FBEM: A FILTER BANK EM ALGORITHM FOR THE JOINT OPTIMIZATION OF FEATURES AND ACOUSTIC MODEL PARAMETERS IN BIRD CALL CLASSIFICATION

Wei Chu, Abeer Alwan

Department of Electrical Engineering,
University of California, Los Angeles,
Los Angeles, 90095
{weichu, alwan}@ee.ucla.edu

ABSTRACT

This paper extends the expectation-maximization (EM) algorithm to estimate not only optimal acoustic model parameters, but also optimal center frequencies and bandwidths of the filter bank used in cepstral feature extraction for bird call classification. The search is done using the gradient ascent method. Filter bank and model parameters are optimized iteratively. Experiments are conducted on a large noisy corpus containing Antbird calls from 5 species. It is shown that features extracted using the optimized filter bank result in a lower classification error rate than those extracted using a Mel-scaled filter bank.

Index Terms— EM algorithm, filter bank, bird call classification

1. INTRODUCTION

In pattern recognition tasks, audio signals are compressed to a sequence of feature vectors. When the distribution of the features is quantitatively modeled, the expectation-maximization (EM) algorithm can be used to estimate acoustic model parameters by iteratively maximizing the expectation of the likelihood from these features [1].

To improve the discriminability of the features, the original feature space can be mapped to new subspaces by certain projections. Different criteria are employed to search for optimal projections. Linear discriminant analysis (LDA) [2] computes the projection by maximizing the Fisher ratio value; heteroscedastic LDA (HLDA) [3] and multiple LDA (MLDA) [4] learn the projection by maximizing the likelihood from the transformed features; while fMPE [5] estimates the projection by minimizing phone error rate.

Changing parameters in feature extraction can also increase the discriminability of the features. The Mel-scaled filter bank is often used for feature extraction in automatic speech recognition (ASR) [6]. Kamm et al. [7] searched a family of optimal warping-scales for ASR through a brute-force data-driven approach, and concluded that the Mel-scale is a member of the family. Graciarena et al [8] manually changed the frequency range, the number of filters, and the frequency scale type of the filter bank for bird song identification.

For bird song classification and recognition applications, researchers also have spent effort in exploring machine learning techniques such as back propagation and multivariate statistics [9], dynamic time warping and hidden Markov models [10] [11], and so on. Denoising filters are also helpful for enhancing bird calls [12].

Supported in part by NSF

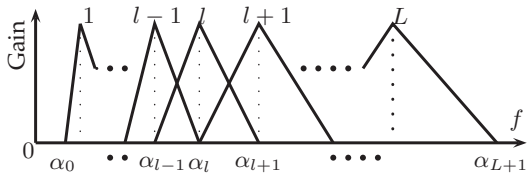

Fig. 1. The frequency response of the filter bank used in feature extraction. L is the number of filters. The letter on top of each filter denotes the filter index.The gain of each filter is the same.

In this paper, the optimal center frequencies and bandwidths of the filter bank are searched in an efficient statistically-based approach. Since the auxiliary function in the EM algorithm is extended for optimizing not only model parameters, but also parameters of the filter bank used in feature extraction, the proposed algorithm is called the filter bank EM (fbEM) algorithm. Note that statistically-based non-uniform DFT analysis/synthesis filter banks are explored to reduce spectral-domain distortion in speech coding [13].

The organization of the paper is as follows: in Section II, joint filter bank and model parameters optimization using the fbEM algorithm is presented; and in Section III, experimental results on an Antbird corpus are analyzed.

