基于感知加权线谱对距离的最小生成误差语音合成模型训练方法

雷 鸣 凌震华 戴礼荣

(中国科学技术大学 电子工程与信息科学系 讯飞语音实验室 合肥 230027)

摘 要 提出一种基于感知加权线谱对(Line Spectral Pair, LSP)距离的最小生成误差(Minimum Generation Error, MGE)模型训练方法,用以改善基于隐马尔科夫模型的参数语音合成系统性能.在采用线谱对参数表征语音频谱特征时,传统 MGE 训练中使用的欧氏距离生成误差计算方法并不能较好地反映生成频谱与自然频谱之间的真实距离,而采用与谱参数无关的对数谱间距(Log Spectral Distortion, LSD)定义的生成误差函数可改善这一问题,但改进后主观效果不明显,且运算复杂度很高.文中先提出基于加权 LSP 距离的 MGE 模型训练方法,并在实验中从主客观对比不同加权方法以及基于 LSD 的 MGE 训练.最后,找到一种感知加权方法,不但具有较好的主观表现,而且在运算复杂度上与传统 MGE 训练相比几乎没有增加.

关键词 语音合成,隐可尔科夫模型(HMM),最小生成误差(MGE),感知加权,线谱对参数

中图法分类号 TN912.33

Minimum Generation Error Training Based on Perceptually Weighted Line Spectral Pair Distance for Statistical Parametric Speech Synthesis

LEI Ming, LING Zhen-Hua, DAI Li-Rong

(*iFly Speech Laboratory, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei* 230027)

ABSTRACT

A Minimum Generation Error (MGE) training method based on perceptually weighted Line Spectral Pair (LSP) distance is proposed to improve the performance of Hidden Markov Model (HMM) based parametric speech synthesis system. The generation error defined by Euclidean distance used in the traditional MGE training, is not eligible in measuring the real gap between generated spectrum and natural spectrum when the speech spectrum is described by LSP. Although using generation error defined by Log Spectral Distortion (LSD) having nothing to do with spectrum parameters manages to deal with this problem, the improvement seems trivial compared to the incurred higher computational complexity. In this paper, an MGE training criterion based on weighted LSP distance is proposed, and this MGE training method is subjectively and objectively contrasted with different weighted methods and LSD based

收稿日期:2009-02-07;修回日期:2009-10-26

作者简介 **雷鸣**,男,1985 年生,博士,主要研究方向为语音合成. E-mail: leiming@mail.ustc.edu.cn. **凌震华**,男,1979 年生,博士后,主要研究方向为语音合成. **戴礼荣**,男,1962 年生,教授,博士生导师,主要研究方向为语音合成、语音识别、语种识别、说话人识别、数字信号处理.

MGE training method. Eventually, a perceptually weighted training method is obtained, which not only achieves the best performance, but also incurs no extra computational complexity compared with the traditional MGE training.

Key Words Speech Synthesis, Hidden Markov Model (HMM), Minimum Generation Error (MGE), Perceptually Weighting, Line Spectral Pair Parameter

1 引　言

语音合成是智能人机交互领域的一个重要研究方向. 近十年来,基于隐马尔可夫模型(Hidden Markov Model, HMM)的统计建模参数语音合成方法被提出,并且得到迅速发展[1]. 在训练阶段,该方法使用统一的 HMM 框架对频谱、基频和时长等声学参数进行建模[2]. 在合成阶段,采用最大似然准则由统计模型预测声学参数[3],最终通过参数合成器重构语音. 该方法可合成平滑流畅的合成语音,具有系统构建自动化程度高、系统尺寸小、灵活性强等优点,体现出相对传统单元挑选与波形拼接语音合成方法的优势[4-5].

传统的基于 HMM 的参数语音合成系统中,声学模型的训练是基于最大似然(Maximum Likelihood, ML)准则进行的. 虽然已能够取得不错的合成效果,但是基于最大似然的模型训练准则却存在两个问题. 第一个问题是 HMM 模型训练算法与语音合成应用的不一致. 一般而言,语音合成的目标就是使生成的语音(参数)与自然语音(参数)尽可能地接近. 而现在采用的基于最大似然准则的 HMM 模型训练算法是从语音识别中借鉴过来的,它并非针对语音合成应用而设计,由此导致 HMM 模型训练算法与语音合成应用的不一致. 另一个问题是在参数生成过程中通过考虑动态和静态参数之间的约束来进行参数平滑. 而现在的训练过程中没有考虑到此约束条件,导致训练得到的 HMM 中静态和动态参数之间存在不一致. 最小生成误差(Minimum Generation Error, MGE)准则的提出正是为了解决 HMM 训练过程中的这两处不足[6]. MGE 训练首先定义一个生成误差函数,通过梯度下降算法最小化训练数据中模型的生成参数相对与自然参数的总生成误差来优化 HMM 参数. 该准则还在语音合成系统的上下文相关 HMM 聚类[7]和 HMM 模型自适应[8]中得到应用.

