ESResNe(X)t-fbsp: Learning Robust Time-Frequency Transformation of Audio

Andrey Guzhov[1,2], Federico Raue[1], Jörn Hees[1], Andreas Dengel[1,2]

[1]DFKI GmbH

[2]TU Kaiserslautern

Kaiserslautern, Germany

firstname.lastname@dfki.de

Abstract—**Environmental Sound Classification (ESC) is a rapidly evolving field that recently demonstrated the advantages of application of visual domain techniques to the audio-related tasks. Previous studies indicate that the domain-specific modification of cross-domain approaches show a promise in pushing the whole area of ESC forward.**

In this paper, we present a new time-frequency transformation layer that is based on complex frequency B-spline (fbsp) wavelets. Being used with a high-performance audio classification model, the proposed fbsp-layer provides an accuracy improvement over the previously used Short-Time Fourier Transform (STFT) on standard datasets. We also investigate the influence of different pre-training strategies, including the joint use of two large-scale datasets for weight initialization: ImageNet and AudioSet. Our proposed model out-performs other approaches by achieving accuracies of 95.20 % on the ESC-50 and 89.14 % on the UrbanSound8K datasets.

Additionally, we assess the increase of model robustness against additive white Gaussian noise and reduction of an effective sample rate introduced by the proposed layer and demonstrate that the fbsp-layer improves the model's ability to withstand signal perturbations, in comparison to STFT-based training. For the sake of reproducibility, our code is made available.

Index Terms—**audio, classification, ESC, Fourier transform, fbsp-wavelet**

I. Introduction

Environmental Sound Classification (ESC) is a challenging task that implies a correct differentiation between sound classes that occur in our everyday life (e.g., "sneezing", "airplane", "jackhammer", "cat", "idling engine", "brushing teeth", "street music"). Widely used datasets, such as ESC-50 [1] and UrbanSound8K [2], provide a reliable basis to compare a variety of approaches on the ESC-task, which allowed to confirm the advantage of using cross-domain techniques [3].

Previously, the general trend in the ESC-community was to design audio-domain-specific architectures. In the last years, however, the focus has shifted to the use of common techniques from other domains, such as the visual one. Both directions are combined usually with either a raw signal or a pre-computed time-frequency transformation, which is more common. Learning of a time-frequency transformation in an end-to-end fashion is a rare exception that, however, is able to provide an increase of accuracy [4]. Also, the usage of

a weight initialization obtained on large-scale datasets is in alignment with the recent tendencies. Such a weight transfer is performed usually in either cross- or intra-domain manner only. Thus, the field of ESC lacks studies on the assessment of effects of a two-stage domain adaptation using large-scale audio datasets. Besides that, a typical accuracy evaluation of models is being accomplished in "ideal" conditions. Measuring of the influence of a perturbed signal on the predictions of best performing models is not quite usual.

In our work, we propose a new time-frequency transformation layer that adjusts its parameters to the data and is based on complex frequency B-spline wavelets (fbsp-wavelets) and contributes to the out-performance of previous models. Also, we introduce an additional pre-training step using a large-scale dataset of audio, namely AudioSet [5], and evaluate its effect on the classification accuracy for randomly and ImageNet-initialized models. Finally, we assess the dependency of prediction accuracy of our best performing models on two types of signal perturbations: additive white Gaussian noise and reduction of the effective sample rate.

The remainder of this paper is organized as follows. In Section II we discuss prior methods and approaches to Environmental Sound Classification. Then, we describe the model that includes our proposed time-frequency transformation layer based on complex frequency B-spline wavelets in Section III, its training and evaluation in Section IV and the obtained results in Section V. Finally, we summarize our work and highlight follow-up research directions in Section VI.

II. Related Work

In this section, we describe previous work done in the field of Environmental Sound Classification (ESC). We highlight approaches that were used to solve the ESC-task, in particular: application of one- and two-dimensional Convolutional Neural Networks (CNN) and the use of pre-computed and trainable transformations.

A. Raw Waveform and 1D-CNN

One-dimensional CNNs use a raw audio signal as an input and provide a more natural way to build an audio-domain-related model, in comparison to 2D-CNNs. Since data pre-processing is not needed in such cases, these models provide an out-of-the box learning of a time-frequency transformation

[6], [7]. Further enhancement of one-dimensional CNNs was performed in two directions. One was the operation on different time scales [8], while the other implied the use of an input layer initialization using gammatone filter banks [9], [10] as a starting point for the training. In this work, we follow a similar direction in application to a two-dimensional CNN, introducing the learning of a time-frequency transformation that is based on the complex fbsp-wavelet filter bank [11].

