# DISCRIMINATIVELY LEARNED FILTER BANK FOR ACOUSTIC FEATURES

*Takumi Kobayashi* *Jiaxing Ye*

National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki, Japan

## ABSTRACT

Filter banks on a frequency domain are widely applied and studied mainly for MFCC and its variant methods on speech recognition tasks. In recent years, other types of acoustic features which are derived from image classification literature have attracted attentions for the tasks regarding environmental sounds. For those features, the filter banks can also be employed mainly to effectively reduce feature dimensionality along the frequency. The filter banks have been designed according to human auditory process and are not necessarily optimal from the viewpoint of distinguishing actual acoustic data. We propose a method to build a filter bank from scratch in a data-driven manner based on the natural properties of filter banks without parametrically modeling them, which thereby more flexibly describes intrinsic characteristics of data. Those filters are optimized by incorporating discriminative criterion so as to provide effective features of high performance even with the smaller-sized filter bank. In the experiments on acoustic scene classification, the proposed method exhibits favorable performance especially on lower dimensional features.

***Index Terms***— Filter bank, discriminative learning, acoustic feature, acoustic scene classification

## 1. INTRODUCTION

Audio-based classification has been applied to various types of application, not limited to human speech recognition [1], but including scene classification [2] and event recognition embedded even in consumer electronic devices [3]. In such classification, auditory features play a key role for improving performance. While in speech recognition mel-frequency cepstral coefficients (MFCC) and its variants [4, 5] have been applied, other types of features which are derived from image classification literature have been gaining keen attention in environmental sound classification [2, 6]. Those types of features are also extracted with preserving frequency information and thus require a filter bank on a frequency domain[1] as in MFCC for adequately reducing feature dimensionality so as to enable computationally efficient classification even in low-end devices. They, however, employ simple filter banks such as a mel-filter bank.

Especially in the framework of MFCC feature extraction, some filter banks are heuristically designed based on the results of psychoacoustic studies by mimicking the human auditory process [1]. On the other hand, there are some works to optimize the filter banks from the viewpoint of classification accuracy, *i.e.*, minimizing classification error, based on actual data [3, 7, 8, 9], which are close to the purpose of this paper. Those works aim to improve discriminative power of MFCC features by parametrically modeling the filters, *e.g.*, regarding control points of filter weights [7], center positions of unimodal filters with fixing weights [3] and vice versa [8]. Although they achieve favorable performance, such parametric model of filters which resembles existing filter banks limits its ability to describe statistical and discriminative characteristics intrinsically contained in data. Without loss of generality, we can naturally assume the following properties on a filter bank:

1. Non-negativity. Filters are composed of non-negative weights.
2. Locality. Support region of the filter weights, namely region of positive weights, is well compact to enable functionality of frequency selection.
3. Disjointness (or less overlapping). Different filters are disjoint, or less overlapped, to each other for effectively and/or selectively encoding the frequency information.
4. Smoothness. Filter weights are smoothly continuous without a lot of fluctuations.

Compared to conventional dimensionality reduction methods such as principal component analysis (PCA) and Fisher discriminant analysis (FDA) [9], the filter bank which works on frequency axis is useful not only for dimensionality reduction but also for facilitating further analysis of acoustic data in terms of frequency; for example, we can analyze which frequency bands are useful for classification.

In this paper, we propose a novel method to build a filter bank from scratch for discriminatively working on the recently proposed acoustic features, such as histogram of oriented gradients (HOG) on spectrogram [2]. For constructing a filter bank on a frequency domain, our method imposes on filter weights the constraint and regularization according to the above-mentioned properties that filter banks inherently pos-

[1]Throughout this paper, we consider filter weights along the *frequency* axis in a *spectrogram* as Mel-filter bank does in MFCC.

sess, instead of parametrically modeling them. We also consider a discriminative criterion for optimizing the filter bank to enhance discriminative power of the resultant features, improving performance in the lower dimensional features.

