基于双耳语音分离和丢失数据技术的鲁棒语音识别算法

周 琳[1]，赵一良[1]，朱竑谕[1]，汤一彬[2]
(1. 东南大学信息与工程学院水声信号处理教育部重点实验室，江苏南京 210096；2. 河海大学物联网学院，江苏常州 213022)

摘要：鲁棒语音识别技术在人机交互、智能家居、语音翻译系统等方面有重要应用。为了提高在噪声和语音干扰等复杂声学环境下的语音识别性能，基于人耳听觉系统的掩蔽效应和鸡尾酒效应，利用不同声源的空间方位，提出了基于双耳声源分离和丢失数据技术的鲁棒语音识别算法。该算法首先根据目标语音的空间方位信息，在双耳声信号的等效矩形带宽(Equivalent Rectangular Bandwidth, ERB)子带内进行混合语音信号的分离，从而得到目标语音的数据流。针对分离后目标语音在频域存在频谱数据丢失的问题，利用丢失数据技术修正基于隐马尔科夫模型的概率计算，再进行语音识别。仿真实验表明，由于双耳声源分离方法得到的目标语音数据去除了噪声和干扰的影响，所提出的算法显著提高了复杂声学环境下的语音识别性能。

关键词：空间听觉；双耳声源分离；丢失数据技术；误识率

中图分类号：H107　　　**文献标识码：**A　　　**文章编号：**1000-3630(2019)-05-0545-09

DOI 编码：10.16300/j.cnki.1000-3630.2019.05.011

Robust speech recognition algorithm based on binaural speech separation and missing data technique

ZHOU Lin[1], ZHAO Yi-liang[1], ZHU Hong-yu[1], TANG Yi-bin[2]
(1. *Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education*, *School of Information Science and Engineering*, *Southeast University, Nanjing* 210096, *Jiangsu*, *China*; 2. *College of Internet of Things Engineering*, *Hohai University*, *Changzhou* 213022, *Jiangsu*, *China)*

Abstract: Robust speech recognition has an important application in human-computer interaction, smart home, voice translation system and so on. In order to improve the speech recognition performance in complex acoustic environment with noise and speech interference, a robust speech recognition algorithm based on binaural speech separation and missing data technique is proposed in this paper. First, according to the azimuth of the target sound source, the algorithm separates the mixed data in the sub-bands of equivalent rectangular bandwidth (ERB), and obtains the data flow of the target sound source. Then, in order to solve the problem that the target source loses spectral data in some ERB sub-bands, the probability calculation based on hidden Markov model is modified by using the missing data technique, and finally the reconstructed spectrum data is utilized for speech recognition. The simulation results show that the proposed algorithm can improve the performance of speech recognition in complex acoustic environment, because the influence of noise and interference on the target sound source data is neglected after binaural speech separation.

Key words: spatial hearing; binaural speech separation; missing data technique.; speech recognition; word error rate (WER)

0 引 言

鲁棒语音信号处理研究认为，实际应用环境和模型训练环境的不匹配是造成了识别系统性能下降的主要原因，因此尽可能减小训练环境和测试环境的失配，是当前鲁棒语音信号处理系统的主要研究方向，常用的方法包括鲁棒特征参数提取、特征补偿和模型自适应等。

人耳听觉系统在实际嘈杂环境下的感知能力是非常强的，BREGMAN[1]结合心理和生理声学研究，分析人耳听觉系统的鸡尾酒效应，指出人耳听觉感知过程可以分为两个阶段：第一，声学信号的切分(segmentation)过程，第二，属于同一声源的感知成分的组合(grouping)过程，从而形成不同声源连贯的数据流(coherent stream)。也就是说听觉系统的感知过程，实际上是听觉场景中不同声源信号的重

收稿日期：2018-09-14；**修回日期：**2018-10-19
基金项目：国家自然科学基金(61571106、61501169、61201345)、中央高校基本科研业务费专项资金(2242013K30010)
作者简介：周琳(1978－)，女，江苏镇江人，副教授，研究方向为语音、声学信号处理。
通讯作者：周琳，E-mail: Linzhou@seu.edu.cn

组织过程，混合声信号中属于同一声源的分量组织到同一个数据流中，得到各个声源对应的数据流，使得人耳听觉系统可以区分不同的声源。由此可见，包含目标声源、噪声和干扰的混合声信号分离和重构是听觉系统声学感知和理解的基础，也为语音信号和声学信号的鲁棒性研究提供了新的方向，因此我们从混合语音分离的角度来研究语音识别系统的鲁棒性。

目前常用的语音分离方法包括：基于基函数的分离方法、基于模型的分离方法和基于计算听觉场景分析(Computational Auditory Scene Analysis, CASA)的方法。基于模型和基于基函数的方法，在实际语音分离中的性能下降都是由于训练环境和测试环境的不匹配导致的。而基于 CASA 的语音分离是根据听觉系统对声学事件的重组织过程实现不同声源的分离[2]，目前认为在 CASA 框架下，引入基于丢失、不可靠声学信息的分类，可以规避训练和测试环境的不匹配问题。通过对混合信号的时频单元(Time-Frequency, TF)估计理想二进制掩蔽(Ideal Binary Mask, IBM)，将其作为各个源信号的标识位，从而形成各个声源对应的时频单元，不仅可以解决欠定语音分离问题，还可以大幅提高噪声环境下分离语音的信噪比、可懂度和识别率，因此基于 CASA 估计 IBM 已经成为 CASA 的主要目标。

CASA 的难点是提取具有感知区分性的分离特征参数用于估计 IBM，常用的分离特征参数包括：基音周期、幅度调制(Amplitude Modulation, AM)、幅度调制谱(Amplitude Modulation Spectrogram, AMS)、Gammatone 频谱倒谱系数(Gammatone Frequency Cepstral Coefficients, GFCC)[3]等。但是以上分离特征参数存在明显不足，首先在噪声环境下，基音周期估计的准确性受到影响，同时基音周期和说话人、说话内容关系密切，因此仅仅依赖于基音、谐波分量来切分和组合感知单元，会严重影响分离的效果。其次，基音周期、谐波可以用于浊音段的分离，但由于语音信号中的清音成分没有谐波结构，且能量较小，更容易受到干扰，因此目前的 CASA 不具备分离清音的能力。