线谱对参数(Line Spectral Pair, LSP)是一种常用的描述语音频谱特征的参数,在语音编码、语音合成等领域有广泛应用[9]. 之前的实验结果也表明,LSP 参数相比 Mel 倒谱(Mel-Cepstrum)参数,在基于 HMM 的参数语音合成中能够获得更好的合成效果[10]. 但是,在传统的 MGE 训练中,生成误差被定义成预测频谱参数和自然频谱参数之间的欧氏距离,这种距离度量方式对 LSP 参数并不适用. 因为,首先,LSP 参数之间的欧氏距离并不能反映实际谱包络的之间的距离;其次,没有考虑人耳对于不同频段、共振峰的感知差异,这种生成误差函数的定义暗含了将 LSP 参数的每一维等价看待,但实际上对于主观听感而言,不同维的 LSP 参数之间应该是不等价的. 为了解决这些问题,在传统的 MGE 训练的基础上,新的生成误差函数定义被提出[11],它采用的是与谱参数无关的自然谱包络和生成谱包络之间的对数谱间距(Log Spectral Distortion, LSD)作为生成误差函数. 这种方法可得到对整个训练数据在谱包络上最逼近自然谱的 HMM 参数. 但由于运算复杂度很大,会大大延长 MGE 训练的时间. 而且这种方法同样没有考虑主观感知的影响,在效果上相对于传统 MGE 训练的提高并不显著.

为此,本文提出一种基于感知加权 LSP 距离的 MGE 模型训练方法. 通过一系列的主客观实验,对比不同加权策略,最终选择一种共振峰范围加权(Formant Bounded Weighting, FBW)的感知加权方法,使得在相对于传统 MGE 训练几乎不增加运算复杂度的情况下,在合成语音的主观测听中获得比基于 LSD 的 MGE 训练更好的主观表现. 同时,由于感知加权距离的生成误差函数定义对应的梯度下降公式形式上更简单,所以实际训练中会更稳定、快捷.

2 最小生成误差模型训练

2.1 参数生成算法

MGE 准则通过最小化模型生成参数相对自然参数的生成误差来指导模型的更新. 对于给定的 HMM 模型 λ,通常使用最大似然准则来进行声学参数的生成,即求出使 $P(\boldsymbol{o}|Q_{\text{opt}},\lambda)$ 最大的特征序列 $\boldsymbol{o}$

$=[o_1',o_2',\cdots,o_T']'$，最优状态序列 Q_{opt} 由时长模型依据待合成句的上下文信息进行预测. 这里，为了保证生成参数的平滑性，特征序列包含静态参数和其一阶、二阶差分的动态参数，如下所示：

$$\boldsymbol{o}_t = [c_t', \Delta^{(1)}c_t', \Delta^{(2)}c_t']'.$$

动态参数和静态参数之间的约束关系由 $\boldsymbol{o} = \boldsymbol{A}\boldsymbol{c}$ 决定，其中 $\boldsymbol{c} = [c_1', c_2', \cdots, c_T']'$，而 $\boldsymbol{A}$ 是一个由差分计算方法决定的系数矩阵[3].

在这种情况下，参数生成的过程等价于求出使 $P(\boldsymbol{o}|q,\lambda)$ 最大的参数序列 $\boldsymbol{c}$. 令

$$\partial P(\boldsymbol{o}|q,\lambda)/\partial \boldsymbol{c} = 0,$$

有

$$\bar{\boldsymbol{c}}_q = \boldsymbol{R}_q^{-1} r_q,$$

其中

$$\boldsymbol{R}_q = \boldsymbol{A}'\boldsymbol{\Sigma}_q^{-1}\boldsymbol{A}, \quad r_q = \boldsymbol{A}'\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}_q.$$

而 $\boldsymbol{\mu}_q = [\mu_1', \mu_2', \cdots, \mu_T']$ 和 $\boldsymbol{\Sigma}_q = diag(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_T)'$ 分别是与状态序列 Q_{opt} 相关的均值矢量和协方差矩阵. $\bar{\boldsymbol{c}}_q$ 就是生成的参数.

2.2 传统最小生成误差训练算法

在传统的 MGE 训练中[6]，生成误差函数被定义成原始声学参数和生成声学参数的欧氏距离：

$$\boldsymbol{D}_c(\boldsymbol{c} - \bar{\boldsymbol{c}}_q) = \sum_{t=1}^{T}\sum_{i=1}^{N}(c_{t,i} - \bar{c}_{t,i})^2.$$

对于最优的 HMM 状态序列 Q_{opt}，总的生成误差为

$$l(C,\lambda) = D(\boldsymbol{C}, \tilde{\boldsymbol{C}}(\lambda, Q_{opt})),$$

其中，λ 为模型参数，可以是均值参数或方差参数.

MGE 训练的目标就是通过最小化 $l(\boldsymbol{C},\lambda)$ 来优化模型参数，使总的生成误差最小，即

$$\bar{\lambda} = \text{argmin}\ l(\boldsymbol{C},\lambda).$$

但如前所述，对于使用 LSP 作为频谱参数的合成系统，基于欧氏距离定义的生成误差既不能反映原始谱包络和预测谱包络的差别，也不能反映主观听感上原始语音和预测语音的差别.