B. Time-Frequency Representation and 2D-CNN

The use of image-domain-related CNNs in conjunction with a pre-computed time-frequency representation is a more common setup to solve audio-related tasks. For the ESC-50 dataset, the baseline was set by a model that is referred to as Piczak-CNN [12]. The architecture of the Piczak-CNN followed its custom design and was combined with Mel-scaled power spectrograms [13] . Later, the follow-up models were based on the Short-Time Fourier Transform (STFT) [14] derived data representations (e.g., [15], [16], [17], [18], [19], [20]) or on sophisticatedly designed filter banks (e.g., [21]).

However, the use of log-power spectrograms without modifications allowed the ESResNet model [3] to achieve state-of-the-art accuracy using a general-purpose visual CNN, suggesting that the Short-Time Fourier Transform itself provides a good representation that can serve as an initial state for a trainable filter bank.

C. Trainable Filter Bank and 2D-CNN

Currently, in the field of Environmental Sound Classification, two-dimensional CNNs that involve a trainable time-frequency transformation represent the smallest subset of the models. This situation was caused mainly by the lack of large-scale audio datasets. However, a successful model was presented that achieved state-of-the-art performance on the ESC-50 dataset [4]. Since the AudioSet dataset [5] was released, it becomes possible to train powerful audio-domain-related neural networks from scratch, including time-frequency transformations.

Finally, we decided to combine advantages of the pre-computed STFT and the ability to fit such a transformation to the data. In details, the proposed transformation layer is described in Section III-C.

III. MODEL

In this section, we will describe the base ESResNet model [3] and the way how it processes its input, the ResNeXt architecture [22] and the proposed trainable time-frequency transformation based on the complex frequency B-spline wavelets [11].

A. ESResNet

The ESResNet model was proposed in [3] and combined commonly used visual domain techniques such as a ResNet-based [23] backbone, Siamese-like [24] multi-channel processing, and depth-wise separable convolutions [25] together with the computation of log-power spectrograms obtained
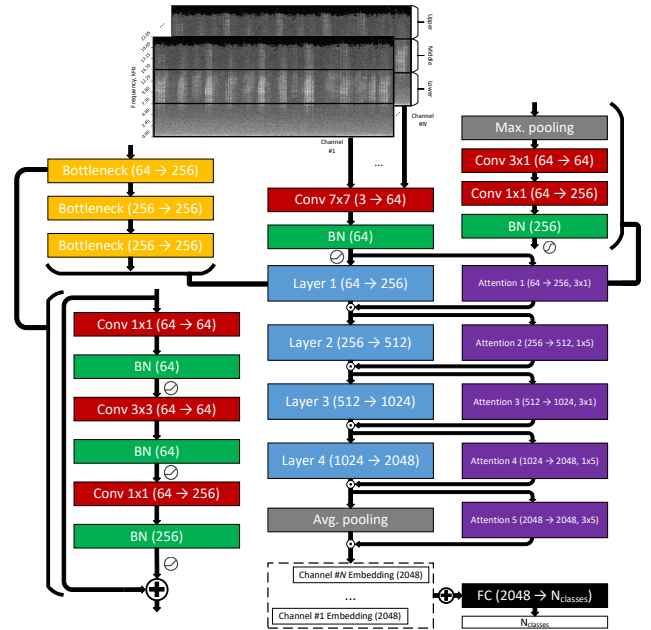

Figure 1: Overview of the ESResNet model. The main branch (central column) consists of the Convolutional layer (red) stacked together with the Batch Normalization layer (green), followed by the residual layers 1 − 4 (blue), the Average pooling (gray), and the Fully-Connected layer (black). On the left, the typical structure of a residual layer is presented. Each residual layer consists of the stack of the bottleneck layers (orange) that include Conv-BN-operations applied sequentially and the skip-connection. Rectified Linear Unit (ReLU) serves as an activation function. On the right, structure of the attention blocks is presented. The attention block (violet) is stacked in parallel to the residual layer 1 to 4 or to the average pooling layer. The attention block includes the Max-pooling operation (gray) followed by the depth-wise separable convolution stacked together with batch normalization. The output of the attention block is given by the logistic function.

using Short-Time Fourier Transform and cross-domain transfer learning in application to the ESC-task. The chosen method allowed to achieve state-of-the-art results on the ESC-50 [1] and UrbanSound8K [2] datasets at the time of publishing. An overview of the ESResNet's processing pipeline is given by Figure 1.

B. ResNeXt

The ResNeXt architecture proposed in [22] is an evolution of the original ResNet model and includes some techniques that were shown successful previously. In particular, it introduces a so-called "cardinality" of the residual layers, which refers to a number of paths in the layer. Given approximately the same number of parameters, the use of the ResNeXt architecture as a backbone for the ESResNet model provides a valuable performance improvement, as described in Section V.

