基于计算听觉场景分析和语者模型信息的语音识别鲁棒前端研究

关 勇 [1,2] 李 鹏 [3] 刘文举 [1] 徐 波 [1,3]

摘 要 传统抗噪算法无法解决人声背景下语音识别 (Automatic speech recognition, ASR) 系统的鲁棒性问题. 本文提出了一种基于计算听觉场景分析 (Computational auditory scene analysis, CASA) 和语者模型信息的混合语音分离系统. 该系统在 CASA 框架下, 利用语者模型信息和因子最大矢量量化 (Factorial-max vector quantization, MAXVQ) 方法进行实值掩码估计, 实现了两语者混合语音中有效地分离出目标说话人语音的目标, 从而为 ASR 系统提供了鲁棒的识别前端. 在语音分离挑战 (Speech separation challenge, SSC) 数据集上的评估表明, 相比基线系统, 本文所提出的系统的语音识别正确率提高了 15.68 %. 相关的实验结果也验证了本文提出的多语者识别和实值掩码估计的有效性.

关键词 计算听觉场景分析, 语音分离, 鲁棒语音识别, 因子最大矢量量化, 语者识别
中图分类号 TP391

A Robust Front-end for Speech Recognition Based on Computational Auditory Scene Analysis and Speaker Model

GUAN Yong[1,2] LI Peng[3] LIU Wen-Ju[1] XU Bo[1,3]

Abstract Conventional noise robust speech recognition system does not work well when human speech is presented in the background. In this paper, a computational auditory scene analysis (CASA) and speaker model based speech segregation system is proposed to solve this problem. By utilizing speaker model and factorial-max vector quantization (MAXVQ) to estimate real-value masks in CASA framework, a robust front-end for speech recognition is constructed. Evaluations on speech separation challenge (SSC) showed that the proposed system won 15.68 % improvement over the baseline system. The results of evaluation also proved the validity of the multi-speaker recognition and the real-value mask estimation module.

Key words Computational auditory scene analysis (CASA), speech segregation, robust speech recognition, factorial-max vector quantization (MAXVQ), speaker recognition

背景噪音环境下的鲁棒性问题是影响语音识别 (Automatic speech recognition, ASR) 走向实际应用的主要挑战之一[1]. 传统的提高噪声环境下语音识别鲁棒性的方法主要有信号增强[2]、特征补偿[3]、模型自适应[4−5] 和训练数据预加噪[6] 等, 这些方法虽然在平稳噪声条件下取得了较好的效果, 但是对非平稳噪声 (如人声背景) 环境, 效果依然较差. 针对上述问题, 本文选取了背景噪声为干扰说话人的两语者混合语音作为主要研究对象, 将语音分离作为 ASR 系统的前端模块, 结合计算听觉场景分析 (Computational auditory scene analysis, CASA) 和语者模型信息, 为解决非平稳噪声环境下的鲁棒语音识别问题提出了新的思路.

与传统的 ASR 系统鲁棒前端相比, 本文具有以下几个特点: 1) 利用语音分离系统作为 ASR 的前端. 不同于已有的 CASA 系统[7] 仅仅利用初始分离得到的掩码信息进行序列组织, 本文直接利用语者模型和因子最大矢量量化 (Factorial-max vector quantization, MAXVQ) 方法[8] 推断二值掩码信息, 在语音分离过程中有效地结合了语者信息. 2) 利用多语者识别模块来识别人声背景下混合语音中存在的语者身份信息, 并利用其选择相应语者模型进行后续分离工作. 3) 针对利用二值掩码重新合成的语音存在特征缺失, 以及重新合成后的语音与 ASR 训

收稿日期 2007-12-18 收修改稿日期 2008-03-12
Received December 18, 2007; in revised form March 12, 2008
国家重点基础研究发展计划 (973 计划) (2004CB318105), 国家自然科学基金 (60675026, 60121302, 90820011), 国家高技术研究发展计划 (863 计划) (20060101Z4073, 2006AA01Z194) 资助
Supported by National Basic Research Program of China (973 Program) (2004CB318105), National Natural Science Foundation of China (60675026, 60121302, 90820011), and National High Technology Research and Development Program of China (863 Program) (20060101Z4073, 2006AA01Z194)
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. 诺基亚中国研究中心 北京 100176 3. 中国科学院自动化研究所数字内容技术研究中心 北京 100190
1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. Nokia Research Center, Beijing 100176 3. Digital Content Technology Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
DOI: 10.3724/SP.J.1004.2008.00410

练语音不匹配的问题, 本文利用语者模型信息和二值掩码进一步估计实值掩码, 得到更符合后续识别目标需要的语音分离结果.

在语音分离挑战 (Speech separation challenge, SSC) 数据集[9] 的双语者子集上的语音识别测试结果表明, 本文所提出的分离系统能够有效地提高语音识别系统的性能, 最终评估结果与基线系统相比, 识别正确率提高了 15.68 %. 同时相关实验结果也验证了本文提供的多语者估计模块和实值掩码估计模块的有效性.

