一．VTS（first order vector Taylor series ）

一种基于矢量泰勒级数(Vector Taylor Series, VTS)的模型自适应算法，利用矢量泰勒级数将环境改变前后模型参数之间的非线性变换关系展开，得到近似的一阶线性关系，代入似然函数，通过 EM(Expectation Maximization)算法计算 噪声的均值和方差

一阶矢量泰勒级数近似去提取噪音补偿ivector

一阶中心统计量在sVTS方法中起到重要作用。

标准ivector是$x^{(i)} \sim \pi_k N(\mu_{xk} + T_k \omega^{(i)}, \Sigma_{xk})$,

一个干净的mfcc向量被新增的卷积噪音扭曲成$y = x+h+g(n-x-h)$, 含噪语音特征向量$\boldsymbol{y}$ ，纯净语音特征向量 $\boldsymbol{x}$ ，信道噪声特征向量$\boldsymbol{h}$ 和加性噪声特征向量$\boldsymbol{n}$ ，g是非线性函数（$g = C\ln(1 + \exp(C^{\dagger}(n - x - h)))$ ，C是离散余弦变换矩阵，$C^{\dagger}$ 是他的伪逆矩阵，。）

在x，h，n均值处一阶矢量泰勒级数展开后有$\mu_{yk} \approx \mu$ xk0 $+\mu_{h0} +g(\mu_{n0}$ $-\mu$ xk0 $-\mu_{h0})$

$+G^{(i)}(\mu_{x} - \mu_{x}\ 0) + G^{(i)}(\mu^{(i)} - \mu^{(i)})$ k k k k h h0

$+F^{(i)}(\mu^{(i)} - \mu^{(i)})$, k n n0

$\boldsymbol{u}_{x,\,im}$ 是纯净语音 HMM 第i 个状态的第m 个高 斯单元的均值向量

对角矩阵。根据式(2)，可以得到对数谱域含噪语音 和纯净语音模型参数之间的变换关系:

$\boldsymbol{u}_{y,\,im} = \boldsymbol{u}_{x,\,im} + \boldsymbol{u}_h + \ln(1+\exp(\boldsymbol{u}_n - \boldsymbol{u}_{x,\,im} - \boldsymbol{u}_h))$

$S_{y,\ im} =(I\ -U_{im}) S_{x,\ im}(I\ -U_{im})\ +U_{im}S_{n}U_{i}$

_ _ _ _

其中S_n 是加性噪声的协方差矩阵，$S_{y,\ im}$ 是含噪语音 的协方差矩阵，式(5)中假设信道噪声为常数，不影 响$S_{y,\ im}$ 。噪声的均值和方差u_h ，u_n 和S_n 都是未知 数，需通过测试环境下的少量自适应数据来估计。

$$\begin{aligned}\boldsymbol{\mu}_{y,im} = \boldsymbol{\mu}_{x,im} + \boldsymbol{\mu}_{h,k-1} + \boldsymbol{C}\ln(1+\exp(\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{n,k-1} \\ - \boldsymbol{\mu}_{x,im} - \boldsymbol{\mu}_{h,k-1})))+(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})(\boldsymbol{\mu}_{h} - \boldsymbol{\mu}_{h,k-1}) \\ + \boldsymbol{U}_{im}^{k-1}(\boldsymbol{\mu}_{n} - \boldsymbol{\mu}_{n,k-1}) \qquad (9)\end{aligned}$$

$$\boldsymbol{\Sigma}_{y,im} = (\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\boldsymbol{\Sigma}_{x,im}(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}} + \boldsymbol{U}_{im}^{k-1}\boldsymbol{\Sigma}_{n}(\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}} \qquad (10)$$

being π_k, μ_{x_k}, and $\mathbf{\Sigma}_{x_k}$, the weight, mean, and covariance, respectively, of Gaussian k of a pre-trained GMM, the universal background model (UBM), $\mathbf{T}_k$ a low-rank matrix spanning a subspace referred to as total variability subspace that describes intersession variability in the space of GMM mean parameters, and $\omega^{(i)}$ a segment-specific low-dimension latent variable with standard normal distributed prior.