针对当前 CASA 的不足，本文对基于空间方位感知的双耳语音分离进行研究，是基于以下考虑：首先，人耳听觉系统的双侧听觉神经系统能够分析和整合同侧、对侧声信号，根据双耳声信号，人耳可以检测最多 5 个声源信号；其次声源的空间方位信息与语音信号内容、说话人无关，即使待分离的源信号基音、谐波特征与训练数据不同，也能依据方位信息进行有效分离。

基于方位信息的语音分离具有以上优势，目前有不少该方向的研究工作。YAO 等[4]将双耳声源定位和盲源分离方法相结合，用于包含语音和噪声的混合双耳声信号的语音分离。ANDRESA 等[5]则在线性约束最小方差框架下实现双耳声信号的波束成形。ZOHOURIAN 等[6]则利用耳间时间差(Inter-aural Time Difference, ITD)、耳间强度差(Inter-aural Level Differences, ILD)特征参数，基于最小均方误差(Minimum Mean Squared Error, MMSE)准则进行双耳声源定位，在此基础上，利用双耳广义旁瓣抵消器(Generalized Sidelobe Canceller, GSC)波束形成方法用于分离目标说话人语音。基于波束形成的不足就是这些方法没有充分利用双耳的空间特征信息。MUROTA 等[7]针对这一问题，提出了对左、右耳声信号利用不同的统计模型进行建模，再基于最小均方误差谱幅度估计(Minimum Mean Square Error-Short Time Spectral Amplitude, MMSE STSA)对混合语音进行分离。

除了基于波束成形的语音分离，基于模式识别的双耳语音分离也是主要的研究方向。KIM 等[8]基于 ITD、ILD 的方差对频点的掩蔽值进行估计，基于频点进行目标声源的分离。由于基于频点的分离方法，容易受到噪声和混响的干扰，会导致频点分类的错误。HARDING 等[9]在听觉分析滤波器 Gammatone 子带内利用 ITD 和 ILD 参数，基于直方图的概率模型实现子带分离。但要求测试声源的角度设置，与训练过程保持一致，否则会造成声源分类的误判。KERONEN 等[10]、ALINAGHI 等[11]利用高斯混合模型(Gaussian Mixed Model, GMM)对混合矢量(Mixing Vector, MV)、ITD、ILD 进行建模，用于解决 TF 单元的分类问题，但混响对该类算法性能的影响较大。WANG 等[12]将深度神经网络(Deep Neural Networks, DNN)引入到语音分离中，将双耳语音分离看成有监督的学习问题，并将空间特征线索从 ITD、ILD 扩展为双耳互相关函数(Cross Correlation Function, CCF)和波束成形后的频谱特征参数[13]，用于训练 DNN。JIANG 等[14]同时提取双耳和单耳特征用于训练每一个频带的 DNN 网络，从而进行二值分类。YU 等[15]则利用 DNN 对 TF 单元的双耳特征线索进行建模，并利用双耳房间脉冲响应(Binaural Room Impulse Responses, BRIR)与单声源信号的卷积结果作为训练样本，这样 DNN 对混响环境下的双耳特征线索进行建模，但如果训练阶段使用的 BRIR 与测试的 BRIR 不匹配，会造成分离语音质量的下降。

基于 GMM、DNN 等模型的 ITD、ILD 子带分

离中，训练和测试需要设置一致的目标声源、干扰声源方位，这一条件限制了算法的应用场合。针对这一不足，本文提出在 Gammatone 子带内基于双耳声信号的相似度实现子带的分离，在此基础上，利用丢失数据技术实现分离后目标语音数据流的识别。本文利用 Gammtone 滤波器组首先对双耳混合声信号进行子带分析，在子带内通过双耳空间特征参数 ITD、ILD，基于双耳间声道的相似度，实现子带目标声源的掩蔽值计算，从而得到了目标声源的数据流。分离后的目标声源流在丢失数据(missing data)技术框架下进行丢失频谱的估计和重建，用于语音识别。本文算法基于耳间声信号的相似度进行目标声源分离，避免了对目标声源、干扰声源方位角度的限制，同时本文仅利用双耳空间特征进行目标声源分离时，当目标声源、干扰声源为语音信号时，也可以实现准确的语音分离和识别。针对不同类型、不同方位的噪声环境下的仿真实验表明，本文算法的识别性能均有明显提升。

1　基于空间分离和丢失数据的语音识别系统结构

本文提出的基于双耳语音分离和丢失数据技术的语音识别算法结构如图 1 所示。针对双耳声信号，算法包括训练和测试两部分。测试阶段，双耳声信号经过 Gammtone 滤波后，根据目标语音的方位角，在每个子带内计算掩蔽值，用于混合双耳声信号的分离，得到目标语音的数据流后，通过丢失数据技术进行目标语音识别。

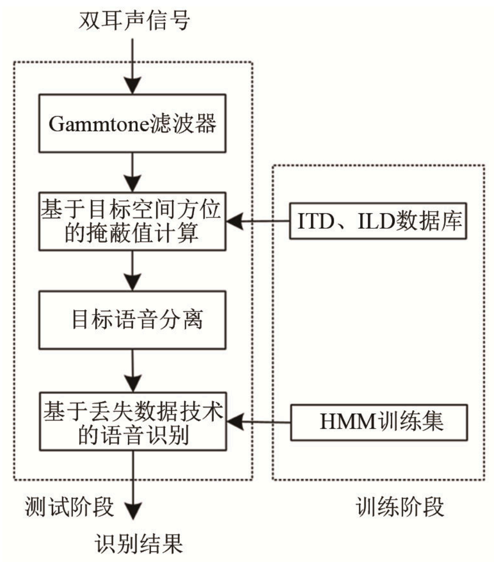


图 1　基于双耳声源分离和丢失数据的语音识别算法结构框图
Fig.1　Framework of speech recognition algorithm based on binaural speech separation and missing data technique