2.3 基于对数谱间距的最小生成误差训练算法

在这样的情况下，Wu 和 Tokuda 提出生成误差函数定义[11]，采用分别由原始谱包络和预测的谱包络之间的间距 LSD 作为生成误差函数. 由于最初的 LSD 是定义在频域上的积分，为了便于操作，将 LSD 写成在频域不同频率处采样和的形式，即生成误差为

$$D_{lsd}(\boldsymbol{c},\bar{\boldsymbol{c}}_q) = \sum_{t=1}^{T}\frac{1}{S}\sum_{s=1}^{S}[\log|A_c(\omega_s)| - \log|A_{\bar{c}}(\omega_s)|]^2,$$

其中，ω_s 是在频域上的采样点，可以是均匀间隔的采样点，也可以是在一些特殊位置的采样点. S 是采样点数. $A_c(\omega_s)$ 是由原始谱参数计算的谱包络，而 $A_{\bar{c}}(\omega_s)$ 是由预测的谱参数计算的谱包络.

由于共振峰通常分布在相邻阶 LSP 参数之间，因此通过在 LSP 频率处采样计算谱间距 LSD，也可以对共振峰起到加强作用，从而在主观听感上好于全频域均匀采样的 LSD 方法.

这种基于 LSD 的 MGE 训练是与谱参数本身无关的，虽然在谱间距上有很好地逼近，但其主观听感的改进效果相对于传统的 MGE 来说并不显著，而且由于更新每一维 LSP 参数时都需要计算多个采样点，因此运算会变得很复杂.

3 线谱对参数的加权距离度量

对于两组 LSP 参数的听感距离度量一直是语音领域的一个重要的研究内容[9]. 考虑了计算复杂度和度量效果，一般的度量方法采用加权欧氏距离的形式[12]，即

$$\boldsymbol{D}_c(\boldsymbol{c} - \bar{\boldsymbol{c}}_q) = \sum_{t=1}^{T}\sum_{i=1}^{N} w_{t,i}(c_{t,i} - \bar{c}_{t,i})^2.$$

但对于不同的度量策略，其度量表现也各有优劣，这里讨论一些常用的距离加权方法.

1) Laroia 和 Phamdo 等提出一种 LSP 参数的感知加权方法 (Inverse Harmonic Mean Weighting, IHMW)[13]. 对于 LSP 参数而言，每个共振峰都会分布在某两个相邻维的 LSP 参数之间，而且可通过这两维 LSP 参数之间的距离来在某种程度上度量对应共振峰的尖锐程度. 另一方面，考虑到人耳对共振峰的感知特性，越尖锐的共振峰对应着越好的主观听感. 因此，对于同一帧的 LSP 参数，当相邻维 LSP 参数越接近时，给予更大的权重就可获得更尖锐的共振峰，从而获得更好的主观听感. 因此其权重为

$$w_{t,i} = s_i^2\left(\frac{1}{c_{t,i} - c_{t,i-1}} + \frac{1}{c_{t,i+1} - c_{t,i}}\right),$$

其中，s_i 是一种人工设置的加权参数，这主要是由于人耳对低频较敏感，因此通过 s_i 来人为降低高频即高维 LSP 参数的权重，详见文献[13].

2) Chang 和 Ann 等提出局部谱近似加权方法 (Local Spectral Approximation Weights, LSAW)[13]. 这种方法将相邻维的 LSP 参数对应的线谱频率之间的谱包络近似为只与这两维 LSP 参数有关，因此简化了谱包络的计算公式，得到一种权重：

$$w_{t,i} = \begin{cases} s_i^2[2(1-\cos\frac{1}{2}(c_{t,2}-c_{t,1}))^{-\alpha}], & i=1 \\ s_i^2[(1-\cos\frac{1}{2}(c_{t,i}-c_{t,i-1}))^{-\alpha}+(1-\cos\frac{1}{2}(c_{t,i+1}-c_{t,i}))^{-\alpha}], & 2\leqslant i\leqslant N-1 \\ s_i^2[2(1-\cos\frac{1}{2}(c_{t,N}-c_{t,N-1}))^{-\alpha}], & i=N \end{cases}$$

其中,s_i 同前,α 是一种缩放因子,α 具体的确定方法可以在文献[13]找到.

3) 与前面的单纯考虑共振峰感知的 IHMW 或单独考虑一组谱包络的 LSAW 不同,Gardner 和 Rao 考虑两组 LSP 参数的谱包络间距 LSD,提出一种新的加权方法 Gardner 加权(Gardner Weighting, GW)[14]:

$$\boldsymbol{D}_c(\boldsymbol{c}) = 4\beta\boldsymbol{J}_c^{\mathrm{T}}(\boldsymbol{c})\hat{\boldsymbol{R}}_A\boldsymbol{J}_c(\boldsymbol{c}),$$

其中,$\hat{\boldsymbol{R}}_A$ 是与 LSP 参数对应的 LPC 参数所表示的 AR 滤波器的自相关矩阵,$\boldsymbol{J}_c(\boldsymbol{c})$ 是 LSP 参数到 LPC 参数的 Jacobian 矩阵,β 是常数,其值见文献[14]. 其权重就是矩阵 $\boldsymbol{D}_c(\boldsymbol{c})$ 的主对角线上的元素.