## 2. PROPOSED METHOD

For efficiently building a filter bank from scratch, we rewrite the above-mentioned requirements (properties) of filters into three conditions, 1) non-negativity, 2) *(near) orthogonality* and 3) smoothness. The combination of non-negativity and orthogonality results in *disjoint* filters since $\sum_f W_{fi}W_{fj} = 0 \wedge W_{fi} \geq 0 \wedge W_{jf} \geq 0 \Leftrightarrow W_{fi} = 0 \vee W_{fj} = 0$ where $W_{fi}$ and $W_{fj}$ are weights for the $f$-th frequency component in the $i$-th and $j$-th filters, respectively; thereby, *near* orthogonality similarly induces *less-overlapping* filters which are not strictly disjoint. In addition, on the orthogonal non-negative weights, smoothness favors *localized* (compactly supported) distribution of weights rather than scattered one; the locality of filter weights is further ensured by the post-processing described in Sec. 2.3.

We formulate the optimization problem by regarding the first two conditions as constraints and the last one as regularization; namely, we consider a problem with *non-negative* and *orthogonal* constraints as well as *smoothness* regularization. In the optimization, a discriminative criterion is introduced according to FDA [10].

### 2.1. Discriminative optimization problem

The typical choice for enhancing discriminativity is to employ Fisher discriminant criterion as a cost function. In this study, however, according to the method [11] which constructs data-driven co-occurrence features from support vector machine (SVM) classifier weights, we optimize a filter bank based on the discriminative projection produced by FDA [10], instead of directly optimizing the Fisher criterion[2].

Given total covariance matrix $\boldsymbol{\Sigma}_T \in \mathbb{R}^{m\times m}$ and between-class one $\boldsymbol{\Sigma}_B \in \mathbb{R}^{m\times m}$ for $m$-dimensional features (corresponding to $m$ frequency components) in $C$ classes, the Fisher discriminant projection $\boldsymbol{P} \in \mathbb{R}^{m\times C-1}$ is obtained by maximizing the cost function,

$$\max_{\boldsymbol{P}} \text{trace}\left\{(\boldsymbol{P}^\top \boldsymbol{\Sigma}_T \boldsymbol{P})^{-1}(\boldsymbol{P}^\top \boldsymbol{\Sigma}_B \boldsymbol{P})\right\}, \tag{1}$$

of which the optimizer is analytically obtained as the first $C-1$ eigenvectors of the generalized eigenvalue problem,

$$\boldsymbol{\Sigma}_B \boldsymbol{P} = \boldsymbol{\Sigma}_T \boldsymbol{P} \boldsymbol{\Lambda}, \tag{2}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. Since any invertible matrix can be applied to $\boldsymbol{P}$ without changing the cost (1), the essential projection is represented as the subspace

[2]Fisher criterion is formulated in a complex form (Rayleigh quotient) due to which we can not empirically obtain favorable convergence of the filter weights. Such optimization approach is our future work.

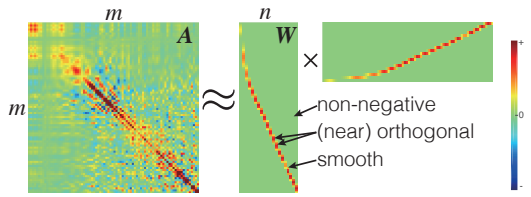



**Fig. 1**. The proposed method optimizes a filter bank $\boldsymbol{W}$ by approximating the FDA subspace $\boldsymbol{A}$ with the orthogonal and non-negative constraints and smoothness regularization.

spanned by the orthonormal matrix $\boldsymbol{Q}$ obtained such as via QR decomposition $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{R}$. Therefore, we aim to optimize a filter bank so as to well approximate the discriminative subspace. The essential form of the subspace is $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^\top$ and thus the optimization formulation of $n$ filters $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ is, as show in Fig. 1,

$$\min_{\substack{\boldsymbol{W}|\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I},\\ W_{ij}\geq 0\,\forall i,j}} \frac{1}{2}\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{A}\|_F^2 + \Omega(\boldsymbol{W}) \tag{3}$$

$$\Leftrightarrow \max_{\substack{\boldsymbol{W}|\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I},\\ W_{ij}\geq 0\,\forall i,j}} \text{trace}(\boldsymbol{W}^\top \boldsymbol{A}\boldsymbol{W}) - \Omega(\boldsymbol{W}), \tag{4}$$

where $\|\cdot\|_F$ indicates Frobenius norm of a matrix and $\Omega(\boldsymbol{W})$ is a smoothness regularization term described in the later.

