Benchmarking Functional Link Expansions for Audio Classification Tasks

Simone Scardapane, Danilo Comminiello, Michele Scarpiniti, Raffaele Parisi and Aurelio Uncini

Abstract Functional Link Artificial Neural Networks (FLANNs) have been extensively used for tasks of audio and speech classification, due to their combination of universal approximation capabilities and fast training. The performance of a FLANN, however, is known to be dependent on the specific functional link (FL) expansion that is used. In this paper, we provide an extensive benchmark of multiple FL expansions on several audio classification problems, including speech discrimination, genre classification, and artist recognition. Our experimental results show that a random-vector expansion is well suited for classification tasks, achieving the best accuracy in two out of three tasks.

Keywords Functional links · Audio classification · Speech recognition

1 Introduction

Music information retrieval (MIR) aims at efficiently retrieving songs of interest from a large database, based on the user's requirements [5]. One of the most important tasks in MIR is automatic music classification (AMC), i.e. the capability of automatically assigning one or more labels of interest to a song, depending on its audio characteristics. Examples of labels are the genre, artist, or the perceived mood.

S. Scardapane (✉) · D. Comminiello · M. Scarpiniti · R. Parisi · A. Uncini
Department of Information Engineering, Electronics and Telecommunications (DIET), "Sapienza" University of Rome, Via Eudossiana 18, 00184 Rome, Italy
e-mail: simone.scardapane@uniroma1.it

D. Comminiello
e-mail: danilo.comminiello@uniroma1.it

M. Scarpiniti
e-mail: michele.scarpiniti@uniroma1.it

R. Parisi
e-mail: raffaele.parisi@uniroma1.it

A. Uncini
e-mail: aurel@ieee.org

S. Bassis et al. (eds.), *Advances in Neural Networks*, Smart Innovation, Systems and Technologies 54, DOI 10.1007/978-3-319-33747-0_13

Clearly, being able to classify a song in a sufficiently large number of categories is extremely helpful in answering a specific user's query.

The problem of AMC can be divided in two components, namely, the choice of a suitable musical feature extraction procedure, and of an appropriate classifier. This latter choice is worsened by the intrinsic difficulties associated to AMC, among which we can list a generally large number of features (to provide a comprehensive overview of the audio content of a signal), an equally large number of class labels, and the necessary subjectivity in assigning such labels to each song. In the literature, classical choices for music classification include support vector machines (SVM) [5], gaussian mixture models (GMM) [16], and multilayer perceptrons (MLP) [11].

A particularly promising line of research stems for the use of functional-link (FL) networks [8]. These are two-layered networks, where the first layer, known as the expansion block, is composed of a given number of *fixed* non-linearities [2, 3, 8]. Due to this, the overall learning problem of an FL network can be cast as a standard linear least-square problem, which can be solved efficiently even in the case of very large datasets. FL-like networks have been shown to achieve comparable results to standard MLP, while requiring a smaller computational time [11, 13, 15]. It is known, however, that the performance of an FL network is dependent on the choice of the expansion block [8]. This in turn depends on the specific application and on the audio signals involved in the processing. Generally speaking, basis functions must satisfy universal approximation constraints and may be a subset of orthogonal polynomials, such as Chebyshev, Legendre and trigonometric polynomials, or just approximating functions, such as sigmoid and Gaussian functions. In this last case, the parameters of the approximating functions are generally assigned stochastically from a predefined probability distribution, and equivalent models have been popularized under the name of extreme learning machine (ELM) [12].

In this paper, we investigate the problem of choosing a suitable expansion block in the case of music classification. To this end, we compare four standard non-linear expansions when applying FL network to three music classification tasks, including genre and artist classification, and music/speech discrimination. This extends and complements our previous work, in which we performed a similar analysis in the case of audio quality enhancement subject to non-linear distortions of a reference signal [1]. Differently from the results analyzed in [1], our experiments show that, in the case of audio classification, random vector expansions provide the best performance, obtaining the lowest classification error in two out of three tasks, and a comparable performance in the last case.