The training of this model is performed via maximum likelihood (ML) in two parts. Firstly, the UBM is pre-trained using the expectation-maximization (EM) algorithm, and π_k, μ_{x_k}, and $\mathbf{\Sigma}_{x_k}$ are obtained for all the Gaussians. Secondly, the sufficient statistics are computed as defined in [2] using fixed Gaussian alignments given by the UBM, and they are used for the training of the $\mathbf{T}_k$ matrices, which is also performed with the EM algorithm [2].

The iVector of utterance *i* is defined as the maximum a posteriori (MAP) point estimate of $\omega^{(i)}$. The posterior probability distribution of $\omega^{(i)}$ is Gaussian with mean, $\langle\omega^{(i)}\rangle$, and covariance, $\mathbf{L}^{(i)}$, and thus the iVector is equal to $\langle\omega^{(i)}\rangle$. The expressions to compute it are

$$\langle\omega^{(i)}\rangle = \mathbf{L}^{(i)} \sum_k \tilde{\mathbf{T}}_k^T \tilde{\mathbf{f}}_k^{(i)} \tag{2}$$

$$\mathbf{L}^{(i)} = (I + \sum_k N_{xk}^{(i)} \tilde{\mathbf{T}}_k^T \tilde{\mathbf{T}}_k)^{-1} \tag{3}$$

where $\mathbf{\Sigma}_{xk} = \mathbf{P}_{xk}\mathbf{P}_{xk}^T$, with $\mathbf{P}_{xk}$ lower triangular by Cholesky decomposition, $\tilde{\mathbf{T}}_k = \mathbf{P}_{xk}^{-1}\tilde{\mathbf{T}}_k$, and

$$N_{xk}^{(i)} = \sum_t \gamma_{xt}^{(i)}(k), \quad \tilde{\mathbf{f}}_k^{(i)} = \mathbf{P}_{xk}^{-1} \sum_t \gamma_{xt}^{(i)}(k)(\mathbf{x}_t^{(i)} - \mu_{xk}) \tag{4}$$

are the zeroth and *whitened* first order sufficient statistics pre-collected using the UBM as proposed in [16]. The first order statistic *whitening* ($\mu_{xk}^{(i)}$ subtraction and multiplication by $\mathbf{P}_{xk}^{-1}$) not only leads to a more efficient implementation, but it also plays an important role in the sVTS approach described in section 2.3.

2.2. VTS-Based iVector System for Noisy Environments

According to the model of the environment presented in [9], a clean MFCC vector affected by additive and convolutional noise is distorted as

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + g(\mathbf{n} - \mathbf{x} - \mathbf{h}), \tag{5}$$

where $\mathbf{y}$, $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{n}$ are the cepstral vectors of the noisy speech, clean speech, channel, and additive noise, respectively, and g is the nonlinear function defined as

$$g = \mathbf{C}\ln(1 + exp(\mathbf{C}^\dagger(\mathbf{n} - \mathbf{x} - \mathbf{h}))), \tag{6}$$

with $\mathbf{C}$ and $\mathbf{C}^\dagger$ the discrete cosine transform matrix and its pseudo-inverse, respectively. The corresponding relationship in the model space for the UBM means [11], assuming that both types of noise follow a Gaussian distribution, is approximated by a first order VTS expansion at $(\mu_{\mathbf{x_k}0}, \mu_{\mathbf{h}0}, \mu_{\mathbf{n}0})$,

