一种改进的基于分析合成框架的语音增强算法

刘斌[1]，陶建华[1]，莫福源[2]
(1. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190；
2. 中国科学院声学研究所 北京 100190）

文 摘： 本文提出了一种基于分析合成框架的语音增强算法；应用一种改进的基于多带梳状滤波方法计算基音周期并判定各个子带的清浊状态；相对于不同基线方法，改进后的算法具有更低基音周期估计误差和更高的清浊判定准确率；引入深层神经网络模型增强线谱对参数，通过该模型重构的线谱对参数误差低于传统方法；将改进基音周期估计方法和线谱对增强方法应用到基于分析合成框架的语音增强算法中，实验结果表明，这种基于分析合成框架的语音增强算法的性能优于各种基线方法，集外测试具有更高的 PESQ 得分。此外上述改进的方法可以直接应用到参数化语音编码算法中，尤其可以改善噪声环境下低速率语音编码的音质。

关键词：分析合成框架；多带梳状滤波；深层神经网络模型；语音增强
中文图书分类号：TP391.

单通道语音增强是语音信号处理领域的一个重要分支，它在语音识别、语音编码、说话人识别等许多系统中有着广泛的应用，通常作为系统的预处理阶段；通过有效的语音增强算法可以改善语音的音质和可懂度。然而从各种复杂噪声环境下对语音信号进行增强处理一直以来都是一个非常具有挑战性的课题，尤其是在低信噪比时、面临非平稳噪声条件下这一问题尤为突现。

一些单通道语音增强方法相继提出，其中最为流行的是谱减法[1]、最小均方误差法[2]和基于听觉场景分析的语音增强算法[3]。谱减法的缺陷是增强的语音信号会带有音乐噪声；在各种语音增强算法中，基于最小均方误差的语音增强方法得到了广泛的关注，它在一些噪声环境下表现出较好的性能。主要原因有两个方面，首先，该方法结合听觉感知重点对幅值谱进行准确的估计，考虑到相位谱对人耳听感较弱，因此并没有对相位谱进行增强；其次，该方法基于贝叶斯准则，充分利用先验知识对幅值谱进行估计；但是这种方法在低信噪比和非平稳噪声条件下性能明显下降。针对这一问题，有人提出了基于分析合成框架的语音增强方法，该方法首先在噪声环境下对语音信号的基频、谱包络和能量进行相对准确的估计，语音信号在参数域进行增强，并通过基频、谱包络和能量重新合成语音，增强的语音信号是通过声码器重构得到的，因此各种噪声可以自动消除。基于分析合成框架的语音增强算法的优势在于它在修复谐波结构的同时可以有效的去除音乐噪声的干扰。为了更加准确的对语音信号进行分析，在提取语音参数前需要对语音信号进行预增强。对于基频和能量增益，即使在很低信噪比的情形下仍然可以从预增强的语音信号中准确的计算，然而大多数预增强算法难以保留谱包络特征；通过增强谱包络信号，可以有效的提高合成语音的质量和可懂度。在噪声环境下提取鲁棒的语音参数对提高语音增强算法性能至关重要。基于分析合成框架的语音增强问题可以分解为噪声环境下的基音周期估计和子带清浊判决、噪声环境下的谱包络增强、噪声环境下语音信号的增益估计和声码器重构增强语音四个子问题。

噪声环境下的基音周期估计算法可以分为三类：基于时域的基音周期估计[4]、基于频域的基音周期估计[5]和基于时频域的基音周期估计[6]。基于时域的基音周期估计直接分析信号在时间轴上的周期性，基于频域的基音周期估计通过在频域上分析语音短时谱的谐波特性来估计基音周期，基于时频域的基音周期估计通常将信号划分成多个子带，然后分别在各个子带中进行时域分析。基于多带梳状滤波的基音周期估计是一种主流的基于时频域的基音周期估计方法，这种方法考虑了人耳的听觉感知特性，对噪声环境具有很好的鲁棒性。基于这种方法进行改进，有人提出基于加权信噪比的多带梳状滤波方法对基音周期进行估计[7]，该方法可以进一步提高噪声环境下基音周期估计的精度。对于浊音度较高的子带，通常会呈现明显的共振峰结构，可以重点利用这些子带更有效的改进基音周期估计算法。