The rest of the paper is organized as follows. Section 2 details the basic mathematical formulation of FL networks. Section 3 presents the 4 functional expansions considered in this work. Then, Sect. 4 introduces the datasets that are used for comparison, together with the basic parameter optimization procedure for the networks. Results are discussed in Sect. 5, while Sect. 6 concludes the paper.

2 Functional Link Neural Networks

Given an input vector $\mathbf{x} \in \mathbb{R}^d$, the output of an FL network is computed as:

$$f(\mathbf{x}) = \sum_{i=1}^{B} \beta_i h_i(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{h}(\mathbf{x}), \tag{1}$$

where each $h_i(\cdot)$ is a fixed non-linear term, denoted as functional-link or base, $\boldsymbol{\beta} = \left[\beta_1, \ldots, \beta_B\right]^T$, while the overall vector $\mathbf{h}(\cdot)$ is denoted as expansion block. Possible choices for $\mathbf{h}(\cdot)$ are detailed in the subsequent section. In a music classification task, the input $\mathbf{x}$ is a suitable representation of a music signal, which may include descriptions of its temporal behavior, frequency and cepstral terms, or meta-information deriving from a user's labeling [5]. We are given a dataset of N pairs song/class for training, denoted as $T = \{\left(\mathbf{x}_1, y_1\right), \ldots, \left(\mathbf{x}_L, y_N\right)\}$. Let:

$$\mathbf{H} = \begin{bmatrix} h_1(\mathbf{x}_1) & \cdots & h_B(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ h_1(\mathbf{x}_N) & \cdots & h_B(\mathbf{x}_N) \end{bmatrix} \tag{2}$$

and $\mathbf{y} = \left[y_1, y_2, \ldots, y_N\right]^{\mathrm{T}}$ be the hidden matrix and the output vector respectively. The optimal weights $\boldsymbol{\beta}^*$ of the FL are obtained as the solution of the overdetermined regularized least-squares problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{H}\boldsymbol{\beta} - \mathbf{y}\|^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2, \tag{3}$$

where $\lambda > 0$ is a regularization factor. Solution to Eq. (3) can be expressed in closed form as:

$$\boldsymbol{\beta}^* = (\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda \mathbf{I})^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{y}. \tag{4}$$

The previous discussion extends straightforwardly to the case of multiple outputs [14]. This is necessary for classification with K classes, where we adopt the standard 1-of-K encoding, associating a K-dimensional binary vector $\mathbf{y}_i$ to each pattern, where $y_{ij} = 1$ if the pattern i is of class j, 0 otherwise. In this case, the predicted class can be extracted from the K-dimensional FL output $\mathbf{f}(\mathbf{x}_i)$ as:

$$\mathrm{g}(\mathbf{x}_i) = \underset{i \in \{1,\ldots,K\}}{\arg\max}\, \mathbf{f}(\mathbf{x}_i). \tag{5}$$

3 Functional Link Expansions

In this section we introduce the most commonly used functional link expansions, that we subsequently compare in our experiments. The first three are deterministic expansions computed from each element of the input, while the fourth is composed of stochastic sigmoid expansions.

3.1 Chebyshev Polynomial Expansion

The Chebyshev polynomial expansion for a single feature x_j of the pattern $\mathbf{x}$ is computed recursively as [8]:

$$h_k\left(x_j\right) = 2x_j h_{k-1}\left(x_j\right) - h_{k-2}\left(x_j\right) , \tag{6}$$

for $k = 0, \ldots, P-1$, where P is the *expansion order*. The overall expansion block is then obtained by concatenating the expansions for each element of the input vector. In (6), initial values (i.e., for $k = 0$) are:

$$\begin{aligned} h_{-1}\left(x_j\right) &= x_j , \\ h_{-2}\left(x_j\right) &= 1. \end{aligned} \tag{7}$$

The Chebyshev expansion is based on a power series expansion, which is able to efficiently approximate a nonlinear function with a very small error near the point of expansion. However, far from it, the error may increases rapidly. With reference to different power series of the same degree, Chebyshev polynomials are computationally cheap and more efficient although, when the power series converges slowly, the computational cost dramatically increases.