本文的组织结构如下: 第 1 节介绍基于语音分离的 ASR 前端系统, 其中主要介绍了多语者识别模块, 基于语者模型和 MAXVQ 的掩码估计模块. 第 2 节系统地评估了本文提出的基于 CASA 和语者模型信息的语音分离系统作为 ASR 鲁棒前端的性能. 最后, 对全文工作进行了总结和展望.

1　基于语音分离的 ASR 前端系统

本文提供了一个语音分离系统作为 ASR 系统的前端, 该系统结合了 CASA 与统计语者模型信息. 图 1 给出了语音分离系统的详细结构, 分为训练和测试两个阶段, 主要由时频分解及特征提取模块、模型训练模块、多语者识别模块和掩码估计及再合成模块构成. 最后重新合成后的语音信号, 输入到 ASR 识别器进行语音识别.

时频分解及特征提取模块是训练和测试阶段共有的处理阶段. 与通常的特征提取方法不同, 本文使用基于听觉感知机理的 Gammatone 听觉滤波器组[10] 来分析输入信号. 再经过分帧处理, 信号被分解为一个二维的时 – 频图, 时 – 频图中的每一个单元称为时 – 频单元, 对应于某个滤波器输出的某个时间帧. 最后, 提取每个时间帧内的时 – 频单元的对数能量构成特征矢量.

在模型训练阶段, 预先准备好的大量说话人的纯净语音信号被首先送入系统的时频分解和特征提取模块. 对每一个说话人, 模型训练模块使用 K-均值聚类算法训练一个具有 K 个混合数的高斯混合模型 (Gaussian mixture model, GMM) 和 K 个码字 (Codeword) 的矢量量化器 (Vector quantizer, VQ). 这些 GMM 模型和矢量量化器, 将作为语者模型信息, 在后续的多语者识别和掩码估计模块中使用, 用于指导语音的分离.

在测试阶段, 待测的混合语音信号, 首先经过时频分解和特征提取模块, 得到二维的时 – 频图以及对数能量特征矢量. 然后经过多语者识别模块, 利用训练好的 GMM 模型, 识别出混合语音中存在的两个语者的身份信息, 供掩码估计模块选取相应的语者模型. 最后在掩码估计及再合成模块, 利用语者模型和 MAXVQ 方法来推断二值掩蔽信号, 并估计最终的实值掩码, 进而利用 Weintraub[11] 的逆滤波方法从二维时 – 频图重新合成语音信号.

本文主要介绍语音分离系统测试阶段的多语者识别模块和掩码估计及再合成模块.

1.1　多语者识别模块

多语者识别模块的目的是判断混合语音中的两个说话人的身份, 为后续的掩码估计模块选取语者模型提供信息. 为实现两语者混合语音的多语者识别, 本文采用了一个两阶段的识别方法[12]. 这里首先介绍两语者混合情况下语音帧似然得分的计算方法, 然后介绍两阶段的多语者识别算法.

1.1.1　语音帧似然得分计算

我们用 GMM 模型来描述每一个具体语者模型信息, 并利用其计算给定语音帧的似然得分

$$p(\boldsymbol{x}_t|m)=\sum_{k=1}^{K}c_m^k p(\boldsymbol{x}_t|\boldsymbol{v}_m^k,\Sigma_m^k) \tag{1}$$

其中, $\boldsymbol{x}_t$ 是第 t 帧的维数为 D 的特征向量, c_m^k 是语者 $m\in\{1,2,\cdots,M\}$ 第 k 个混合成分的混合系数, $\boldsymbol{v}_m^k$ 是均值向量, Σ_m^k 是协方差矩阵 (取对角阵), K 是 GMM 的混合数个数.

考虑两语者混合语音 (多种不同信噪比组合) 的

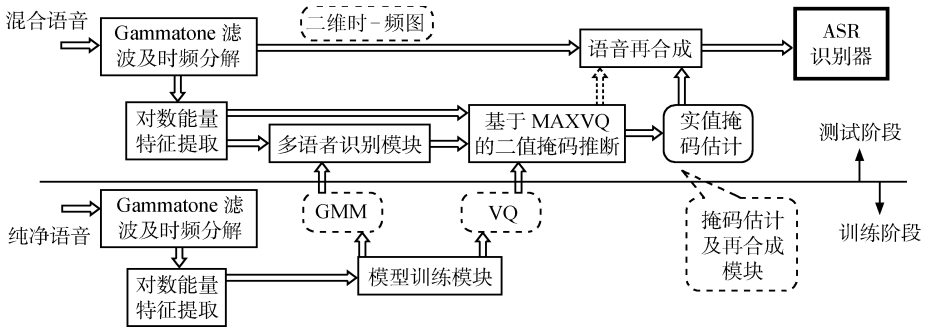


图 1　语音分离系统示意图

Fig. 1　Schematic diagram of speech separation system

测试特征与用于模型训练的纯净语音特征可能存在的不匹配现象, 我们引入了增益信息[12−13], 则式 (1) 可以改写为

$$p(\boldsymbol{x}_t|m)=\sum_r\sum_k \pi_r c_m^k p(\boldsymbol{x}_t|\boldsymbol{v}_m^k+\boldsymbol{g}_r,\Sigma_m^k) \quad (2)$$