2. OPTIMIZING THE FILTER BANK IN FEATURE EXTRACTION

The procedure and parameters of cepstral feature extraction are the same as the Mel-frequency cepstral coefficients (MFCCs) extraction except for the parameters of the filter bank. In the new filter bank shown in Fig. 1, it is assumed that the number of filters is fixed as L, the shape of each filter is triangular, the gain of each filter is the same, and the center frequency of each filter is equal to the low and high cut-off frequencies of its right and left filters, respectively. $\boldsymbol{\alpha} = [\alpha_0, \cdots, \alpha_{L+1}]^T$ is used to represent the parameters of the filter bank, where $\alpha_l, l = 1 \cdots L$, denotes the center frequency of filter l, α_0 and α_{L+1} denote the low and high cut-off frequencies of the filter bank, respectively. Audio signals denoted by $\mathbf{x}$ is compressed to a sequence of column feature vectors denoted by $\mathbf{Y}$ which can be represented as $\{\mathbf{y}_1, \cdots, \mathbf{y}_T\}$, where T is the number of the frames. The procedure of feature extraction can be viewed as a function denoted by $f_{\boldsymbol{\alpha}}$ from $\mathbf{x}$ to $\mathbf{Y}$, i.e. $\mathbf{Y} = f_{\boldsymbol{\alpha}}(\mathbf{x})$.

If the feature sequence is assumed to be independent and identically distributed within each class, Gaussian mixture model (GMM) can be used to model the distribution of the homogeneous data.

The proposed fbEM algorithm is shown in Algorithm 1.

Algorithm 1. fbEM: joint filter bank and model parameter optimization using EM algorithm

Step 1: Initialization: initialize the filter bank parameters $\boldsymbol{\alpha}$; extract feature $\mathbf{Y}$, i.e. $f_{\boldsymbol{\alpha}}(\mathbf{x})$ from acoustic signals $\mathbf{x}$; train an initial model $\mathcal{M}$ from $\mathbf{Y}$ using the conventional EM algorithm.

Step 2: Constrained Filter bank optimization without updating the model $\mathcal{M}$:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\bar{\boldsymbol{\alpha}}} \mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\})$$
$$\text{s.t.} \quad \alpha_{\min} \leq \bar{\alpha}_0 < \cdots < \bar{\alpha}_{L+1} \leq \alpha_{\max} \tag{1}$$

$\hat{\boldsymbol{\alpha}}$ is solved as follows: initialize $\bar{\boldsymbol{\alpha}}$ to be $\boldsymbol{\alpha}$, and $\bar{\mathbf{Y}}$ to be $f_{\boldsymbol{\alpha}}(\mathbf{x})$, then update $\bar{\boldsymbol{\alpha}}$:

$$\bar{\boldsymbol{\alpha}} \leftarrow \bar{\boldsymbol{\alpha}} + \eta \frac{\partial \mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\})}{\partial \bar{\boldsymbol{\alpha}}}$$

where η denotes the step size, $\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\})$ is an auxiliary function defined in Eq. 5. Extract $\bar{\mathbf{Y}}$ using the updated $\bar{\boldsymbol{\alpha}}$, i.e. $\bar{\mathbf{Y}} = f_{\bar{\boldsymbol{\alpha}}}(\mathbf{x})$, repeat until the increment of $\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\})$ falls below a certain threshold. Then let $\hat{\boldsymbol{\alpha}} = \bar{\boldsymbol{\alpha}}$, $\hat{\mathbf{Y}} = \bar{\mathbf{Y}}$.

Step 3: Estimate model parameters without updating the filter bank $\hat{\boldsymbol{\alpha}}$ and feature $\hat{\mathbf{Y}}$, i.e. $f_{\hat{\boldsymbol{\alpha}}}(\mathbf{x})$:

$$\hat{\mathcal{M}} = \arg\max_{\bar{\mathcal{M}}} \mathcal{Q}(\{\hat{\boldsymbol{\alpha}}, \mathcal{M}\}, \{\hat{\boldsymbol{\alpha}}, \bar{\mathcal{M}}\}) \tag{2}$$

which is the same as the conventional EM algorithm [1].