谱包络增强问题可以描述为在噪声环境下对安静语音的谱包络进行估计；主要的方法包括维纳滤波[8]、卡尔曼滤波[9]、高斯混合模型、人工神经网络等。维纳滤波跟线性预测相关，安静语音的谱包络能够在噪声环境下通过迭代计算估计；此外，卡尔曼滤波在语音增强中得到了广泛的应用，可以利用卡尔曼滤波器对线谱对参数的时间轨迹进行

跟踪，增强的线谱对参数可以直接应用到基于分析合成框架的语音增强算法中，提高谱包络估计的性能，进而提高增强语音的音质。高斯混合模型是一种在语音转换[10]和人工带宽扩展[11]中广泛应用的模型，这种数据驱动的方法可以有效的对谱包络信号进行重构。深度学习方法是目前在机器学习领域中非常主流的方法[12]，它能够发现不同特征之间的潜在规则，相对于浅层模型，它具有更强的泛化能力。训练多层神经网络的基本策略是首先逐层进行无监督预训练，预训练完成后再进行多层联合有监督训练。深层神经网络模型可以直接应用到谱包络增强中，通过该模型建立带噪语音谱包络和安静语音谱包络之间的映射关系。Denoising autoencoder（DAE）是一种典型的多层神经网络结构[13]，可以用于谱包络增强，该模型是基于最小重构误差准则训练得到的。

本文介绍一种基于分析合成框架的语音增强技术对带噪语音进行增强。带噪语音首先执行预增强，通过预增强可以初步滤除一些噪声信号，预增强后的信号更适合分析合成框架的语音增强算法。在文献[14]的基础上，对基于多带梳状滤波的基音周期检测算法（MBSC）进行改进。改进后的算法只考虑具有明显共振峰特性的子带进行加权，同时低频带的线性预测残差信号也用于基音周期计算；采用 DAE 模型对反应谱包络特性的线谱对参数进行增强处理，从而降低了语音信号谱包络的重构误差，提高了语音的音质和可懂度；所提算法利用分析合成框架可以有效的消除音乐噪声，这种方法在带噪语音和安静语音之间通过深层神经网络结构建立映射模型，在噪声得到抑制的同时，谱包络可以得到有效修复。本文第一部分重点阐述所提出的算法，第二部分分析实验结果，结论将在第三部分阐述。

1　基于分析合成框架的语音增强算法

基于分析合成框架的语音增强算法流程如图 1 所示。

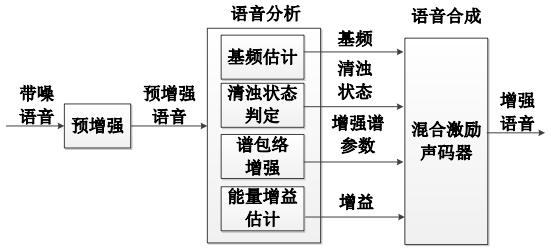

图 1　基于分析合成框架语音增强算法原理框图

算法共包括六个部分：预增强、基音周期计算、子带清浊判决、谱参数增强、谱增益估计和声码器重构语音。在预增强阶段，通过 logMMSE 算法估计带噪语音信号的短时幅值谱得到预增强后的语音信号。然后通过改进的 MBSC 算法估计基音周期并判定各个子带的清浊状态。通过 DAE 模型对反映语音谱包络特征的线谱对参数进行增强处理。在对预增强的语音计算谱增益时，同时考虑基音周期同步问题。将输出的基音周期、子带清浊状态、线谱对参数和能量增益等语音参数通过混合激励声码器重构增强语音。以下介绍算法的细节流程。

