Computer Science and Application 计算机科学与应用, 2018, 8(4), 546-552
Published Online April 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.84061

Hans汉斯

Supervised Speech Enhancement Algorithm Based on Phase Spectrum Estimation

Baoming Li, Xiaoning Fu

College of Electromechanical Engineering, Xidian University, Xi'an Shaanxi
Email: 1215918085@qq.com

Received: Apr. 6th, 2018; accepted: Apr. 23rd, 2018; published: Apr. 30th, 2018

Abstract

In order to solve the problem that the traditional speech enhancement algorithms only estimate the speech amplitude spectrum, but make phase spectrum remain unchanged, the supervision of speech separation algorithm based on phase spectrum estimation is proposed. Firstly, after an analysis of the traditional phase compensation, an ideal combination of mask (ICM) considering amplitude spectrum and phase spectrum is proposed and applied to supervised speech enhancement algorithm. The simulation experiment proves the algorithm proposed can not only suppress background noise effectively, but also improve the intelligibility and automatic recognition rate of the speech significantly.

Keywords

Phase Spectrum Estimation, Phase Compensation, Phase Complemental Mask, Supervised Speech Enhancement

基于理想组合掩蔽的监督性语音增强算法

李保明，付小宁

西安电子科技大学机电工程学院，陕西 西安
Email: 1215918085@qq.com

收稿日期：2018年4月6日；录用日期：2018年4月23日；发布日期：2018年4月30日

摘 要

为了解决传统的语音增强算法只对语音幅值谱进行估计，而让语音相位谱保持不变的问题，提出了基于相位谱估计的监督性语音分离算法。首先，对传统的相位补偿理论进行分析，提出了一种同时考虑语音

文章引用: 李保明, 付小宁. 基于理想组合掩蔽的监督性语音增强算法[J]. 计算机科学与应用, 2018, 8(4): 546-552.
DOI: 10.12677/csa.2018.84061

幅值谱和相位谱的理想组合掩码(ICM)，并将其应用到监督性语音增强算法中。经过仿真实验，证实该算法能够有效地抑制背景噪声，并且能够显著地提高语音的可懂性和自动识别率。

关键词

相位谱估计，相位补偿，理想组合掩码，监督性语音增强

1. 引言

在实际的语音通信系统中，语音信号总是不可避免的受到噪声的干扰。这些噪声的存在不仅极大地损害了语音的可懂性，还对原始语音的数学模型造成破坏，使得语音质量下降。因此，为了得到纯净语音，一些语音增强算法相继提出。从信号处理的角度来看，许多方法提出估计噪音的功率谱或者理想维纳滤波器，比如谱减法、维纳滤波法、最小均方误差估计法和子空间法等[1]。但传统的单通道语音增强算法都聚集在语音幅度谱的估计，而忽略了相位谱估计。这是因为有研究表明，人耳对信号相位信息并不敏感[2]。但是，最近一些研究显示，相位信息对于提高语音的感知质量具有重要的作用[3]。文献[4]提出一种非接触语音检测增强算法，通过对语音信号振幅谱保持不变，通过相位谱来对信号进行重构增强。这种方法虽然能够对背景噪声起到一定的抑制作用，但对语音的整体结构没有较好的还原。文献[5]提出了一种改进的相位谱补偿算法。该算法对相位补偿函数进行改进，通过语音存在概率算法估计噪声功率谱密度，取得了较好的增强效果。本文对传统的相位补偿算法进行分析，提出了一种同时考虑幅值和相位信息的分离目标，即理想组合掩码(Ideal Compositional Mask, ICM)，并应用到监督性语音分离算法中。

2. 相位补偿理论

2.1. 传统相位谱补偿算法[6]