Step 4: Convergence or keep iterating: if

$$\frac{|\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\hat{\boldsymbol{\alpha}}, \hat{\mathcal{M}}\}) - \mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\boldsymbol{\alpha}, \mathcal{M}\})|}{|\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\boldsymbol{\alpha}, \mathcal{M}\})|} \geq \epsilon \tag{3}$$

where ϵ denotes the threshold, then $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$, $\mathcal{M} = \hat{\mathcal{M}}$, go to Step 2; else stop and exit.

As shown in Algorithm 1, the auxiliary function $\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \bar{\mathcal{M}}\})$ of the fbEM algorithm has both feature extraction and model parameters as variables. In conventional EM algorithm, the auxiliary function only has model parameters as variables. In fbEM algorithm, since $\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\boldsymbol{\alpha}, \mathcal{M}\}) \leq \mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\hat{\boldsymbol{\alpha}}, \mathcal{M}\}) \leq \mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\hat{\boldsymbol{\alpha}}, \hat{\mathcal{M}}\})$, which is illustrated in Step 2 and 3, the increase of the auxiliary function is guaranteed.

The details of the Algorithm 1 are shown in the following.

2.1. Filter bank α and model $\mathcal{M}$ initialization

In Step 1, it is important to choose a good initial guess to the solution for an iterative method like the EM algorithm. Graciarena et al [8] showed that a Mel-scaled filter bank results in a higher bird call verification accuracy compared to the linear-scaled counterpart. In this paper, the parameters of a Mel-scaled filter bank are used as the initial guess for $\boldsymbol{\alpha}$.

Note that the parameters of the initial GMMs are trained from the MFCC features using the conventional EM algorithm [1].

2.2. Compute the auxiliary function $\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\})$

Because there is no closed-form solution for $\hat{\boldsymbol{\alpha}}$ in Eq. 1, the gradient ascent method is employed in Step 2.

Let $\mathbf{y}_t^{(r)}$ denote the features extracted using the filter bank $\boldsymbol{\alpha}$ at frame t. The current $\boldsymbol{\alpha}$ is either initialized in Step 1, or obtained from Step 2 of the previous iteration. The probability of $\mathbf{y}_t^{(r)}$ belonging to mixture m of class r denoted by $\gamma_m^{(r)}(t)$ can be calculated as:

$$\gamma_m^{(r)}(t) = \frac{\omega_m^{(r)} \mathcal{N}(\mathbf{y}_t^{(r)}; \boldsymbol{\mu}_m^{(r)}, \boldsymbol{\Sigma}_m^{(r)})}{\sum_{m'=1}^{M} \omega_{m'}^{(r)} \mathcal{N}(\mathbf{y}_t^{(r)}; \boldsymbol{\mu}_{m'}^{(r)}, \boldsymbol{\Sigma}_{m'}^{(r)})} \tag{4}$$

where M denotes the number of Gaussians in each GMM, $\omega_m^{(r)}$, $\boldsymbol{\mu}_m^{(r)}$, and $\boldsymbol{\Sigma}_m^{(r)}$ are the weight, mean, and covariance matrix of the Gaussian mixture m of class r, obtained from the initialization or Step 3 of the previous iteration, $\mathcal{N}(\cdot)$ means Gaussian distribution.

Assuming that the discrete cosine transform (DCT) in feature extraction eliminates the dependencies among features from different dimensions, the covariance matrix of each Gaussian is a diagonal matrix. Suppose static (s), derivative (d), and acceleration (a) cepstral features are extracted, i.e. $\bar{\mathbf{y}}_t^T = [\bar{\mathbf{y}}_t^{s^T} \bar{\mathbf{y}}_t^{d^T} \bar{\mathbf{y}}_t^{a^T}]^T$.