1.1 语音预增强

首先通过 logMMSE 对原始带噪语音信号进行预增强处理。设原始带噪语音信号为 *y(t)*，所对应的安静环境下的语音信号为 *x(t)*；logMMSE 算法是在已知带噪语音 *y(t)*的条件下对安静语音的傅里叶幅值谱进行估计。算法中假设语音信号和噪声信号在不同频点处的傅里叶幅值均服从高斯分布。设 $X_k = A_k e^{j\alpha k}$， D_k 和 $Y_k = R_k e^{j\beta k}$ 分别表示安静语音信号、噪声信号和带噪语音信号的第 k 个傅里叶幅值系数，算法的目标是使得安静语音幅值谱的估计值 $\hat{A}_k$ 尽可能接近真实值，通过最小化对数幅值谱的误差准则进行优化。相对于传统的 MMSE，logMMSE 算法具有更低的残差噪声，噪声得到抑制的同时语音信号的损伤并没有增加[15]。原始带噪语音首先通过 logMMSE 进行预增强处理，此时语音信号可以得到近似的重构，但是通过这种方法输出得到的增强语音仍然有不小的损伤，尤其是在非平稳噪声条件下。修复这些受损语音对提高音质和可懂度来说都是非常重要的。

1.2 基音周期估计和子带清浊判定

基音周期估计和子带清浊判定的流程如图 2 所示。

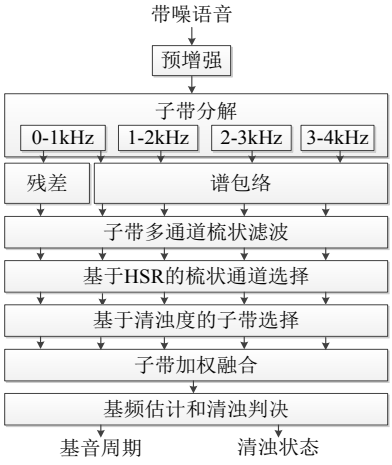

图 2　基音周期估计和子带清浊判决算法流程

输入语音信号首先通过 32 阶 FIR 滤波器分解为 4 个子带，即 0-1000Hz、1000-2000Hz、2000-3000Hz、3000-4000Hz。各个子带的带宽为 1kHz，此时每个子带至少可以捕获到两个谐波成分。对于四个子带信号分别进行希尔伯特变换，计

算子带的希尔伯特包络；针对第一个子带(0-1000Hz)，本文同时考虑希尔伯特包络和线性预测残差，线性残差的引入可以有效的提高基音周期检测的性能，尤其是针对包含有低频带噪声的语音，这一优势更为明显。

对于各个子带分别进行多通道梳状滤波，梳状滤波器的作用是在增强主谐波谱能量的同时抑制子谐波谱能量。主谐波和子谐波的能量比（HSR）是非常重要的参数，通过该参数可以在各个子带中选择更加有效的梳状滤波通道。子带 s 的第 k 个通道的 HSR 可以表示为

$$q_{s,t}(k)=\sum_{f}|X_{k,s,t}(f)c_k(f)|^2/\sum_{f}|X_{k,s,t}(f)(1-c_k(f))|^2 \quad (1)$$

其中 $X_{k,s,t}(f)$ 和 $c_k(f)$ 分别表示傅里叶系数和梳状滤波系数。

各个子带中的梳状滤波通道选择分为三个阶段，基于 HSR 和自相关系数（ACR）进行通道选择，通道选择的详细过程参照文献 14。完成通道选择后，对所选通道基于 HSR 进行加权处理；这种加权机制的引入，使得权重更高的通道中的 ACR 在相应子带中作用更大；所对应的 ACR 在基音周期估计中起到了更加主导的作用。

为了提高噪声环境下的基音周期估计和子带清浊状态判决的性能，本文对原有算法进行改进，只保留浊音强度较高的子带进行基音周期估计。各个子带的清浊强度可以根据相应子带中所选通道计算得到；子带的清浊决策根据固定阈值确定。本文针对三个子带进行子带选择（1-2k，2-3k，3-4k）。由于 0-1k 子带中包含重要的共振峰结构，因此在基音周期检测和子带清浊判决过程中一直保留。通过这种子带选择机制可以有效的消除周期性噪声的干扰。