In this study, we generally assume that the local features, such as HOG [12], are extracted along the frequency axis in the spectrogram as in [2] and thus the input signal is represented by a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ where $m$ is the number of frequency component and $d$ denotes the local feature dimensionality. A filter bank $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n] \in \mathbb{R}^{m\times n}$ works on the frequency components mainly to reduce dimensionality by $\boldsymbol{W}^\top \boldsymbol{X}$. In this case of matrix features, to produce the projection $\boldsymbol{P}$ (eventually $\boldsymbol{A}$), we follow the procedure of 2D-FDA [13] which applies FDA twice. First, we compute the FDA projection $\hat{\boldsymbol{P}} \in \mathbb{R}^{d\times k}$ for the $d$-dimensional local features based on the following covariance matrices;

$$\hat{\boldsymbol{\Sigma}}_T = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i - \boldsymbol{M})^\top(\boldsymbol{X}_i - \boldsymbol{M}), \tag{5}$$

$$\hat{\boldsymbol{\Sigma}}_B = \sum_{c=1}^{C}\frac{N_c}{N}(\boldsymbol{M}_c - \boldsymbol{M})^\top(\boldsymbol{M}_c - \boldsymbol{M}), \tag{6}$$

where $\boldsymbol{M}$ and $\boldsymbol{M}_c$ are the mean feature matrices on all $N$ samples and on $N_c$ samples of the $c$-th class, respectively. Then, the FDA projection $\boldsymbol{P} \in \mathbb{R}^{m\times k}$ of our interest is obtained via

$$\boldsymbol{\Sigma}_T = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i - \boldsymbol{M})\hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top(\boldsymbol{X}_i - \boldsymbol{M})^\top, \tag{7}$$

$$\boldsymbol{\Sigma}_B = \sum_{c=1}^{C}\frac{N_c}{N}(\boldsymbol{M}_c - \boldsymbol{M})\hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\top(\boldsymbol{M}_c - \boldsymbol{M})^\top. \tag{8}$$

Here, $k$ indicates the subspace dimensionality which is usually set as $k = C - 1$.

On the other hand, as to $\Omega(\boldsymbol{W})$ in (4), smooth weights are obtained via minimizing the Laplacian cost,

$$l(w) = \int |\Delta w(f)|^2 df \approx \sum_f |-w_{f-1}+2w_f-w_{f+1}|^2. \quad (9)$$

The above Laplacian cost is written by the quadratic form $l(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{L}\boldsymbol{w}$ where the matrix $\boldsymbol{L}$ is determined according to (9). Thus, the regularization term $\Omega(\boldsymbol{W})$ is represented with the regularization parameter $\eta$ by $\Omega(\boldsymbol{W}) = \eta\,\mathrm{trace}(\boldsymbol{W}^\top \boldsymbol{L}\boldsymbol{W})$; thereby, the optimization problem is finally described as

$$\max_{\substack{\boldsymbol{W}|\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I},\\ W_{ij}\geq 0\,\forall i,j}} \mathrm{trace}\left\{\boldsymbol{W}^\top(\boldsymbol{A}-\eta\boldsymbol{L})\boldsymbol{W}\right\}. \quad (10)$$

### 2.2. Optimization with nearly orthogonal and non-negative constraints

We employ the approach proposed in [14] to optimize (10) which contains the constraints regarding orthogonality and non-negativity. Let us generally consider to maximize a function $\mathtt{f}(\boldsymbol{W})$ with respect to $\boldsymbol{W}$ on which those constraints are imposed;

$$\max_{\boldsymbol{W}} \mathtt{f}(\boldsymbol{W}),\ \text{s.t. } \boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{I},\ W_{ij} \geq 0\ \forall i,j. \quad (11)$$