4) 在 GW 方法考虑客观距离 LSD 以及 IHMW 考虑共振峰感知的基础上,Lee 和 Kim 等进一步考虑共振峰感知的影响,提出另一种 LSP 参数的 FBW 方法[12]. 对于同一帧语音数据而言,其 N 阶 LSP 参数的谱包络表示和 $N-1$ 阶 LSP 参数的谱包络表示与共振峰的分布有一定的联系,单个共振峰一般分布在 N 阶 LSP 参数的某一维和 $N-1$ 阶 LSP 参数的某一维之间[15]. 在此基础上,用 N 阶 LSP 参数和对应的 $N-1$ 阶 LSP 参数就可进一步局限共振峰的分布范围,进而进行更好的感知加权,定义权重为

$$w_{t,i} = s_i'^2\left(\frac{1}{c_{t,i}^{(N)}-c_{t,i-1}^{(N-1)}}+\frac{1}{c_{t,i}^{(N-1)}-c_{t,i}^{(N)}}\right),$$

$$s_i' = \begin{cases} \alpha_i, & i=1,2 \\ s_i, & \text{otherwise} \end{cases}$$

其中,s_i 与 IHMW 中的 s_i 相同,而 α_i 的确定是选择使对于训练数据库而言 FBW 方法的权重均值和 GW 方法的权重均值得到

$$\frac{\bar{\boldsymbol{w}}^{fbw}\cdot\bar{\boldsymbol{w}}^{gw}}{\|\bar{\boldsymbol{w}}^{fbw}\|\,\|\bar{\boldsymbol{w}}^{gw}\|}$$

最大的一组 α_i. 这里加入 α_i 的意义在于 N 阶 LSP 参数和 $N-1$ 阶 LSP 参数的 1、2 维很接近,如果单纯考虑共振峰容易产生较大的不稳定权重,因此需要参考 GW 权重进行平衡.

4　基于加权距离的最小生成误差训练算法

为了在不降低效率的情况下实现谱包络距离上的生成误差计算,同时使得训练过程更加稳定、可控,我们将 LSP 参数的加权欧氏距离引入 MGE 训练中,提出基于加权 LSP 距离的 MGE 模型训练方法. 我们希望可通过某种加权的 MGE 训练,不但可以像基于 LSD 的 MGE 训练那样对原始谱包络逼近,还可以拥有较好的效率以及在主观听感上有不错的表现.

对于参数为 LSP 参数的 MGE 训练,如果定义其生成误差函数为 LSP 参数的加权欧氏距离,即

$$\boldsymbol{D}_{wc}(\boldsymbol{c}-\bar{\boldsymbol{c}}_q) = \sum_{t=1}^{T}\sum_{i=1}^{N} w_{t,i}(c_{t,i}-\bar{c}_{t,i})^2,$$

其中 N 是 LSP 参数的阶数. 就可以将 LSP 参数的加权欧氏距离度量引入到 MGE 训练中.

由于两组 LSP 参数之间的加权欧氏距离中的权重一般是由其中某一组 LSP 参数来计算. 对于 MGE 训练,如果我们选择用原始 LSP 参数计算,会使得权重更精确、合理. 另一方面,权重对生成参数即模型而言就相当于常数. 此时,有总生成误差

$$l(\boldsymbol{C},\lambda) = \boldsymbol{D}_{wc}(\boldsymbol{C},\hat{\boldsymbol{C}}(\lambda,Q_{\mathrm{opt}})).$$

采用 GPD 算法进行模型参数的更新,梯度为

$$\frac{\partial l(\boldsymbol{C},\lambda)}{\partial\lambda} = 2(\hat{\boldsymbol{C}}-\boldsymbol{C})'\boldsymbol{W}_{wc}\frac{\partial\hat{\boldsymbol{C}}}{\partial\lambda},$$

其中,$\boldsymbol{C} = [c_1',c_2',\cdots,c_T']'$ 是 $NT\times 1$ 的矢量. 而 $\boldsymbol{W}_{wc}$ 是 $NT\times NT$ 的权重矩阵,有

$$\boldsymbol{W}_{wc} = diag\{\boldsymbol{w}_{wc,1},\boldsymbol{w}_{wc,2},\cdots,\boldsymbol{w}_{wc,T}\},$$

$$\boldsymbol{w}_{wc,i} = diag\{w_{i,1},w_{i,2},\cdots,w_{i,N}\},\ 1\leqslant i\leqslant T,$$

其中 $diag\{\cdots\}$ 是生成对角阵.

对于均值参数,记 $\mu_{i,j}$ 为当前句中第 i 帧对应的 HMM 模型的第 j 维均值参数,则

$$\frac{\partial 1(\boldsymbol{C},\lambda)}{\partial\mu_{i,j}} = 2(\hat{\boldsymbol{C}}-\boldsymbol{C})'\boldsymbol{W}_{wc}\frac{\partial\hat{\boldsymbol{C}}}{\partial\mu_{i,j}},$$

根据参数生成算法,有

$$\frac{\partial \tilde{C}}{\partial \mu_{i,j}} = \boldsymbol{R}^{-1}\boldsymbol{A}'\boldsymbol{U}^{-1}\boldsymbol{Z}_{\mu},$$

其中，$\boldsymbol{Z}_{\mu}$ 是一个列矢量，元素为1时表示该帧数据在所有上下文相关模型中对应的模型为当前更新的模型[6].

