基于不确定性感知的语音分离方法*

涂斌炜　吕俊

（广东工业大学自动化学院，广东 广州 510006）

摘要：为抵御噪声的干扰，提出一种基于不确定性感知的语音分离方法。在训练阶段，采用双链路架构分别学习噪声和语音源成分的编解码子网和分离子网；在测试阶段，以闭式解的形式自适应更新噪声编码子网，减小训练与测试噪声在特征空间的均值偏移，降低认知不确定性，并尽量保持重要参数不变，间接限制语音分离的经验误差。在公开数据集 LibriSpeech, NoiseX 和 NonSpeech 上的实验结果表明：本文提出的方法能够快速有效地提高噪声干扰下语音分离的尺度不变信噪比。

关键词：语音分离；噪声干扰；不确定性感知

中图分类号：TN912　　**文献标识码：**A　　**文章编号：**1674-2605(2021)01-0008-06

DOI：10.3969/j.issn.1674-2605.2021.01.008

0　引言

语音分离一词最初源于“鸡尾酒会问题[1]”，是指从混合的两个或多个说话人的声音中得到想要的目标说话人（一人或多人）的语音信号，广泛应用于语音识别、情感识别或翻译等任务的前端处理。按信号输入的通道数划分，语音分离可分为单通道语音分离和多通道语音分离 2 种。本文主要讨论单通道语音分离技术。

单通道语音分离技术又分为有背景噪声和无背景噪声 2 类。无背景噪声的单通道语音分离技术发展较早，常见方法包括基于听觉场景分析[2]、基于非负矩阵分解[3-4]和基于深度神经网络的语音分离方法[5-6]。这些方法推动了单通道语音分离技术的发展，但没有考虑噪声干扰的影响，与真实使用场景相差较大。

近年，许多专家学者逐渐关注有背景噪声的单通道语音分离技术。文献[7]~文献[9]通过串联方法将语音降噪网络和语音分离网络结合起来，该方法已被证明能够改善嘈杂环境下的语音识别性能；文献[10]通过多场景训练方法将语音降噪和语音分离结合在一起，2 个任务共用 1 个网络。上述方法改善了语音分离技术在噪声环境下的分离效果，但没有考虑异常噪声带来的分布差异问题。由于噪声具有较强的多样性，因此测试信号中难免会出现与训练集噪声相差较大的噪声信号，这些异常噪声会严重影响语音分离效果。

为抵御噪声的干扰，本文提出一种基于不确定性感知的语音分离方法（speech separation based on uncertainty perception, SSUP）。该方法采用变换域特征的均值偏移来度量预测不确定性，采取双链路网络结构，通过自适应更新噪声编码网络的参数，减小噪声带来的均值偏移，同时采用弹性权重固化（elastic weight consolidation, EWC）策略[11]，间接保持较小的训练集经验误差。

1　分离网络

1.1　问题描述

在背景噪声下，单通道输入信号 $y(t)$ 由语音信号 $x(t)$ 和噪声信号 $n(t)$ 叠加而成，本文只讨论 2 个说话人的情况，因此 $y(t)$ 建模为

$$y(t)=x_1(t)+x_2(t)+n(t) \tag{1}$$

单通道语音分离的目标是从混合信号 $y(t)$ 中估计得到 $x_1(t)$ 和 $x_2(t)$。

1.2　网络结构

现有的单通道语音分离方法主要采用单链路架构[12-13]。但由于噪声与语音信号的分布不一样，采用不同的表达方式更合理。本文提出的 SSUP 采用双链路网络架构，如图 1 所示。