Oja and Yang [14] proposed an efficient optimization approach based on the following multiplicative update of $\boldsymbol{W}$:

$$W_{ij} \leftarrow W_{ij}\frac{(\boldsymbol{\nabla}^+ + \boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\nabla}^-)_{ij}}{(\boldsymbol{\nabla}^- + \boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\nabla}^+)_{ij}}, \quad (12)$$

where $\boldsymbol{\nabla}^+$ and $\boldsymbol{\nabla}^-$ are elementwise non-negative matrices such that $\frac{\partial}{\partial \boldsymbol{W}}\mathtt{f}(\boldsymbol{W}) = \boldsymbol{\nabla}^+ - \boldsymbol{\nabla}^-$ and $\nabla^+_{ij} \geq 0, \nabla^-_{ij} \geq 0\ \forall i,j$. This approximately optimizes the Lagrangian function derived from the constraints, and thus produces the non-negative and *nearly orthogonal* $\boldsymbol{W}$ of which the column vectors (filters in our case) are less overlapped to each other[3]. Practically speaking, such nearly orthogonal optimizer is preferable for our task of filter bank learning since less overlapped filters enhance robustness to frequency shift by mitigating boundary effects, compared to strictly orthogonal filters.

We can actually optimize (10) by

$$W_{ij} \leftarrow W_{ij}\frac{(\boldsymbol{S}^+\boldsymbol{W} + \boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{S}^-\boldsymbol{W})_{ij}}{(\boldsymbol{S}^-\boldsymbol{W} + \boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{S}^+\boldsymbol{W})_{ij}}, \quad (13)$$

where $S^+_{ij} = A_{ij} - \eta L_{ij} + |A_{ij} - \eta L_{ij}|$, $S^-_{ij} = |A_{ij} - \eta L_{ij}| - (A_{ij} - \eta L_{ij})$ and we substitute $\boldsymbol{\nabla}^+ = \boldsymbol{S}^+\boldsymbol{W}$ and $\boldsymbol{\nabla}^- = \boldsymbol{S}^-\boldsymbol{W}$ in (12).

[3]In the paper [14], the authors pointed out that the orthogonality constraint is approximately satisfied in the approximated optimization approach.

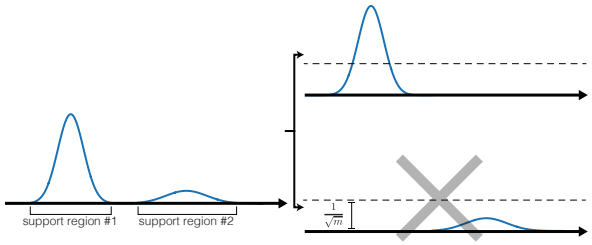



**Fig. 2**. Filter weight decomposition. Multi-modal filter weights are decomposed, and trivial weight modes are removed during optimization.

### 2.3. Technical tips

**Initialization**. The optimization problem (10) is vulnerable to local optima, and thus for obtaining the better optimizer, we apply warm startup by first optimizing (10) with $\eta = 0$ (excluding smoothness regularization) and then passing the optimizer as an initial $\boldsymbol{W}$ of the target optimization problem with smoothing ($\eta > 0$).

**Locality of filters**. The optimized filter $\boldsymbol{w}_i$ occasionally contains more than one mode especially in the case of smaller regularization parameter $\eta$. Since filter weights are supposed to be locally distributed as described in Sec. 1, such a filter that consists of multi-modal weights are decomposed into multiple filters each of which has only one mode as shown in Fig. 2. Consequently, the size of the resultant filter bank is greater than or equal to $n$. In addition, during the optimization, we remove a trivial mode of which local maximum is less than $\frac{1}{\sqrt{m}}$ in order to produce a meaningful filter bank.