其中, $r\in\{1,2,\cdots,R\}$ 为一个对应于总共 R 个信噪比的离散变量, 每个增益参数 $\boldsymbol{g}_r$ 对应一个先验概率 π_r.

同时我们发现, 如果特征向量的某维 d 严重偏离均值 v_{md}^k, 似然得分 $p(\boldsymbol{x}_t|m)$ 将由这些受噪声干扰的维所主导, 结果导致它会远小于真实值, 因而我们限制每一维 $(x_{td}-v_{md}^k-g_{rd})^2/(\sigma_{md}^k)^2$ 的得分. 当其取值大于某一阈值 λ 时, 将其取值固定为 λ.

实验中, 限制参数 λ 取 7.5, 增益参数取 $g_{rd}\in\{6\,\text{dB},3\,\text{dB},0\,\text{dB},-3\,\text{dB},-6\,\text{dB},-9\,\text{dB}\}$.

如文献 [13], 增益参数 $\boldsymbol{g}_r$ 及限制参数 λ 的引入, 降低了由于训练与测试特征不匹配造成的识别误差, 提高了语者识别的正确率.

1.1.2 两阶段多语者识别算法

考虑混合语音的特征受到干扰语者语音的影响, 因而我们采取两阶段的多语者识别算法.

在第一阶段, 系统利用相对纯净的语音帧判断混合语音中可能存在的语者, 并给出一个候选语者列表. 这里, 首先计算每一帧语音在各个语者模型上的似然得分, 然后计算其在所有语者模型上的归一化得分作为置信度, 以确定这一帧语音属于某一语者的可能性, 最后逐帧累加得到总的得分. 具体实现方法如下[12]:

1) 给定一帧语音 $\boldsymbol{x}_t$, 计算其对模型 m 的归一化的似然值

$$b_{\boldsymbol{x}_t}(m)=\frac{p(\boldsymbol{x}_t|m)}{\sum\limits_{m'=1}^{M}p(\boldsymbol{x}_t|m')} \quad (3)$$

2) 计算整个测试语音在模型上的得分

$$p(\boldsymbol{x}|m)=\sum_{t=1}^{T}\psi(b_{\boldsymbol{x}_t}(m))\cdot b_{\boldsymbol{x}_t}(m) \quad (4)$$

其中, $\psi(b_{\boldsymbol{x}_t}(m))$ 为与 $b_{\boldsymbol{x}_t}(m)$ 相关的置信权重, 即

$$\psi(b_{\boldsymbol{x}_t}(m))=\begin{cases}1, & \max(b_{\boldsymbol{x}_t})>\gamma\\ 0, & \text{否则}\end{cases} \quad (5)$$

由此可以得到第一阶段的处理结果, 即一个后验概率由高到低的语者列表. 实验中参数 γ 取 0.9.

经过第一阶段的处理, 我们发现, 结果列表中的前两个语者与正确语者组合的符合程度 (即识别正确率) 并不足够好, 尤其是在低信噪比的情况下. 因而我们改进算法, 进行第二阶段处理. 首先, 对第一阶段得到的候选列表, 取其排第一位的模型 (认为其正确), 与列表中前十位的模型两两组合得到新的组合语者模型 (包括同一语者的组合), 然后利用组合后的语者模型, 应用传统的语者识别算法对混合语音重新进行识别, 按最大似然原理, 得到最终识别结果, 即一个语者组合.

这里, 每个组合语者模型最终为 $K\times K$ 个混合数的混合高斯模型, 新的语者模型的参数可按以下规则确定: 组合后混合数的权重为组合前两个高斯混合权重的乘积; 组合后的均值和方差, 每一维单独处理, 比较分别来自两个语者模型待混合的高斯成分的均值, 取其较大的高斯成分的参数作为组合模型的均值和方差.

经过第二阶段处理, 最终的多语者识别结果由一对语者组合构成, 为后续的掩码估计模块选取语者模型提供两个语者的身份信息.

1.2 掩码估计及再合成模块

掩码估计及再合成模块的主要功能是利用多语者识别模块的处理结果, 选取相应的语者模型, 借助 MAXVQ 推断时频单元的掩码, 从而合成出分离后的目标语音信号. 本节将从二值掩码推断、实值掩码估计和再合成三个方面, 进行详细论述.

