Trainable wavelet-like transform for feature extraction to audio classification

I D Ilyashenko, R S Nasretdinov, Y A Filin and A A Lependin

Altay State University, 61, Lenina ave., Barnaul, 656049, Russia

E-mail: andrey.lependin@gmail.com

Abstract. A method of learning optimal orthonormal filters for feature extraction from 1-D signal based on learning wavelet-like transform was proposed. Filters had been learned by using backpropagation simultaneously with neural network, which was used as a classifier. Orthonormality of filters during the learning process was provided by several quadratic regularization terms that follow from the orthogonality of the scaling functions. The proposed method was evaluated on the environmental sound classification task. We used the trainable wavelet-like transform and wavelet transform with different bases as feature extraction methods with fixed architecture of the neural network. The proposed method obtained the best results. The spectrum characteristics of learned filters of wavelet-like transform were compared with the corresponding characteristics of reverse biorhogonal wavelet basis rbior1.5 that obtained the closest accuracy results.

1. Introduction
Feature extraction is one of the most important phases of signal processing. It should extract relevant information about signals. Traditionally transforms with fixed parameters of some basis are used in this phase. Fast Fourier and wavelet transforms are examples of such feature extraction methods.

Neural networks are effectively used for classification of signals using extracted feature vectors. Its training involves learning parameters of the network by the backpropagation method. This becomes possible due to the fact that each element of the neural network is a differentiable transformation. Therefore its output’s gradient, which is used for tuning the parameters, can be estimated. Neural nets can effectively adapt to the structure of training data and to requirements of a task determined by a cost function. Feature extraction is traditionally the only element that doesn’t change during the learning process.

Different adaptive feature extraction methods have recently been presented in a great number of papers. In [1] authors presented SincNet architecture to speaker identification and authentication, where the selection of certain frequency bands in the signal was made using trained filters. In [2] the neural net with mel-cepstral-like adaptive filters was used for speech recognition and reached state-of-art results on Wall Street journal corpus. In [3-6] wavelet filters were learned using various additional restrictions.

This paper proposes a new approach to one-dimensional signal classification and processing based on feature extraction layer that consists of differentiable transformations. This layer can be trained by backpropagation similar to other neural net components. The proposed transform is a generalization of the discrete wavelet transform. It has a small number of parameters that represents coefficients of two filters: high- and low-pass that adapt during the training process.
2. Discrete wavelet transform

A scheme of the discrete wavelet transform calculation [7] is presented in figure 1a. Coefficients $cD_i$ and $cA_i$ are i-th level coefficients of details and approximations. They are a result of applying a high- and low-pass filters to approximation coefficients of a previous (i-1)-th level respectively.

![Diagram of discrete wavelet transform](image)

Figure 1. Discrete wavelet transform: a) as an application of a set of low-pass and high-pass filters; b) as an application of filters details and approximation.

From a computational point of view, coefficients of detail and approximation can be computed not only hierarchically (figure 1a), but also by applying a set of pre-computed filters (figure 1b). Detail filter $D_i$ and approximation filter $A_i$ are applied to the signal for extracting detail coefficients $cD_i$ and approximation coefficients $cA_i$ respectively. A first level detail and approximation filters match up with high-pass $H$ and low-pass $L$ filters respectively. Filters of next levels can be computed iteratively: previous level filter $A_i$ is convolved with stretched by $2^i$ times filters $L$ and $H$.

Low- and high-pass filters are represented by sets of coefficients $L = \{l_k, k = 1, ..., N\}$ and $H = \{h_k, k = 1, ..., N\}$, where $N$ is a filter size. Additional conditions [6] $Z_1 - Z_5$ are imposed on filters coefficients.

Condition $Z_1$ follows from the orthonormality of wavelet scaling functions [7]:

$$Z_1(H) = \sum_k h_k h_{k+2m} = \delta_{m,0}. \quad (1)$$

Restriction of the filter’s energy is a particular case of equation (1):

$$\sum_k h_k^2 = 1. \quad (2)$$

Condition $Z_2$ also follows from the orthonormality of wavelet scaling functions [7]:

$$Z_2(L) = \sum_k l_k l_{k+2m} = \delta_{m,0}. \quad (3)$$

Condition $Z_3$ ensures that the area of $H$ is zero:

$$Z_3(H) = \sum_k h_k = 0. \quad (4)$$

Condition $Z_4$ follows from the wavelet scaling function definition [6]:

$$Z_4(L) = \sum_k l_k = \sqrt{2}. \quad (5)$$

Condition $Z_5$ follows from the orthogonality of $H$ and $L$:

$$Z_5(H, L) = \sum_k h_k h_{k+2m} = 0. \quad (6)$$

Coefficients of $L$ and $H$ can be computed by:

$$h_k = (-1)^k * l_{N-1-k}. \quad (7)$$
l_k = (-1)^{k+1} * h_{N-1-k}, \quad (8)
for \( k = 1, \ldots, N \). In this case condition \( Z_5 \) is met.

