde{\mathbf{\Phi}}_{\mathbf{W}}\boldsymbol{x}|^2[f,t] - \widetilde{\mathbf{Y}}[f,t]\big|^2, \end{aligned} \quad (1)$$

where F is the number of filters. We seek to minimize the quantity above by gradient-based optimization on $\mathbf{W}$, on a real-world dataset of audio signals $\{\boldsymbol{x_1} \dots \boldsymbol{x_N}\}$, and with no prior knowledge on $\mathbf{\Lambda}$.

2. NEURAL AUDIO MODELS

2.1. Learnable time-domain filterbanks (Conv1D)

As a baseline, we train a 1-D convnet $\mathbf{\Phi}_{\mathbf{W}}$ with F kernels of the same length $2L$. With a constant stride of 2^J, $\mathbf{\Phi}_{\mathbf{W}}\boldsymbol{x}$ writes as

$$\mathbf{\Phi}_{\mathbf{W}}\boldsymbol{x}[f,t] = (\boldsymbol{x} * \boldsymbol{\phi_f})[2^J t] = \sum_{\tau=-L}^{L-1} \boldsymbol{x}\big[2^J t - \tau\big]\boldsymbol{\phi_f}[\tau], \quad (2)$$

where $\boldsymbol{x}$ is padded by L samples at both ends. Under this setting, the trainable weights $\mathbf{W}$ are the finite impulse responses of $\boldsymbol{\phi_f}$ for all f, thus amounting to $2LF$ parameters.We initialize $\mathbf{W}$ as Gaussian i.i.d. entries with null mean and variance $1/\sqrt{F}$.

2.2. Gabor 1-D convolutions (Gabor1D)

As a representative of the state of the art (i.e., LEAF [14]), we train a Gabor filtering layer or Gabor1D for short. For this purpose, we parametrize each FIR filter $\boldsymbol{\phi_f}$ as Gabor filter; i.e., an exponential sine wave of amplitude a_f and frequency η_f which is modulated by a Gaussian envelope of width σ_f. Hence a new definition:

$$\boldsymbol{\phi_f}[\tau] = \frac{a_f}{\sqrt{2\pi}\sigma_f}\exp\left(-\frac{\tau^2}{2\sigma_f^2}\right)\exp(2\pi \mathrm{i}\eta_f\tau). \quad (3)$$

Under this setting, the trainable weights $\mathbf{W}$ amount to only $3F$ parameters: $\mathbf{W} = \{a_1, \sigma_1, \eta_1, \dots, a_F, \sigma_F, \eta_F\}$. Following LEAF, we initialize center frequencies η_f and bandwidths σ_f so as to form a mel-frequency filterbank [17] and set amplitudes a_f to one. We use the implementation of Gabor1D from SpeechBrain v0.5.14 [18].

2.3. Multiresolution neural network (MuReNN)

As our original contribution, we train a multiresolution neural network, or MuReNN for short. MuReNN comprises two stages, multiresolution approximation (MRA) and convnet; of which only the latter is learned from data. We implement the MRA with a dual-tree complex wavelet transform (DTCWT) [19]. The DTCWT relies on a multirate filterbank in which each wavelet $\boldsymbol{\psi_j}$ has a null average and a bandwidth of one octave. Denoting by ξ the sampling rate of $\boldsymbol{x}$, the wavelet $\boldsymbol{\psi_j}$ has a bandwidth with cutoff frequencies $2^{-(j+1)}\pi$ and $2^{-j}\pi$. Hence, we may subsample the result of the convolution $(\boldsymbol{x} * \boldsymbol{\psi_j})$ by a factor of 2^j, yielding:

$$\forall j \in \{0, \dots, J-1\},\ \boldsymbol{x_j}[t] = (\boldsymbol{x} * \boldsymbol{\psi_j})[2^j t], \quad (4)$$