假设 $x(t)$ 为纯净语音， $v(t)$ 为加性噪声，且 $x(t)$ 与 $v(t)$ 相互独立，则加噪语音可表示为

$$y(t)=x(t)+v(t) \tag{1}$$

经过短时傅里叶变换变换到频域，可表示为

$$Y(n,k)=\sum_{m=-\infty}^{\infty} y(m)w(n-m)\exp(-j2\pi km/N) \tag{2}$$

其中，k 表示频率，n 表示帧数，N 表示离散傅里叶变换长度，$w(n)$ 为分帧窗函数，一般为汉宁窗。信号经过傅里叶变换，都可通过幅值谱和相位谱表示。$Y(n,k)$ 可表示为位极坐标形式，即

$$Y(n,k)=|Y(n,k)|\exp(j\angle Y(n,k)) \tag{3}$$

其中，$|Y(n,k)|$ 表示短时幅值谱，$\angle Y(n,k)$ 表示短时相位谱。在传统的相位补偿算法中[7]，定义一个相位谱补偿函数，其表达式为

$$\Lambda(n,k)=\lambda\psi(n,k)\left|\hat{D}(n,k)\right| \tag{4}$$

其中，λ 为补偿因子，$\psi(n,k)$ 为判决因子，其表达式如式(6)，$\left|\hat{D}(n,k)\right|$ 为是噪声短时幅度谱的估计值

$$\psi(k)=\begin{cases}1 & 0<k/N<0.5\\ -1 & 0.5<k/N<1\\ 0 & \text{其他}\end{cases} \tag{5}$$

将相位谱补偿函数与混合语音的频谱相叠加，得到补偿后的频谱表达式:

$$Y_{\Lambda}(n,k)=Y(n,k)+\Lambda(n,k) \tag{6}$$

则得到增强的语音频谱表达式为:

$$S_{\Lambda}(n,k)=\left|S(n,k)\right|\exp\left(j\angle S_{\Lambda}(n,k)\right) \tag{7}$$

2.2. 相位谱估计

在传统相位谱补偿算法中，利用加噪语音经过短时傅里叶变换是共轭对称的性质，通过相位补偿函来实现相位的增强。但因为 λ 是一个经验常数，对语音增强增益是固定不变的，而实际希望可以根据不同的信噪比来实现不同的增益。其次，在相位补偿函数中，传统算法是直接应用带噪语音的幅度谱代替噪声幅度谱估计来实现相位谱补偿，这样，会使得语音信号严重失真，降低语音增强效果。针对以上两个问题，提出理想组合掩码(ICM)，该分离目标表达式为

$$\mathrm{ICM}=\left[\begin{array}{c}\left(\dfrac{X^2(n,k)}{Y^2(n,k)}\right)^{\alpha}\\ \left(\dfrac{\left|Y(n,k)\right|-\left|X(n,k)\right|}{\left|Y(n,k)\right|}\right)\end{array}\right] \tag{8}$$

其中，α 为可调因子，一般取值为 0.5。本文通过监督性语音分离算法实现分离目标的估计。然后，将估计出的补偿因子代入补偿函数中，即

$$\Lambda(n,k)=c\exp\left(\frac{\left|Y(n,k)\right|-\left|X(n,k)\right|}{\left|Y(n,k)\right|}-1\right)\psi(n,k)\frac{\left|Y(n,k)\right|-\left|X(n,k)\right|}{\left|Y(n,k)\right|} \tag{9}$$

其中，C 为经验常数。估计的相位谱表达式为

$$\angle Y_{\Lambda}(n,k)=\angle Y(n,k)+\Lambda(n,k) \tag{10}$$

估计的幅值谱表达式为

$$\left|Y_{\Lambda}(n,k)\right|=\left(\frac{X^2(n,k)}{Y^2(n,k)}\right)^{\alpha}\left|Y(n,k)\right| \tag{11}$$

将估计的相位谱和估计的幅值谱相结合，得到增强后的频谱表达式为

$$S_{\Lambda}(n,k)=\left|Y_{\Lambda}(n,k)\right|\exp\left(j\angle Y_{\Lambda}(n,k)\right) \tag{12}$$