所以，均值参数更新公式为

$$\mu_{i,j}(n+1) = \mu_{i,j}(n) - 2\varepsilon_n(\tilde{\boldsymbol{C}} - \boldsymbol{C}_n)'\boldsymbol{W}_{wc}\boldsymbol{R}^{-1}\boldsymbol{A}'\boldsymbol{U}^{-1}\boldsymbol{Z}_{\mu},$$

其中 ε_n 是步长.

对于方差参数，记 $V_{i,j} = (\sigma_{i,j}^2)^{-1}$ 为当前句中第 i 帧对应的HMM模型的第 j 维方差参数，由参数生成算法

$$\boldsymbol{R}\frac{\partial \tilde{C}}{\partial v_{i,j}} + \frac{\partial \boldsymbol{R}}{\partial v_{i,j}}\tilde{C} = \frac{\partial r}{\partial v_{i,j}},$$

因此

$$\frac{\partial \tilde{C}}{\partial v_{i,j}} = \boldsymbol{R}^{-1}\left(\frac{\partial r}{\partial v_{i,j}} - \frac{\partial \boldsymbol{R}}{\partial v_{i,j}}\tilde{C}\right),$$

有

$$\frac{\partial \tilde{C}}{\partial \boldsymbol{v}_{i,j}} = \boldsymbol{R}^{-1}\boldsymbol{A}'\boldsymbol{Z}_v(\boldsymbol{\mu} - \boldsymbol{A}\tilde{C}),$$

所以，方差参数更新公式为

$$v_{i,j}(n+1) = v_{i,j}(n) - 2\varepsilon_n(\tilde{\boldsymbol{C}}_n - \boldsymbol{C}_n)'\boldsymbol{W}_{wc}\boldsymbol{R}^{-1}\boldsymbol{A}'\boldsymbol{Z}_v(\boldsymbol{\mu} - \boldsymbol{A}\tilde{C}).$$

对于FBW加权方法，其权重中的 α_i 可以提高加权度量时的第1、2维稳定性[12]，实际上就是在GW的基础上结合共振峰的思想进行进一步的加权. 因此，为了进一步考虑GW加权方法所代表的谱包络逼近特性，将FBW权重中的 s_i' 修改为 $s_i' = \alpha_i$，$i = 1, 2, \cdots, N$，其中，α_i 的计算方法与原FBW中一样，但需要计算的 α_i 从1、2维的 α_i 扩大到所有 N 维的 α_i.

考虑到这种加权距离度量对数据更具有依赖性，容易造成过训练的情况，有可能当MGE训练在训练数据集上收敛的时候，在集外数据上反而不是最优的. 为了避免这种情况的发生，在MGE训练中加入与训练数据集无关的开发集进行检测，用总的谱间距作为一种客观度量标准，即不断的进行MGE迭代训练，当开发集上原始谱包络和预测谱包络的总谱间距的降低值小于一定门限时，将此时的模型作为最终的MGE训练结果. 训练流程如图1所示.

另外，相对于传统的MGE训练，各种方法都有不同程度的运算复杂度的增加. 对于训练数据库总帧数为 M 帧，LSP参数为 N 阶，进行 K 次迭代的不同加权方法来说，增加的运算复杂度如下：IHMW/LSAW 为 $O(NMK)$，GW 为 0，FBW 为 $O(NMK)$，LSD 为 $O(SMN^2K)$. GW方法的权重是提前离线计算. FBW方法的权重中的常数项采用提前离线计算出的GW权重计算，也是离线的，而另外的 $N-1$ 阶LSP参数也是对训练数据直接离线计算，都不增加运算复杂度. LSD方法作为对比，其中的 S 是采样点数. LSD方法是相对于模型参数求偏导，所以只能在线计算.

从结果可以看出，基于加权距离的MGE训练相对于传统MGE训练增加的运算度远小于基于LSD的MGE训练. 实际上，带来的运算复杂度增加相对于传统MGE训练而言几乎可以忽略不计，而LSD方法的 $O(SMN^2K)$ 的运算复杂度增加却会大大延长训练时间.

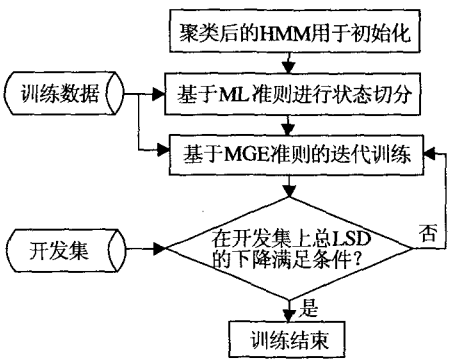

图1 基于感知加权LSP距离的MGE训练流程

Fig. 1 MGE training procedure based on perceptually weighted LSP distance

5 实 验

5.1 实验配置

实验中采用中文女声1 000句音库，其中随机选择50句作为开发集，剩余950句作为训练数据. 原始数据为16kHz采样率，建模采用5状态无跳转的HMM模型，使用的频谱特征包括STRAIGHT分析的40阶LSP参数和增益，以及对应的一阶二阶差分参数，分析帧长为5ms.