C. Proposed Layer

As mentioned in Section III-A, in the ESResNet model [3], the Short-Time Fourier Transform was applied to transform the input audio to the corresponding time-frequency representation. This operation can be decomposed into two independent steps: signal framing and a subsequent Discrete Fourier Transform (DFT) of each frame. In this work, we propose an approach to train a such time-frequency transformation layer that evolves from a sub-optimal state represented by the inverse DFT to an optimal transform in a data-driven manner.

STFT belongs to the family of Fourier-related transforms and is used to determine magnitude and phase of basis sinusoidal frequencies f_c at different time points τ in a time-domain signal x using kernel function $K(n)$.

$$X(x, \tau) = \sum_{n=-\infty}^{\infty} x[n]w[n-\tau]K_{DFT}^{f_c}(n) \qquad (1)$$

1) Discrete Fourier Transform: Discrete Fourier Transform is used mainly to analyze the frequency content of discretized continuous-time signals [26]. The DFT kernel function shown in Equation 2 defines the corresponding time-frequency transformation.

$$K_{DFT}^{f_c}(n) = e^{-2i\pi f_c n} \qquad (2)$$

2) Complex Frequency B-Spline Wavelets (fbsp): The complex frequency B-spline wavelets are compactly defined in the frequency domain and described in terms of order m, bandwidth f_b and central frequency f_c [11].

$$K_{fbsp}^{m-f_b-f_c}(n) = \sqrt{f_b}\left(\frac{f_b n}{m}\right)^m e^{2i\pi f_c n} \qquad (3)$$

One can notice that the DFT-kernel defined in Equation 2 represents an inverted in time limiting case of the fbsp-kernel (Equation 3). Thus, assigning the fbsp-layer's parameters $m = 0$ and $f_b = 1$, it's initialization becomes identical to the inverse DFT, which makes it possible to start the network's training from a good enough state.

Finally, the transform performed by the proposed fbsp-layer is defined by Equation 4.

$$X(x, \tau) = \sum_{n=-\infty}^{\infty} x[n]w[n-\tau]K_{fbsp}^{m-f_b-f_c}(n) \qquad (4)$$

3) fbsp-Specific Loss Term: In order to regularize weights of the fbsp-filter bank and to preserve the overall signal's energy, we decided to introduce an additional loss term that is specific to the fbsp-layer and, for the N-point fbsp-transform, is described by Equation 5.

$$\mathcal{L}^{fbsp} = \frac{1}{N}\sum_{f_c=0}^{N}(\|K_{fbsp}^{m-f_b-f_c}\|^2 - 1)^2 \qquad (5)$$

IV. EXPERIMENTAL SETUP

In this section, we describe the datasets that were used, training of the models, including our approach to transfer from the visual to the audio domain, and the procedure of robustness' evaluation.

A. Datasets

In this work, we used three audio-domain-related datasets: AudioSet [5], ESC-50 [1] and UrbanSound8K [2]. The AudioSet dataset was used for pre-training of models, whereas the others served for the performance evaluation and assessment of model robustness against additive white Gaussian noise and reduction of an effective sample rate. Also, the ImageNet [27] dataset was used as a source of weight initialization in the cross-domain transfer learning.

1) AudioSet: AudioSet was proposed and described in [5]. It is a large-scale sound dataset that provides $\sim 1.8\,\mathrm{M}$ sound clips (each 10 s) organized into 527 classes in a multi-label manner. The amount of training data allows to use the AudioSet dataset for the initialization of deep neural networks, thus, providing a better initial state for fine-tuning on the ESC-50 and UrbanSound8K datasets.

2) ESC-50: The ESC-50 dataset consists of 2,000 monaural samples belonging to 50 classes that can be divided into 5 groups, such as *animal* sounds, *natural and water* sounds, *non-speech human* sounds, *interior* and *exterior* sounds [1]. Samples are distributed equally among classes, thus each category consists of 40 recordings. Each track has length of 5 s, the native sample rate is $44.1\mathrm{kHz}$. The dataset was divided into 5 folds by its authors that we used in current work to perform our evaluation.

3) UrbanSound8K: The US8K dataset consists of 8,732 samples (both mono and stereo) belonging to 10 classes: "air conditioner", "car horn", "children playing", "dog bark", "drilling", "engine idling", "gun shot", "jackhammer", "siren", and "street music" [2]. The classes are not balanced in terms of overall recording lengths per class. Each track has variable length up to 4 s, the native sample rate varies from $16\,\mathrm{kHz}$ to $48\,\mathrm{kHz}$. The dataset was divided into 10 folds by its authors that we used in current work to perform our evaluation.