* 基金项目：广东省自然科学基金（2018A030313306）

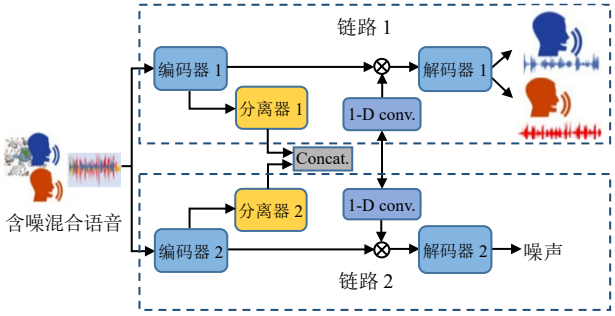

图 1　SSUP 双链路网络架构

SSUP 双链路网络包括网络结构相同的 2 个链路，每个链路皆包含编码器、分离器和解码器 3 个主要部分。编码器和解码器分别为一维卷积和一维逆卷积网络；分离器由多个双路循环神经网络（dual-path RNN, DPRNN）模块组成[12]。其中，链路 1 的输出为 2 个说话人的语音信号，链路 2 的输出为噪声信号。首先，在训练集中训练得到初始模型；然后，根据每条测试信号，有针对性地更新链路 2 中编码器的参数，并保持其他参数不变。

依据验证集的分离性能，SSUP 双链路网络的参数设置如表 1 所示。模型训练采用的优化器为 Adam，迭代步长为 10^{-3}，迭代次数为 100。

表 1　SSUP 双链路网络参数设置

	参数	数量/个
编码器	卷积核个数	256
	卷积核的窗宽	16
	卷积核的步长	8
分离器	分离器中 DPRNN 模块的个数	5
	DPRNN 模块中 BLSTM 的隐含状态单元	128
解码器	卷积核个数	256
	卷积核的窗宽	16
	卷积核的步长	8

1.3　训练目标

网络最终输出是估计信号的时域波形。本文采用的训练目标为最大化尺度不变信噪比（scale-invariant source-to-noise ratio, SI-SNR）[14]。在单通道语音分离中，标准的信号失真比（source-to-distortion, SDR）可能出现误导性结果，即在感知上并没有改变估计信号的情况下，仅依靠缩放估计信号便能提高 SDR 值，然而这种提高没有实际意义[14]。为避免这种情况，SI-SNR 取代 SDR 作为语音分离的评价指标[12,15]，其定义为

$$SI\text{-}SNR = 10\log_{10}\left(\left\|\frac{\langle\hat{\boldsymbol{s}},\boldsymbol{s}\rangle\boldsymbol{s}}{\langle\boldsymbol{s},\boldsymbol{s}\rangle}\right\|^2 \Big/ \left\|\hat{\boldsymbol{s}}-\frac{\langle\hat{\boldsymbol{s}},\boldsymbol{s}\rangle\boldsymbol{s}}{\langle\boldsymbol{s},\boldsymbol{s}\rangle}\right\|^2\right) \tag{2}$$

式中，$\hat{\boldsymbol{s}}$ 为估计信号；$\boldsymbol{s}$ 为目标信号。为确保尺度不变性，$\hat{\boldsymbol{s}}$ 和 $\boldsymbol{s}$ 做 0 均值归一化处理。

2　基于不确定性感知的语音分离

2.1　不确定性感知

由于噪声往往是非平稳的，且具有较强的多样性，训练集和测试集的噪声分布差异给语音分离模型带来了认知不确定性[16]。因此，需要对这种认知不确定性进行度量，并自适应地调整网络参数，克服分布差异带来的影响。目前，大多数获取预测不确定性的方法基于贝叶斯神经网络[17-18]，需要对网络参数进行大量采样，计算量大，且优化效率低。针对该问题，本文采用变换域特征的均值偏移来度量预测不确定性，不确定性 D 的计算公式为

$$D = \left\|\frac{1}{C}\boldsymbol{B}_0\boldsymbol{X}_{\text{test}}\boldsymbol{l} - \boldsymbol{m}_0\right\|_2^2 \tag{3}$$

$$\boldsymbol{m}_0 = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{C}\boldsymbol{B}_0\boldsymbol{X}_j\boldsymbol{l}\right) \tag{4}$$

式中，$\boldsymbol{B}_0 \in R^{K\times L}$ 为更新前链路 2 中的编码器参数；$\boldsymbol{l} \in R^{C\times 1}$ 和 N 分别为元素为 1 的列向量和训练集的样本个数；$\boldsymbol{X}_{\text{test}} \in R^{L\times C}$ 为输入信号按卷积核滑动窗切割后堆叠的矩阵，其中 L 和 C 分别代表编码器卷积核的窗宽和滑动窗个数。