3. 监督性语音分离算法设计

典型的监督性语音分离算法是通过监督性学习算法训练分离模型，从而实现从带噪语音特征到分离

目标的映射函数[8]。设计的语音分离算法主要框图如图 1 所示。该算法主要由时频分解、特征提取、分离目标、模型训练，相位补偿和波形合成组成。

通过时频分解，可以将输入的一维语音信号分解为二维的时频信号。目前常见时频分解方法有 gammatone 听觉滤波和短时傅里叶变换。本文采用短时傅里叶变换进行时频分解，短时傅里叶变换表达式如下：

$$S_x(\tau, f)=\int_{-\infty}^{+\infty} x(t)w(t-\tau)\exp(-j2\pi ft)\mathrm{d}t . \tag{13}$$

其中，$x(t)$ 为一维时域信号，$w(t-\tau)$ 为是实对称窗函数，可以选取汉宁窗作为分析窗函数，$S_x(\tau, f)$ 为信号在第 τ 个时间帧第个 f 频带的 STFT 系数。

训练模型选择深层神经网络 DNN，设置一个输入层，三个隐层和一个输出层，其中三个隐层都包含 1024 个节点。激活函数采用 Relu 函数，表达式为 $R(x)=\begin{cases}x & x>0\\ 0 & x\le 0\end{cases}$。输出层采用 sigmoid 的线性激活函数实现线性分类，sigmoid 表达式为 $s(x)=\dfrac{1}{1+\mathrm{e}^{-x}}$。深层神经网络层与层之间的神经元是全连接的，因此随着每层神经元的个数和层数增加，网络的结构会变得复杂，所以网络训练采用网络训练采用标准反向传播算法与 Dropout [9]技术，提高神经网络的学习效率和性能。分离特征选择最优组合特征[10]，即 AMS+RASTA_PLP+MFC。分离目标选择相位补偿掩码 PCM。经过模型训练，即可得到加噪语音最优组合特征到分离目标 PCM 的映射。

4. 实验仿真及结果对比

为了证实本文算法的增强效果，实验过程中使用传统的谱减法、维纳滤波法和最小均方误差估计法作为对比算法。实验采用 NOIZEUS 语音库中的 600 句语音作为训练阶段的纯净语音，另外的 120 句语音作为测试阶段的纯净语音，实验噪声来自某种旋翼直升机的旋翼噪声。实验是在信噪比分别为−6 dB、−3 dB、0 dB、3 dB、6 dB 的情况下进行测试的。

