MCE(吳俊德實驗室提供)

The minimum classification error (MCE) criterion can also be used to optimize the temporal filter coefficients.

In the general formulation of MCE analysis, a classification error function $d_j(\cdot)$ is defined for a certain class *j*, an observation feature *X* that belongs to this class *j*, and a model set $\Lambda = \{\lambda_i, i = 1, 2, \ldots\ldots, J\}$, where λ_i is the model representing class i

$$\begin{aligned} &d_j(X, \Lambda) = -g(X, \lambda_j) + h(g(X, \lambda_i)), \\ &i = 1, 2, \ldots, J, i \neq j), X \in classj \end{aligned} \tag{1}$$

where *J* is the total number of classes or speech models ; $g(X, \lambda_j)$ is usually related to the class-conditioned likelihood $P(X \mid \lambda_j)$; $h(\cdot)$ is a function defining how the class-conditioned likelihoods $g(X, \lambda_i)$ for the competing models $\lambda_i, i = 1, 2, \ldots, J, i \neq j,$ are counted in the classification error function. This classification error function is often smoothed by a sigmoid function

$$l(d) = \frac{1}{1 + \exp(-\alpha(d - \beta))} \tag{2}$$

where α (here we set α=1) and β (β=0) define the slope and center of the sigmoid. As a result, in MCE, a total loss function defined as the smoothed classification error averaged over all the training data in all different classes is minimized

$$R_{MCE} = \sum_{j=1}^{J} \sum_{X \in classj}^{N_j} l(d_j(X, \Lambda)) = \min \tag{3}$$

where N_j is the number of patterns which belongs to class *j*.

In the temporal filtering problem discussed here, for the *k*th time trajectory we seek to derive a temporal filter impulse response $W_{k,MCE}$ that generates an optimal representation of the windowed segments $X_k(n)$ for the *k*th time trajectory, $x_k(n) = W_{k,MCE}^T X_k(n)$, which minimizes the above loss function $R_{k,mce}$ as defined in (3)

$$W_{k,MCE} = \arg\min_{W} R_{k,MCE} = \arg\min_{W} \sum_{j=1}^{J} \sum_{n=1}^{N_j} l\left(d_j\left(W_k^T z_k^{(j)}(n), \Lambda_k\right)\right) \quad (4)$$

$$\lambda_{j,k} = N\left(W_k^T \mu_k^{(j)}, W_k^T \Sigma_k^{(j)} W_k\right) \quad (5)$$

Which are those models in used in (4).

- Feature-Based MCE Filter

In this case, the classification error function in (1) is defined as

$$d_j\left(W_k^{(j)}X_k^{(j)}(n),\Lambda_k\right)=-\ln P\left(W_k^{(j)}X_k^{(j)}(n)\mid\lambda_{j,k}\right)+\ln\left\{\frac{1}{J-1}\sum_{\substack{m=1\\m\neq j}}^{J}P\left(W_k^{(j)}X_k^{(j)}(n)\mid\lambda_{m,k}\right)\right\} \tag{6}$$

$$R_{k,MCE}=\sum_{j=1}^{J}\sum_{n=1}^{N_j}l\left(d_j\left(W_k^TX_k^{(j)}(n),\Lambda_k\right)\right)$$

$$=\sum_{j=1}^{J}\sum_{n=1}^{N_j}l\left\{-\ln N\left(W_k^TX_k^{(j)}(n);W_k^T\mu_k^{(j)},W_k^T\Sigma_k^{(j)}W_k\right)\right.$$

$$\left.+\log\left[\frac{1}{J-1}\sum_{\substack{m=1\\m\neq j}}^{J}N\left(W_k^TX_k^{(j)}(n);W_k^T\mu_k^{(m)},W_k^T\Sigma_k^{(m)}W_k\right)\right]\right\} \tag{7}$$

Taking the (*)derivative of (7) with respect to w_k, we have

$$\frac{\partial R_{k,MCE}}{\partial W_k}=\sum_{j=1}^{J}\sum_{n=1}^{N}\frac{\partial l}{\partial d_j}\frac{\partial d_j\left(W_k^TX_k^{(j)}(n),\Lambda_k\right)}{\partial W_k} \tag{8}$$

Where

$$\frac{\partial l}{\partial d_j}=\alpha l(d_j)\left(1-l(d_j)\right) \tag{9}$$

$$\frac{\partial}{\partial W_k}d_j\left(W_k^TX_k^{(j)}(n),\Lambda_k\right)=\frac{\sum_{\substack{m=1\\m\neq j}}^{J}\left\{P\left(W_k^TX_k^{(j)}(n)\mid\lambda_{m,k}\right)\left[\frac{\partial}{\partial W_k}\ln P\left(W_k^TX_k^{(j)}(n)\mid\lambda_{m,k}\right)-\frac{\partial}{\partial W_k}\ln P\left(W_k^TX_k^{(j)}(n)\mid\lambda_{j,k}\right)\right]\right\}}{\sum_{\substack{m=1\\m\neq j}}^{J}P\left(W_k^TX_k^{(j)}(n)(n)\mid\lambda_{m,k}\right)} \tag{10}$$