2.2　参数更新方法

测试信号与训练集的编码特征分布应尽量接近，以减小分离模型的认知不确定性。与此同时，采用弹性权重固化策略[11]，间接保持较小训练集经验误差，自适应地学习有利于目标信号实现语音分离的变换域。因此，设计代价函数为

$$\min_{\boldsymbol{B}} J=\left\|\frac{1}{C}\boldsymbol{B}\boldsymbol{X}_{\text{test}}\boldsymbol{l}-\boldsymbol{m}_0\right\|_2^2+\alpha\left\|\sqrt{\boldsymbol{F}}\odot\left(\boldsymbol{B}-\boldsymbol{B}_0\right)\right\|_F^2 \quad (5)$$

式中，$\boldsymbol{B}$ 为更新后链路 2 中的编码器参数；$\boldsymbol{F}$ 是 $\boldsymbol{B}_0$ 的费雪信息矩阵；$\odot$ 代表点乘运算。

因为费雪信息越大的参数对网络的输出越重要，所以尽量不要对其做太大调整。设 $\boldsymbol{B}_i\in R^{1\times L}$ 为 $\boldsymbol{B}$ 的第 i 行，即第 i 个卷积核的参数。式(5)等号两边对 $\boldsymbol{B}_i$ 求导：

$$\frac{\partial J}{\partial \boldsymbol{B}_i}=\frac{2}{C}\left(\frac{1}{C}\boldsymbol{B}\boldsymbol{X}_{\text{test}}\boldsymbol{l}-\boldsymbol{m}_0\right)^T\frac{\partial\left(\boldsymbol{B}\boldsymbol{X}_{\text{test}}\boldsymbol{l}\right)}{\partial \boldsymbol{B}_i}+2\alpha\boldsymbol{F}_i\odot\left(\boldsymbol{B}_i-\boldsymbol{B}_{0i}\right) \quad (6)$$

令 $\frac{\partial J}{\partial \boldsymbol{B}_i}=0$，得

$$\frac{1}{C}\left(\frac{1}{C}\boldsymbol{B}\boldsymbol{X}_{\text{test}}\boldsymbol{l}-\boldsymbol{m}_0\right)^T\frac{\partial(\boldsymbol{B}\boldsymbol{X}_{\text{test}}\boldsymbol{l})}{\partial \boldsymbol{B}_i}+\alpha\boldsymbol{F}_i\odot(\boldsymbol{B}_i-\boldsymbol{B}_{0i})=0 \quad (7)$$

由于代价函数(5)是一个加权最小二乘优化问题，因此可求得其闭式解的第 i 行为

$$\boldsymbol{B}_i=\left(\frac{1}{C}\boldsymbol{m}_{0i}\boldsymbol{a}^{\mathrm{T}}+\alpha\boldsymbol{F}_i\odot\boldsymbol{B}_{0i}\right)\left(\frac{1}{C^2}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}+\alpha\boldsymbol{A}\right)^{-1} \quad (8)$$

式中，$\boldsymbol{a}=\boldsymbol{X}_{\text{test}}\boldsymbol{l}$；$\boldsymbol{A}=diag(\boldsymbol{F}_{i,1},\boldsymbol{F}_{i,2},\cdots,\boldsymbol{F}_{i,L})$。

对于每一条测试信号，都可通过式(8)快速地求得唯一解 $\boldsymbol{B}$，式(8)的时间复杂度为 $O(L^3+L\times C)$。因此，本文方法可在不进行反向传播的基础上快速更新编码器参数。

若不引入费雪信息，式(5)的最后一项是 Frobenius 范数正则化约束，此时式(5)可改写为

$$\min_{\boldsymbol{B}} J=\left\|\frac{1}{C}\boldsymbol{B}\boldsymbol{X}_{\text{test}}\boldsymbol{l}-\boldsymbol{m}_0\right\|_2^2+\alpha\left\|\boldsymbol{B}-\boldsymbol{B}_0\right\|_F^2 \quad (9)$$

其闭式解的第 i 行为

$$\boldsymbol{B}_i=\left(\frac{1}{C}\boldsymbol{m}_{0i}\boldsymbol{a}^{\mathrm{T}}+\alpha\boldsymbol{B}_{0i}\right)\left(\frac{1}{C^2}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}+\alpha\boldsymbol{I}\right)^{-1} \quad (10)$$