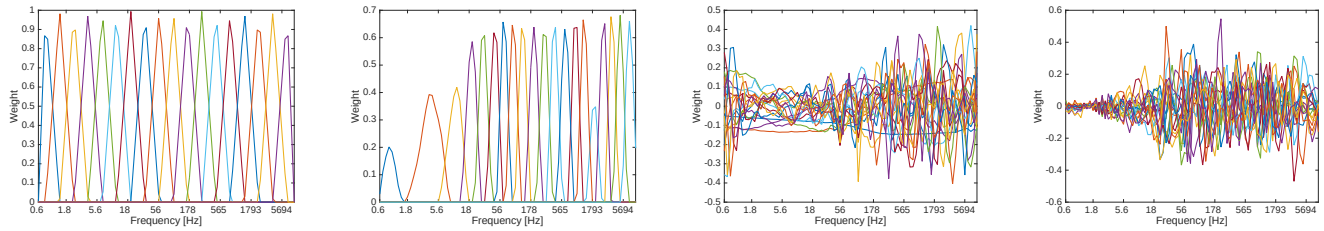
## 3. EXPERIMENTAL RESULTS

We test the proposed method on acoustic scene classification using Rouen dataset [2]. This dataset contains 3,026 sound clips (in 30 seconds) of 19 scene categories, such as *cafe* and *market*. The authors [2] also provided 20-fold training/test splits on which the averaged classification accuracy is reported. For feature extraction, we follow the procedure in [2] which extracts HOG features [12] with cells of $8\times8$ pixels on the spectrogram produced by constant Q-transform [15] of 48 bins per octave. The HOG features are averaged along the time axis, resulting in the feature matrix $\boldsymbol{X} \in \mathbb{R}^{84\times31}$ where the HOG feature dimensionality is $d = 31$ as in [16] and the number of frequency component is $m = 84$.

For comparison, we apply the following three alternative methods.

**Uniform filter bank (Mel filter bank)**. As a simple yet commonly used filter bank, we apply filters of which band widths are exponentially increased along the frequency axis; that is, they are equally distributed on log-frequency axis as in mel filter bank (Fig. 3a).

**FDA projector bank**. We can directly employ the FDA projection $\boldsymbol{P}$ described in Sec. 2.1 (Fig. 3d); this is close to the




(a) Uniform (b) Ours (c) PCA (d) FDA

**Fig. 3**. Filter/projector banks that we used in the experiment. While the uniform filter bank (a) is obtained a priori, the others (b,c,d) are learned from data.

work [9] which employs FDA projection for (raw) spectrogram features. Note that the projection is not regarded as a *filter* bank since it does not satisfy the before-mentioned conditions for filters. Such projector could not provide any interpretable information regarding frequency.

**PCA projector bank**. As in the FDA projection, we can apply PCA projection (Fig. 3c) to reduce the frequency dimensionality. This projector is learned from the generative, not discriminative, perspective. Note again that this is not a *filter*.

According to the convention in FDA, we apply $n = C - 1 = 18$ filters/projectors to reduce the frequency dimensionality, while setting $\eta = 10$ in the proposed method. The performance results are shown in Table 1. Even though the dimensionality is reduced to only a quarter, the proposed method exhibits high performance in comparison with the generative methods, uniform filter and PCA projector; the FDA projector from which the proposed method stems is competitive with ours. This result shows that the constraints and regularization introduced for building filters do not degrade performance, reasonably characterizing the physical properties of frequency. Furthermore, the optimized filters can provide us with interpretable information for further analysis of frequency; for example, by looking at the obtained filter bank (Fig. 3b), lower frequencies are less important and do not require high resolution for distinguishing scene categories.

In the above experiment, the proposed method actually produces 20 filters on average, which are slightly larger than $n = 18$, due to the decomposition of multi-modal filter weights as described in Sec. 2.3. We investigate the performance of the proposed method in detail from the viewpoint of filter bank size by varying the parameters $\eta$ and $n$. Fig. 4 shows the performance results on various sizes of filter bank. The proposed method produces better performance even in smaller size, exhibiting superiority over the uniform filter bank. While the performance of the uniform filter is gradually approaching that of full features (91.64%), the proposed method outperforms it at about half of that dimensionality (91.92% at the size of 47 (averaged) for $n = 18$). On the cases of $n \neq 18$, we can also observe similar tendency of performance; the best is 92.01% at $\eta = 0.1$ and $n = 40$, producing 60 filters on average, which is also superior to 91.7 reported in [2]. These experimental results validate the effectiveness of the proposed method which discriminatively learns a filter bank to produce better performance in a smaller-sized filter bank.