4) ImageNet: The ImageNet dataset was proposed and described in [27]. It is a large-scale visual dataset that provides more than $1\,\mathrm{M}$ images divided into 1000 classes. As the use of the ImageNet pre-training is beneficial from the performance point of view (e.g., [3], [18]), we use ImageNet-trained networks and evaluate the influence of a such initialized models w.r.t. the follow-up training.

B. Hyper-Parameters

In the experiments, we performed training on the AudioSet, ESC-50 and UrbanSound8K datasets from scratch and after the initialization using ImageNet-weights. The training on the AudioSet dataset was used as an intermediate step, while the two later datasets served as a target for the final performance assessment.

The training on the ESC-50 and UrbanSound8K datasets was derived from the one used for the ESResNet model [3]. According to it, the model was trained for 300 epochs using the Adam optimizer [28] with the learning rate varied from $2.5e-4$ (training from scratch) to $2.5e-5$ (fine-tuning phase), the exponential decay $\gamma = 0.985$ and the *weight decay* set to $5e-4$. Other hyper-parameters such as β_1, β_2 and ϵ were set to the default values. The fbsp-variant of our model demonstrated a preference to lower learning rate values, which substantiated the decision to reduce it to $1e-5$ and set $\gamma = 0.99$ during the fine-tuning process.

The pre-training stage on the AudioSet dataset is a modification of the original fine-tuning schema, those optimizer is replaced by Stochastic Gradient Descent (SGD) [29] with Nesterov's momentum [30]. In comparison to the Adam optimizer, it introduces a smaller number of hyper-parameters, which eased finding a well working combination of them. For the AudioSet pre-training, the *weight decay* was used the same way as for the fine-tuning phase, as well as the momentum, which was equal to the β_1 parameter of the Adam optimizer. The huge amount of training samples in the AudioSet dataset allowed us to reduce the number of training epochs to 5. The choice of the learning rate was determined by its maximum value that allowed to perform the training successfully, and the value varied from $1.6e-3$ (training from scratch) to $4e-4$ (after ImageNet initialization).

For both setups, cross-entropy served as a loss function.

C. Data Augmentation

In order to prevent overfitting and improve the prediction accuracy, we utilized the following data augmentation techniques (for both the AudioSet pre-training and the ESC-50 / UrbanSound8K fine-tuning):

1) Time Scaling: This method can be considered as a combination of time stretching and pitch shift (see [15], [7]). While the first one changes the duration of an audio file keeping its spectral characteristic untouched, the later one, in opposite to time stretching, allows to manipulate spectral characteristics and preserve the duration of the track. Both methods rely on computationally expensive operations, which makes it inefficient to apply them in an on-the-fly manner. The time scaling was chosen to be applied as an augmentation step because of its computational cheapness and effectiveness [3]. In this work, the scaling factor was distributed uniformly in the range $[-1.5, 1.5]$.

2) Time Inversion: Time inversion that was applied in [7] is an effective data augmentation technique that is related to random flip of images during the training on the visual datasets. Probability of the inversion was set to 0.5.

3) Random Crop: For the scaled audio tracks, there was a requirement to align their lengths in order to process the input through the model. The use of the random cropping instead of the center one allowed to increase the diversity of the data samples even more, thus, acting as an additional augmentation step. Audio tracks were cropped to the duration of $10/5/4$ s (AudioSet / ESC-50 / UrbanSound8K,

respectively) if they were longer. Otherwise, no cropping was performed. During the evaluation phase, the random cropping was replaced by the center one.

4) Random Padding: The rationale behind random padding is the same as for the random cropping. Audio tracks were padded to the duration of $10/5/4$ s (AudioSet / ESC-50 / UrbanSound8K, respectively) if they were shorter. Otherwise, no padding was performed. During the evaluation phase, the random padding was replaced by the center one.

D. Cross- and Intra-Domain Transfer Learning

In order to evaluate the role of the AudioSet weight initialization, we performed a series of experiments that included the AudioSet dataset as a pre-training step. The ESResNet model as well as its fbsp-variant was evaluated on the ESC-50 and UrbanSound8K datasets after the training from scratch, ImageNet initialization, AudioSet pre-training from scratch or after the ImageNet weight transfer. The ESResNeXt-based models were evaluated after the ImageNet weight initialization and the two-stage transfer learning that included the AudioSet intermediate training after the ImageNet initialization.

Additionally, as we noticed that a completely random initialization of all network components could result in mode collapse due to the additional freedom introduced by the early fbsp-filter bank, in such cases we employed a late unfreeze strategy: We froze the fbsp-layer's parameters for the first three epochs. This results in the later parts of the network to be trained based on an STFT-like first layer for the first three epochs (similar to the previous [3]), before then updating all parameters in later epochs based on more meaningful gradients.