1.2.1 基于 MAXVQ 的二值掩码推断模块

二值掩码推断模块利用语者模型信息, 引入因子最大矢量量化模型 (MAXVQ) 算法[8], 进行二值掩码估计[14]. 在掩码估计模块, 我们选用事先训练好的 VQ 作为语者模型. 首先对于选定的每组两个 VQ 模型, 利用元素维最大化的思想, 组合成新的语者模型; 然后根据最大似然的原则, 选择与测试语音特征最匹配的码字组合, 最后根据选定的两个码字的组合, 推断二值掩码信号.

二值掩码的思想源于听觉掩蔽现象, 即在一个临界频带内, 较弱的信号会被较强的信号所掩蔽[7]. 利用二值掩蔽的思想对混合信号进行再合成构成了许多 CASA 系统实现语音分离的基础[15−16]. 在某种意义上, 基于 CASA 的语音分离问题可以看作是一个以正确划分对应于掩蔽信号 1 的前景流和对应于掩蔽信号 0 的背景流为目标的处理过程, 因而可以简化为二值掩蔽信号的估计问题.

本文使用的 MAXVQ 算法, 语者模型由 $M=2$ 个 VQ 组成, 每个 VQ 有 K 个码字 (均值矢量 $\boldsymbol{v}_m^k$, 协方差矩阵 $\{\sigma_m^k\}$, 这里 $m\in\{1,2\}$, $k\in\{1,2,\cdots,K\}$). 隐变量 $z_m\in\{1,2,\cdots,K\}$ 控制每一个 VQ 进行码字的选择. 这里假定每个 VQ 以固定概率完

全独立地选择它的码字[8], 即

$$p(z_m = k|\boldsymbol{\pi}) = \pi_m^k \tag{6}$$

那么, 选择一组隐变量 $\boldsymbol{z}$ 的概率可以按照下式进行计算

$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{m=1}^{M} p(z_m|\boldsymbol{\pi}), \quad \boldsymbol{z} = (z_1, \cdots, z_M) \tag{7}$$

给定 $M = 2$ 个语者的 VQ 模型以及隐变量, 组合后的新的语者模型的码字是通过对选中的码字使用元素最大化的方式得到的. 具体而言, 可以首先确定最终输出的第 d 维选中的 VQ 的序号 a_d[8] 为

$$a_d = \arg\max_m(v_{md}^{z_m}), \quad d \in \{1, \cdots, D\} \tag{8}$$

这样, 全部 D 维内选中的 VQ 的序号便构成了矢量 $\boldsymbol{a} = (a_1, a_2, \cdots, a_D)^{\mathrm{T}}$, 从而最终输出的均值矢量和协方差矩阵可以表示为[8]

$$\boldsymbol{v_a} = (v_{a_1 1}^{z_{a_1}}, v_{a_2 2}^{z_{a_2}}, \cdots, v_{a_D D}^{z_{a_D}})^{\mathrm{T}} \tag{9}$$

$$\Sigma_{\boldsymbol{a}} = \mathrm{diag}\{\Sigma_{a_1 1}^{z_{a_1}}, \Sigma_{a_2 2}^{z_{a_2}}, \cdots, \Sigma_{a_D D}^{z_{a_D}}\} \tag{10}$$

这样就完成了元素维最大化并建立新的语者 VQ 模型的码字的过程. 由于假定特征的每一维相互独立, 我们可以计算选定一组隐变量 $\boldsymbol{z}$ 的条件下观测到 $\boldsymbol{x}_t$ 的概率[8]

$$p(\boldsymbol{x}_t|\boldsymbol{z}, \boldsymbol{v_a}, \Sigma_{\boldsymbol{a}}) = N(\boldsymbol{x}_t|\boldsymbol{v_a}, \Sigma_{\boldsymbol{a}}) \tag{11}$$

其中, N 是高斯分布. 在此基础上, 在模型参数和先验概率 $\boldsymbol{\pi}$ 的条件下, 隐变量 $\boldsymbol{z}$ 与特征矢量 $\boldsymbol{x}_t$ 的联合概率可以按照下式进行计算[8]

$$p(\boldsymbol{x}_t, \boldsymbol{z}|\boldsymbol{v}, \Sigma, \boldsymbol{\pi}) = p(\boldsymbol{z}|\boldsymbol{\pi}) p(\boldsymbol{x}_t|\boldsymbol{z}, \boldsymbol{v_a}, \Sigma_{\boldsymbol{a}}) \tag{12}$$

这样, 利用最大似然准则, 通过比较在所有可能的隐变量 $\boldsymbol{z}$ 的选择 ($K \times K$ 种组合) 情况下观测到的特征矢量 $\boldsymbol{x}_t$ 的概率, 我们可以找到使得特征矢量 $\boldsymbol{x}_t$ 的观测概率最大的那组最可能的 $\boldsymbol{z}$, 即得到两个语者模型各自选定的码字[14].