In Step 2, the auxiliary function can be expressed as:

$$\begin{aligned}
&\mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\}) \\
&= \sum_{r=1}^{R} \sum_{m=1}^{M} \sum_{t=1}^{T^{(r)}} \gamma_m^{(r)}(t) \mathcal{N}(\bar{\mathbf{y}}_t^{(r)}; \boldsymbol{\mu}_m^{(r)}, \boldsymbol{\Sigma}_m^{(r)}) \\
&= -\frac{1}{2} \sum_{g \in \{s,d,a\}} \sum_{r=1}^{R} \sum_{m=1}^{M} \sum_{t=1}^{T^{(r)}} \big[\gamma_m^{(r)}(t) (\bar{\mathbf{y}}_t^{g^{(r)}} - \boldsymbol{\mu}_m^{g^{(r)}})^T \\
&\qquad \boldsymbol{\Sigma}_m^{g^{-1}(r)} (\bar{\mathbf{y}}_t^{g^{(r)}} - \boldsymbol{\mu}_m^{g^{(r)}})\big] + C
\end{aligned} \tag{5}$$

where R denotes the number of classes, $T^{(r)}$ denotes the number of frames in class r, $\bar{\mathbf{y}}_t^{(r)}/\bar{\mathbf{y}}_t^{s^{(r)}}/\bar{\mathbf{y}}_t^{d^{(r)}}/\bar{\mathbf{y}}_t^{a^{(r)}}$ denotes the whole/static/derivative/acceleration features at frame t extracted using the filter bank $\bar{\boldsymbol{\alpha}}$, C denotes a term that is invariant to $\bar{\boldsymbol{\alpha}}$.

2.3. Compute $\partial \mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\}) / \partial \bar{\boldsymbol{\alpha}}$

By using the chain rule, we have

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\{\boldsymbol{\alpha}, \mathcal{M}\}, \{\bar{\boldsymbol{\alpha}}, \mathcal{M}\})}{\partial \bar{\boldsymbol{\alpha}}} = &- \sum_{g \in \{s,d,a\}} \sum_{r=1}^{R} \sum_{m=1}^{M} \sum_{t=1}^{T^{(r)}} \big[\gamma_m^{(r)}(t) \\
&\frac{\partial \bar{\mathbf{y}}_t^{g^{(r)}}}{\partial \bar{\boldsymbol{\alpha}}} \boldsymbol{\Sigma}_m^{g^{-1}(r)} (\bar{\mathbf{y}}_t^{g^{(r)}} - \boldsymbol{\mu}_m^{g^{(r)}})\big]
\end{aligned} \tag{6}$$

At frame t, let $\bar{e}_{t_l}$ denote the energy out of the l_{th} filter, and $\bar{\mathbf{e}}_t = [\bar{e}_{t_1} \cdots \bar{e}_{t_L}]$ denote the energy output of the filter bank. When $\bar{\mathbf{e}}_t$ is taken as input, the static cepstral coefficient $\bar{\mathbf{y}}_t^s$ is the output of the three cascaded feature extraction sub-procedures: logarithm, DCT and cepstral liftering:

$$\bar{\mathbf{y}}_t^s = \mathbf{M}_{\text{CEP\_LFT}}^T \mathbf{M}_{\text{DCT}}^T \log \bar{\mathbf{e}}_t \tag{7}$$

where $\mathbf{M}_{\text{CEP\_LFT}}$ is a $D \times D$ diagonal matrix:

$$[\mathbf{M}_{\text{CEP\_LFT}}]_d = \quad 1 + \frac{d-1}{2} \sin \frac{\pi(d-1)}{N}, \quad d = 1 \cdots D \tag{8}$$

where d denotes the diagonal index, D denotes the dimension of the static features, N denotes the cepstral liftering coefficient; $\mathbf{M}_{\text{DCT}}$ is an $L \times D$ matrix:

$$\begin{aligned}
[\mathbf{M}_{\text{DCT}}]_{l,d} = &\sqrt{\frac{2}{L}} \cos \frac{\pi(l-0.5)(d-1)}{L} \\
&l = 1 \cdots L, \quad d = 1 \cdots D
\end{aligned} \tag{9}$$

where l denotes the row index, d denotes the column index, L denotes the number of the filters.