The transfer of a model to another dataset was done by replacing its last fully-connected layer (the model's linear classifier) by a randomly initialized one, which output shape suited to the task.

E. Evaluation of Robustness

In order to evaluate robustness of trained models to perturbations in the input signal, we conducted experiments that included the addition of noise to the signal and reduction of information in it.

1) Robustness Against Additive White Gaussian Noise: To assess the model's robustness to additive noise, white Gaussian noise at desired Signal-to-Noise Ratios (SNR) was generated and mixed-up to the audio tracks before performing the forward pass through the model.

2) Robustness Against Reduction of an Effective Sample Rate: The ability of model to deal with the reduced effective sample rate was tested using low-pass filtering at different cutoff frequencies.

Signal filtering implies attenuation of unwanted frequency components in the signal while preserving amplitude and phase of desired ones without changes. A low-pass filter passes frequency components that are lower than a chosen cutoff frequency. The frequency components that lie in a higher band

Table I: Evaluation Results of the ESResNe(X)t Model on STFT- and fbsp-Spectrograms on the AudioSet dataset. We can see that fbsp improves over STFT on Audioset.

Model	Input Type	ImageNet Initialized	Mean Average Precision
ESResNet	STFT		0.1892
		✓	0.2514
	fbsp		0.2394
		✓	0.2616
ESResNeXt	STFT	✓	0.2514
	fbsp	✓	0.2817

are being suppressed, and, thus, the unwanted part of the input signal is being weakened. The exact properties of a digital filter depend on its design and include but are not limited to: cutoff frequency (defines filter's passband), passband ripple, slope, width of transition band, stopband, etc.

In this work, we decided to use a 5th order Butterworth low-pass filter, as it provides maximally flat passband response [31] and a quick roll-off around its cutoff frequency. The filter was applied before feeding audio samples to the model.

V. RESULTS

A. Pre-Training on AudioSet

In Table I, we present mean Average Precision (mAP) obtained by the variants of our proposed model on the evaluation subset of the AudioSet. The results include scores of the ESResNet model after the training from scratch as well as after the ImageNet initialization for both STFT- and fbsp-based transformations. Additionally, the effect of the backbone replacement from ResNet-50 to ResNeXt-50 is evaluated for the two best performing setups, namely STFT- and fbsp-based models in conjunction with the ImageNet weight transfer. One can observe that the initialization using ImageNet weights is beneficial for the evaluated ESResNet model, as it provides an steady increase of mAP.

The STFT-based model demonstrated a low sensitivity to the replacement of the backbone from ResNet-50 to ResNeXt-50, as the corresponding mAP value does not change (0.2514). At the same time, the fbsp-layer provided a valuable increase of the mAP from 0.2616 to 0.2871.

Apart from this, we also computed the frequency responses of the trained on AudioSet fbsp-layers and compared them to the frequency response of an STFT-filter bank. The frequency response of a filter describes the dependency of the output gain on the frequencies of an input signal. The DFT-matrix of the STFT-filter bank and its frequency response is shown in Figure 2a. In the top marginal we can observe that the gain is almost flat for the entire frequency band, up to the Nyquist frequency.

In contrast, the frequency responses of the trained fbsp-filter banks (Figure 2b – Figure 2d) consist of distinguishable peaks and valleys in the frequency domain. This fact may

(a) STFT DFT-Matrix (b) ESResNet-fbsp @ AudioSet (trained from scratch)

(c) ESResNet-fbsp @ AudioSet (ImageNet initialization) (d) ESResNeXt-fbsp @ AudioSet (ImageNet initialization)
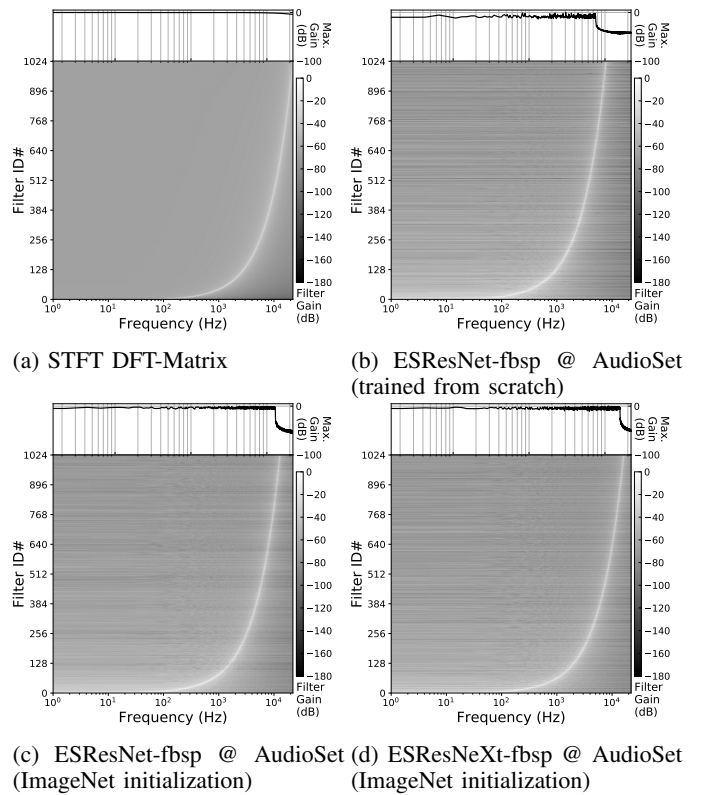
Figure 2: Frequency response of filter banks. Frequency response shows how the magnitude of each output filter (vertical axis) depends on the frequency of an input signal (horizontal axis). A log-scaled plot on top of the corresponding heat map provides an overview of the maximum gain over the frequencies. We notice that the fbsp-layers dampen the high frequencies.