Also

$$P\left(W_k^TX_k^{(j)}(n)\middle|\lambda_m\right)=N\left(W_k^TX_k^{(j)}(n);W_k^T\mu_k^{(m)},W_k^T\Sigma_k^{(m)}W_k\right) \tag{11}$$

(*)Proof will be shown in APPENDIX A

$$
\begin{aligned}
&{}^{(**)}\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n)\mid \lambda_{m,k}\right)\\
&=-\frac{\Sigma_k^{(m)}W_k}{W_k^T\Sigma_k^{(m)}W_k}-\frac{1}{\left(W_k^T\Sigma_k^{(m)}W_k\right)^2}((W_k^T\Sigma_k^{(m)}W_k)\\
&\times((z_k^{(j)}(n)-\mu_k^{(m)})(z_k^{(j)}(n)-\mu_k^{(m)})^T W_k)\\
&-(W_k^T(z_k^{(j)}(n)-\mu_k^{(m)})\\
&\times(z_k^{(j)}(n)-\mu_k^{(m)})^T W_k)\Sigma_k^{(m)}W_k)
\end{aligned}
\tag{12}
$$

Starting with an initial guess of, and with the help of (8)-(12), the gradient-descent algorithm can be used to obtain a better estimate of the temporal filter W_k for the $(t+1)$th iteration, $W_k(t+1)$, based on its estimate obtained from the tth iteration $W_k(t)$

$$
\overline{W_k}(t+1)=W_k(t)-\eta_t\frac{\partial R_{k,MCE}}{\partial W_k}|_{W_k=W_k(t)} \tag{13}
$$

Where η_t is the learning rate at the tth iteration, and

$$
W_k(t+1)=\frac{\overline{W_k}(t+1)}{|\overline{W_k}(t+1)|} \tag{14}
$$

Equation (14) is used here to normalize the norm of the vector representing the temporal filter to unity in order to be consistent with the eigenvectors used in LDA or PCA. The gradient-descent procedure terminates when there is no substantial difference between $W_k(t)$ and $W_k(t+1)$.

(*)Proof will be shown in APPENDIX B

STEP IN FEATURE-BASED MCE

Step1: Set initial $W=\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$

Step2: Update W

$$\overline{W}_k(t+1)=W_k(t)-\eta_t\frac{\partial R_{k,MCE}}{\partial W_k}\Big|_{W_k=W_k(t)}$$

where $$\frac{\partial R_{k,MCE}}{\partial W_k}=\sum_{j=1}^{J}\sum_{n=1}^{N}\frac{\partial l}{\partial d_j}\frac{\partial d_j\left(W_k^T X_k^{(j)}(n),\Lambda_k\right)}{\partial W_k}$$

$$\frac{\partial l}{\partial d_j}=\alpha l(d_j)\left(1-l(d_j)\right)$$

$$\frac{\partial}{\partial W_k}d_j\left(W_k^T X_k^{(j)}(n),\Lambda_k\right)=\frac{\sum_{\substack{m=1\\ m\neq j}}^{J}\left\{P\left(W_k^T X_k^{(j)}(n)\mid\lambda_{m,k}\right)\left[\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n)\mid\lambda_{m,k}\right)-\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n)\mid\lambda_{j,k}\right)\right]\right\}}{\sum_{\substack{m=1\\ m\neq j}}^{J}P\left(W_k^T X_k^{(j)}(n)(n)\mid\lambda_{m,k}\right)}$$

Since *numerator*=0 at *m*=*j*, index *m* in *numerator* is $\sum_{m=1}^{J}$ in the program.

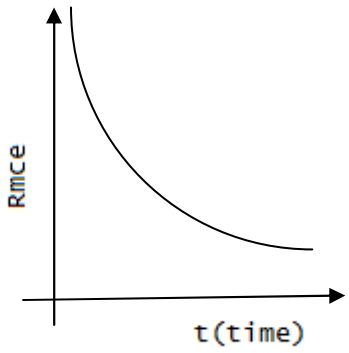
Step3: $W_k(t+1)=\frac{\overline{W}_k(t+1)}{|\overline{W}_k(t+1)|}$

Step4: Check

$$R_{k,MCE}=\sum_{j=1}^{J}\sum_{n=1}^{N_j}l\left(d_j\left(W_k^T X_k^{(j)}(n),\Lambda_k\right)\right)$$

$$=\sum_{j=1}^{J}\sum_{n=1}^{N_j}l\left\{-\ln N\left(W_k^T X_k^{(j)}(n);W_k^T\mu_k^{(j)},W_k^T\Sigma_k^{(j)}W_k\right)\right.$$

$$\left.+\log\left[\frac{1}{J-1}\sum_{\substack{m=1\\ m\neq j}}^{J}N\left(W_k^T X_k^{(j)}(n);W_k^T\mu_k^{(j)},W_k^T\Sigma_k^{(j)}W_k\right)\right]\right\}$$