得到最大可能的 $\boldsymbol{z}$ 后, 可以按如下的方法来估计二值掩码信号: 对输入混合语音的每一帧, 将对应于目标语者模型码字的均值为最大 (对于两语者情况, 即为目标语者模型码字均值大于干扰语者模型码字均值) 的那些频带上的掩蔽信号设为 1, 而把其他频带的掩蔽信号设为 0. 由此, 结合语者模型信息和 MAXVQ 算法, 可以推断出二值掩码信号[14].

1.2.2　实值掩码估计

由于语音分离的目标是提供一个 ASR 的鲁棒前端, 如果我们按照通常的做法, 在重新合成语音的过程中, 保留掩码为 1 的时频单元的能量, 丢弃所有掩码为 0 的时频单元的能量, 这将导致重新合成后的语音频谱损失很多, 因而在后续 ASR 特征处理的过程中造成特征缺失, 并导致识别性能的下降.

为了解决以上问题, 我们对重新合成的过程做了修改. 一个简单直接的方式就是保留所有掩码为 1 的单元的能量, 同时以一个固定比率 θ 的掩码保留原始掩码为 0 的单元的能量 ($[1, \theta]$ 掩码). 本文选择保留的能量比率 $\theta = 25\,\%$, 实验表明保留掩码为 0 的单元的能量比例在 $10\,\% \sim 40\,\%$ 范围内, 识别的性能并不发生明显的变化[17].

然而, 这种简单的处理并未充分利用语者模型信息. 本文提出一个实值掩码估计来替代二值掩码, 以确定相应时频单元中的能量有多少被保留. 具体地, 如果我们忽略相位损失, 时频单元的实值掩码正比于混合前信号的幅度比. 由于特征向量由对数能量听觉谱构成, 因此我们可以采用以下公式近似计算时频单元的实值掩码

$$FR_i^c = \frac{\mathrm{e}^{\frac{v_l^c}{2}}}{\mathrm{e}^{\frac{v_l^c}{2}} + \mathrm{e}^{\frac{v_n^c}{2}}} \tag{13}$$

这里, FR_i^c 为频率通道 c, 第 i 帧构成的时频单元的实值掩码, v_l^c 和 v_n^c 分别是两个语者 VQ 模型中的选定码字的均值参数.

1.2.3　再合成

计算得到实值掩码信号后, 可以使用 Weintraub 提出的再合成方法[11] 合成出分离后的语音波形. 合成后的语音更加倾向于来自同一个声源, 从而为语音识别系统提供了更理想的前端输入.

2　实验设计与评估

2.1　实验数据集

本文采用 SSC 数据集评估语音分离作为 ASR 前端的效果. SSC 数据集是由英国谢菲尔德大学 Cooke 等提供的, 在 2006 年国际口语语音处理学术会议 (ICSLP 2006) 上作为语音分离专题的标准数据集[9], 同时提供的还有对语音分离结果进行评估的 ASR 测试平台.

SSC 数据集中的训练集由 34 个人 (16 个女性和 18 个男性), 每人 500 句纯净语音构成. 数据集中语音文件均为单声道 "wav" 格式的文件, 采样率为 25 kHz. 训练集中的语句均由六个词构成, 分关键词和填充词, 其中关键词有字母, 数字和颜色.

本文使用 SSC 数据集中的两语者混合语音测试集作为测试语料, 这部分测试集提供了 6 种不同的目标语音和干扰语音信噪比 (Signal to noise ratio, SNR) (包括 $6, 3, 0, -3, -6$ 和 -9 dB) 条件下的混合语音数据, 每种条件下有 600 个测试语句. 这里

目标语音和干扰语音均来自训练集中的 34 个说话人, 混合前的语句格式与训练语料相同, 当然, 具体内容与训练语句不同.

需要强调的是, 在全部两语者混合语音数据中, 约有 1/3 的数据由同一语者所说的句子组成 (ST 子集), 另有约 1/3 的数据由同一性别的不同语者所说的句子组成 (SG 子集), 剩余约 1/3 的数据由不同性别的语者所说的句子组成 (DG 子集).

2.2 语音分离系统具体实现

2.2.1 特征提取的实现

在分解和特征提取模块, 首先采用一组由 128 个 Gammatone 滤波器组成的滤波器组对输入信号进行滤波, 滤波器的中心频率从 80 Hz 到 10 kHz. 之后, 相应的滤波结果按照 20 ms 帧长、10 ms 帧移进行分帧处理以得到信号的二维时–频图. 然后再对各频带滤波器输出信号的能量取对数, 每一帧内 128 个频带的对数能量构成了这一帧的特征矢量.