在完成子带选择后，采用同样的方法对所选子带进行加权，这种加权机制可以有效的减少带噪语音最大峰值幅度的波动。10 个最高峰值点处的基频作为候选，每个峰值和它毗邻的峰值通过抛物线固定；确定幅值和抛物线最大点的滞后位置，作为基音周期的度量值。对于子带清浊度判定，通过阈值完成初始判决后，在利用前后四帧进行平滑得到最终的清浊判决结果。

1.3 线谱对参数增强

谱参数增强的流程如下图 3 所示。

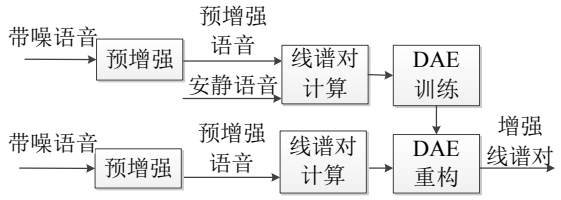

图 3　线谱对参数增强流程

本文应用 DAE 模型对反映谱包络特征的线谱对参数进行增强。设 X' 和 Y' 分别为归一化的安静语音线谱对参数和预增强语音线谱对参数；对于每一维线谱对参数，通过均值和方差进行归一化处理，使其满足零均值、单位方差的高斯分布。线谱对参数增强过程如下所示。

$$X'=\eta(Y') \quad (2)$$

其中 η 表示基于 DAE 模型重构线谱对参数的增强处理过程，线谱对参数增强的目标函数为

$$\varphi=\arg\min E_x[\|\eta(Y')-X'\|_2^2] \quad (3)$$

训练的任务是确定最优的 φ 实现对 η 的准确估计。算法中 DAE 模型的输入和输出参数分别为预增强的线谱对参数和所对应的安静环境下的线谱对参数；对于模型中的每一个隐藏层，包括一个非线性编码过程和一个线性解码过程。

$$\begin{aligned} h(y_i)&=\sigma(W_1y_i+b)\\ x_i&=W_2h(y_i)+c \end{aligned} \quad (4)$$

式中 W_1 和 W_2 分别表示神经网络连接权值的编码矩阵和解码矩阵，在 DAE 模型中 $W_1=W_2^T$ 作为模型训练的约束。参数 b 和 c 分别作为输入层和输出层的偏置矢量，$\sigma(x_i)=(1+\exp(-x_i))^{-1}$ 是 sigmoid 激活函数，h 表示隐藏层的激活函数。

算法中应用了多隐藏层的 DAE 模型训练预增强线谱对参数和安静语音线谱对参数的映射关系，训练过程采用贪婪的逐层预训练加多层联合训练；第一个 DAE 训练对是 X' 和 Y'，接下来的训练对是 $h(X')$ 和 $h(Y')$。在完成逐层贪婪预训练后，各个层进行堆栈生成深层神经网络结构，再进行多层联合训练时，将预训练过程中获取的网络参数作为深层神经网络模型的初始参数，并通过 BP 方法不断优化模型参数。相对于浅层模型，深层神经网络模型具有更强的泛化能力，采用这种 DAE 模型可以有效的修复受损的谱参数，在噪声环境下实现对线谱对参数的增强处理。