2.3 噪声信号在特征空间上的均值偏移

为探究噪声信号在特征空间上的均值偏移，本文从 Nonspeech 数据集中选取 8 种不同的噪声数据[19]，与语音信号生成 8 个测试集，每个测试集的样本个数和所采用的语音信号皆相同。计算每个测试集的噪声特征至训练集噪声特征中心的平均偏差为

$$\boldsymbol{d}=\frac{1}{M}\sum_{j=1}^{M}\left\|\frac{1}{C}\boldsymbol{B}_0\boldsymbol{X}_j\boldsymbol{l}-\boldsymbol{m}_0\right\|_2^2 \quad (11)$$

式中，M 为测试集的样本个数。

8 种不同噪声特征至训练集噪声特征中心的平均偏差如图 2 所示。

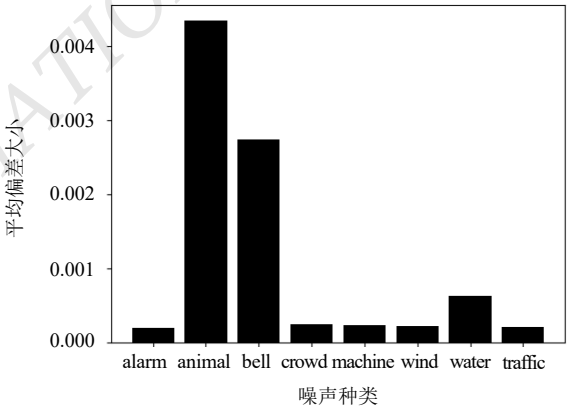

图 2　8 种不同噪声特征至训练集噪声特征中心的平均偏差

由图 2 可知：animal 和 bell 这 2 种噪声的编码特征偏离训练数据均值中心 $\boldsymbol{m}_0$ 的程度非常明显，给语音分离模型带来较大的泛化风险；而另外 6 种噪声的编码特征偏离均值中心比较小，可见并非所有的噪声都会在特征空间上带来严重的均值偏差。因此，需要设置 1 个阈值，只有满足阈值要求的测试信号才会触发参数更新。

2.4 参数更新触发条件

本文采用变换域特征的均值偏移来度量预测不确定性。针对不确定性较大的测试数据，将进行参数的动态调整。因此，设置了 1 个不确定性阈值 TH，计算公式为

$$TH = mean(\{D\}_{tr}) + n \times std(\{D\}_{tr}) \quad (12)$$

式中，n 为超参数；$\{D\}_{tr}$ 为训练样本 D 值的集合。

当测试信号的 D 值大于 TH，通过式(8)或式(10)对编码器 2 的参数进行更新。

3 实验及参数分析

3.1 实验设置

实验采用的深度学习框架为 Pytorch，服务器 CPU 为 8 核 3.90 GHz AMD Ryzen 3700X，内存为 32 GB，GPU 为 Nvidia RTX 2080 Ti。

本文采用公开的语音数据集 LibriSpeech[20]，噪声数据集 NoiseX[21]和 Nonspeech[19]进行实验。为方便网络训练，所有数据统一采样率为 8 kHz。本文的语音数据全部来自于 LibriSpeech 数据集中的“train-clean-100”子集，该子集包含了 100 h 来自 251 个不同个体的语音数据。首先，取任意 2 个不同说话人的语音以-2.5 dB~2.5 dB 的任意比例混合，得到干净的 2 个说话人的混合数据；然后，选取 NoiseX 数据集中的 10 种噪声生成训练集数据，同时将 Nonspeech 数据集中的 8 种噪声生成测试集数据，详情如表 2 所示。其中，噪声与说话人声按-5 dB~10 dB 的任意信噪比混合，训练集的样本个数为 8000，测试集中每种噪声数据的样本个数为 3000。

表 2 噪声数据集

数据集	噪声类型	用途
NoiseX[21]	babble、buccaneer2、destroyerengine、f16、destroyerops、factory2、hfchannel、white、machinegun、factory1	构建训练集
Nonspeech[19]	alarm、animal、bell、crowd、machine、traffic、water、wind	构建测试集