$$\begin{aligned} \mu_{y_k}^{(i)} \approx \mu_{x_k0} + \mu_{h0}^{(i)} + g(\mu_{n0}^{(i)} - \mu_{x_k0} - \mu_{h0}^{(i)}) \\ + \mathbf{G}_k^{(i)}(\mu_{x_k} - \mu_{x_k0}) + \mathbf{G}_k^{(i)}(\mu_h^{(i)} - \mu_{h0}^{(i)}) \\ + \mathbf{F}_k^{(i)}(\mu_n^{(i)} - \mu_{n0}^{(i)}), \end{aligned} \tag{7}$$

where $\mathbf{G}_k$ is the Jacobian of g with respect to $\mathbf{x_k}$, and with respect to $\mathbf{h}$, and $\mathbf{F}_k$ with respect to $\mathbf{n}$. They are defined as

$$\mathbf{G}_k^{(i)} = \mathbf{C} \cdot diag(\frac{1}{1 + exp(\mathbf{C}^\dagger(\mu_{n0}^{(i)} - \mu_{x_k} - \mu_{h0}^{(i)}))}) \cdot \mathbf{C}^\dagger, \tag{8}$$

$$\mathbf{F}_k^{(i)} = \mathcal{I} - \mathbf{G}_k^{(i)}. \tag{9}$$

To compute the means of the noise-adapted UBM, $\mu_{\mathbf{y_k}0}$, the VTS is evaluated at $(\mu_{\mathbf{x_k}} = \mu_{\mathbf{x_k}0}, \mu_{\mathbf{h}} = \mu_{\mathbf{h}0}, \mu_{\mathbf{n}} = \mu_{\mathbf{n}0})$,

$$\mu_{y_k0}^{(i)} \approx \mu_{x_k0} + \mu_{h0}^{(i)} + g(\mu_{n0}^{(i)} - \mu_{x_k0} - \mu_{h0}^{(i)}) \tag{10}$$

The relationship of the UBM covariances [11], following the same reasoning as for the mean, is

$$\mathbf{\Sigma}_{y_k} \approx \mathbf{G}_k^{(i)}\mathbf{\Sigma}_{x_k}\mathbf{G}_k^{(i)T} + \mathbf{F}_k^{(i)}\mathbf{\Sigma}_n^{(i)}\mathbf{F}_k^{(i)T}, \tag{11}$$

where $\mathbf{\Sigma}_n^{(i)}$ is the additive noise covariance matrix, and $\mathbf{\Sigma}_h^{(i)}$ is set to zero since the channel is considered to be fixed. Finally, the mean and covariance of the model for the noisy MFCC first derivative (Δ) are calculated with the continuous-time approximation also used in [11]. That is,

$$\mu_{\Delta y_k}^{(i)} \approx \mathbf{G}_k^{(i)}\mu_{\Delta x_k}^{(i)} \tag{12}$$

$$\mathbf{\Sigma}_{\Delta y_k}^{(i)} \approx \mathbf{G}_k^{(i)}\mathbf{\Sigma}_{\Delta x_k}\mathbf{G}_k^{(i)T} + \mathbf{F}_k^{(i)}\mathbf{\Sigma}_{\Delta n}^{(i)}\mathbf{F}_k^{(i)T}, \tag{13}$$

and identically for the MFCC second derivative (Δ^2), substituting Δ by Δ^2.

One important role of the VTS approximation is to make the EM objective function of the noise-adapted UBM differentiable, so closed form update formulae of the model parameters are obtained. As per [8] the objective function becomes

$$\begin{aligned} Q = \sum_i \sum_t \sum_k \gamma_{yt}^{(i)}(k)[-\frac{1}{2}\ln|\mathbf{\Sigma}_{y_k}^{(i)}| \\ -\frac{1}{2}(\mathbf{y}_t^{(i)} - \mu_{y_k0}^{(i)})^T(\mathbf{\Sigma}_{y_k}^{(i)})^{-1}(\mathbf{y}_t^{(i)} - \mu_{y_k0}^{(i)})], \end{aligned} \tag{14}$$

In order to include the total variability subspace in the model of the noisy MFCC of every utterance, $\mathbf{y}^{(i)}$, μ_{x_k} is substituted by $\mu_{x_k0} + T_k\omega^{(i)}$ in (7), and also considering (11), it can be shown that

$$\mathbf{y}^{(i)} \sim \sum_k \pi_k \mathcal{N}(\mu_{y_k0}^{(i)} + \mathbf{G}_k^{(i)}\mathbf{T}_k\omega^{(i)}, \mathbf{\Sigma}_{y_k}^{(i)}). \tag{15}$$

This model is trained using the EM algorithm and the equations are detailed in [8].