1.4 增益估计

针对预增强后的语音帧，计算语音增益时考虑了基音同步周期，每帧语音信号需要进行两次增益计算[16]。计算增益时窗长的确定方式如下：针对清音帧，窗长是 120 个采样点，针对浊音帧，窗长是超过 120 个采样点时的最小基音周期倍数。第一次增益计算的中心点位于当前帧最后一个采样点之前的 80 点处，第二次增益计算的中心点位于当前帧最后一个采样点处，帧长设置为是 160 个采样点，增益参数通过 dB 进行度量。

$$G_i=10\log_{10}(0.01+\frac{1}{L}\sum_{n=1}^{L}s_n^2) \quad (5)$$

其中 s_n 是语音采样点，L 表示窗长。

1.5 混合激励语音合成

声码器模块选择混合激励模型，相对于二元激励模型，混合激励可以有效的消除合成语音的嗡嗡声。根据子带清浊状态确定激励信号的类型（脉冲激励或白噪声激励）。各个子带的清浊状态可以根据 3.2 节中改进的基音周期估计算法计算得到；各个子带的激励信号叠加生成最终的激励信号。

将自适应谱增强滤波器应用到混合激励信号，它是一个带有一阶斜坡补偿的十阶零极点滤波器；通过自适应谱增强模块可以有效的弥补 LPC 滤波器只包含极点的缺点，同时增强合成语音共振峰的结构；以每个基音同步周期为单元，对增益进行调节；最后通过脉冲扩散滤波器进行后处理，它是一个 65 阶 FIR 滤波器，它的作用是将激励信号的能量在一个基音周期中进行扩散，从而减少合成语音中的刺耳的成分。

2 实验和结果分析

2.1 实验数据和评估方法

本节对基于分析合成框架的语音增强算法进行性能评估，从基音周期估计和子带清浊度判决性能、线谱对参数增强误差、语音增强和低速率语音编码的音质四个方面进行评估。语音数据从 TIMIT 数据库中抽取，该数据库中的语音全部是在安静环境下进行录制的[17]。Noisex-92 数据库作为噪声数据，从中选取四种噪声进行模型训练，所选择的噪声类型包括：粉噪声、工厂噪声、车载噪声和 buccaneer 噪声[18]。语音数据和噪声数据需要降采样至 8kHz 后再进行后续处理。原始安静语音和噪声信号以不同信噪比进行叠加生成带噪语音数据，本节中生成三种信噪比的实验数据：0dB、5dB 和 10dB。对于基于分析合成框架的语音增强，帧长设为 160 个采样点，帧移设为 80 个采样点。对于低速率语音编码，根据 MELP 标准逐帧进行处理，帧长和帧移均为 180 个采样点。从 TIMIT 数据库中抽取 3000 个音频样本以三种不同的信噪比叠加四种类型噪声生成训练集；同时从 TIMIT 数据库中分别抽取 300 个音频样本作为测试集和 500 个样本作为发展集并以同样的方式叠加噪声；此外，为了评估算法对噪声环境的鲁棒性，我们选择白噪声和 babble 噪声以 0dB、5dB 和 10dB 三种信噪比叠加到 500 个测试样本中，用于噪声不匹配测试。

针对基音周期估计和子带清浊状态判决，三种度量机制用于评估改进算法的性能，分别是清浊状态判定准确率（VDE）、基频估计误差（GPE）和 FFE，其中 FFE 同时考虑了 VDE 和 GPE 两种特征[19]。

$$VDE = \frac{N_{V\to U} + N_{U\to V}}{N} * 100\% \quad (6)$$

$$GPE = \frac{N_{F0E}}{N_{VV}} * 100\% \quad (7)$$

$$FFE = \frac{N_{VV}}{N} * GPE + VDE \quad (8)$$

针对线谱对参数增强，本节选择高斯混合模型作为基线方法，这种模型广泛应用到线谱对参数的特征转换。通过计算安静语音的线谱对参数和所对应的噪声环境下增强的线谱对参数距离评估算法性能。

$$d(lsf_s, lsf_t) = \sum_{i=1}^{10} | lsf_{si} - lsf_{ti} | \quad (9)$$

针对语音增强，本节选择 logMMSE[15]和子空间方法[22]作为基线方法。通过 PESQ 得分对各种语音增强算法的性能进行评估[20]。PESQ 是一种最为常用的语音质量客观评估方法，相对于其它客观评估方法，该方法与主观评估结果更接近。