Re-applying the chain rule, $\partial \bar{\mathbf{y}}_t^s / \partial \bar{\boldsymbol{\alpha}}$ can be expressed as:

$$\frac{\partial \bar{\mathbf{y}}_t^s}{\partial \bar{\boldsymbol{\alpha}}} = \frac{\partial \bar{\mathbf{e}}_t}{\partial \bar{\boldsymbol{\alpha}}} \frac{\partial \log \bar{\mathbf{e}}_t}{\partial \bar{\mathbf{e}}_t} \mathbf{M}_{\text{DCT}} \mathbf{M}_{\text{CEP\_LFT}} \tag{10}$$

$\partial \bar{\mathbf{e}}_t / \partial \bar{\boldsymbol{\alpha}}$ is solved as follows. Let $H_l[f]$ denote the frequency response of the triangular filter l in the filter bank shown in Fig. 1, we have:

$$H_l[f] = \begin{cases} \dfrac{f - \bar{\alpha}_{l-1}}{\bar{\alpha}_l - \bar{\alpha}_{l-1}} & \bar{\alpha}_{l-1} \le f < \bar{\alpha}_l \\ \dfrac{f - \bar{\alpha}_{l+1}}{\bar{\alpha}_l - \bar{\alpha}_{l+1}} & \bar{\alpha}_l \le f < \bar{\alpha}_{l+1} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where f denotes the frequency. Let $S_t[f]$ denote the power spectrum at frame t, the energy output of l_{th} filter can be expressed as: $\bar{e}_{t_l} = \sum_{f=0}^{F_s/2} H_l[f] S_t[f]$, where F_s is the sampling frequency. $\partial \log \bar{\mathbf{e}}_t / \partial \bar{\mathbf{e}}_t$ in Eq. 10 is an $L \times L$ diagonal matrix:

$$\Big[\frac{\partial \log \bar{\mathbf{e}}_t}{\partial \bar{\mathbf{e}}_t}\Big]_l = \frac{1}{\bar{e}_{t_l}}, \quad l = 1 \cdots L \tag{12}$$

where l denotes the diagonal index. $\partial \bar{\mathbf{e}}_t / \partial \bar{\boldsymbol{\alpha}}$ in Eq. 10 is an $(L+2) \times L$ band matrix:

$$\Big[\frac{\partial \bar{\mathbf{e}}_t}{\partial \bar{\boldsymbol{\alpha}}}\Big]_{p,l} = \begin{cases} \sum\limits_{f=\bar{\alpha}_{l-1}}^{\bar{\alpha}_l} \dfrac{f - \bar{\alpha}_l}{(\bar{\alpha}_{l-1} - \bar{\alpha}_l)^2} S_t[f] & p = l \\ \Big[\sum\limits_{f=\bar{\alpha}_l}^{\bar{\alpha}_{l+1}} \dfrac{f - \bar{\alpha}_{l+1}}{(\bar{\alpha}_l - \bar{\alpha}_{l+1})^2} & p = l+1 \\ - \sum\limits_{f=\bar{\alpha}_{l-1}}^{\bar{\alpha}_l} \dfrac{f - \bar{\alpha}_{l-1}}{(\bar{\alpha}_l - \bar{\alpha}_{l-1})^2} \Big] S_t[f] & \\ \sum\limits_{f=\bar{\alpha}_l}^{\bar{\alpha}_{l+1}} \dfrac{f - \bar{\alpha}_l}{(\bar{\alpha}_{l+1} - \bar{\alpha}_l)^2} S_t[f] & p = l+2 \\ 0 & \text{otherwise} \end{cases}$$
$$p = 1 \cdots L+2, \quad l = 1 \cdots L \tag{13}$$

where p denotes the row index, l denotes the column index.