2.3. Simplified VTS

The major drawback of the VTS approach presented in previous section is the computational cost of the EM training algorithm for the total variability subspace $\mathbf{T}_k$ of (15). In particular, in the *M step* the computation of the Kronecker product and large matrix inversion given in equation (18) of [8] is several orders of magnitude more computationally and memory demanding than the calculations required for training the standard model of (1). The main differences between the two techniques are that in the VTS approach the UBM mean and covariance are utterance-dependent, and that the total variability subspace is adapted to noise differently for each utterance through the term $\mathbf{G}_k^{(i)}\mathbf{T}_k$ in (15).

In [12], a new approach is proposed that largely simplifies the equations and reduces the computational cost, the sVTS. In the sVTS, first, the UBM is adapted to each file as described in section 2.2. Then, the zeroth and *whitened* first order sufficient statistics of utterance *i* are collected over its noise-adapted UBM as

$$N_{yk}^{(i)} = \sum_t \gamma_{yt}^{(i)}(k), \quad \tilde{\mathbf{f}}_{yk}^{(i)} = \mathbf{P}_{yk}^{(i)-1} \sum_t \gamma_{yt}^{(i)}(k)(\mathbf{y}_t^{(i)} - \mu_{y_k}^{(i)}), \tag{16}$$

SimplifiedVTS

VTS 与 ivector 主要不同在于在 VTS 中计算 ubm 均值和方差是基于话语的，对每段话的噪音的适应是不同的。

之前的 VTS 的问题是 T 计算量大，于是提出 sVTS。在 ubm 部分与上面一样，但一阶统计量与一阶中心统计量发生变化，用含噪语音代替它是一种统计量层面的补偿，而 VTS 是模型层面的补偿

两个MAP

对一个新说话人，eigenvoice MAP 在语音识别之前先进行无监督说话人适应。类似的，eigenchannel MAP 用信道补偿的先验分布使说话人的GMM适应测试语句，用这些自适应模型来计算似然得分来进行说话人识别。

本征音模型可以用相当少数量的本征音描述大多数说话人之间的变异，本征信道模型以用相当少数量的本征信道描述大多数说话人内部的变异

Whereas the number of eigenvoices that can be estimated from a given training set is bounded by the number of training speakers (which is generally insufficient), the number of eigenchannels that can be estimated is bounded only by the number of conversation sides (which is probably more than enough).

For each let be the speaker independent mean vector associated with the mixture compo- nent and, for each speaker , let denote the correspond- ing speaker-dependent mean vector. The MAP approach to speaker modeling assumes that for each mixture component and speaker , there is an unobservable offset vector such that

compensations ought to be chosen in such a way as to permit adaptation at recognition time to channels that have not previ- ously been seen. Note that eigenvoice modeling lends itself eas- ily to adapting previously unseen speakers at recognition time because it uses a prior in which speakers are statistically inde- pendent and identically distributed. So for purposes of channel compensation it is natural to use a prior in which the channel compensations for all speakers and channels are independent and identically distributed.

To be more specific, let denote the mean vector cor- responding to a speaker , a recording and a mixture compo- nent and let be an (unobservable) vector such that

公式(1)

and that the prior distribution of the matrix Point estimates of the speaker-dependent mean vector can be obtained by calculating the mode of the posterior distribution (that is, the distribution obtained by conditioning on the training data) of the matrix .