3.2 实验结果

为验证本文提出方法的有效性，在测试集中比较以下 4 种方法的分离性能：1） 文献[12]提出的单链路网络方法；2） 编码参数更新前（before parameter update, BPU）的双链路网络方法；3） 在方法 2 的基础上，采取 Frobenius 范数正则化（Frobenius norm regularization, FNR）的参数更新方法；4） 在方法 3 的基础上，采取费雪信息加权的 FNR （Fisher information weighted FNR, FIW-FNR）的参数更新方法。实验的结果如表 3 所示。（实验中 n 和 α 分别取 0.5 和 10^{-8}）

表 3 4 种方法的分离性能比较

方法	单链路[12]	BPU	FNR	FIW-FNR
SI-SNR /dB	0.11	0.53	0.66	0.85

由表 3 可知：1） BPU 取得了比单链路更好的分离性能，说明双链路网络方法是有效的；2） FNR 和 FIW-FNR 方法获得的 SI-SNR 指标高于 BPU，其中 FIW-FNR 是 4 种方法中分离性能最好的，可见本文提出的参数更新方法可以改善模型的分离性能。

3.3 参数分析

3.3.1 折中系数 α

式(5)中，α 越大意味着代价函数对编码器参数更新的惩罚力度越大。本文在测试集中进行了实验，当 n = 0.5 时，对比不同 α 对模型分离效果的影响，结果如表 4 所示。

表 4 α 取不同值时，3 种方法的 SI-SNR 指标

α	方法		
	BPU/dB	FNR/dB	FIW-FNR/dB
10^{-9}	0.53	0.66	0.48
10^{-7}		0.64	**0.80**
10^{-5}		0.63	0.65
10^{-3}		0.64	0.64
10^{-1}		0.61	0.64

由表 4 可知：1） 当 α =10^{-7} 时，FIW-FNR 取得最好的分离效果，比 BPU 提高了 0.27 dB；2） 当 α >10^{-9} 时，本文提出的 FNR 和 FIW-FNR 方法都优于 BPU 方法，可见在 α 相当大的取值范围内，本文提出的参数更新方法都是有效的。

3.3.2 阈值系数 n

由 2.4 节可知：不确定性阈值 TH 随着 n 的增大而增大。为探究阈值系数 n 对 FNR 和 FIW-FNR 两种方法的影响，本文在测试集中进行对比实验，实验中 $\alpha=10^{-7}$。对比结果如表 5 所示。

表 5 n 取不同值时，3 种方法的 SI-SNR 指标

阈值系数	需要更新参数的测试样本数	方法	SI-SNR/dB
$n=0$	1569	BPU	1.21
		FNR	1.24
		FIW-FNR	**1.36**
$n=0.5$	1056	BPU	0.53
		FNR	0.64
		FIW-FNR	**0.80**
$n=1$	696	BPU	−0.31
		FNR	−0.11
		FIW-FNR	**0.16**

由表 5 可知：随着 n 不断增大，阈值相应提高，需要更新编码器参数的测试样本数也不断减少；在 3 种不同的阈值条件下，FNR 和 FIW-FNR 方法都优于 BPU，当 $n=1$ 时，FIW-FNR 方法相较于 BPU 在 SI-SNR 指标上提高了 0.47 dB。

3.4 运行效率

针对每一条测试信号，本文提出的基于不确定性感知的语音分离方法都可以通过式(8)或式(10)闭式更新噪声编码网络参数，而无需经过反向梯度传播，从而保证了模型的运行效率。经过测试 1000 条数据，FIW-FNR 方法平均处理一条测试信号的时间约为 (0.15 ± 0.01) s（每条数据长度为 5 s）。

4 结语

为减小噪声的干扰，本文提出一种基于不确定性感知的语音分离方法。针对每一条测试信号，自适应更新噪声编码网络的参数，减小噪声带来的均值偏移，并尽量保持重要参数不变，间接限制语音分离的经验误差。该方法具有闭式解，执行效率高，能够快速调整编码网络参数，增强语音分离模型对环境噪声的泛化能力。

参考文献

[1] BELL A J, SEJNOWSKI T J. An information-maximization approach to blind separation and blind deconvolution[J]. Neural Computation, 1995,7(6):1129-1159.