where J is the number of multiresolution levels. We take $J = 9$ in this paper, which roughly coincides with the number of octaves in the hearing range of humans. The second stage in MuReNN consists in defining convnet filters $\boldsymbol{\phi_f}$. Unlike in the Conv1D setting, those filters do not operate over the full-resolution input $\boldsymbol{x}$ but over one of its MRA levels $\boldsymbol{x_j}$. More precisely, let us denote by $j[f]$ the decomposition level assigned to filter f, and by $2L_j$ the kernel size for that decomposition level. We convolve $\boldsymbol{x}_{j[f]}$ with $\boldsymbol{\phi_f}$ and apply a subsampling factor of $2^{J-j[f]}$, hence:

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{W}}\boldsymbol{x}[f,t] &= (\boldsymbol{x_{j[f]}} * \boldsymbol{\phi_f})[2^{J-j[f]}t] \\ &= \sum_{\tau=-L_j}^{L_j-1} \boldsymbol{x_{j[f]}}\big[2^{J-j[f]}t - \tau\big]\boldsymbol{\phi_f}[\tau] \end{aligned} \quad (5)$$

The two stages of subsampling in Equations 4 and 5 result in a uniform downsampling factor of 2^J for $\mathbf{\Phi}_{\mathbf{W}}\boldsymbol{x}$. Each learned FIR filter $\boldsymbol{\phi}_f$ has an effective receptive field size of $2^{j[f]+1}L_{j[f]}$, thanks to the subsampling operation in Equation 4. This resembles a dilated convolution [20] with a dilation factor of $2^{j[f]}$, except that the DTCWT guarantees the absence of aliasing artifacts.

Besides this gain in frugality, as measured by parameter count per unit of time, the resort to an MRA offers the opportunity to introduce desirable mathematical properties in the non-learned part of the transform (namely, $\boldsymbol{\psi_f}$) and have the MuReNN operator $\mathbf{\Phi}_{\mathbf{W}}$ inherit them, without need for a non-random initialization nor regularization during training. In particular, $\mathbf{\Phi}_{\mathbf{W}}$ has at least as many vanishing moments as $\boldsymbol{\psi}_f$. Furthermore, the DTCWT yields quasi-analytic coefficients: for each j, $\boldsymbol{x_j} = \boldsymbol{x_j^{\mathbb{R}}} + \mathrm{i}\boldsymbol{x_j^{\mathbb{I}}}$ with $\boldsymbol{x_j^{\mathbb{I}}} \approx \mathcal{H}\left(\boldsymbol{x_j^{\mathbb{R}}}\right)$, where the exponent $\mathbb{R}$ (resp. $\mathbb{I}$) denotes the real part (resp. imaginary part) and $\mathcal{H}$ denotes the Hilbert transform. Since $\boldsymbol{\phi_f}$ is real-valued, the same property holds for MuReNN: $\mathbf{\Phi}^{\mathbb{I}}\boldsymbol{x} = \mathcal{H}(\mathbf{\Phi}^{\mathbb{R}}\boldsymbol{x})$.

We implement MuReNN on GPU via a custom implementation of DTCWT in PyTorch[1]. Following [19], we use a biorthogonal wavelet for $j = 0$ and quarter-shift wavelets for $j \geq 1$. We set $L_j = 8M_j$ where M_j is the number of filters f at resolution j. We refer to [21] for an introduction to deep learning in the wavelet domain, with applications to image classification.

3. KNOWLEDGE DISTILLATION

3.1. Target auditory filterbanks

For each of the three different domains, speech, music and urban environmental sounds, we use an auditory filterbank $\mathbf{\Lambda}$ that is tailored to its respective spectrotemporal characteristics.