3.2 Legendre Polynomial Expansion

The Legendre polynomial expansion is defined for a single feature x_j as:

$$h_k\left(x_j\right) = \frac{1}{k}\left\{(2k-1)\, x_j h_{k-1}\left(x_j\right) - (k-1)\, h_{k-2}\left(x_j\right)\right\} \tag{8}$$

for $k = 0, \ldots, P-1$. Initial values in Eq. (8) are set as (7). In the case of signal processing, the Legendre functional links provide computational advantage with respect to the Chebyshev polynomial expansion, while promising better performance [9].

3.3 Trigonometric Series Expansion

Trigonometric polynomials provide the best compact representation of any nonlinear function in the mean square sense [10]. Additionally, they are computationally cheaper than power series-based polynomials. The trigonometric basis expansion is given by:

$$h_k(x_j) = \begin{cases} \sin\left(p\pi x_j\right), & k = 2p-2 \\ \cos\left(p\pi x_j\right), & k = 2p-1 \end{cases}, \tag{9}$$

where $k = 0, \ldots, B$ is the functional link index and $p = 1, \ldots, P$ is the expansion index, being P the expansion order. Note that the expansion order for the trigonometric series is different from the order of both Chebyshev and Legendre polynomials. Cross-products between elements of the pattern $\mathbf{x}$ can also be considered, as detailed in [3].

3.4 Random Vector Expansion

The fourth functional expansion type that we consider is the random vector (RV) functional link [6, 8]. The RV expansion is parametric with respect to a set of internal weights, that are stochastically assigned. A RV functional link (with sigmoid nonlinearity) is given by:

$$h_k(\mathbf{x}) = \frac{1}{1 + e^{(-\mathbf{a}\mathbf{x}+b)}}, \tag{10}$$

where the parameters $\mathbf{a}$ and b are randomly assigned at the beginning of the learning process. It is worth noting that the sigmoid function is just one of the possible choices to apply a nonlinearity to the vector $\mathbf{x}$. Unlike the previous expansion types, the RV expansion does not involve any expansion order. Additionally, the overall number B of functional links is a free parameter in this case, while in the previous expansions it depends on the expansion order. Convergence properties of the RVFL model are analyzed in [6].

4 Experimental Setup

We tested the FL expansions described in the previous section on three standard audio classification benchmarks, whose characteristics are briefly summarized in Table 1. To compute the average misclassification error, we perform a 3-fold cross-validation on the available data. We optimize the models by performing a grid search procedure, using an inner 3-fold cross-validation on the training data to compute the validation performance. In particular, we search the following intervals:

Table 1 General description of the datasets

Dataset name	Features	Instances	Task	Classes	Reference
Garageband	49	1856	Genre recognition	9	[7]
Artist20	30	1413	Artist recognition	20	[4]
GTZAN	13	120	Speech/Music discrimination	2	[16]

- The exponential interval $2^i, i \in \{-5, -4, \ldots, 4, 5\}$ for the regularization coefficient λ in Eq. (4).
- The set $\{1, 2, \ldots, 8, 9\}$ for the expansion order P in Eqs. (6)-(8)-(9).
- The set $\{50, 100, \ldots, 300\}$ for the expansion block size L when using the random-vector expansion.

Additionally, we extract the parameters $\mathbf{a}$ and b from the uniform probability distribution in $[-1, +1]$. In all cases, input features were normalized between -1 and $+1$ before the experiments. Finally, we repeat the 3-fold cross-validation 10 times to average out statistical effects due to the initialization of the partition and the random-vector expansion. The code is implemented using MATLAB R2013a on an Intel Core2 Duo E7300, @2.66 GHz and 2 GB of RAM.