2.2.2 模型训练的实现

在模型训练模块, 利用训练集中说话人的独立、纯净数据, 学习特定说话人的模型. 这里, 我们为 34 个语者, 分别训练了用于多语者识别模块的 GMM 模型和用于掩码估计模块的 VQ 模型. 每一个特定说话人对应一个 VQ; 每一个 VQ 由一个码本组成, 包含 $K = 256$ 个码字. 码本的获取是利用 K-均值聚类的方法对特定说话人的所有特征矢量进行局部最优聚类得到的. 这些预先训练好的 VQ 模型在分离过程中被结合到系统中. 同时, 每一个特定说话人对应一个 $K = 256$ 个混合数的 GMM 模型, 模型同样经过 K-均值聚类的方法进行训练, 训练好的 GMM 模型用于多语者识别.

2.2.3 测试流程

每一个待测试的混合语音, 首先经多语者识别, 得到语句中存在的语者信息, 即判断语句是由哪两个语者的语音混合而成; 然后利用得到的两个语者的 VQ 模型和 MAXVQ 方法推断时频掩码并重新合成语音完成分离工作; 最后将重新合成后的语音送入 ASR 识别器进行语音识别.

2.3 ASR 系统设计及评价准则

ASR 实验中, 我们使用了一个随 SSC 数据集提供的标准的 ASR 测试平台对系统性能进行评估. 这是一个基于 HTK 3.1 版本构建的的语音识别器. 该 ASR 识别器使用 39 维的 MFCC 特征 (包括 12 维 Mel 倒谱系数特征, 1 维对数能量特征以及相应的一阶、二阶差分特征, 即 MFCC_E_D_A 特征); 识别器采用整词建模的方式, 每个词的状态数由组成该词的音素数决定. 其中每个音素对应两个状态, 每个状态的输出建模为 32 高斯混合 (对角协方差矩阵). ASR 识别器是利用 SSC 数据集全部训练语料完成的模型参数训练.

语音识别结果按照如下的方式自动进行评分: 测试语音根据其对字母和数字这两个关键词的识别正误分别给以 0 分、1 分或 2 分, 再计算平均得分, 即得到最终的识别结果. 考虑到多语者识别模块没有提供识别得到的两个语者中哪一个是目标语者, 因此识别结果通过选择分离后语音中结果较好的一个作为目标语者. 同时, 对于使用真实语者信息的系统, 识别结果中只对 ST 子集做优选.

2.4 多语者识别结果

在多语者识别模块, 使用训练阶段得到的混合数为 $K = 256$ 的 GMM 模型, 利用两阶段识别方法, 得到测试语句中包含的两个语者的身份信息.

表 1 是两阶段处理后的多语者识别结果, 表中数据为正确率: 即两语者均识别正确的语句数占总的测试语句的百分比.

表 1 多语者识别结果

Table 1 Multi-speaker SID results

SNR (dB)	ST (%)	SG (%)	DG (%)	平均正确率 (%)
6	71.5	98.3	98.0	88.3
3	70.1	100	99.0	88.7
0	77.4	99.4	99.5	91.3
−3	75.6	99.4	99.0	90.5
−6	80.1	98.9	97.5	91.5
−9	72.9	94.4	97.0	87.3
AVE	74.6	98.4	98.3	89.6

从表 1 中可以看出, 平均识别正确率为 89.6 %. 但对于不同语者的混合语音 (SG 和 DG 子集), 两阶段识别算法均得到平均超过 98 % 的识别正确率. 需要强调的是, 虽然对于来自同一个语者的混合语音 (ST 子集), 系统识别结果的说话人组合的正确率只有 74.6 %, 但是算法可以保证最后识别结果的两语者组合之中包含正确语者的概率为 100 %, 即 ST 子集中的 74.6 % 的语音能够被正确识别为由同一个语者的语句构成的混合语音, 而 ST 子集中另外 25.4 % 的混合语者被识别成是正确语者与另一语者的混合.

2.5 语音分离结果在 ASR 实验上的评估

本节将评估语音分离作为鲁棒语音识别前端的效果, 同时证明本文提出的实值掩码模块和多语者识别模块的有效性.

表 2 和表 3 (见下页) 列出了一组系统在不同信噪比 (Signal to noise ratio, SNR) 和不同混合类型子集条件下的语音识别结果. 其中, 系统 0 为基线系

统, 该系统将测试语音直接输入到 ASR 系统中进行识别. 系统 I 为本文所提出的系统, 该系统采用多语者识别模块选取用于掩码估计的语者模型, 再利用实值掩码合成分离语音后, 送入 ASR 系统中进行识别. 系统 II 为采用二值掩码替代系统 I 中的实值掩码所构成的系统 (这里, 采用固定比例 $\theta = 25\,\%$ 来保留掩码为 0 的时频单元内的能量)[17]. 系统 III 为采用真实语者信息来替代系统 I 中的多语者识别模块所构成的系统 (真实语者信息可从数据集中相应混合语音文件名中得到). 与系统 I 相比, 系统 II 可以有效地评估二值掩码与实值掩码对识别性能的影响, 系统 III 可以验证多语者识别模块对识别性能的影响.