HMM的基于ML准则的训练过程与文献[16]中的ML训练一致. 我们使用传统基于欧氏距离度量生成误差的MGE训练[6]作为基线系统.

5.2 客观实验结果

分别计算使用不同LSP距离加权方法进行MGE训练得到的最终模型在开发集上的生成频谱参数所表征的谱包络和自然谱包络的平均谱间距，

同时与 ML 训练以及基线系统训练结果进行对比,如图 2 所示.

从图 2 看出,GW 方法在开发集谱间距上的表现是最好的,FBW 方法次之,LSAW 比基线系统略好,而 IHMW 是加权方法中最差的. IHMW 比基线系统略差,其原因是考虑共振峰的参数更新方向与整体谱间距下降的方向可能并不一致,因此会在总的谱间距上有所损失.

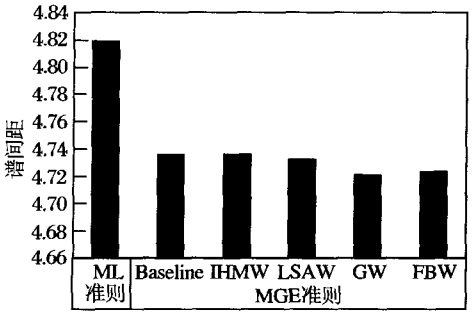

图 2　不同 MGE 训练方法对比

Fig. 2　Comparison among different MGE training methods

5.3　FBW 与 GW 对比

FBW 同 IHMW 一样,也考虑共振峰感知,其参数更新方向不一定是向着整体谱间距下降的方向,因此在总的谱间距上不如单纯考虑谱间距的 GW 方法. 但 FBW 是在 GW 方法的基础上,其总的谱间距有一定的保障,而且会形成更尖锐的共振峰,因此会对浊音表现较好,而单纯的谱间距的逼近会对浊音和清音等同对待.

为了确定 FBW 的共振峰感知加强对主观听感的实际作用,选择训练集和开发集外的 20 句进行合成,对 FBW 和 GW 方法进行主观测听实验. 由 5 名测听人员进行,对每组内打乱顺序的由两种方法合成的同一句语音,由测听人员选择出更自然的一个. FBW 方法和 GW 方法的主观倾向性测听结果如图 3 所示. 图中有端点的线覆盖的范围对应着统计均值的 95% 置信区间.

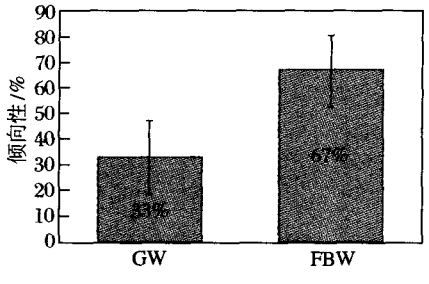

图 3　2 种方法主观倾向性测试结果对比

Fig. 3　Comparison of subjective preference test result between FBW and GW

从图中可以看出,FBW 方法明显好于 GW 方法,而且具有一致性,这种差别完全体现共振峰对主观听感的影响. 由此可见,单纯的基于谱包络的逼近并不能获得最好的主观听感,而在谱包络逼近的基础上再对共振峰进行加强,就会获得较好的主观听感. 在基于 LSD 的 MGE 训练的实验中[11]可看到与此类似的结论.

5.4　与基线系统的对比

对于 IHMW 方法和 LSAW 方法,其开发集上平均谱间距与基线系统相差不大. 而且虽然 IHMW 方法考虑共振峰的影响,但较粗糙. 而 FBW 方法是所有加权距离方法中既对原始谱包络有不错的逼近效果,也考虑共振峰感知的方法,理应具有最好的主观表现,所以 IHMW 方法和 LSAW 方法不会有比 FBW 方法更好的主观表现.

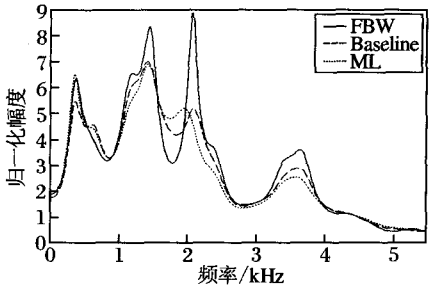

图 4　FBW 方法预测频谱包络与基线系统以及 ML 训练系统的对比

Fig. 4　Comparison of predicted spectrum envelopes among FBW, baseline and ML

图 4 是由 FBW 方法训练得到的模型对元音/a/预测的某一帧谱包络,同时画出基线系统和 ML 准则预测的同一帧的谱包络作为对比. 从图中可以看出,基线系统与 ML 相比并没有很明显的变化,而 FBW 预测的谱包络起伏更明显,共振峰更尖锐,明显优于基线系统和 ML 方法.

在此基础上,重新选择 20 句训练集和开发集外的句子分别用 ML 准则训练出的模型、基线系统训练出的模型以及 FBW 加权的 MGE 方法训练出的模型合成,由 5 名测听人员对合成的句子做两两之间的主观倾向性测试,结果见图 5.