indicate that the input signal is redundant w.r.t. sample rate, i.e. the networks learned some sort of signal's decimation. Also, the trained fbsp-filter banks obtain stopbands in the higher frequency band, which suggests that the filter banks are able to suppress high-frequency components of an input signal's noise.

B. Model Comparison

In this section, we will discuss the influence of the backbone and the chosen time-frequency transformation on the model's accuracy.

1) Backbone – ResNet vs. ResNeXt: The influence of the backbone on the model performance was evaluated for two training setups, both of them included the initialization using ImageNet weights (Table II). For the first one, the models were fine-tuned on the target datasets without intermediate training steps. The second one included also an AudioSet pre-training. The use of the ResNeXt instead of the ResNet model as a backbone provided a steady improvement of the prediction accuracy for both setups using STFT for the audio transformation.

Table II: Evaluation Results of the ESResNe(X)t Model on STFT- and fbsp-Spectrograms, accuracy (%). We can see that (i) fbsp in general improves over STFT, (ii) pre-training on both ImageNet and AudioSet improves results, and (iii) ESResNeXt improves over ESResNet.
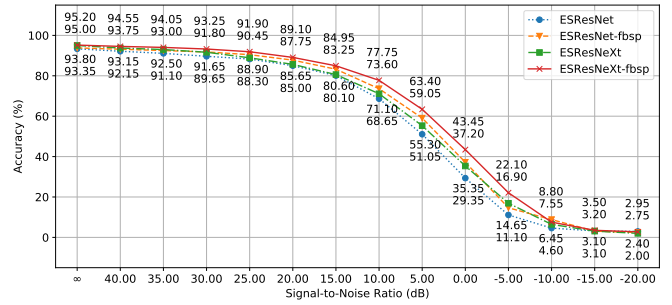
Model	Input Type	ImageNet Initialized	AudioSet Pre-Trained	ESC-50	US8K
[3]				83.15	82.76
[3]		✓		91.50	85.42
	STFT		✓	92.45	87.74
		✓	✓	93.35	88.03
ESResNet				86.25	83.20
	fbsp	✓		91.25	85.92
			✓	92.40	88.47
		✓	✓	93.80	88.38
	STFT	✓		91.60	86.02
ESResNeXt		✓	✓	95.00	89.02
	fbsp	✓		91.30	85.47
		✓	✓	**95.20**	**89.14**

Figure 3: Dependency of model performance on the addition of white Gaussian noise. We can see that fbsp-based models are more robust against such a noise.

Addition of the intermediate AudioSet training also provided a monotonic increase of the performance. At the same time, when not further pre-training the fbsp-layer with audio data, we can see a minor deterioration of the result for US8K going from ESResNet to ESResNeXt. However, given such audio-data, we see a strong performance boost of the same setup.
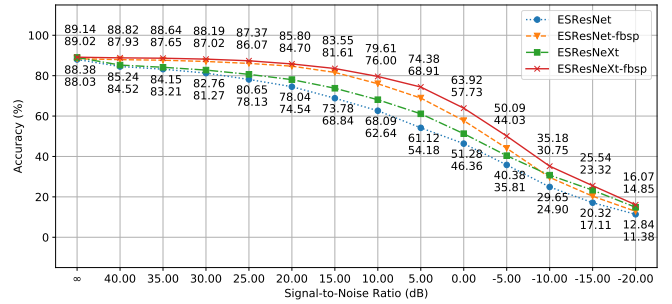
ResNeXt-based variants of the model also demonstrated higher sensitivity to the reduction of the effective sample rate, as will be detailed in Section V-C.