此外，本文将改进的基音周期估计算法和线谱对参数增强方法应用到低速率语音编码算法中，改进低速率下语音的音质。MELP 是一种最为成熟的低速率语音编码算法，本节中选择这种方法作为基线，评估改进后性能。通过 PESQ 得分评估改进前后低速率语音编码的性能。

2.2 模型参数优化

本节介绍用于谱参数增强的 GMM 和 DAE 两种模型参数的优化过程。对于 DAE 模型，在一些参数范围内进行搜索，确定最优隐层个数（1、2 和 3）和各个隐层所包含的最优节点个数（30、50、70 和 90）；对于 GMM 模型，同样在一些参数范围内进行搜索，确定最优高斯个数（16、32、64 和 128）。本文最终确定的最优 DAE 模型参数如下：输入层和输出层的特征维度是 10，每个隐藏层的节点个数是 50，最终形成一个 10-50-50-50-10 的深层神经网络，每次批量处理的尺寸是 100，每层预训练迭代次数是 20，多层联合训练的迭代次数是 100；在 0.005 到 0.05 范围内优化学习率（步长 0.005），学习率最终设置为 0.02。本文最终确定的 GMM 模型高斯成分个数为 64。

2.3 基音周期和清浊判定性能评估

本节对基音周期的估计误差和子带清浊状态判定的准确率进行评估。采用三种基线方法与本文提出的改进方法进行比较：GetF0[21]、MELP[16]和原始的 MBSC 方法[14]。不同方法所得到的实验结果如表 1 所示。表 1 中展示了在 0dB、5dB 和 10dB 三种信噪比下的平均性能。

实验结果表明在不同噪声环境下相对于各种基线方法，本文所提的方法具有更低的 GPE 和 FFE。针对白噪声、babble 噪声等宽带噪声，文献 14 中的 MBSC 方法的 VDE 略低于本文方法；针对汽车噪声这类窄带噪声，本文所提方法的 VDE 低于各种基线方法。本文提出的改进的 MBSC 方法在基音周期估计和子带清浊判定时具有更好的性能，这是由于通过子带选择机制保留浊音强度较高的子带进行基音估计和清浊判决，同时在基音周期计算时考虑了低频带线性预测残差的特性，由此改善了算法的性能，在低频带噪声环境下这一优势更加明显。

表 1　在三种信噪比条件下的平均 GPE、VDE 和 FFE

噪声类型	实验方法	VDE(%)	GPE(%)	FFE(%)
Babble 噪声	GetF0	29.4364	26.3463	42.5632
	MELP	30.4676	28.0446	45.8182
	MBSC	26.4846	14.4925	33.9186
	Proposed	26.7041	10.8447	31.9626
白噪声	GetF0	19.4524	16.3642	28.7864
	MELP	18.5931	15.2937	27.0854
	MBSC	16.4207	9.6668	21.3271
	Proposed	17.5618	4.7345	19.9083
车载噪声	GetF0	17.4265	7.3425	20.4353
	MELP	23.4122	9.1026	29.2582
	MBSC	10.8934	3.2069	13.0459
	Proposed	10.1208	1.9618	11.2918

2.4 线谱对参数增强性能评估

本节介绍线谱对参数增强的实验结果，通过线谱对参数的重构误差评估 GMM 和 DAE 两种模型下参数增强的效果。其中粉噪声、工厂噪声、汽车噪声和 buccaneer 噪声用于匹配环境下的测试；白噪声和 babble 噪声用于失配环境下的测试。实验结果在图 4 中展示，从图 4 中可以发现，无论是在匹配噪声环境下，还是在失配噪声环境下，DAE 模型具有更小的重构误差；将 DAE 模型用于噪声环境下的谱参数增强具有很好的鲁棒性，甚至可以对训练模型中未包含的一些噪声进行增强处理。这是由于 DAE 模型具有深层的神经网络结构，相对于浅层模型具有更强的泛化能力。