从图中可以看出,基线系统和直接用 ML 准则训练的模型的对比符合[6]的实验结果,而 FBW 方法明显比基线系统即以 LSP 参数的欧氏距离定义

生成误差的 MCE 方法好，而且具有一致性. 这种差别不但体现共振峰的影响，也体现两种方法对于原始谱包络的逼近程度. 这说明加权距离的 MGE 训练相对于传统的 MGE 训练确实有较大的提升.

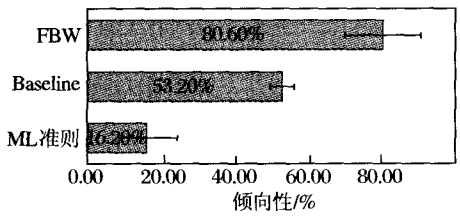

图 5 FBW 方法的主观评测

Fig. 5 Subjective evaluation of FBW

5.5 与基于对数谱间距的最小生成误差训练对比

在基于 LSD 的 MCE 方法中，如果选择在全频域均匀采样，就相当于只考虑谱间距，参数更新的方向与整体谱间距下降的方向一致，这对应着加权距离方法中的 GW 方法. 而如果选择在 LSP 频率处采样，就相当于对 LSP 频率附近的共振峰进行加强[11]，而且还部分考虑谱间距，这就对应着加权距离方法中的感知加权方法 FBW.

对于全频域采样的 LSD 方法和 GW 方法，MGE 训练结束时在开发集上的平均谱间距如图 6 所示. 图中 LSD512 就是在全频域均匀采样的 LSD 方法($S=512$)[11]. 从图中可以看出，两种方法在开发集上的平均谱间距很接近. 而在训练过程中，这两种方法在开发集上的平均谱间距关于迭代次数的变化情况如图 7.

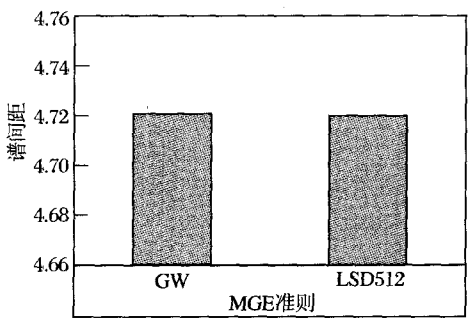

图 6 GW 与 LSD512 方法客观对比

Fig. 6 Objective comparison between GW and LSD512

这两种方法在开发集上的谱间距变化曲线几乎完全一样. 通过一些谱包络的对比和主观对比，可得知，这两种方法取得的效果是一致的. 更进一步来说，加权欧氏距离可达到对自然谱包络逼近的效果，而且跟直接用谱间距作为生成误差来进行逼近效果相当.

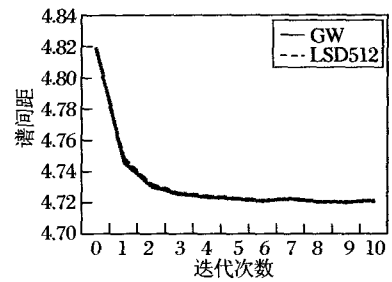

图 7 GW 与 LSD512 方法开发集谱间距变化

Fig. 7 Reduction of LSD on development set for GW and LSD512

同 FBW 方法与 GW 方法的对比类似，由于对共振峰进行加强，在 LSP 频率处采样的 LSD 方法会获得比在全频域均匀采样的 LSD 方法更好的主观听感[11]. 为了对比 FBW 方法和在 LSP 频率处采样的 LSD 方法的主观听感，再选择 20 句训练集和开发集外的句子，分别用两种方法合成，由 5 名测听人员做主观倾向性测听，结果如图 8 所示. 图中，LSD_LSP 就是在 LSP 频率处采样的 LSD 方法，由于实验中 LSP 参数为 40 维，因此采样点就是 40 个 LSP 频率点.

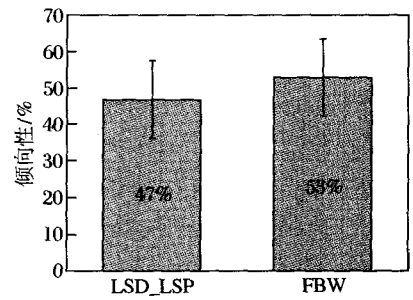

图 8 FBW 与在 LSP 频率处采样的 LSD 方法对比

Fig. 8 Comparison between FBW and LSD_LSP

从图中可以看出，FBW 方法比在 LSP 频率处采样的 LSD 方法略好，从置信区间看，存在整体分布的差别. 因此，FBW 方法比 LSD_LSP 方法主观听感要好一些. 这主要与 FBW 更精细地考虑了共振峰的分布范围，以及训练过程更简单、稳定有关.

5.6 讨论

单纯考虑谱间距时，所有加权方法中 GW 方法是最优的，FBW 方法次优. 而且 GW 方法与全频域均匀采样的 LSD 方法客观表现一样. 因此对于 MGE 训练，加权欧氏距离定义的生成误差函数可与直接

用谱间距定义的生成误差函数在谱间距上获得相同的效果.