5 Results and Discussion

The average misclassification error and training time are reported in Table 2, together with one standard deviation. Best results in each dataset are highlighted in boldface. Out of three datasets, the Legendre expansion obtains a lower error in the

Table 2 Final misclassification error and training time for the four functional expansions, together with standard deviation

Dataset	Algorithm	Error	Time [s]
Garageband	TRI-FL	0.415 ± 0.013	0.156 ± 0.030
	CHEB-FL	0.407 ± 0.0126	$\mathbf{0.055 \pm 0.001}$
	LEG-FL	$\mathbf{0.404 \pm 0.0140}$	0.072 ± 0.026
	RV-FL	0.411 ± 0.017	0.090 ± 0.015
Artist20	TRI-FL	0.410 ± 0.016	0.084 ± 0.013
	CHEB-FL	0.401 ± 0.021	$\mathbf{0.037 \pm 0.001}$
	LEG-FL	0.442 ± 0.020	0.040 ± 0.001
	RV-FL	$\mathbf{0.375 \pm 0.018}$	0.070 ± 0.014
GTZAN	TRI-FL	0.316 ± 0.073	$\mathbf{0.001 \pm 0.002}$
	CHEB-FL	0.317 ± 0.066	0.004 ± 0.001
	LEG-FL	0.334 ± 0.071	0.004 ± 0.001
	RV-FL	$\mathbf{0.222 \pm 0.062}$	0.005 ± 0.002

Best results are shown in boldface

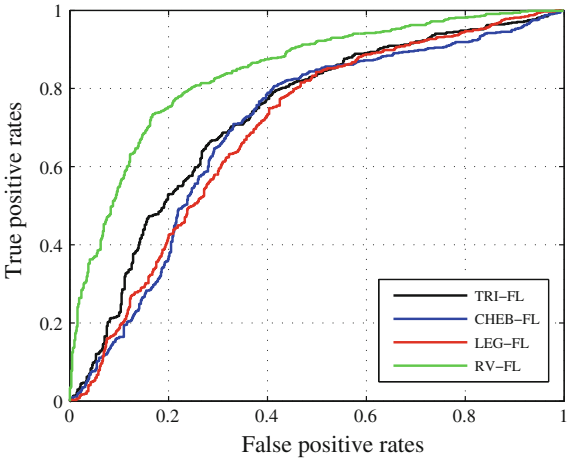

Fig. 1 ROC curve for the GTZAN dataset

Garageband dataset, while the random-vector expansion has the the lowest error in the remaining two datasets. It is interesting to note, however, that in the first case the differences between the four algorithms are small (roughly 1 % of error), while in the second and third case RV-FL strongly outperforms the other two, with a 2.5 % and 9 % decrease in error respectively, with respect to the second best. Additionally, despite its stochastic nature, the variance of RV-FL is always comparable to the other three models. To comment more on this discrepancy, we show in Fig. 1 the ROC curve for the GTZAN dataset, which is the only binary classification dataset in our experiments. It can be seen that RV-FL (shown with a green line) strongly dominates the other three curves, showing the superior performance of the random-vector expansion in this case.

We show in Table 3 the optimal parameters found by the grid search procedure. As a rule of thumb, we can see that for large datasets (Garageband and Artist20), the Legendre expansion requires a larger expansion order with respect to the trigonometric and Chebyshev expansions. Similarly, the random vector FL requires a large expansion block (around 250 hidden nodes), while for smaller datasets (the GTZAN), the optimal expansion decreases to around 100 hidden nodes. The trigonometric and Chebyshev expansions also requires a large regularization coefficient, which can be kept much smaller for the random-vector expansion (in the larger datasets), and for the Legendre one.

With respect to training time, all four expansions are performing comparably. In particular, the gap between the slowest algorithm and the fastest one never exceeded more than 0.1 seconds. Generally speaking, the trigonometric expansion is the most expensive one to compute, except for datasets with a low number of features, such as GTZAN. In the other cases, the Chebyshev expansion is the fastest one, followed closely by the Legendre expansion, with the random-vector in the middle.