1.1　基于空间方位的双耳语音分离

根据图 1 的算法结构，训练阶段利用头相关脉冲响应函数 (Head Related Impulse Response, HRIR)与单声道白噪声信号进行卷积，得到[−90°, 90°]方位角范围内间隔为 5°的方向性双耳声信号，这里−90°表示正左方，0°表示正前方，90°表示正右方。训练时采用的方向性双耳声信号只包含特定方位的单个声源，用于建立每个方位角对应的 ITD 和 ILD 数据库，其中 ITD 定义为双耳声信号互相关函数最大值对应的延迟：

$$D_{\mathrm{ITD},\tau}=\arg\max_{k}\left(\sum_{m=0}^{N-|k|-1}x_{\mathrm{left}}(\tau,m)x_{\mathrm{right}}(\tau,m+k)\right) \tag{1}$$

其中，$x_{\mathrm{left}}(\tau,m)$、$x_{\mathrm{right}}(\tau,m)$ 分别表示给定方位角的训练数据对应的左、右耳声信号；τ 表示帧号。

所有帧 ITD 的均值，设定为该方位角 θ 的训练值，记为 $D_{\mathrm{ITD}}(\theta)$。

同样通过训练数据得到各个方位角对应的 ILD。假设 $X_{\mathrm{left}}(\tau,\omega)$、$X_{\mathrm{right}}(\tau,\omega)$ 分别表示 $x_{\mathrm{left}}(\tau,m)$、$x_{\mathrm{right}}(\tau,m)$ 的频域信号，则第 τ 帧信号的耳间强度差 ILD 计算如下：

$$D_{\mathrm{ILD},\tau}(\omega)=20\lg\frac{|X_{\mathrm{right}}(\tau,\omega)|}{|X_{\mathrm{left}}(\tau,\omega)|} \tag{2}$$

所有帧 D_{ILD} 的均值设置为对应方位角 θ 的 D_{ILD} 训练值，记为 $D_{\mathrm{ILD}}(\theta,\omega)$。由此可见，ITD 参数与频率 ω 无关，而 ILD 参数与频率 ω 相关。

训练完成后，讨论实际测试中的双耳声源分离过程，设 $x_{\mathrm{L}}(t)$、$x_{\mathrm{R}}(t)$ 分别表示测试混合双耳声信号的左、右耳信号，与训练阶段不同，测试双耳声信号包含不同方位的两个或者两个以上声源，因此称之为混合双耳声信号。由于 Gammatone 滤波器组能够模拟听觉神经元的脉冲响应，从而能对人耳基底膜的分频感知特性进行建模，因此采用 Gammatone 滤波器组对混合双耳声信号进行子带划分。Gammtone 滤波器组的脉冲响应函数为：

$$g_i(t)=Ct^{n-1}\mathrm{e}^{-2\pi b_i t}\cos(2\pi f_i t+\varphi_i)u(t) \tag{3}$$

其中，i 代表滤波器的序号；t 为时间；C 表示滤波器增益；n 代表滤波器的阶数；b_i 为滤波器的衰减因子；f_i 表示中心频率；$u(t)$表示阶跃信号。

利用 Gammtone 滤波器分别对 $x_{\mathrm{L}}(t)$、$x_{\mathrm{R}}(t)$ 进行子带滤波，得到第 i 个 Gammtone 滤波器的输出 $x_{\mathrm{L}}(i,t)$、$x_{\mathrm{R}}(i,t)$：

$$\begin{cases}x_{\mathrm{L}}(i,t)=x_{\mathrm{L}}(t)*\mathrm{g}_i(t)\\ x_{\mathrm{R}}(i,t)=x_{\mathrm{R}}(t)*\mathrm{g}_i(t)\end{cases} \tag{4}$$

由于语音信号在频域具有良好的稀疏性，并且人耳听觉系统具有掩蔽效应，我们将不同声源信号

的频点离散正交性[16]扩展到子带正交性，即用 $S_{k_1}(i,\omega)$ 表示第 j 个声源、第 i 个子带信号的傅里叶变换，则在第 i 个子带内，不同声源的声信号满足：

$$S_{k_1}(i,\omega)\cdot S_{k_2}(i,\omega)=0,\quad \forall i,\omega \quad \forall k_1\neq k_2 \tag{5}$$

其中，k_1、k_2 代表两个不同声源的标号。

根据子带正交性条件，在任意一个子带内，至多只有一个声源信号占主导。以右耳信号为例，子带内的混合信号可做近似为

$$S_{J(i)}(i,\omega)=X_{\mathrm{R}}(i,\omega) \tag{6}$$

其中，$J(i)$ 表示第 i 个子带上占主导的声源索引号；$X_{\mathrm{R}}(i,\omega)$ 表示 $x_{\mathrm{R}}(i,\omega)$ 的傅里叶变换。

对第 k 个声源建立二值掩码：

$$M_k(i,\omega)=\begin{cases}1, & J(i)=k\\ 0, & 其他\end{cases} \tag{7}$$

这样，可以从混合信号中得到第 k 个声源在第 i 个子带的频域信号 $S_k(i,\omega)$：

$$S_k(i,\omega)=M_k(i,\omega)X_{\mathrm{R}}(i,\omega),\ \forall i,\omega \tag{8}$$

双耳声信号包含不同声源的方位信息，我们利用双耳 ITD、ILD 估计掩蔽值 $M_k(i,\omega)$。这里假定目标语音、干扰声源的方位已知，实际上，双耳声源定位算法可以用于估计双耳混合声信号中声源的个数和方位信息，由于双耳声源定位算法不是本文研究重点，因此这里我们假设目标声源方位角已知。