FBW 方法在 GW 方法的基础上,不但对原始谱包络进行逼近,而且可对共振峰进行加强,是一种感知加权方法. 经过修改的 FBW 方法可获得更好的稳定性,虽然在开发集上的谱间距较 GW 方法略差,但主观评测表现明显好于其他所有方法,包括传统 MGE 方法、基于 LSD 的全频域采样和共振峰加强采样的 MGE 方法. 其原因一方面是 FBW 方法在谱包络上对原始谱包络的逼近效果较好,另一方面是共振峰感知考虑地更精细,而且训练简单、稳定.

6　结束语

本文将不同的加权 LSP 距离计算方法应用在最小生成误差语音合成模型训练中,通过对比开发集上的谱间距,发现加权 LSP 距离最小生成误差语音合成模型训练方法在客观上完全可达到基于 LSD 的最小生成误差模型训练方法的效果. 通过进一步感知加权,并与传统最小生成误差方法以及基于 LSD 的共振峰加强最小生成误差方法对比,发现感知加权的 FBW 方法在主观听感上比传统最小生成误差方法以及在 LSP 频率处采样的基于 LSD 的最小生成误差方法都要好. 而且通过实验及分析,所有加权距离方法包括 FBW 方法的运算复杂度与传统最小生成误差方法相当,明显少于基于 LSD 的最小生成误差模型训练方法,可大大减少训练时间. 因此,最终找到一种在客观度量上较好,但在主观表现上更好,而且运算复杂度相对于传统 MGE 训练几乎没有增加的感知加权方法,即 FBW 方法. 这种加权距离度量方法可较容易地用到其他使用 LSP 参数的合成模型训练过程中.

参　考　文　献

[1] Masuko T, Tokuda K, Kobayashi T, *et al*. Speech Synthesis Using HMMs with Dynamic Features // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Atlanta, USA, 1996, Ⅰ: 389 - 392

[2] Yoshimura T, Tokuda K, Masuko T, *et al*. Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Phoenix, USA, 1999, Ⅴ: 2347 - 2350

[3] Tokuda K, Kobayashi T, Imai S. Speech Parameter Generation from HMM Using Dynamic Features // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Detroit, USA, 1995, Ⅰ: 660 - 663

[4] Ling Zhenhua, Qin Long, Lu Heng, *et al*. The USTC and iFlytek Speech Synthesis Systems for Blizzard Challenge 2007 // Proc of the Blizzard Challenge Workshop. Bonn, Germany, 2007: 17 - 21

[5] Zen H, Toda T. An Overview of Nitech HMM-Based Speech Synthesis System for Blizzard Challenge 2005 // Proc of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal, 2005: 93 - 96

[6] Wu Yijian, Wang Renhua. Minimum Generation Error Training for HMM-Based Speech Synthesis // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse, France, 2006, Ⅰ: 889 - 892

[7] Wu Yijian, Guo Wu, Wang Renhua. Minimum Generation Error Criterion for Tree-Based Clustering of Context Dependent HMMs // Proc of the 9th International Conference on Speken Language Processing. Pittsburgh, USA, 2006: 2046 - 2049

[8] Qin Long, Wu Yijian, Ling Zhenhua, *et al*. Minimum Generation Error Linear Regressión Based Model Adaptation for HMM-Based Speech Synthesis // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA, 2008: 3953 - 3956

[9] McLoughlin I V. Line Spectral Pairs. Signal Processing Journal, 2008, 88(3): 448 - 467

[10] Wu Yijian, Wang Renhua. HMM-Based Trainable Speech Synthesis for Chinese. Journal of Chinese Information Processing, 2006, 20(4): 75 - 81 (in Chinese)
(吴义坚,王仁华. 基于 HMM 的可训练中文语音合成. 中文信息学报, 2006, 20(4): 75 - 81)

[11] Wu Yijian, Tokuda K. Minimum Generation Error Training with Direct Log Spectral Distortion on LSPs for HMM-Based Speech Synthesis // Proc of the 9th Annual Conference of the International Speech Communication Association. Brisbane, Australia, 2008: 577 - 580

[12] Lee M S, Kim H K, Lee H S. A New Distortion Measure for Spectral Quantization Based on the LSF Intermodel Interlacing Property. Speech Communication, 2001, 35(3/4): 191 - 202

[13] Laroia R, Phamdo N, Farvardin N. Robust and Efficient Quantization of Speech LSP Parameters Using Structured Vector Quantizers // Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada, 1991, Ⅰ: 641 - 644

[14] Gardner W R, Rao B D. Theoretical Analysis of the High-Rate Vector Quantization of LPC Parameters. IEEE Trans on Speech and Audio Processing, 1995, 3(5): 367 - 381

[15] Kim H K, Lee H S. Interlacing Properties of Line Spectrum Pair Frequencies. IEEE Trans on Speech and Audio Processing, 1999, 7(1): 87 - 91

[16] Ling Zhenhua, Wu Yijian, Wang Yuping, *et al*. USTC System for Blizzard Challenge 2006 an Improved HMM-Based Speech Synthesis Method [EB/OL]. [2006-09-16]. http://citeseerx.ist.psu.edu/viewdoc/download? doi = 10.1.1.130.7143&rep = rep1&type = pdf