3. Trainable wavelet-like transform layer

The trainable parameters in the one-dimensional wavelet-like transform layer are low- and high-pass filters represented as two \( N \)-dimensional sets of coefficients \( L = \{ l_k \} \) and \( H = \{ h_k \} \) respectively. Coefficients \( M \) of detail filters \( D_i \), where \( i = 1 \ldots M \) and one approximation filter \( A_M \) were calculated using these values for given number layers \( M \) of decomposition. The raw waveforms to be classified were convolved with computed filters to obtain a set of details and approximation coefficients. For an input signal of \( S \) output matrix was \( M + 1 \times S \).

The gradient propagated through the neural network and further through the coefficients of the wavelet-like filters during training process. Restrictions on filter coefficients \( Z_1 - Z_5 \) were imposed by introducing a regularization term of the form:

\[
R(H, L) = R_1(H) + R_2(L) + R_3(H) + R_4(L) + R_5(L, H),
\]

where \( R_i, \ i = 1,5 \) were formed by L2-method. Every condition:

\[
Z_i(H, L) = b_i.
\]

(10)
corresponded to:

\[
R_i(H, L) = [Z_i(H, L) - b_i]^2.
\]

(11)
Therefore \( R_i \) were given by:

\[
R_1(H) = \left( \sum_m \sum_k h_k h_{k+2m} - \delta_{m,0} \right)^2,
\]

(12)
\[
R_2(L) = \left( \sum_m \sum_k l_k l_{k+2m} - \delta_{m,0} \right)^2,
\]

(13)
\[
R_3(H) = \left( \sum_{k=1}^N h_k \right)^2,
\]

(14)
\[
R_4(L) = \left( \sum_{k=1}^N l_k - 2 \sqrt{2} \sum_{m=1}^L \right)^2,
\]

(15)
\[
R_5(L, H) = \left( \sum_m \sum_k l_k h_{k+2m} \right)^2.
\]

(16)

The objective function was a sum of a data loss function \( E(y, p) \) and the regularization term \( R(H, L) \) for the wavelet-like transform parameters:

\[
C(y, p, H, L) = E(y, p) + \lambda R(H, L),
\]

(17)
where \( \lambda \) is a constant that defines regularization impact, \( y \) is a class number, \( p \) is a predicted class probability.

Initialization of high-pass filter coefficients \( H \) was made by Xavier uniform method [8]. Sum of these coefficients had to be equal to zero as in equation (4). Mean subtraction operation was applied to enforce this restriction.

\[
h_k = h_k - \frac{1}{N} \sum_{i=1}^N h_i, \text{ for } k = 1, \ldots, N.
\]

(18)

Energy limitation (equation (2)) of initialized filters also needs to be enforced to ensure the stability of calculations. Thus initial scaling was made for high-pass filter coefficients:

\[
h_k = h_k / \sqrt{\sum_{i=1}^N h_i^2}, \text{ for } k = 1, \ldots, N.
\]

(19)

Initial approximation of filter low-pass filter coefficients \( L \) was computed from high-pass filter coefficients \( H \) according to the equation (8). The resulting filter was orthogonal to \( H \) and preserved the normality condition.
4. Experiment

The method was tested on a “toy” sound classification task (figure 2). The raw input waveform was processed by the feature extraction layer. The features were fed to a convolution neural network that can predict the class number.

![Figure 2. Classification scheme.](image)

We used wavelet transform with different fixed bases and our learnable wavelet-like layer as a feature extraction method. Architecture of the classification neural network was fixed (table 1). Output shape of the features was $B \times N_l + 1 \times S_l$, where $B$ is a batch size. Gradients of the wavelet-like layer were calculated by backpropagation simultaneously with the neural network. Additional restrictions were provided by regularization.

| Layer         | Parameters                               |
|---------------|------------------------------------------|
| 1-D conv      | Input shape 6, number of filters 15, stride 3 |
| Max pool      | Kernel size 2                             |
| 1-D conv      | Input shape 15, number of filters 20, stride 3 |
| Maxpool       | Kernel size 2                             |
| 1-D conv      | Input shape 15, number of filters 20, stride 3 |
| Average pool  | -                                        |
| Fullyconnected| Input shape 25, output shape 120         |
| ReLU[9]       | -                                        |
| Fullyconnected| Input shape 120, output shape 50         |

Table 1. Architecture of the classification neural network.