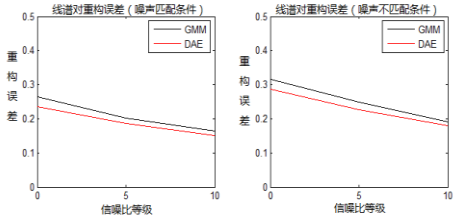

图 4　各种方法线谱对参数重构误差对比

2.5 语音增强性能评估

本节对比基于分析合成框架的语音增强方法和其它两种的语音增强方法：基于 logMMSE 的语音增强方法和基于子空间的语音增强；通过 PESQ 得分评估不同语音增强方法的性能；实验结果如表 2 所示。相对于各种基线方法，所提方法具有更好的性能；在各种环境下，这种改进的基于分析合成框架的语音增强方法的平均 PESQ 得分高于最好基线方法 0.3 分；本节提出的语音增强方法比基线方法具有更好的语音音质，这是由于这种方法可以有效的消除音乐噪声同时谱包络结构得到了很好的修复。

表 2　各种语音增强方法的 PESQ 得分

噪声类型	实验方法	0dB	5dB	10dB
匹配噪声	带噪语音	1.7487	2.0719	2.4017
	Logmmse	2.0993	2.4659	2.7979
	子空间	1.9174	2.3058	2.6720
	本文方法	2.4098	2.6810	2.8663
不匹配噪声	带噪语音	1.7313	2.0462	2.3713
	Logmmse	2.0525	2.4304	2.7713
	子空间	1.9527	2.3237	2.6770
	本文方法	2.4066	2.6801	2.8624

2.6 低速率语音编码性能评估

本节将所提的改进的基音周期估计算法和线谱对参数增强方法应用到低速率语音编码中，对 MELP 语音编码算法框架进行改进。带噪语音信号首先通过 Logmmse 算法进行预增强处理，基音周期和子带清浊状态通过改进的 MBSC 算法进行计算，通过 DAE 模型对原始带噪语音的线谱对参数进行增强处理后再进行矢量量化。语音参数量化模块和语音合成模块与传统的 MELP 标准相同。通过 PESQ 得分评估改进前后语音编码算法的性能；实验结果如表 3 所示。

表 3　各种语音编码方法的 PESQ 得分

噪声类型	实验方法	0dB	5dB	10dB
匹配噪声	原始 MELP	1.4657	1.7999	2.1480
	增强 MELP	1.8634	2.2340	2.5347
	本文方法	2.2920	2.5547	2.7354
不匹配噪声	原始 MELP	1.4679	1.7828	2.1198
	增强 MELP	1.9019	2.2678	2.5604
	本文方法	2.3618	2.6014	2.7634

由表 3 所示，改进后的语音编码算法在不同环境下具有更高的 PESQ 得分；它比 MELP 标准具有更好的语音音质，这是由于所提出的算法中改进了

噪声环境下基音周期和线谱对参数的分析方法。

3 结论和展望

本文介绍了一种基于分析合成框架的单通道语音增强方法，所提出的方法通过分析合成框架有效的消除了音乐噪声，算法中通过深层神经网络建立预增强语音和安静语音线谱对参数之间的映射关系，从而实现对谱包络特征的增强处理，在噪声得到抑制的同时，可以有效的修复受损的语音信号。实验结果表明，在不同噪声环境下，所提方法比各种基线方法具有更好的性能。

后续，我们会进一步优化现有算法，提高算法在不同噪声环境下的自适应能力；此外，我们会尝试利用先验知识优化神经网络结构；同时我们会改进声码器模型，并将现有语音增强算法扩展到宽带语音信号。

参 考 文 献

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” in Acoustics, Speech and Signal Processing, IEEE Transactions on, 1979, 27(2): 113-120.

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimummean square error short-time spectral amplitude estimator,” in Acoustics, Speech and Signal Processing, IEEE Transactions on, 1984, 32(6): 1109-1121.

[3] 张卫强，郭璁，张乔，等. 一种基于计算听觉场景分析的语音增强算法[J]. 天津大学学报：自然科学与工程技术版，2015,48（8）：663-669.