[1] https://github.com/kymatio/murenn

Domain	Dataset	Teacher	Conv1D	Gabor1D	MuReNN
Speech	NTVOW	Gammatone	2.12 ± 0.05	10.14 ± 0.09	$\mathbf{2.00 \pm 0.02}$
Music	TinySOL	VQT	8.76 ± 0.2	16.87 ± 0.06	$\mathbf{5.28 \pm 0.03}$
Urban	SONYC-UST	ANSI S1.11	3.26 ± 0.1	13.51 ± 0.2	$\mathbf{2.57 \pm 0.2}$
Synth	Sine waves	CQT	11.54 ± 0.5	22.26 ± 0.9	$\mathbf{9.75 \pm 0.4}$

Table 1: Mean and standard deviation of test loss after knowledge distillation over five independent trials. Each column corresponds to a different neural audio model $\mathbf{\Phi_W}$ while each row corresponds to a different auditory filterbank and audio domain. See Section 4.2 for details.

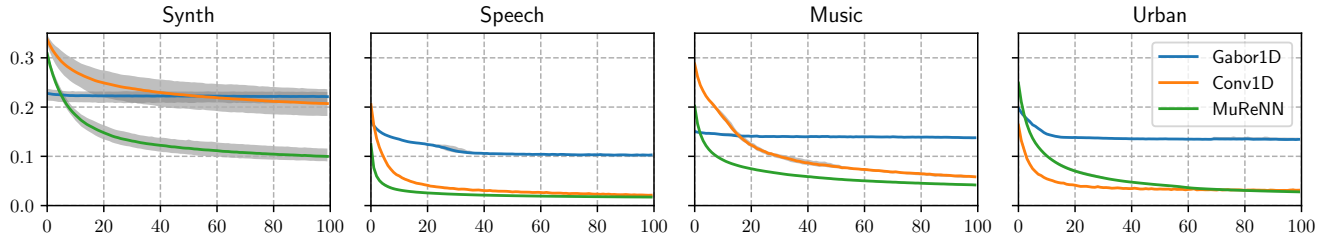

Figure 2: Left to right: evolution of validation losses on different domains with Conv1D (green), Gabor1D (blue), and MuReNN (orange), as a function of training epochs. The shaded area denotes the standard deviation across five independent trials. See Section 4.2 for details.

Synth A constant-Q filterbank with $Q = 8$ filters per octave, covering eight octaves with Hann-modulated sine waves.

Speech A filterbank with 4-th order Gammatone filters tuned to the ERB-scale, a frequency scale which is adapted to the equivalent rectangular bandwidths of the human cochlea [22]. In psychoacoustics, Gammatone filters provide a good approximation to measured responses of the filters of the human basilar membrane [3]. Unlike Gabor filters, Gammatone filters are asymmetric, both in the time domain and frequency domain.We refer to [23] for implementation details.

Music A variable-Q transform (VQT) with $M_j = 12$ frequency bins per octave at every level. The VQT is a variant of the constant-Q transform (CQT) in which Q is decreased gradually towards lower frequencies [24], hence an improved temporal resolution at the expense of frequency resolution.

Urban A third-octave filterbank inspired by the ANSI S1.11-2004 standard for environmental noise monitoring [25]. In this filterbank, center frequencies are not exactly in a geometric progression. Rather, they are aligned with integer Hertz values: 40, 50, 60; 80, 100, 120; 160, 200, 240; and so forth.

We construct the Synth teacher via nnAudio [26], a PyTorch port of librosa [27]; and Speech, Music, and Urban using the Large Time–Frequency Analysis Toolbox (LTFAT) for MATLAB [28].