2) Time-Frequency Transformation – STFT vs. fbsp-wavelets: The evaluation results of the proposed fbsp-layer in comparison to the STFT demonstrate that it in general improves results (Table II). Pre-training on AudioSet as an intermediate stage is desirable for the fbsp-based models if transfer learning is performed. Absence of a such smooth transition between domains restricts the model performance on the target datasets. Thus, the best performing setup includes our proposed fbsp-layer and an intermediate pre-training on AudioSet after the ImageNet initialization.

3) Model – ESResNe(X)t-fbsp vs. Others: The proposed ESResNeXt-fbsp model achieves an outstanding accuracy on both datasets ESC-50 (95.20 %) and UrbanSound8K (89.14 %). In comparison to other approaches, it does not require the use of meta-learning [17] or ensembling [18] techniques in order to perform the best. Moreover, the proposed model provides the highest single-model accuracy among the models that were fine-tuned on both target datasets (Table III). Unlike many other models, our proposed fbsp-layer provides also insights on the desired by models representation of an input signal.
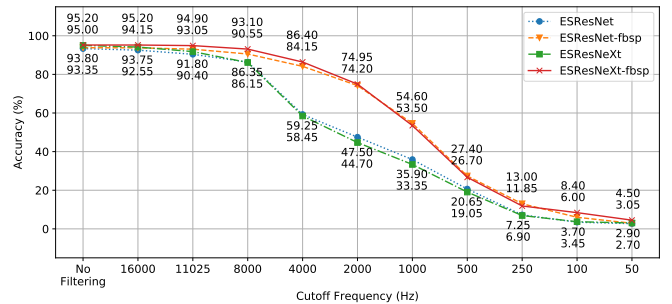

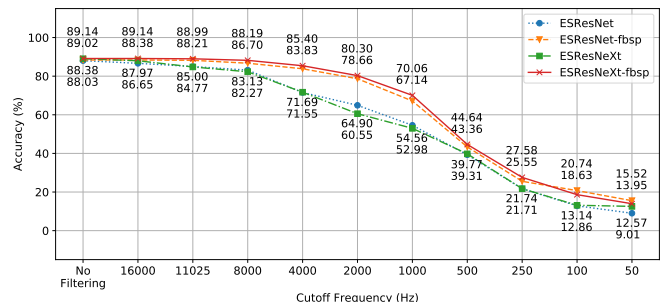Figure 4: Dependency of model performance on the effective sample rate of the input signal. We can see that fbsp-based models are more robust against lower frequency cutoffs.

Table III: Evaluation Results, accuracy (%)

Model	Source	Representation	ImageNet Initialized	AudioSet Pre-Trained	ESC-50	US8K
Human (2015)	[1]	–	–	–	81.30	–
Raw waveform and 1D-CNN						
EnvNet (2017)	[6]	raw			74.10	71.10
EnvNet v2 (2017)	[7]	raw			84.70	78.30
Multiresolution 1D-CNN (2018)	[8]	raw			75.10	–
Time-frequency representation and 2D-CNN						
Piczak-CNN (2015)	[12]	Mel-spec			64.50	73.70
SB-CNN (2017)	[15]	Mel-spec			–	79.00
Piczak-CNN (2017)	[21]	(PE)FBE			84.15	–
VGG-like CNN + mix-up (2018)	[19]	Mel-, GT-spec			83.90	83.70
VGG-like CNN + Bi-GRU + attention (2019)	[20]	GT-spec			86.50	–
CNN10 (2020)	[16]	Mel-spec		✓	90.00	86.10
WEANET N^0 (2020)	[17]	Mel-spec		✓	92.60	–
WEANET N^4 (2020)	[17]	Mel-spec		✓	94.10	–
DenseNet-201 (2020)	[18]	Mel-spec			72.50	76.32
DenseNet-201 (2020)	[18]	Mel-spec	✓		91.16	85.14
DenseNet-201 × 5, ensemble (2020)	[18]	Mel-spec	✓		92.89	87.42
Trainable filter bank and 2D-CNN						
Piczak-CNN + ConvRBM (2017)	[4]	FBE			86.50	–
Time-frequency representation and 2D-CNN						
ESResNet		STFT-spec	✓	✓	93.35	88.03
ESResNeXt		STFT-spec	✓	✓	95.00	89.02
Learnable filterbank and 2D-CNN						
ESResNet-fbsp		fbsp-spec	✓	✓	93.80	88.38
ESResNeXt-fbsp		fbsp-spec	✓	✓	**95.20**	**89.14**

Abbreviations: FBE: FilterBank Energies [4];   spec: spectrogram;   GT: GammaTone [9];   (PE)FBE: (Phase-Encoded) FBE [21]; STFT: Short-Time Fourier Transform;   fbsp: (complex) Frequency B-SPline (wavelets).