假设目标语音和干扰声源的方位分别为 θ_1、θ_2。通过左、右耳通道间的最大相似度，估计每个子带相似度最高的方位，即第 i 个子带的声源索引号 $J(i)$：

$$J(i)=\arg\min_k \sum_{\omega}\frac{\left|D_{\mathrm{ILD}}(\theta_k,\omega)\cdot \mathrm{e}^{\mathrm{j}D_{\mathrm{ILD}}(\theta_k)\cdot\omega}X_{\mathrm{L}}(i,\omega)-X_{\mathrm{R}}(i,\omega)\right|^2}{1+D_{\mathrm{ILD}}(\theta_k,\omega)^2}\quad k=1,2 \tag{9}$$

其中，$X_{\mathrm{L}}(i,\omega)$、$X_{\mathrm{R}}(i,\omega)$ 分别表示测试双耳声信号 $x_{\mathrm{L}}(i,t)$、$x_{\mathrm{R}}(i,t)$ 的频谱。

根据式(9)，计算每个子带相对于目标语音和干扰声源的方位 θ_1、θ_2 的耳间相似度，如果耳间相似度越高，则对应式(9)的差值越小，以此进行子带的分类，实现目标语音的分离。

根据每个子带的声源标识 $J(i)$ 和式(7)、(8)，可以得到目标语音对应的子带频域信号 $S_1(i,\omega)$。

1.2 基于丢失数据的语音识别

由于本文利用 Gammtone 滤波组对双耳声信号进行子带划分，得到目标语音对应各个 Gammtone 子带的频域信号，因此选择基于 Gammtone 子带频谱的 RateMap 参数作为 HMM 语音识别的特征参数，RateMap 定义为每个子带信号 $S_1(i,\omega)$的均值组成的向量。

基于 HMM 的语音识别利用 GMM 模型对每个状态的 RateMap 参数进行建模，假设 GMM 包含 M 个高斯分量，协方差矩阵为对角阵，则某一状态下 RateMap 的概率密度函数表示为

$$b(y)=\sum_{m=1}^{M}\omega_m f_m(y)=\sum_{m=1}^{M}\omega_m\frac{1}{(2\pi)^{p/2}|\boldsymbol{C}_m|^{1/2}}\exp\{\frac{1}{2}(y-\boldsymbol{\mu}_m)\boldsymbol{C}_m^{1/2}(y-\boldsymbol{\mu}_m)^{\mathrm{H}}\} \tag{10}$$

其中，y 表示 RateMap 特征参数；M 表示高斯混合分量数目；ω_m 表示高斯混合分量的权重；$f_m(y)$ 表示第 m 个高斯分量的概率分布；$\boldsymbol{\mu}_m$、$\boldsymbol{C}_m$ 分别表示高斯分量的均值向量、协方差阵；p 表示 RateMap 参数的维数。

根据式(8)，目标语音的子带频域信号 $S_1(i,\omega)$ 是根据二进制掩码 $M_1(i)$ 计算得到，这样当 $M_1(i)$ 的取值为 0 时，$S_1(i,\omega)$ 也为 0，则表示该子带的频域数据是丢失的。因此分离后，需要把 RateMap 特征参数 y 分为可靠分量 y_r 和不可靠分量 y_u，其中 y_u 表示丢失的子带特征参数，即 $y=(y_r,y_u)$。

由于 RateMap 特征参数存在丢失的问题，直接利用丢失数据技术[17]对式(10)进行修正，其中边缘概率方法直接忽略丢失的特征参数，则式(10)可改写为

$$b(y_r)=\sum_{m=1}^{M}\omega_m f_m(y_r) \tag{11}$$

丢失数据技术中的数据估计方法则将条件概率 $b(y_u|y_r)$ 中概率最大值对应的特征参数作为丢失特征 y_u 的估计值。对于高斯分布，即为均值：

$$E_{y_u|y_r}\{y_u\}=\int b(y_u|y_r)y_u\mathrm{d}y_u \tag{12}$$

根据 Bayes 准则：

$$b(y_u|y_r)=\sum_{m=1}^{M}P(m|y_r)f_m(y_u) \tag{13}$$

其中：

$$P(m|y_r)=\frac{\omega_m f_m(y_r)}{b(y_r)}=\frac{\omega_m f_m(y_r)}{\sum_{m=1}^{M}\omega_m f_m(y_r)} \tag{14}$$

$P(m|y_r)$ 表示式(13)混合分量的权值。

将式(13)代入到式(12)，得到：

$$E_{y_u|y_r}\{y_u\}=\sum_{m=1}^{M}P(m|y_r)\int f_m(y_u)y_u\mathrm{d}y_u=\sum_{m=1}^{M}P(m|y_r)\cdot\boldsymbol{\mu}_{u|m} \tag{15}$$

数据估计方法将均值 $E_{yu|yr,i}\{y_u\}$ 替代特征向量

中丢失的特征 y_u，从而构成完整的 RateMap 参数，使用常规的 HMM 解码方法即可得到最后的识别结果。

整体而言，本文算法利用 Gammtone 滤波器，对包含干扰信号的混合双耳声信号进行子带分析，在频域上根据目标语音的方位信息，基于式(9)，获取目标语音的子带数据。根据目标语音的子带信号，计算对应的 RateMap 参数，并针对 RateMap 存在特征参数丢失的情况，根据式(11)对 HMM 的概率计算进行修正，或者利用式(15)对丢失特征进行估计，再通过常规的 HMM 方法进行识别，得到最终的识别结果。