3.2. Gradient-based optimization

For all four "student" models, we initialize the vector $\mathbf{W}$ at random and update it iteratively by empirical risk minimization over the training set. We rely on the Adam algorithm for stochastic optimization with default momentum parameters. Given the definition of spectrogram-based cosine distance in Equation 1, we perform reverse-mode automatic differentiation in PyTorch to obtain

$$\nabla \mathcal{L}_{\boldsymbol{x}}(\mathbf{W})[i] = \sum_{f=1}^{F} \sum_{t=1}^{T/2^J} \frac{\partial |\widetilde{\mathbf{\Phi}}_{\mathbf{W}} \boldsymbol{x}|^2[f,t]}{\partial \mathbf{W}[i]}(\mathbf{W}) \times \left(|\widetilde{\mathbf{\Phi}}_{\mathbf{W}} \boldsymbol{x}|^2[f,t] - \widetilde{\mathbf{Y}}[f,t] \right) \quad (6)$$

for each entry $\mathbf{W}[i]$. Note that the gradient above does not involve the phases of the teacher filterbank $\mathbf{\Lambda}$, only its normalized magnitude response $\mathbf{Y}$ given the input $\boldsymbol{x}$. Consequently, even though our models $\mathbf{\Phi_W}$ contain a single linear layer, the associated knowledge distillation procedure is nonconvex, and thus resembles the training of a deep neural network.

4. RESULTS AND DISCUSSION

4.1. Datasets

Synth As a proof of concept, we construct sine waves in a geometric progression over the frequency range of the target filterbank.

Speech The North Texas vowel database (NTVOW) [29] contains utterances of 12 English vowels from 50 American speakers, including children aged three to seven as well as male and female adults. In total, it consists of 3190 recordings, each lasting between one and three seconds.

Music The TinySOL dataset [30] contains isolated musical notes played by eight instruments: accordion, alto saxophone, bassoon, flute, harp, trumpet in C, and cello. For each of these instruments, we take all available pitches in the tessitura (min = B_0, median = E_4, max = $C\sharp_8$) in three levels of intensity dynamics: *pp*, *mf*, and *ff*. This results in a total of 1212 audio recordings.

Urban The SONYC Urban Sound Tagging dataset (SONYC-UST) [31] contains 2803 acoustic scenes from a network of autonomous sensors in New York City. Each of these ten-second scenes contains one or several sources of urban noise pollution, such as: engines, machinery and non-machinery impacts, powered saws, alert signals, and dog barks.

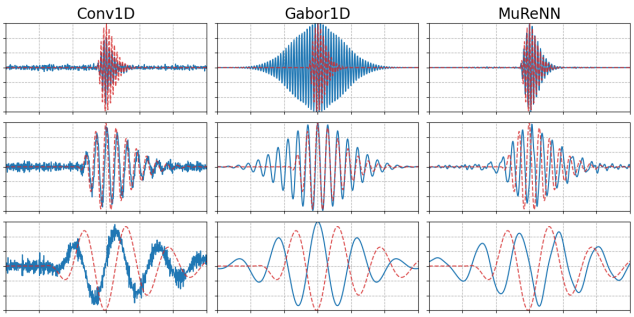

Figure 3: Compared impulse responses of Conv1D (left), Gabor1D (center), and MuReNN (right) with different center frequencies after convergence, with a Gammatone filterbank as target. Solid blue (resp. dashed red) lines denote the real part of the impulse responses of the learned filters (resp. target). See Section 4.3 for details.

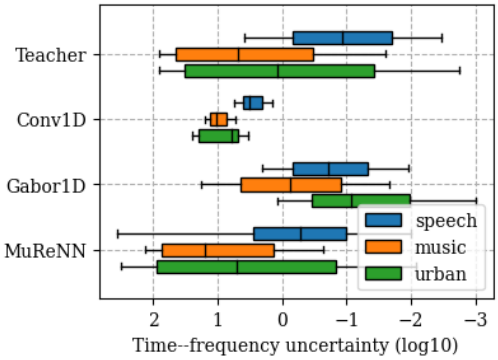

Figure 4: Distribution of Heisenberg time–frequency ratios for each teacher–student pair (lower is better). See Section 4.3 for details.

4.2. Benchmarks

For each audio domain, we randomly split its corresponding dataset into training, testing and validation subsets with a 8:1:1 ratio. During training, we select 2^{12} time samples from the middle part of each signal, i.e., the FIR length of the filters in the teacher filterbank. We train each model with 100 epochs with an epoch size of 8000.