We used python 3.6 as a programming language and PyTorch as a deep learning library. We trained the neural net using stochastic gradient descent with momentum. Training parameters are presented in table 2. We used the cross-entropy loss as a data loss function. Our model was trained on an Nvidia GTX 1080 TI GPU.

| Parameter                  | Value                          |
|----------------------------|--------------------------------|
| Learning rate scheduler    | 0.4-0.0079                     |
| Momentum                   | ReduceLROnPlateau (step 0.7)   |
| $B$                        | 250                            |
| $N_l$                      | 5                              |

Table 2. Training hyperparameters.

We trained out model on the ESC-50 dataset [10]. It consists of 50 classes of different sound noises like door knock, bark, applause, etc. The dataset contained 2000 audio recording (with 40 samples per class) of 5 seconds each. We used accuracy for testing the proposed method since the dataset is balanced:

$$\text{Accuracy} = \frac{T}{T+F},$$

where $T$ and $F$ are the numbers of true and false classified samples, respectively.

5. Results

Classification results for the proposed feature extraction method are presented in table 3. Trainable wavelet-like transform shows the best results in comparison with non-adaptive feature extraction.
methods with fixed classification neural net architecture. Reverse biorthogonal wavelet filters rbior1.5 has the closest results to our method.

**Table 3.** Classification results for different feature extraction methods.

| Feature extraction method                          | Accuracy |
|---------------------------------------------------|----------|
| **Trainable wavelet-like transform**              | **0.99** |
| Wavelet decomposition based on db4                | 0.95     |
| Wavelet decomposition based on sym4               | 0.93     |
| Wavelet decomposition based on coif1              | 0.97     |
| Wavelet decomposition based on bior1.5            | 0.82     |
| Wavelet decomposition based on rbior1.5           | 0.98     |

We built spectral characteristics of learned details and approximation filters of wavelet-like transform (figure 3). The dashed line represents the spectral characteristics of the best wavelet-basis rbior1.5 for respective decomposition layers. The wavelet decomposition at each level extracts a certain narrow frequency band. This band shifts to the low-frequency region with increasing of the level of decomposition. As can be seen from the figure, the learned wavelet-like transform extracted substantially different details from the signal. A wide frequency band at each level of decomposition was used. Main information about the signal contained in low-, high- and relatively narrow middle-pass frequency region judging by amplitude-frequency power characteristic of 4-5 decomposition levels.

![Graph](a.png)

**Figure 3.** Spectral characteristics of the trainable wavelet-like transformation filter (solid line) and wavelet transform with the rbior1.5 basis (dashed line). Frequencies are normalized by $f_d/2$, where $f_d$ is the signal’s sampling rate. a) - e) – detail filters of 1-5 levels, f) – approximation filter of 5 level.

6. **Conclusions**

This paper presents a new method for extracting significant informative features of one-dimensional signals that can be obtained using the trainable wavelet-like feature extraction layer. We also provided an additional regularization function that restricts its parameters. The proposed approach was compared with other feature extraction methods based on fixed wavelet bases. Our trainable wavelet-like transform shows best performance on the “toy” sound classification task. This suggests its potential appliance in the 1-D sound processing methods design.
References

[1] Revanelli M and Bengio Y 2018 Preprint arXiv:1808.00158 [eess.AS]
[2] Zeghidour N, Xu Q, Liptchinsky V, Usunier N, Synnaeve G and Collobert K 2018 Preprint arXiv:1812.06864 [cs.CL]
[3] Hurley N, Rickard S, Curran P and Drakakis K 2007 15th International Conference on Digital Signal Processing (Cardiff: Cardiff School of Engineering) pp 631–4
[4] Zhuang Y and Baras J S 1996 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing Conf. Proc. (Atlanta, GA, USA) vol 4 pp 2351–4
[5] Zhuang Y and Baras J S 1994 Proc. of SPIE (Wavelet Applications) 2242 200–11
[6] Sogaard A 2017 Preprint arXiv:1706.03041 [cs.NE]
[7] Lee Fugal D 2009 Conceptual Wavelets in digital signal processing (San Diego: Space & Signals Technologies) p 22
[8] Glorot X and Bengio Y 2010 Proc. of the 13th Int. Conf. on Artificial Intelligence and Statistics (AISTATS) pp 249–56
[9] Nair V and Hinton G E 2010 Proc. of the 27th Int. Conf. on Machine Learning ed J Furnkranz and T Joachims pp 807–15
[10] Karol J 2015 Proc. of the 23rd ACM Int. Conf. on Multimedia (New York: ACM Publications) pp 1015–8