C. Model Robustness

Apart from the model comparison, we also evaluated the robustness against two types of signal perturbations: additive white Gaussian noise and reduction of an effective sample rate.

1) Additive White Gaussian Noise: The evaluation of model performance on the ESC-50 and UrbanSound8K datasets given different values of SNR shows clearly (Figure 3) that the use of the fbsp-layer improves the model's robustness to the presence of additive white Gaussian noise. As indicated in Figure 2b – Figure 2d, the higher frequency band is being suppressed by trained fbsp-layers, in comparison to the DFT-filter bank. This allows to reduce the amount of the added noise partially, thus, improving the signal-to-noise ratio for the obtained spectrograms. As shown in Figure 3, fbsp-equipped models provide higher classification accuracy given decreasing SNR on both datasets, ESC-50 and UrbanSound8K.

2) Reduction of an Effective Sample Rate: In order to quantify the influence of the effective sample rate reduction on the model accuracy, several experiments were performed. The obtained results support the point that fbsp-equipped models are able to extract the information that is relevant for classification more effectively. In particular, the significant performance drop at the cutoff frequency 4 kHz occurred in the case of the STFT-based models on both datasets (Figure 4) indicates

that the use of the fbsp-equipped models is beneficial for the applications that imply the use of low-bandwidth channels. Also, the ResNet backbone demonstrated less sensitivity to lower frequency cutoffs, in comparison to ResNeXt (Figure 4).

VI. CONCLUSION

In this paper, we proposed a new fbsp-layer that is based on complex frequency B-spline wavelets and tailored towards effective and robust time-frequency representation.

Based on the fbsp-layer, and a common ResNeXt, our ESResNeXt-fbsp model achieves new state-of-the-art results on two datasets: ESC-50 (95.20 %) and UrbanSound8K (89.14 %). To ease reproducibility, detailed code and settings of our approach are published[1].

We also evaluated the influence of the additional pre-training on the AudioSet dataset, which is beneficial for the model performance, as well as the improvement obtained by the change of the model's backbone from ResNet-50 to ResNeXt-50.

Further, we found that the proposed fbsp-layer allows to obtain ESC models that are more robust against additive white Gaussian noise and a possible reduction of the sample

[1]https://github.com/AndreyGuzhov/ESResNeXt-fbsp

rate, in comparison to models that were trained using STFT-based spectrograms as an input. Additionally, the frequency responses of the trained fbsp-filter banks provide insights into the importance of specific frequencies for the audio classification, making it possible to understand the models' predictions and behavior better.

In the future, we would like to further investigate the influence of the internal cardinality of the fbsp-layer, as an increased number of internal parameters could potentially further improve model performance. Also, changing the current split of input spectrograms according to RGB-channels to a full-frame representation could influence positively on the prediction accuracy, so we would like to quantify its effect.

REFERENCES

[1] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1015–1018. [Online]. Available: https://doi.org/10.1145/2733373.2806390

[2] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1041–1044. [Online]. Available: https://doi.org/10.1145/2647868.2655045

[3] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Esresnet: Environmental sound classification based on visual domain models," in *25th International Conference on Pattern Recognition (ICPR)*, January 2021, pp. 4933–4940.

[4] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification." in *INTERSPEECH*, 2017, pp. 3107–3111.

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[6] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2721–2725.

[7] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," 2017. [Online]. Available: https://arxiv.org/abs/1711.10282

[8] B. Zhu, K. Xu, D. Wang, L. Zhang, B. Li, and Y. Peng, "Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 528–537.

[9] M. Slaney *et al.*, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[10] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[11] A. Teolis and J. J. Benedetto, *Computational signal processing with wavelets*. Springer, 1998, vol. 182.

[12] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2015, pp. 1–6.

[13] J. Volkmann, S. S. Stevens, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 208–208, 1937. [Online]. Available: https://doi.org/10.1121/1.1901999

[14] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, June 1977.

[15] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[16] A. Arnault, B. Hanssens, and N. Riche, "Urban sound classification : striving towards a fair comparison," 2020.

[17] A. Kumar and V. Ithapu, "A sequential self teaching approach for improving generalization in sound event recognition," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5447–5457.

[18] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking cnn models for audio classification," 2020.

[19] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 356–367.

[20] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Learning attentive representations for environmental sound classification," *IEEE Access*, vol. 7, pp. 130 327–130 339, 2019.

[21] R. N. Tak, D. M. Agrawal, and H. A. Patil, "Novel phase encoded mel filterbank energies for environmental sound classification," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2017, pp. 317–325.

[22] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2017.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[26] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[29] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM journal on control and optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[30] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k^ 2) o (1/k2)," in *Sov. Math. Dokl*, vol. 27, no. 2, 1983.

[31] S. Butterworth *et al.*, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.