4.1. 语音谱分析

从图 2 可以看到，谱减法和维纳滤波法都能够抑制背景噪声，但语音信号中仍然残留大量噪声。最

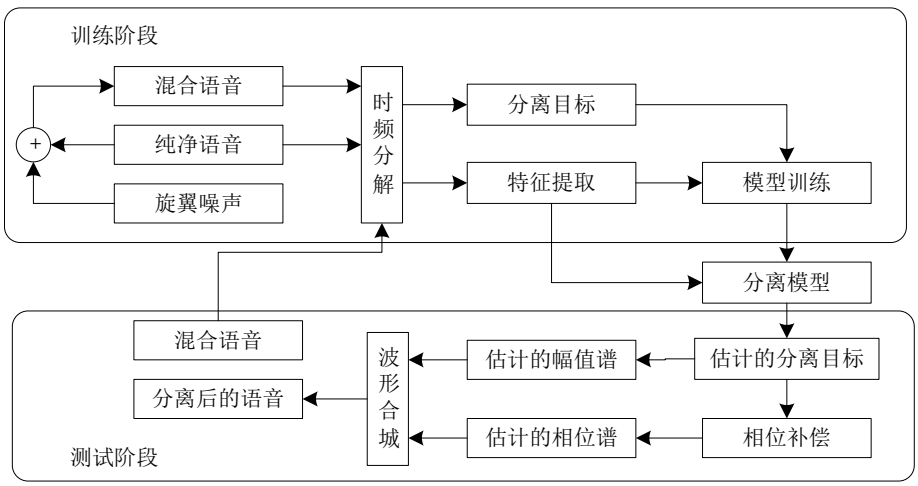

Figure 1. The emblem speech separation algorithm block diagram
图 1. 监督性语音分离算法框图

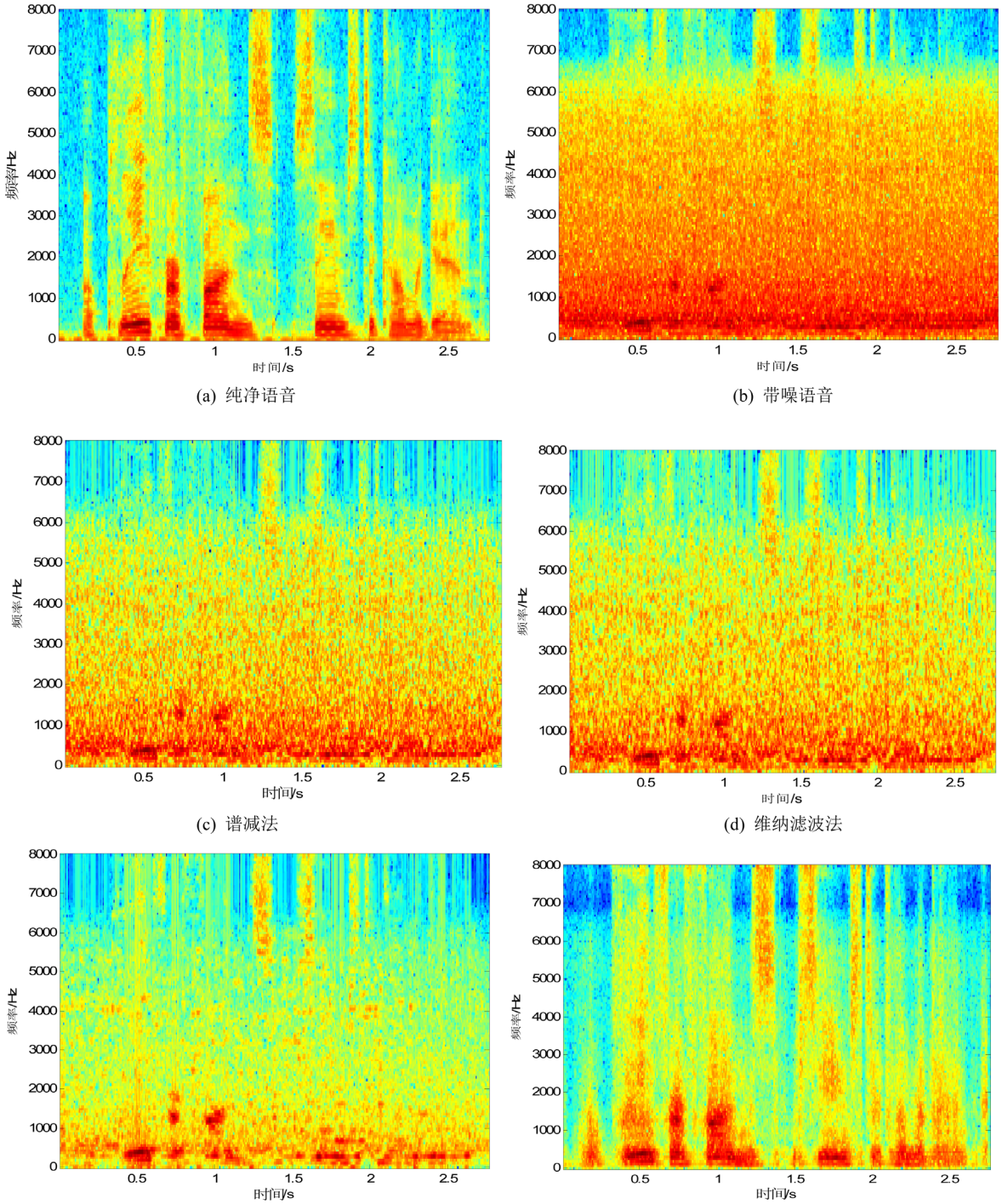

Figure 2. Language spectra
图 2. 语谱图

小均方差误差估计法在有效地抑制背景噪声的同时，也引起了严重的语音失真。相比于传统算法，监督性语音分离算法在有效地去除背景噪声的同时，能够保留语音的整体结构，增强的语音在保留较清晰的端点和较分明的频谱过渡的基础上，在语音细节上也有了较明显的提高。