[2] WANG D L, BROWN G J. Computational auditory scene analysis: principles, algorithms, and applications[J]. IEEE Trans. Neural Networks, 2008,19(1):199.

[3] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755):788-791.

[4] 李煦,屠明,吴超,等.基于 NMF 和 FCRF 的单通道语音分离[J].清华大学学报(自然科学版),2017,57(1):84-88.

[5] WANG D L, CHEN J. Supervised speech separation based on deep learning: an overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018,26(10):1702-1726.

[6] 刘文举,聂帅,梁山,等.基于深度学习语音分离技术的研究现状与进展[J].自动化学报,2016,42(6):819-833.

[7] MA C, LI D, JIAN X. Two-stage model and optimal SI-SNR for monaural multi-speaker speech separation in noisy environment[J]. arXiv preprint arXiv: 2004.06332, 2020.

[8] LIU Y, DELARIA M, WANG D L. Deep casa for talker-independent monaural speech separation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 6354-6358.

[9] WANG X, DU J, CRISTIAN A, et al. A study of child speech extraction using joint speech enhancement and separation in realistic conditions[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7304-7308.

[10] WU Y K, TUAN C I, LEE H Y, et al. SADDEL: Joint Speech separation and denoising model based on multitask learning[J]. arXiv preprint arXiv: 2005.09966, 2020.

[11] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521-3526.

[12] LUO Y, CHEN Z, YOSHIOKA T. Dual-Path RNN: efficient long sequence modeling for time-domain single-channel

speech separation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020:46-50.

[13] LUO Y, MESGARANI N. Conv-tasnet: surpassing ideal time-frequency magnitude masking for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[14] ROUX J L, WISDOM S, ERDOGAN H, et al. SDR half-baked or well done[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: 626-630.

[15] LUO Y, CHEN Z, MESGARANI N. Speaker-independent speech separation with deep attractor network[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2018, 26(4):787-796.

[16] TAGASOVSKA N, LOPEZ-PAZ D. Single-model uncertainties for deep learning[C]. In Advances in Neural Information Processing Systems, 2019: 6414-6425.

[17] WELLING M, YEE W T. Bayesian learning via stochastic gradient Langevin dynamics[C]. Proceedings of the International Conference on Machine Learning (ICASSP), 2011: 681-688.

[18] GAL Y, GHAHRAMANI Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning[C]. Proceedings of the International Conference on Machine Learning (ICML), 2016: 1050-1059.

[19] HU G, WANG D L. A tandem algorithm for pitch estimation and voiced speech segregation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010,18(8): 2067-2079.

[20] PANAYIOTOU V, CHEN G, POKEY D, et al. LibriSpeech: an ASR corpus based on public domain audio books[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 5206-5210.

[21] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: Ii.noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993,12(3): 247-251.

Speech Separation Method Based on Uncertainty Perception

Tu Binwei　Lü Jun

(School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: In order to resist the disturbances of noises, we proposed a speech separation method based on uncertainty perception. In the training phase, a two-link architecture is adopted to learn the codec subnet and separate subnet of noise and speech source components respectively. In the testing phase, the noise coding subnet is updated adaptively in the form of closed solution, so as to reduce the mean deviation of training and testing noises in the feature space, reduce cognitive uncertainty, keep the important parameters unchanged as far as possible, and indirectly limit the empirical error of speech separation. Experimental results on the public datasets LibriSpeech, NoiseX and NonSpeech show that the proposed approach can rapidly and effectively improve the scale-invariant source-to-noise ratio of speech separation under the interferences of unknown noises.

Key words: speech separation; noise interference; uncertainty perception

作者简介：

涂斌炜，男，1995 年生，硕士研究生，主要研究方向：机器学习，语音分离。E-mail: tubinwei@mail2.gdut.edu.cn

吕俊（通信作者），男，1979 年生，博士，副研究员，主要研究方向：生物信号检测与识别。E-mail: lujun.rylj@gmail.com

（上接第 34 页）

Key words: rales detection; signal processing; ResNet; attention mechanism

作者简介：

杨淋坚，男，1994 年生，硕士研究生，主要研究方向：模式识别、机器学习、生物信号处理。E-mail: 429667439@qq.com

张宇，男，1992 年生，硕士研究生，主要研究方向：模式识别、机器学习、生物信号处理。