Since the derivative features are calculated as:

$$\bar{\mathbf{y}}_t^d = \frac{\sum_{\theta=1}^{\Theta_d} \theta (\bar{\mathbf{y}}_{t+\theta}^s - \bar{\mathbf{y}}_{t-\theta}^s)}{2 \sum_{\theta=1}^{\Theta_d} \theta^2} \tag{14}$$

where Θ_d denotes the coefficient for computing the derivative features, after calculating $\partial \bar{\mathbf{y}}_t^s / \partial \bar{\boldsymbol{\alpha}}$, $\partial \bar{\mathbf{y}}_t^d / \partial \bar{\boldsymbol{\alpha}}$ in Eq. 6 can be computed as:

$$\frac{\partial \bar{\mathbf{y}}_t^d}{\partial \bar{\boldsymbol{\alpha}}} = \frac{\sum_{\theta=1}^{\Theta_d} \theta \big(\frac{\partial \bar{\mathbf{y}}_{t+\theta}^s}{\partial \bar{\boldsymbol{\alpha}}} - \frac{\partial \bar{\mathbf{y}}_{t-\theta}^s}{\partial \bar{\boldsymbol{\alpha}}} \big)}{2 \sum_{\theta=1}^{\Theta_d} \theta^2} \tag{15}$$

Since the acceleration features are obtained from the derivative features in the same way as obtaining the derivative features from the static features, $\partial \bar{\mathbf{y}}_t^a / \partial \bar{\boldsymbol{\alpha}}$ in Eq. 6 is calculated as:

$$\frac{\partial \bar{\mathbf{y}}_t^a}{\partial \bar{\boldsymbol{\alpha}}} = \frac{\sum_{\theta=1}^{\Theta_a} \theta \big(\frac{\partial \bar{\mathbf{y}}_{t+\theta}^d}{\partial \bar{\boldsymbol{\alpha}}} - \frac{\partial \bar{\mathbf{y}}_{t-\theta}^d}{\partial \bar{\boldsymbol{\alpha}}} \big)}{2 \sum_{\theta=1}^{\Theta_a} \theta^2} \tag{16}$$

where Θ_a denotes the coefficient for computing the acceleration features.

3. EXPERIMENTS

The Antbird call corpus contains 3366 bird calls from 5 species: Barred Antshrike (BAS), Dusky Antbird (DAB), Great Antshrike (GAS), Mexican Antthrush (MAT), Dot-winged Antwren (DWA) [11]. The training set is 85 minutes long and the testing set is 42 minutes long. The calls are 0.5 - 5.0 seconds long. Examples of bird calls are shown in [11]. The frequency range of the bird calls is from 500 to 6000 Hz. The signal is downsampled from 44.1 kHz to 16 kHz. The low and high cut-off frequencies of the filter bank, $\alpha_{\min}$ and $\alpha_{\max}$, are set to 360 and 6500 Hz, respectively, to remove irrelevant frequency components for bird call classification [12].

Two feature extraction methods are compared: the standard MFCC extraction with a Mel-scaled filter bank, and the improved MFCC extraction with an optimized filter bank obtained from Algorithm 1. The number of filters in the filter bank, L, is set to 26. The cepstral liftering coefficient, N, is set to 22. The dimension of the static, derivative, and acceleration features, D, is set to 13. The coefficients for computing the derivative and acceleration features, Θ_d and Θ_a, are both set to 2. The frame step size is 10 ms, and the frame length is 25 ms. In the GMM classifier, the number of Gaussians in each species' model, M, is set to 256. In the filter bank optimization, the convergence threshold, ϵ, is set to 10^{-3}.

The baseline system using MFCC features has a classification error rate of 8.7%. By using the new features extracted using the optimal filter bank obtained from Algorithm 1, the error rate is reduced to 6.2%. The p-value of significance test is 0.024, which means that the proposed method is statistically significant for a significance level of 0.05. The optimization converges at the 6-th iteration, while the lowest classification error rate is achieved at the 4-th iteration. Model overfitting can be the explanation.

The confusion matrix of results obtained by using Mel-scaled and optimized filter bank are shown in Table 1. The calls of BAS, MAT, and DWA are less likely to be misclassified as other species compared to those of DAB and GAS. The optimized filter bank effectively reduced the DAB and GAS classification errors by 1.0% and 0.4%, respectively.