Table 1 summarizes our findings. On all three benchmarks, we observe that MuReNN reaches state-of-the-art performance, as measured in terms of cosine distance with respect to the teacher filterbank after 100 epochs. The improvement with respect to Conv1D is most noticeable in the Synth benchmark and least noticeable in the Speech benchmark. Furthermore, Figure 2 indicates that Gabor1D barely trains at all: this observation is consistent with the sensitivity of LEAF with respect to initialization, as reported in [32]. We also notice that MuReNN trains faster than Conv1D on all benchmarks except for Urban, a phenomenon deserving further inquiry.

4.3. Error analysis

The mel-scale initialization of Gabor1D filters and the inductive bias of MuReNN enabled by octave localization gives a starting advantage when learning filterbanks on log-based frequency scales, as used for the Gammatone and VQT filterbank. Expectedly, this advantage is absent with a teacher filterbank that does not follow a geometric progression of center frequencies, as it is the case in the ANSI scale. Figure 2 reflects these observations.

To examine the individual filters of each model, we take the speech domain as an example and obtain their learned impulse responses. Figure 3 visualizes chosen examples at different frequencies learned by each model together with the corresponding teacher Gammatone filters. In general, all models are able to fit the filter responses well. However, it is noticeable that the prescribed envelope for Gabor1D impedes it from learning the asymmetric target Gammatone filters. This becomes prominent especially at high frequencies. From the strong envelope mismatches at coinciding frequency we may deduce that center frequencies and bandwidths did not play well together during training. On the contrary, MuReNN and Conv1D are flexible enough to learn asymmetric temporal envelopes without compromising its regularity in time. Although the learned filters of Conv1D are capable of fitting the frequencies well, they suffer from noisy artifacts, especially outside their essential supports. Indeed, through limiting the scale and support of the learned filters, MuReNN restrains the potential introduction of high-frequency noises of a learned filter of longer length. The phase misalignment at low frequencies is a natural consequence of the fact that the gradients are computed from the magnitudes of the filterbank responses.

Finally, we measure the time–frequency localization of all filters by computing the associated Heisenberg time–frequency ratios [33]. From theory we know that Gaussian windows are optimal in this sense [34]. Therefore, it is not surprising that Gabor1D yields the best localized filters, even outperforming the teacher, see Figure 4. Expectedly, the localization of the filters from Conv1D is poor and appears independent of the teacher. MuReNN roughly resembles the localization of the teachers but has some poorly localized outliers in higher frequencies, deserving further inquiry.

5. CONCLUSION

Multiresolution neural networks (MuReNN) have the potential to advance waveform-based deep learning. They offer a flexible and data-driven procedure for learning filters which are "wavelet-like": i.e., narrowband with compact support, vanishing moments, and quasi-Hilbert analyticity. Those experiments based on knowledge distillation from three domains (speech, music, and urban sounds) illustrate the suitability of MuReNN for real-world applications. The main limitation of MuReNN lies in the need to specify a number of filters per octave M_j, together with a kernel size L_j. Still, a promising finding of our paper is that prior knowledge on M_j and L_j suffices to finely approximate non-Gabor auditory filterbanks, such as Gammatones on an ERB scale, from a random i.i.d. Gaussian initialization. Future work will evaluate MuReNN in conjunction with a deep neural network for sample-efficient audio classification.

6. ACKNOWLEDGMENT

V.L. thanks Fergal Cotter and Nick Kingsbury for maintaining the dtcwt and pytorch_wavelets libraries; LS2N and ÖAW staff for arranging research visits; and Neil Zeghidour for helpful discussions. D.H. thanks Clara Hollomey for helping with the implementation of the filterbanks. V.L. and M.L. are supported by ANR MuReNN; D.H., by a DOC Fellowship of the Austrian Academy of Sciences (A 26355); P.B., by FWF projects LoFT (P 34624) and NoMASP (P 34922); and M.E., by WWTF project CHARMED (VRG12-009).