where $l(d)=\frac{1}{1+\exp(-d_j)}$

Step5: Repeat step 2,3 until converge

- APPENDIX A

$$d_j\left(W_k^T X_k^{(j)}(n), \Lambda_k\right) = -\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{j,k}\right) + \ln\left\{\frac{1}{J-1}\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\right\}$$

$$= -\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{j,k}\right) + \left\{\ln\left(\frac{1}{J-1}\right) + \ln\left(\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\right)\right\}$$

$$\frac{\partial}{\partial W_k} d_j\left(W_k^T X_k^{(j)}(n), \Lambda_k\right) = -\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{j,k}\right) + \cancel{\frac{\partial}{\partial W_k}\ln\left(\frac{1}{J-1}\right)} + \frac{\partial}{\partial W_k}\ln\left(\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\right)$$

$$= -\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{j,k}\right) + \frac{\sum_{\substack{m=1 \\ m\neq j}}^{J} P'\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}{\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}$$

$$= -\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{j,k}\right) + \frac{\sum_{\substack{m=1 \\ m\neq j}}^{J}\left[P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\frac{P'\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}{P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}\right]}{\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}$$

where $\frac{P'\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}{P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)} = \frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)$

$$= \frac{\sum_{\substack{m=1 \\ m\neq j}}^{J}\left\{P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\left[\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\right]\right\} - \left\{\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\right\}\left[\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{j,k}\right)\right]}{\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}$$

$$= \frac{\sum_{\substack{m=1 \\ m\neq j}}^{J}\left\{P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)\left[\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right) - \frac{\partial}{\partial W}\ln P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{j,k}\right)\right]\right\}}{\sum_{\substack{m=1 \\ m\neq j}}^{J} P\left(W_k^T X_k^{(j)}(n) \mid \lambda_{m,k}\right)}$$

Since *numerator*=0 at *m=j*, index *m* in *numerator* is $\sum_{m=1}^{J}$ in the program.

- APPENDIX B

$$P\left(W_k^T X_k^{(j)}(n)\middle|\lambda_{m,k}\right)=N\left(W_k^T X_k^{(j)}(n);W_k^T\mu_k^{(m)},W_k^T\Sigma_k^{(m)}W_k\right)$$

$$\frac{\partial}{\partial W_k}\ln P\left(W_k^T X_k^{(j)}(n)\,|\,\lambda_{m,k}\right)$$

$$=\frac{\partial}{\partial W_k}\ln\left\{\frac{1}{\sqrt{2\pi W_k^T\Sigma_k^{(m)}W_k}}\times e^{-\frac{1}{2}\times\frac{\left(W_k^T X_k^{(j)}(n)-W_k^T\mu_k^{(m)}\right)^2}{W_k^T\Sigma_k^{(m)}W_k}}\right\}$$

$$=\frac{\partial}{\partial W_k}\left(-\frac{1}{2}\ln 2\pi\right)+\frac{\partial}{\partial W_k}\left(-\frac{1}{2}\ln\left(W_k^T\Sigma_k^{(m)}W_k\right)\right)-\frac{1}{2}\frac{\partial}{\partial W_k}\left\{\frac{\left[W_k^T\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\right]^2}{W_k^T\Sigma^{(m)}W_k}\right\}$$

$$=-\frac{1}{2}\frac{2\Sigma_k^{(m)}W_k}{W_k^T\Sigma_k^{(m)}W_k}-\frac{1}{2}\frac{\partial}{\partial W_k}\left[\frac{W_k^T\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)^T W_k}{W_k^T\Sigma_k^{(m)}W_k}\right]$$

$$=-\frac{\Sigma_k^{(m)}W_k}{W_k^T\Sigma_k^{(m)}W_k}-\frac{1}{2}\frac{1}{\left(W_k^T\Sigma_k^{(m)}W_k\right)^2}\left\{\left[W_k^T\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)^T W_k\right]'\times\left(W_k^T\Sigma_k^{(m)}W_k\right)-\right.$$

$$\left.\left[W_k^T\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)^T W_k\right]\left(W_k^T\Sigma_k^{(m)}W_k\right)'\right\}$$

$$=-\frac{\Sigma_k^{(m)}W_k}{W_k^T\Sigma_k^{(m)}W_k}-\frac{1}{2}\frac{1}{\left(W_k^T\Sigma_k^{(m)}W_k\right)^2}\left\{2\left[\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)^T W_k\right]\times\left(W_k^T\Sigma_k^{(m)}W_k\right)-\right.$$

$$\left.\left[W_k^T\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)^T W_k\right]\left(2\Sigma_k^{(m)}W_k\right)\right\}$$

$$=-\frac{\Sigma_k^{(m)}W_k}{W_k^T\Sigma_k^{(m)}W_k}-\frac{1}{\left(W_k^T\Sigma_k^{(m)}W_k\right)^2}\left\{\left[\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)^T W_k\right]\times\left(W_k^T\Sigma_k^{(m)}W_k\right)-\right.$$

$$\left.\left[W_k^T\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)\left(X_k^{(j)}(n)-\mu_k^{(m)}\right)^T W_k\right]\left(\Sigma_k^{(m)}W_k\right)\right\}$$