Let $\alpha_l^0 / \hat{\alpha}_l$ and $B_l^0 / \hat{B}_l$ denote the center frequency and bandwidth of l_{th} filter in the Mel-scaled/optimal filter bank, respectively. Note that in the triangular filter bank shown in Fig. 1, we have:

$$B_l^0 = \alpha_{l+1}^0 - \alpha_{l-1}^0 \tag{17}$$

$$\hat{B}_l = \hat{\alpha}_{l+1} - \hat{\alpha}_{l-1} \tag{18}$$

To show the percentages of the center frequencies and bandwidths of optimal filter bank being shifted from the corresponding ones in the Mel-scaled filter bank, two difference measures regarding l_{th} filter denoted by Δ_l^α and Δ_l^B are defined as follows:

$$\Delta_l^\alpha = (\hat{\alpha}_l / \alpha_l^0 - 1) \times 100\% \tag{19}$$

$$\Delta_l^B = (\hat{B}_l / B_l^0 - 1) \times 100\% \tag{20}$$

In Mel-scaled filter bank, the distances of the center frequency of l_{th} filter to its left and right counterparts are $\alpha_l^0 - \alpha_{l-1}^0$ and $\alpha_{l+1}^0 - \alpha_l^0$. The smaller the distances are, the higher frequency resolution at frequencies near α_l^0 is [6]. Since $B_l^0 = [\alpha_{l+1}^0 - \alpha_l^0] + [\alpha_l^0 - \alpha_{l-1}^0]$, the bandwidth of the filter can be used as a measure of the frequency resolution at frequencies near the center frequency of the filter. The same conclusion can be drawn from the optimal filter bank.

A comparision of frequency parameters of the Mel-scaled and optimal filter banks are shown in Table 2. In the optimal filter bank, the bandwidth sequence $\{\hat{B}_0, \cdots, \hat{B}_L\}$ is no longer monotonically

Table 1. The confusion matrix of the species classification results on the test set. The numbers without parentheses are obtained by using Mel-scaled filter bank. The numbers in parentheses denote the changes after using optimized filter bank. For example, GAS was confused as MAT 32 times with Mel-scaled filter bank, but the confusion times reduced by 11 after the optimization.

		Classified (#)				
		BAS	DAB	GAS	MAT	DWA
Classes (#)	BAS	118(+1)	0(0)	1(-1)	0(0)	1(+1)
	DAB	2(0)	415(0)	13(-4)	13(-2)	1(+2)
	GAS	9(-5)	7(+2)	127(+3)	32(-11)	0(0))
	MAT	0(0)	0(-2)	3(-1)	301(+2)	0(0)
	DWA	1(+2)	9(-2)	3(-2)	2(0)	62(0)

increasing compared to the Mel-scaled filter bank. As mentioned before, the shifting of the center frequencies and changing of the bandwidths compared to their counterparts in the Mel-scaled filter bank cause the frequency resolutions at different frequencies to change. In the fbEM algorithm, the maximum likelihood criterion is used to raise or lower the frequency resolutions at certain frequencies such that more discriminative information for classification can be extracted from spectra. Therefore, a lower classification error rate can be achieved.

The bandwidths of the filters in both filter banks are small at low frequencies, which means more discriminative information for classification resides at low frequencies. The bandwidths of 1st, 2nd, 9th, 10th, and 15th filters in the optimal filter bank are small compared to other adjacent filters. The bandwidths of these filters are also significantly less ($> 25\%$) than their counterparts in the Mel-scaled filter bank. Thus, more discriminative information for classification may reside between 360 - 532, 1176 - 1458, and 2227 - 2552 Hz compared to other frequencies in the filter bank.

4. CONCLUSIONS

The fbEM algorithm offers an approach to jointly estimate filter bank parameters in feature extraction, and model parameters. Using the fbEM algorithm, the bird species classification accuracy on a large noisy corpus is increased by optimizing the center frequencies and bandwidths of the filter bank used in cepstral feature extraction. In the future, we will attempt to expand the work to speech recognition.