Based on this analysis, we can conclude that, for audio classification tasks, the random expansion seems to outperform other common choices, such as the trigono-

Table 3 Optimal parameters found by the grid search procedure, averaged across the runs

Dataset	Algorithm	Reg. Coeff. λ	Exp. ord. p / B
Garageband	TRI-FL	15.52	1.53
	CHEB-FL	12.14	3
	LEG-FL	9.48	4.27
	RV-FL	0.75	263.3
Artist20	TRI-FL	8.92	1
	CHEB-FL	6.89	3
	LEG-FL	0.30	3.4
	RV-FL	0.33	270
GTZAN	TRI-FL	16.55	3.8
	CHEB-FL	17.93	3.07
	LEG-FL	1.18	7.4
	RV-FL	10.13	113.3

The value in the fourth column is the expansion order p for TRI-FL, CHEB-FL and LEG-FL, and the hidden layer's size B for RV-FL

metric one. This is an interesting result, since it shows a difference between the performance of FL networks with random vector expansions in dynamic audio modeling tasks [1], and static classification tasks. We hypothesize this is due to its higher capacity of extracting non-linear features from the original signal.

6 Conclusions

FL networks are common learning models when considering audio classification tasks. In this paper, we presented an analysis of several functional expansion blocks, considering three different tasks, including genre and artist recognition. Our experimental results suggest that the random vector expansion outperforms other common choices, while requiring a comparable training time.

References

1. Comminiello, D., Scardapane, S., Scarpiniti, M., Parisi, R., Uncini, A.: Functional link expansions for nonlinear modeling of audio and speech signals. In: Proceedings of the International Joint Conference on Neural Networks (2015)
2. Comminiello, D., Scardapane, S., Scarpiniti, M., Parisi, R., Uncini, A.: Online selection of functional links for nonlinear system identification. In: Smart Innovation, Systems and Technologies, Springer International Publishing AG, **37**, pp. 39–47 (2015)
3. Comminiello, D., Scarpiniti, M., Azpicueta-Ruiz, L.A., Arenas-García, J., Uncini, A.: Functional link adaptive filters for nonlinear acoustic echo cancellation. IEEE Trans. Acoust. Speech Signal Process. **21**(7), 1502–1512 (2013)

4. Ellis, D.P.W.: Classifying music audio with timbral and chroma features. In: Proceedings of the 8th International Conference on Music Information Retrieval, pp. 339–340. Austrian Computer Society (2007)
5. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Trans. Multimedia **13**(2), 303–319 (2011)
6. Igelnik, B., Pao, Y.H.: Stochastic choice of basis functions in adaptive function approximation and the functional-link net. IEEE Trans. Neural Netw. **6**(6), 1320–1329 (1995)
7. Mierswa, I., Morik, K.: Automatic feature extraction for classifying audio data. Mach. Learn. **58**(2–3), 127–149 (2005)
8. Pao, Y.H.: Adaptive Pattern Recognition and Neural Networks. Addison-Wesley, Reading, MA (1989)
9. Patra, J.C., Chin, W.C., Meher, P.K., Chakraborty, G.: Legendre-FLANN-based nonlinear channel equalization in wireless communication system. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), Singapore, pp. 1826–1831, Oct 2008
10. Patra, J.C., Pal, R.N., Chatterji, B.N., Panda, G.: Identification of nonlinear dynamic systems using functional link artificial neural networks. IEEE Trans. Syst. Man Cybern. Part B **29**(2), 254–262 (1999)
11. Scardapane, S., Comminiello, D., Scarpiniti, M., Uncini, A.: Music classification using extreme learning machines. In: 8th International Symposium on Image and Signal Processing and Analysis (ISPA), Trieste, Italy, pp. 377–381, Sep 2013
12. Scardapane, S., Comminiello, D., Scarpiniti, M., Uncini, A.: Online sequential extreme learning machine with kernels. IEEE Trans. Neural Netw. Learn. Syst. **26**(9), 2214–2220 (2015). doi:10.1109/TNNLS.2014.2382094
13. Scardapane, S., Fierimonte, R., Wang, D., Panella, M., Uncini, A.: Distributed music classification using random vector functional-link nets. In: Proceedings of the International Joint Conference on Neural Networks (2015)
14. Scardapane, S., Wang, D., Panella, M., Uncini, A.: Distributed learning for random vector functional-link networks. Inf. Sci. **301**, 271–284 (2015)
15. Turnbull, D., Elkan, C.: Fast recognition of musical genres using RBF networks. IEEE Trans. Knowl. Data Eng. **17**(4), 580–584 (2005)
16. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002)