表 2　不同信噪比下的语音识别结果比较

Table 2　ASR results in different SNRs

SNR (dB)	系统 0 (%)	系统 I (%)	系统 II (%)	系统 III (%)
6	63.58	64.75	54.33	64.25
3	45.75	62.33	49.25	61.83
0	31.92	54.33	42.00	54.17
−3	19.42	42.67	29.58	42.67
−6	11.75	29.83	22.08	30.33
−9	6.75	19.33	14.58	18.67
AVE	29.86	45.54	35.30	45.32

系统 0: 基线系统, 系统 I: 多语者识别 – 实值掩码, 系统 II: 多语者识别 – 二值掩码 ($\theta = 25\,\%$), 系统 III: 真实语者 – 实值掩码

表 3　不同混合类型子集下的语音识别结果比较

Table 3　ASR results in different subsets

混合类型	系统 0 (%)	系统 I (%)	系统 II (%)	系统 III (%)
ST	28.62	25.79	22.74	24.92
SG	30.68	50.28	38.73	50.42
DG	30.50	63.13	46.13	63.29
AVE	29.86	45.54	35.30	45.32

系统 0: 基线系统, 系统 I: 多语者识别 – 实值掩码, 系统 II: 多语者识别 – 二值掩码 ($\theta = 25\,\%$), 系统 III: 真实语者 – 实值掩码

从表 2 和表 3 可以看出, 本文所提出的系统 (系统 I) 识别正确率比作为基线系统的系统 0 提高了 15.68 %. 这一结果表明本文提出的系统可以有效地改善两个语者同时发音的混和语音情况下的语音识别性能. 进一步比较表 2 列出的不同信噪比条件下的结果可以看出, 系统 I 在不同信噪比条件下能够稳定地提高语音识别正确率, 具有一定的鲁棒性. 此外, 从表 3 中列出的不同混合类型子集情况下的结果可以发现, 系统 I 在 SG 子集和 DG 子集上性能分别提高了 19.60 % 和 32.63 %. 上述结果也进一步证实了本文所提出的方法在不同语者混合子集 (SG 和 DG 子集) 上的鲁棒性. 而在 ST 子集 (相同语者的混合语音) 上, 系统 I 甚至低于基线系统的识别率, 分析其原因, 本文提出的语音分离方法是基于语者模型信息的, 而在同一语者混合语音情况下, 语者模型失去了区分力, 无法有效地分离出目标语音, 从而影响识别性能. 而在实际应用中, 几乎不存在同一语者混合语音的情况, 我们更关心不同语者组成的混合语音.

为了分析实值掩码估计模块的作用, 我们比较了系统 I 和系统 II 的识别结果. 由表 2 和表 3 可看出, 在不同信噪比和不同混合语音子集情况下, 采用实值掩码的系统 I 的识别结果均较采用二值掩码的系统 II 有一定程度的提高, 平均识别正确率提高了 10.24 %, 这证明了实值掩码相对于二值掩码可以稳定提高语音识别性能, 更符合作为语音识别前端的需要; 同时也从侧面验证了实值掩码估计的有效性.

为了验证多语者识别模块的效果, 我们比较了系统 I 和系统 III 的识别结果. 由表 2 可发现, 对于不同信噪比条件下的平均识别正确率, 系统 I 较利用真实语者模型的系统 III 反而高出 0.22 %, 这似乎是不合理的. 再由表 3, 我们发现了原因, 系统 I 在 ST 子集上的识别正确率较系统 III 提高了 0.87 %. 而 ST 子集的语者识别结果存在 25.4 % 的识别错误 (即将同一个说话人的混合语音, 识别成目标说话人和另一干扰说话人的组合), 减少了因采用同一说话人模型所带来的二值掩码推断的混淆, 反而提高了掩码估计的准确程度. 在我们关心的由不同语者混合语音构成的 SG 和 DG 子集上, 系统 I 较系统 III 的识别正确率仅分别下降了 0.14 % 和 0.18 %, 这也从侧面验证了本文提出的多语者识别模块的有效性.

3　结论

本文提出了一个基于 CASA 和语者模型信息的语音分离系统, 作为语音识别的前端处理模块, 其中集成了多语者识别算法. 在 SSC 数据集双语者子集上的 ASR 测试结果表明, 本文提供的系统能够很好地提高 ASR 系统的鲁棒性. 最终的实值掩码系统与基线系统相比, 识别正确率提高 15.68 %, 成功地将语者信息引入到基于 CASA 的语音分离算法中, 提供了一个两语者混合语音情况下鲁棒的语音识别前端. 另外, 本文对 CASA 中的掩码估计问题提出了新的思路. 针对二值掩码分离后的语音与识别的目标不匹配的问题, 提出了有效的实值掩码估计算法, 相对于二值掩码, 识别结果正确率进一步提高 10.24 %, 能够提供更加鲁棒的语音识别前端. 因此, 本文提出的实值掩码估计算法比传统二值掩码更适于作为语音识别系统的前端模块.