4.2. 评价指标分析

语音分离系统主要针对人耳和语音信号处理设备两个目标受体，以提高语音可懂度和语音感知质量为目的。目前，国外普遍使用短时客观可懂度评分(STOI) [11]和语音识别率评估(PESQ) [7]作为实验中的评价指标。

1) PESQ(Perceptual evaluation of speech quality),即主观语音质量评估，是 ITU-TP.862 建议书提供的客观 MOS 值评价方法。PESQ 的取值范围是[−0.5 4.5]，取值越高说明语音质量越好。PESQ 计算框图 3 如所示。

2) STOI (Short-Time Objective Intelligibility)，即短时客观可懂性，是用来评估在时域上经过掩蔽或经过短时傅里叶变换且频域上加权的带噪语音的可懂性。计算 STOI 时，用时间对其的纯净与混合语音信号来计算每个音频通道 $k\left(k=1,\cdots,K\right)$ 与 400ms 短时分段 $m\left(m=1,\cdots,M\right)$ 的中间值 $d\left(k,m\right)$。首先，对纯净和带噪语音信号进行短时傅里叶变换，得到第 j 个频段第 n 个时间侦的短时能量谱 $\left|X\left(j,n\right)\right|^2$ 和 $\left|Y\left(j,n\right)\right|^2$。将 j 个跨越 1/3 倍频带间隔的 $\left|X\left(j,n\right)\right|^2$ 和 $\left|Y\left(j,n\right)\right|^2$ 相加得到第 k 个音频通道的能量谱 $\left|X\left(k,n\right)\right|^2$ 和 $\left|Y\left(k,n\right)\right|^2$。带噪语音能量谱 $\left|Y\left(k,n\right)\right|^2$ 被限制为信号失真比不能低于−15dB。中间值 $d\left(k,m\right)$ 是 $\left|X\left(k,m\right)\right|^2$ 和第 k 通道 m 分段的带噪语音能量谱 $\left|Y\left(k,n\right)\right|^2\left(n=1,\cdots,N\right)$ 的相关指数。STOI 评分 d 是带噪语音每个频带可懂性的平均值，表达式如下：

$$d=\frac{1}{KM}\sum_{k,m}d\left(k,m\right)$$

STOI 通过对纯净语音和待评价的语音进行比较从而得到评分，取值范围为 0-1。取值越高语音质量越好。

3) 指标分析

从表 1 和表 2 可以得到，在低信噪比的情况下，传统的语音分离算法并不能有效地提高语音的可懂性和自动识别率。本文提出的算法在不同的信噪比下，对分离指标 PESQ 和 STIO 都有较明显的提高。

Table 1. Comparison of PESQ indicators.
表 1. PESQ 指标对比

信噪比 (dB)	混合语音	谱减法	维纳滤波法	最小均方差估计法	本文算法
−6	1.303	1.462	1.291	1.336	1.856
−3	1.411	1.582	1.360	1.520	2.059
0	1.588	1.770	1.789	1.931	2.266
3	1.587	1.720	1.753	1.940	2.264
6	1.918	2.105	2.164	2.497	2.695

Table 2. Comparison of STIO indicators
表 2. STIO 指标对比

信噪比 (dB)	混合语音	谱减法	维纳滤波法	最小均方差估计法	本文算法
−6	0.567	0.505	0484.	0.472	0.764
−3	0.632	0.567	0.553	0.551	0.813
0	0.702	0.645	0.642	0.677	0.860
3	0.701	0.631	0.623	0.632	0.859
6	0.830	0.765	0.763	0.779	0.928