2　基于 MATLAB 平台的双耳声源分离和丢失数据的孤立词识别性能分析

2.1　仿真实验参数设置

基于 HMM 的孤立词识别系统，本节详细分析基于双耳声源分离和丢失数据技术的鲁棒语音识别系统性能。本文选择 TIMIT 语音库[18]中的 21 个单词作为孤立词，将每个孤立词的 144 个样本用于训练，71 个样本用于测试，这样共有 3 024 个样本用于训练，1 491 个样本用于测试。样本采样率为 16 kHz，Gammatone 滤波器组通道数为 64，对应的中心频率取值范围为 50～8 000 Hz，滤波器的阶数取 4。每个孤立词的 RateMap 参数采用对应的 HMM 模型，状态数为 10，每个状态包含 8 个高斯分量。

由于本文算法是基于混合声信号中不同声源的空间方位，实现目标语音的分离，因此需要得到不同空间方位的双耳声信号，这里通过将单声道的源信号和对应方位的 HRIR 进行卷积，得到对应方位的左、右耳双耳声信号，生成过程如图 2 所示。

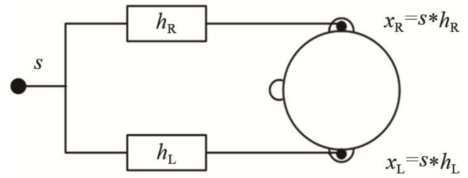


图 2　方向性双耳声信号生成过程
Fig.2　The generation of binaural signal

图 2 中，s 表示单声道的源语音信号，h_R 与 h_L 分别为对应方位角的 HRIR 数据，x_L、x_R 分别表示生成的左、右耳声信号，*表示卷积运算。

基于 HMM 的孤立词识别系统对方位角为 0°的目标语音进行识别，这样测试声信号的特征参数包含了 0°方位角的空间信息。相对应的，图 1 中训练阶段，HMM 模型集采用的同样是方位角为 0°的训练样本。

本节的仿真实验主要分为两部分，首先分析干扰声源为噪声信号时，本文基于双耳声源分离和数据丢失的语音识别系统性能，采用 NoiseX92[19]中的白噪声、工厂噪声和粉红噪声。其次分析干扰声源为语音信号时，本文所提算法的性能，干扰语音选择了 CHAINS Speech Corpus[20]语音库 SOLO 中的一段女声语音。测试时按不同的信噪比将测试样本集中的目标语音与干扰噪声、干扰语音进行混合，信噪比(Signal Noise Ratio, SNR)取 0、5、10、15、20 dB。两类仿真测试过程中，目标语音的方位角均为 0°，干扰噪声、干扰语音的方位角分别设置为 30°、60°、90°，目标语音和干扰信号的方位角分布如图 3 所示。

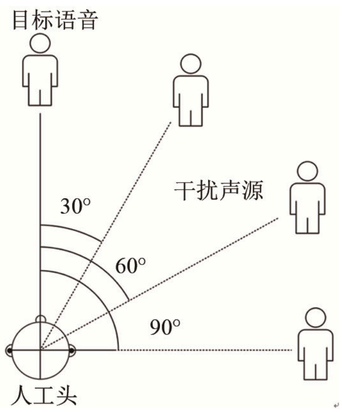


图 3　目标语音与干扰声源的方位示意图
Fig.3　The distribution of target source and interference sources

由于丢失数据技术分为基于边缘概率的方法和基于数据估计的方法，因此本文分别对基于双耳声源分离和边缘概率的孤立词识别算法、基于双耳声源分离和数据估计的孤立词识别算法性能进行分析，采用误识率(Word Error Rate, WER)作为识别系统性能指标。同时我们为了考虑系统性能的上限，给出子带分离的理想掩蔽值(也称为理想掩膜)。理想掩蔽值直接根据每个 Gammatone 子带内的目标语音和干扰声源的能量，计算对应的信噪比获得局部判决值(Local Criterion, LC)，通过设定 LC 阈值对每个子带进行目标语音的分类：

$$T_{LC} = 10\lg\frac{\sum_{\omega}|S(i,\omega)|^2}{\sum_{\omega}|N(i,\omega)|^2} \quad (16)$$

其中，$|S(i,\omega)|^2$ 表示目标语音的能量，$|N(i,\omega)|^2$ 表示干扰声源的能量。

LC 阈值一般取在−5～5 dB 之间，本文取 LC 阈值为−5 dB，理想掩蔽值 M_{mask} 表示为

$$M_{\mathrm{mask}}(i)=\begin{cases}1, & L_{\mathrm{C}}(i)>-5\\ 0, & \text{其他}\end{cases} \tag{17}$$

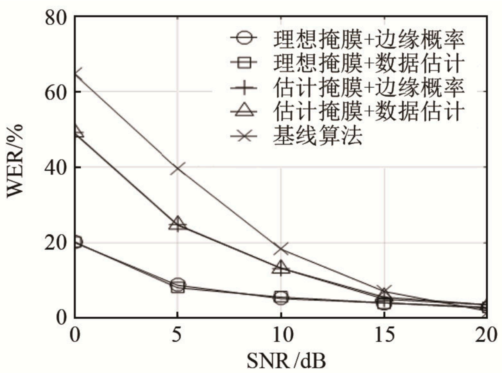
$M_{\mathrm{mask}}(i)$ 取值为 1，表示该子带属于目标语音，取值为 0，表示该子带属于干扰声源。

我们将基于 MFCC 系数和 HMM 模型的识别系统作为基线系统，因此本文比较三种识别算法的性能：基线系统、基于双耳声源分离和丢失数据技术的识别系统、基于理想掩蔽值的双耳声源分离和丢失数据技术的识别系统。

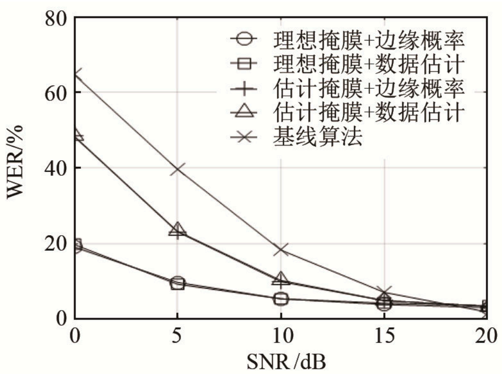
2.2 实验一：干扰为噪声的仿真结果

首先给出干扰分别为白噪声、粉红噪声和工厂噪声时，不同信噪比(Signal Noise Ratio, SNR)时，不同识别算法的误识率结果，如图 4～6 所示。针对每一类噪声，图示自上而下分别表示噪声在 30°、60°、90°的误识率 WER 比较结果。