5. REFERENCES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[2] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annuals of Eugenis*, vol. 7, pp. 179–188, 1936.

[3] N. Kumar, "Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition," *PhD thesis*, 1997.

[4] M.J.F. Gales, "Maximum likelihood multiple subspace projections for hidden Markov models," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 37–47, 2002.

[5] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," *ICASSP*, 2005, vol. 1, pp. 961–964.

Table 2. Center frequencies (α_l^0 and $\hat{\alpha}_l$) and bandwidths (B_l^0 and $\hat{B}_l$) of the Mel-scaled and optimized filter bank, where $l = 1 \cdots L$. $L = 26$. Δ_l^α and Δ_l^B are percentage change as defined in Eqs. 19 and 20. The cut-off frequencies of the filter banks are: $\alpha_0^0 = \hat{\alpha}_0 =$ 360 Hz, $\alpha_{L+1}^0 = \hat{\alpha}_{L+1} = 6500$ Hz.

l	α_l^0 (Hz)	$\hat{\alpha}_l$ (Hz)	Δ_l^α (%)	B_l^0 (Hz)	$\hat{B}_l$ (Hz)	Δ_l^B (%)
1	438	415	-5.3	162	112	-30.4
2	522	472	-9.5	174	118	-31.8
3	611	532	-12.8	186	177	-4.8
4	708	650	-8.2	200	270	35.2
5	811	803	-1.0	215	250	16.3
6	923	899	-2.5	230	213	-7.6
7	1042	1016	-2.5	247	277	11.9
8	1170	1176	0.5	266	257	-3.2
9	1307	1273	-2.6	285	187	-34.5
10	1455	1363	-6.3	306	185	-39.6
11	1614	1458	-9.6	329	286	-12.9
12	1784	1649	-7.5	353	315	-10.8
13	1966	1772	-9.9	379	577	52.4
14	2162	2227	-3.0	407	595	46.4
15	2373	2368	-0.2	436	325	-25.5
16	2599	2552	-1.8	469	399	-14.9
17	2841	2767	-2.6	503	607	20.7
18	3102	3159	1.8	540	639	18.4
19	3381	3406	0.7	580	508	-12.4
20	3682	3667	-0.4	622	676	8.7
21	4004	4083	2.0	668	700	4.8
22	4350	4367	0.4	717	667	-7.0
23	4721	4750	0.6	770	829	7.6
24	5120	5196	1.5	827	788	-4.7
25	5547	5537	-0.2	887	886	-0.1
26	6007	6082	1.3	953	963	1.1

[6] S.S. Stevens, J. Volkman, and E.B. Newman, "A scale for the measurement of the psychological magnitude of pitch," *JASA*, vol. 8, no. 3, pp. 185–190, 1937.

[7] A.G. Andreou T. Kamm, H. Hermansky, "Learning the mel-scale and optimal VTN mapping," *Technical report, JHU/CLSP Workshop*, 1997.

[8] M. Graciarena, M. Delplanche, E. Shriberg, A. Stolcke, and L. Ferrer, "Acoustic front-end optimization for bird species recognition," *ICASSP*, 2010, pp. 293–296.

[9] A.L. McIlraith and H.C. Card, "Bird song recognition using back propagation and multivariate statistics," *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2740–2748, 1997.

[10] J.A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *JASA*, vol. 103, no. 4, pp. 2185–2196, 1998.

[11] V. Trifa, A. Kirschel, and C. E. Taylor, "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models," *JASA*, vol. 123, no. 4, pp. 2424–2431, 2008.

[12] W. Chu and A. Alwan, "A correlation-maximization denoising filter used as an enhancement frontend for noise robust bird call classification," *Proc. of Interspeech*, 2009, pp. 2831–2834.

[13] A. Satt and D. Malah, "Design of uniform DFT filter banks optimized for subband coding of speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1672–1679, 1989.