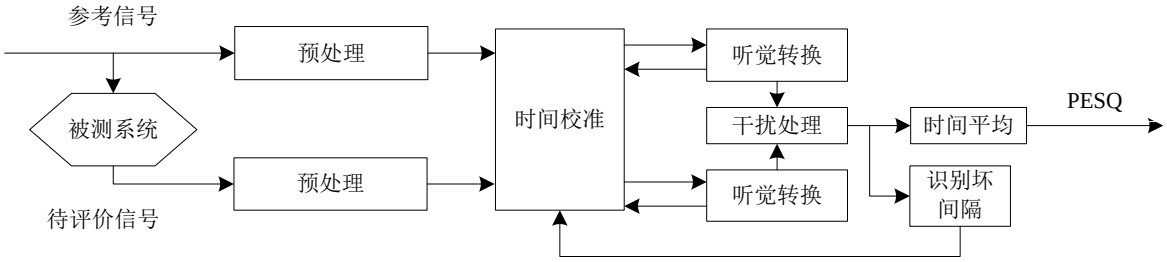

Figure 3. PESQ calculation flow chart
图 3. PESQ 计算流程图

5. 总结

传统语音增强算法的优势是计算容易，操作简单，具有很好的实时性。但在低信噪比下，并不能有效地抑制背景噪声。同时传统算法是用加噪语音相位信息直接代替纯净语音相位信息，并没有实现相位信息的增强。针对上述问题，提出理想组合掩码分离目标，并应用监督性语音分离算法进行估计，实现了语音幅值增强和相位谱同时增强。经过仿真实验证实该算法在不同的信噪比下，能够有效地抑制背景噪声和恢复相位信息，并且能够显著地提高语音的可懂性和自动识别率。

参考文献

[1] 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, 2016, 42(6): 819-833.

[2] 高银秋, 邓宗元, 杨震. 数字音频产品中基于人耳听觉感知特性的水印嵌入系统设计[J]. 南京邮电大学学报, 2006, 26(5): 56-64.

[3] Paliwal, K., Wqjcicki, K. and Shannon, B. (2011) The Importance of Phase in Speech Enhancement. *Speech Communication*, **53**, 465-494. https://doi.org/10.1016/j.specom.2010.12.003

[4] 薛慧君, 李盛, 路国华, 等. 基于短时相位谱补偿的非接触语音检测增强算法研究[J]. 中国医疗设备报, 2013, 11(28): 12-14.

[5] 孟祥彩, 王中训, 刘伟, 等. 基于相位影响的单声道语音增强改进算法[J]. 通信电声, 2016, 11(40): 59-61.

[6] 王栋, 贾海蓉. 改进相位谱补偿的语音增强算法[J]. 西安电子科技大学学报, 2017, 3(44): 83-88.

[7] Rix, A.W., Beerends, J.G., Hollier, M.P., *et al*. (2001) Perceptual Evaluation of Speech Quality(PESQ)—A New Method for Speech Quality Assessment of Telephone Networks and Codes. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001. *Proceedings*, 749-752.

[8] 王燕南. 基于深度学习的说话人无关单通道语音分离[D]. 合肥: 中国科学技术大学, 2017.

[9] Hinton, G.E., Srivastava, N., Krizhevsky, A., *et al*. (2012) Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *Computer Science*, **3**, 212-213.

[10] Wang, Y., Han, K. and Wang, D.L. (2013) Exploring Monaural Features for Classification-Based Speech Segregation. *IEEE Transaction on Audio Speech & Language Processing*, **21**, 270-279. https://doi.org/10.1109/TASL.2012.2221459

[11] Taal, C.H., Hendriks, R.C., Heusdens, R., *et al*. (2011) An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech. *IEEE Transactions on Audio Speech & Language Processing*, **19**, 2125-2136. https://doi.org/10.1109/TASL.2011.2114881

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org