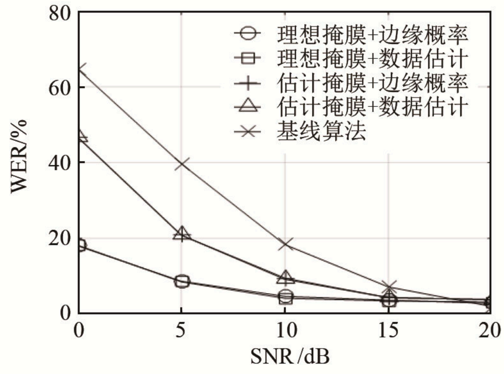
首先根据图 4～6 的仿真结果，不同类型噪声条件下，本文算法的 WER 均低于基线系统。同时，信噪比越低，本文算法的性能改善越明显，这是由于信噪比越低，基线系统的语音特征参数受到的噪声干扰越严重，而本文算法利用方位信息分离出的目标语音数据，受到噪声的干扰明显减少。



(a) 干扰噪声位于 30°



(b) 干扰噪声位于 60°



(c) 干扰噪声位于 90°

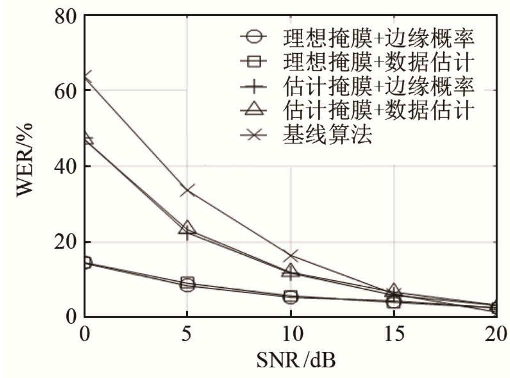
图 4 白噪声环境下不同算法的误识率比较

Fig.4 WER comparison of different algorithms in white noise environment

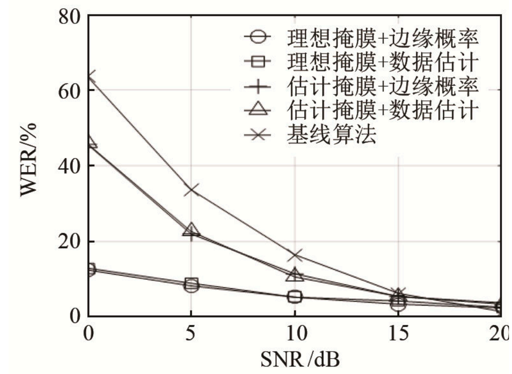
其次，对于同一类型的噪声，目标语音与干扰噪声的空间方位相隔越大，即干扰噪声的方位从 30°变化到 60°、90°时，本文算法的性能改善越明显。这是由于目标语音和干扰噪声的空间方位间隔越大，各个子带内，不同方位声源的双耳声信号 ITD、ILD 的差异性逐步增加，从而提高了子带分类的正确率，进而改善了目标语音数据流的识别率。

再者，不管是基于理想的掩蔽值，还是基于估计的掩蔽值，掩蔽值分离后的目标语音采用边缘化概率方法和采用数据估计方法，进行孤立词识别时的误识率基本一致。相比较而言，边缘化算法略优于数据估计算法的识别性能，这是由于边缘概率算法避免利用单一估计值来表示不可靠分量，相反，它只考虑丢失特征参数的分布，而数据估计技术更依赖于特征参数的统计概率分布而不是数据的可靠性，其优势在于可以得到完整的语音特征向量，从而可以采用常规的识别算法。

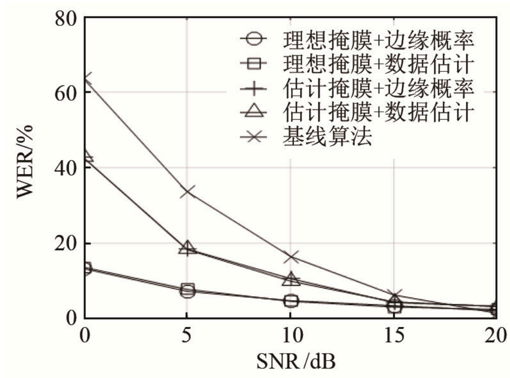
在干扰为噪声时，基于理想掩蔽值的孤立词识别性能要优于基于估计掩蔽值的识别性能。这是由于理想掩蔽值对目标语音子带的判决更加准确，分离后的目标语音数据流基本只包含目标语音成分，因此理想掩蔽值对应的识别系统性能可以作为基于空间分离的识别系统上限。根据图 4～6，本文算法的识别性能和基于理想掩蔽值的识别系统性能存在一定的差距，根据我们目前的研究发现，当混合双耳声信号中包含了两个以上的声源时，ITD、ILD 的联合分布与单声源的 ITD、ILD 分布有较大的差异，混合双耳声信号仅仅利用欧式距离进行分离，其子带分类的准确性受到限制。因此提高混合双耳声信号中子带分类的正确率，可以显著提高后端识别系统的性能，即基于双耳声源分离和丢失数据的孤立词识别系统性能还有较大的提升空间。