7. REFERENCES

[1] R. F. Lyon, *Human and machine hearing: Extracting meaning from sound*. Cambridge University Press, 2017.

[2] R.-A. Knight and J. Setter, *The Cambridge Handbook of Phonetics*. Cambridge University Press, 2021.

[3] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[4] J. C. Brown, "Calculation of a constant-Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, 1991.

[5] M. Dörfler, T. Grill, R. Bammer, and A. Flexer, "Basic filters for convolutional neural networks applied to music: Training or design?" *Neural Comput. Appl.*, vol. 32, pp. 941–954, 2020.

[6] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, 2022.

[7] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[8] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[9] P. Bizopoulos and D. Koutsouris, "Deep learning in cardiology," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 168–193, 2018.

[10] F. J. Bravo Sanchez, M. R. Hossain, N. B. English, and S. T. Moore, "Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[11] M. Faiß, "Adaptive representations of sound for automatic insect recognition," Master's thesis, Naturalis Biodiversity Center, 2022.

[12] F. Lluís, J. Pons, and X. Serra, "End-to-end music source separation: Is it possible in the waveform domain?" *arXiv preprint arXiv:1810.12187*, 2018.

[13] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE SLT*, 2018.

[14] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *Proc. ICML*, 2021.

[15] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. ICML*, 2013, pp. 1139–1147.

[16] J. Schlüter and G. Gutenbrunner, "EfficientLEAF: A faster learnable audio frontend of questionable use," in *Proc. EUSIPCO*. IEEE, 2022, pp. 205–208.

[17] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve, and E. Dupoux, "Learning filterbanks from raw speech for phone recognition," in *Proc. IEEE ICASSP*. IEEE, 2018.

[18] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[19] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Proc. Mag.*, vol. 22, no. 6, pp. 123–151, 2005.

[20] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018, pp. 3918–3926.

[21] F. Cotter, "Uses of complex wavelets in deep convolutional neural networks," Ph.D. dissertation, University of Cambridge, 2020.

[22] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, 09 1983.

[23] T. Necciari, N. Holighaus, P. Balazs, Z. Průša, P. Majdak, and O. Derrien, "Audlet filter banks: A versatile analysis/synthesis framework using auditory frequency scales," *Applied Sciences*, vol. 8, no. 1, 2018.

[24] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Proc. AES*, 2014.

[25] J. Antoni, "Orthogonal-like fractional-octave-band filters," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. 884–895, 2010.

[26] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "nnAudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks," *IEEE Access*, vol. 8, pp. 161 981–162 003, 2020.

[27] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen, *et al.*, "librosa/librosa: 0.10.0.post2," Mar. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7746972

[28] Z. Průša, P. Søndergaard, P. Balazs, and N. Holighaus, "LTFAT: A Matlab/Octave toolbox for sound processing," in *Proc. CMMR*, 2013, pp. 299–314.

[29] P. F. Assmann and W. F. Katz, "Time-varying spectral change in the vowels of children and adults," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1856–1866, 2000.

[30] C. E. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "OrchideaSOL: A dataset of extended instrumental techniques for computer-aided orchestration," in *Proc. ICMC*, 2020.

[31] M. Cartwright, A. E. Mendez Mendez, G. Dove, J. Cramer, V. Lostanlen, H.-H. Wu, J. Salamon, O. Nov, and J. P. Bello, "SONYC Urban Sound Tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network," in *Proc. DCASE*, 2019.

[32] M. Anderson, T. Kinnunen, and N. Harte, "Learnable frontends that do not learn: Quantifying sensitivity to filterbank initialisation," in *Proc. IEEE ICASSP*, 2023.

[33] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.

[34] K. Gröchenig, *Foundations of time-frequency analysis*, ser. Appl. Numer. Harmon. Anal. Boston, MA: Birkhäuser, 2001.