**Table 1**. Performance comparison (%) with $n = C - 1 = 18$.

| full | Ours | Uniform | FDA | PCA |
|---|---|---|---|---|
| 91.64 | 89.22 | 85.52 | 88.98 | 86.80 |

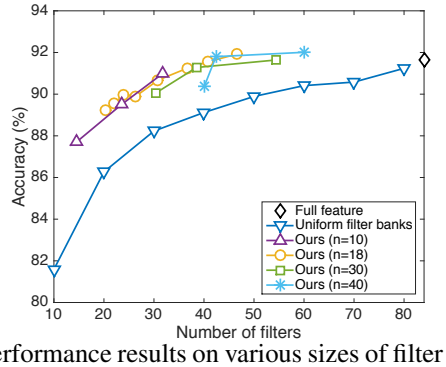



**Fig. 4**. Performance results on various sizes of filter bank. We apply a uniform filter bank of $n \in \{10, 20, \cdots, 80\}$, the proposed method with $n = 18$ of $\eta \in \{10, 5, 4, 3, 2, 1, 0.5, 0.1\}$ and that with $n = 10, 30, 40$ of $\eta \in \{10, 1, 0.1\}$ for gradually increasing filter bank size.

## 4. CONCLUSION

We have proposed a novel method to discriminatively learn a filter bank for effectively reducing frequency dimensionality of acoustic features. The method is formulated by introducing constraints regarding *non-negativity* and *(near) orthogonality* as well as *smoothness* regularization for filter weights in order to form a reasonable filter bank. In addition, we incorporate discriminative optimization based on the Fisher criterion via approximating the FDA projection. Thus, the proposed method builds an effective filter bank from scratch without assuming any parametric model on the filters. In the experiments on acoustic scene classification using Rouen dataset, the method exhibits favorable performance even in smaller size of filter bank.

## 5. REFERENCES


[1] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[2] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of timefrequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2015.

[3] S. Park, W. Choi, D. K. Han, and H. Ko, "Acoustic event filterbank for enabling robust event recognition by cleaning robot," *IEEE Transactions on Consumer Electronics*, vol. 61, no. 2, pp. 189–196, 2015.

[4] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 4574–4577.

[5] M. J. Alam, P. Kenny, and D. OShaughnessy, "Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique," *Digital Signal Processing*, vol. 29, pp. 147–157, 2014.

[6] T. Kobayashi and J. Ye, "Acoustic feature extraction by statistics based local binary pattern for environmental sound classification," in *International Conference on Acoustic, Speech and Signal Processing*, 2014, pp. 3076–3080.

[7] B. K.-W. Mak, Y.-C. Tam, and P. Q. Li, "Discriminative auditory-based features for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 27–36, 2004.

[8] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 299–304.

[9] L. Burget and H. Hemansk, "Data driven design of filter bank for speech recognition," in *Text, Speech and Dialogue*, 2001, pp. 299–304.

[10] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research*, vol. 6, pp. 483–502, 2005.

[11] T. Kobayashi, "Higher-order co-occurrence features based on discriminative co-clusters for image classification," in *British Machine Vision Conference*, 2012, pp. 64.1–64.11.

[12] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[13] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in Neural Information Processing Systems 17*, 2004.

[14] E. Oja and Z. Yang, "Orthogonal nonnegative learning for sparse feature extraction and approximate combinatorial optimization," *Frontiers of Electrical and Electronic Engineering in China*, vol. 5, no. 3, pp. 261–273, 2010.

[15] C Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, 2014.

[16] P. F. Felzenszwalb, R. B. Grishick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.