(a) 干扰噪声位于 30°



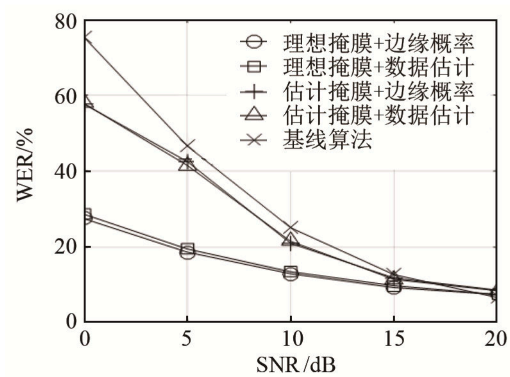
(b) 干扰噪声位于 60°



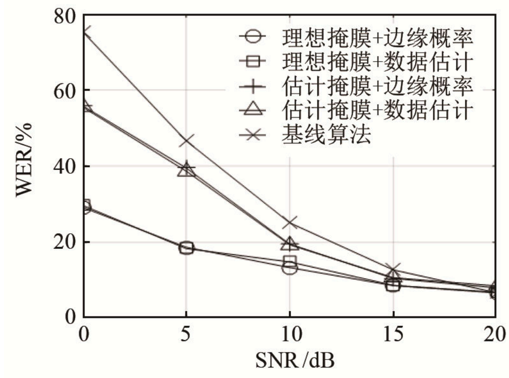
(c) 干扰噪声位于 90°

图 5　粉红噪声环境下不同算法的误识率比较

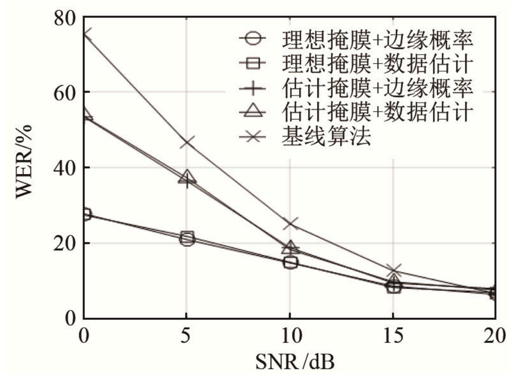
Fig.5　WER comparison of different algorithms in pink noise environment



(a) 干扰噪声位于 30°



(b) 干扰噪声位于 60°



(c) 干扰噪声位于 90°

图 6　工厂噪声环境下不同算法的误识率比较

Fig.6　WER comparison of different algorithms in factory noise environment

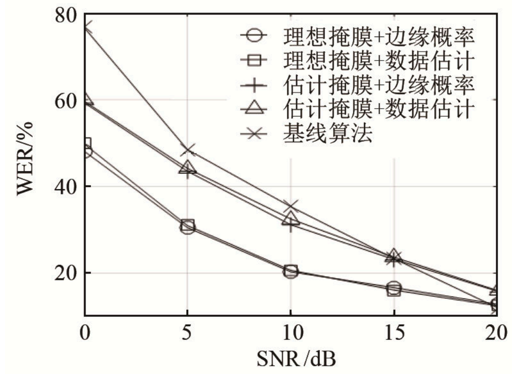
2.3　仿真实验二：干扰为语音的仿真结果

目前常用的鲁棒语音识别算法如特征补偿、模型自适应等，对提取的纯净语音特征参数采用 HMM、GMM 等进行建模，测试环境中的噪声影响了 HMM、GMM 等模型参数分布，例如概率模型的均值向量和协方差矩阵，通常采用线性方法对噪声干扰后的模型参数分布进行预测。但当干扰为语音信号时，例如有两个或者多个说话人时，由于干扰语音的特征参数与目标语音的特征参数分布相似度高，那么干扰语音对目标语音特征参数模型的影响就不能简单地利用线性模型来模拟，因此常规的鲁棒语音识别算法对干扰语音的鲁棒性较弱。本文所提算法从空间分离的角度，在空间域进行混合双耳声信号的分离，不同声源的区分特征更为明显，因此本文进一步分析在干扰为语音时，不同算法的识别性能差异。

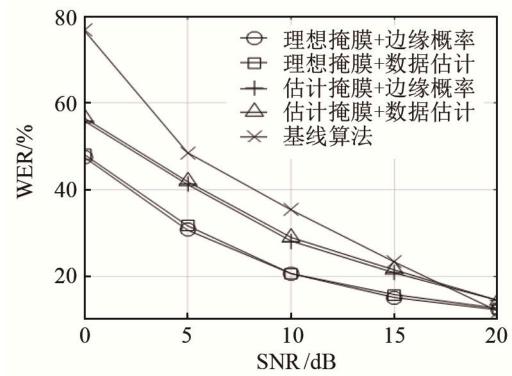
同样，目标说话人的方位固定在方位 0°，另一说话人的方位为 30°、60°和 90°，语音干扰的信干比取值范围与噪声的信噪比取值方位相同，不同算法的误识率结果如图 7 所示。

根据图 7 结果，相对于图 4～6 的识别结果，在干扰语音条件下，基线系统的误识率要远远高于

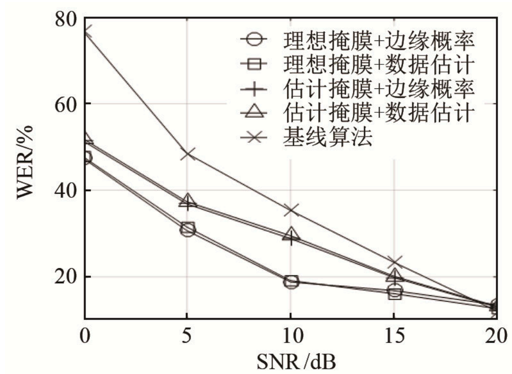
在噪声环境下的误识率，这是由于在相同的信噪比下，由于语音干扰和目标语音特征参数的相似度较高，从而无法从混合语音信号中提取有效的目标语音信号特征参数。而本文算法利用不同声源的空间方位，在空间域实现不同说话人语音信号的分离，因此本文算法相对于基线系统的性能提高，在语音干扰环境下，要比噪声环境下的性能改善明显。同时根据图 7，语音干扰和目标语音的角度分隔越大，则本文算法的性能提高越明显。