[4] A. Cheveigne´, H. Kawahara. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111, 1917– 1930, 2002

[5] A. Camacho, J. Harris. A sawtooth waveform inspired pitch estimator for speech and music. J. Acoust. Soc. Am. 124, 1638– 1652, 2008.

[6] J. Rouat, Y. Liu, D. Morissette. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. Speech Communication. 21, 191–207, 1997.

[7] L. N. Tan, A. Alwan. Noise-robust F0 estimation using SNR weighted summary correlograms from multi-band comb filters. In: Proc. IEEE ICASSP, pp. 4464–4467, 2011

[8] T. V. Sreenivas and P. Kirnapure, “Codebook constrained wiener filtering for speech enhancement,” IEEE Trans. Speech Audio Process., vol. 4, no. 5, pp. 383–389, Sep. 1996

[9] R. F. Chen, C. F. Chan and H. C. So, “Model-Based Speech Enhancement With Improved Spectral Envelope Estimation via Dynamics Tracking,” IEEE Trans. Speech Audio Process, vol. 20, no. 4, pp. 1324–1336, Sep. 2012

[10] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum, In Proc of ICASSP, 2001,pp.941-944

[11] K. Y. Park, H. S. Kim. Narrowband to wideband conversion of speech using GMM based transformation. Proceeding of IEEE International Conference on Acoustics, Speech, Signal Processing, 2000:4: 1843-1846

[12] G. E. Hinton, S. Osindero, Y. W. Teh, “A fast learning algorithm for deep belief nets,” in Neural computation, 2006, 18(7): 1527-1554

[13] Y. Bengio, L. Yao, G. Alain, et al, “Generalized denoising autoencoders as generative models” in Advances in Neural Information Processing Systems, USA, pp: 899-907, 2013

[14] L. N. Tan, A. Alwan. Multi-band summary correlogram-based pitch detection for noisy speech. In: Speech communication, pp. 841–856, 2013.

[15] Y. ephraim and D. malah, “Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator,” IEEE Trans. Speech Audio Process., vol. 33, no. 2, pp. 443–445, Sep. 1985

[16] L. M. Supplee, R. P. Cohn,J. S. Collura and A. V. McCree, “MELP: the new federal standard at 2400bps” In Acoustics Speech and Signal Processing, Germany, 1591-1594,1997

[17] J. S. Garofolo, “TIMIT: Acoustic-phonetic Continuous Speech Corpus,” Linguistic Data Consortium, 1993.

[18] Rice University, NOISEX-92 Database, [Online] Available:http://spib.rice.edu/spib/select noise.html.

[19] W. Chu, A. Alwan, “reducing F0 frame error of F0 tracking algorithm under noisy condition with an unvoiced/voiced classification frontend” In Acoustics Speech and Signal Processing, Germany,3969-3972,2009.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, et al, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in Acoustics, Speech, and Signal Processing (ICASSP), USA, pp. 749-752, 2001

[21] D. Talkin. Speech Coding and Synthesis, Elsevier, pp. 497–518, 1995.

An Improved Speech Enhancement Based on Analysis Synthesis Framework

LIU Bin[1]，TAO Jianhua[1]，MO Fuyuan[2]

(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190；

2. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190）

Abstract: This paper presents a speech enhancement approach based on analysis–synthesis framework. An improved multi-band summary correlogram (MBSC) algorithm is proposed for pitch estimation and voiced/unvoiced (V/UV) detection. The proposed pitch detection algorithm achieves a lower pitch detection error compared with the reference algorithm. The denoising autoencoder (DAE) is applied to enhance the line spectrum frequencies (LSFs). The reconstruction loss could be decreased compare with the swallow model. The experimental results show that the proposed approach improves the performance of speech enhancement compared with the conventional speech enhancement approach. In addition, it could be applied to parametric speech coding even at low bit rate.

Key words: analysis-synthesis framework, multi-band summary correlogram, denoising autoencoder, speech enhancement.