(a) 干扰噪声位于 30°



(b) 干扰噪声位于 60°



(c) 干扰噪声位于 90°

图 7　语音干扰环境下不同算法的误识率比较

Fig.7　WER comparison of different algorithms in speech interference environment

同时，在相同的信噪比下，语音干扰条件下的本文算法识别性能不如在噪声环境下的性能。特别的，理想掩蔽值对应的识别系统性能在低信噪比下的性能不如图 4～6 在噪声干扰下的识别性能，这并不是由于理想掩蔽值不能实现正确的目标声源分离，而是由于式(16)、(17)在进行子带分类时，是基于式(3)给出的两个声源信号在子带内的正交性，根据子带内目标语音和干扰语音的能量比实现分类的，这实际上是从听觉系统的感知机制得到的结论，即在一个临界频带内，人耳听觉系统由于掩蔽效应，只能感知一个声源信号，但该子带内实际包含两个或者两个以上声源信号。这样子带分类后，虽然某一子带归为目标语音，但该子带内实际上也包含了干扰语音成分，并且干扰语音对该子带内占主导的目标语音的影响，比噪声对目标语音的影响大。因此如果要进一步提高基于空间分离的识别系统在干扰语音下的识别性能，不能简单的利用式(5)计算二进制掩蔽值，用于目标声源、干扰声源的分离，而是可以采用软判决值的方法用于子带内目标语音的分离，这也是我们下一步研究的方向。

3　结 论

本文基于人耳听觉系统的掩蔽效应和鸡尾酒效应，利用不同声源信号的空间方位进行语音信号的分离，实现目标语音的数据重构，再基于丢失数据技术，进行语音识别，从而提高了语音识别系统的鲁棒性。在不同噪声环境、不同空间方位条件下的仿真实验结果表明，本文算法显著提高了识别系统的性能。

同时本文研究表明，基于空间分离和丢失数据的识别算法性能取决于子带分类的准确性，如果要提高目标语音分离的正确率，需要从两个方面入手，其一是子带分类的方法，目前课题组正在进行基于深度神经网络的子带分离算法研究，初步的仿真结果表明了该方法的有效性；其二需要对式(3)的感知正交性进行建模，不使用二进制进行子带的硬判决方法，而是利用软判决的方法实现子带的分类，从而能够准确地重构目标语音的数据流，这也是我们目前正在开展的研究工作。

参 考 文 献

[1] BREGMAN S. Auditory scene analysis: The perceptual organization of sound[M]. Cambridge, MA,US: The MIT Press, 1994.

[2] WANG D L, BROWN G. Computational auditory scene analysis: Principles, algorithms, and applications[M]. New York: Hoboken NJ: Wiley & IEEE Press, 2006.

[3] WANG Y, HAN K, WANG D L. Exploring monaural features for classification-based speech segregation[J]. IEEE Transactions on

Audio, Speech and Language Processing, 2013, **21**(2): 270-279.
[4] YAO J, XIANG Y, QIAN S, et al. Noise source separation of diesel engine by combining binaural sound localization method and blind source separation method[J]. Mechanical Systems & Signal Processing, 2017, **96**: 303-320.
[5] KOUTROUVELIS A I, HENDRIKS R C, HEUSDENS R, et al. Relaxed Binaural LCMV Beamforming[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017, **25**(1): 137-152.
[6] ZOHOURIAN M, MARTIN R. Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 2016: 430-434.
[7] MUROTA Y, KITAMURA D, KOYAMA S, et al. Statistical modeling of binaural signal and its application to binaural source separation[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, Queensland, Australia, 2015: 494-498.
[8] KIM Y I, AN S J, KIL R M. Zero-crossing based binaural mask estimation for missing data speech recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 2006: 89-92.
[9] HARDING S, BARKER J, BROWN G J. Mask estimation for missing data speech recognition based on statistics of binaural interaction[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, **14**(1): 58-67.
[10] KERONEN S, KALLASJOKI H, REMES U. Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment[J]. Computer Speech & Language, 2013, **27**(3): 798-819.
[11] ALINAGHI A, JACKSON P J B, LIU Q, et al. Joint mixing vector and binaural model based stereo source separation[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, **22**(9): 1434-1448.
[12] WANG Y, WANG D L. Towards scaling up classification-based speech separation[J]. IEEE Trans. Audio, Speech, Lang. Process. 2013, **21**(7): 1381-1390.
[13] ZHANG X, WANG D L. Deep Learning Based Binaural Speech Separation in Reverberant Environments[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017, **25**(5): 1075-1084.
[14] JIANG Y, WANG D, LIU R, et al. Binaural classification for reverberant speech segregation using deep neural networks[J]. IEEE/ACM Trans. Audio, Speech, Lang. Process. 2014, **22**(12): 2112-2121.
[15] YU Y, WANG W, HAN P. Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks[J]. EURASIP J. Audio Speech Music Proc. 2016(1): 1-18.
[16] JOURJINE A, RICKARD S, YILMAZ O. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul Turkey, 2000: 2985-2988.
[17] COOKE M, GREEN P, JOSIFOVSKI L, et al. Robust automatic speech recognition with missing and unreliable acoustic data[J]. Speech Communication, 2001, **34**(3): 267-285.
[18] FISHER W, DODDINGTON G, and GOUDIE-MARSHALL K. The DARPA speech recognition research database: Specifications and status[C]//DARPA Speech Recognition Workshop, Palo Alto, CA, 1986: 93-99.
[19] VARGA A, STEENEKEN H, TOMLINSON M, et al. The NOISEX-92 study on the effect of additive noise on automatic speech recognition[R]. Speech Research Unit, Defense Research Agency, Malvern, UK, 1992.
[20] CUMMINS F, GRIMALDI M, LEONARD T, et al. The chains speech corpus: Characterizing individual speakers[C]//11th International Conference Speech and Computer, St. Petersburg, Russia, 2006: 1-6.