本文提出的基于语音分离的鲁棒语音识别前端系统还存在一些需要改进之处. 首先, 本文对于计算听觉场景分析框架的运用, 只采用了其中的时频

分解模块、重新合成模块以及时频组织的思想, 并没有充分利用已有的研究成果, 如基于基音的同时组织、基于多基音跟踪的序列组织等算法; 其次, 考虑到语音分离的目的是为语音识别提供鲁棒的前端模块, 因此在估计出实值掩码信号以后, 并不需要重新合成出语音信号, 而可以采用缺失特征语音识别的方法, 直接进行语音识别. 上述工作将会在后续研究中陆续开展.

References

1 Gong Y. Speech recognition in noisy environments: a survey. *Speech Communication*, 1995, **16**(3): 261−291

2 Boll S F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1979, **27**(2): 113−120

3 Sanches I. Noise-compensated hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 2000, **8**(5): 533−540

4 Gauvain J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 1994, **2**(2): 291−298

5 Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 1998, **12**(2): 75−98

6 Das S, Bakis R, Nadas A, Nahamoo D, Picheny M. Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system. In: Proceedings of International Conference of Acoustics, Speech, and Signal Processing. Minneapolis, USA: IEEE, 1993. 71−74

7 Srinivasan S, Shao Y, Jin Z Z, Wang D L. A computational auditory scene analysis system for robust speech recognition. In: Proceedings of the 9th International Conference on Spoken Language Processing. Pittsburgh, Pennsylvania, USA: ISCA, 2006. 73−76

8 Roweis S T. Factorial models and refiltering for speech separation and denoising. In: Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva, Switzerland: ISCA, 2003. 1009−1012

9 Cooke M P, Lee T W. Speech separation challenge [Online], available: http://www.dcs.shef.ac.uk/martin/ SpeechSeparationChallenge.htm, 2006

10 Moore B C J. *An Introduction to the Psychology of Hearing*. San Diego: Academic Press, 1997

11 Weintraub M. A Theory and Computational Model of Monaural Auditory Sound Separation [Ph. D. dissertation], Stanford University, USA, 1985

12 Kristjansson T, Hershey J, Olsen P, Rennie S, Gopinath R. Super-human multi-talker speech recognition: the IBM 2006 speech separation challenge system. In: Proceedings of the 9th International Conference on Spoken Language Processing. Pittsburgh, Pennsylvania, USA: ISCA, 2006. 97−100

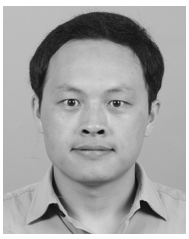
13 Guan Y, Li P, Zhang X L, Liu W J, Xu B. Applying restrained likelihood and floating TMR to multi-speaker identification for co-channel speech. In: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China: IEEE, 2007. 459−462

14 Li P, Guan Y, Liu W J, Xu B. Combining machine learning and computational auditory scene analysis to separate monaural speech of two talkers. In: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China: IEEE, 2007. 280−284

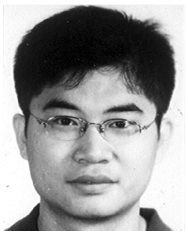
15 Cooke M P. Modeling Auditory Processing and Organization [Ph. D. dissertation], University of Sheffield, UK, 1991

16 Hu G N, Wang D L. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Network*, 2004, **15**(5): 1135−1150

17 Li Peng. Monaural Mixture Speech Separation Based on Computational Auditory Scene Analysis [Ph. D. dissertation], Institude of Automation, Chinese Academy of Sciences, China, 2007
(李鹏. 基于计算听觉场景分析的单声道混合语音分离研究 [博士学位论文], 中国科学院自动化研究所, 中国, 2007)

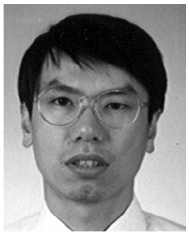


关 勇 诺基亚 (中国) 研究中心博士后. 2008 年获中国科学院自动化研究所模式识别与智能系统专业博士学位, 主要研究方向为计算听觉场景分析, 语音识别, 说话人识别和基于 HMM 的语音合成. 本文通信作者.
E-mail: ext-yong.guan@nokia.com
(**GUAN Yong** Post doctor researcher in Nokia (China) Research Center. He received his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2008. His research interest covers computational auditory scene analysis (CASA), speech recognition, speaker recognition, and HMM based speech synthesis. Corresponding author of this paper.)



李 鹏 中国科学院自动化研究所助理研究员. 2007 年获中国科学院自动化研究所模式识别与智能系统专业博士学位. 主要研究方向为语音信号处理, 计算听觉场景分析和语音识别.
E-mail: pengli@hitic.ia.ac.cn
(**LI Peng** Assistant professor at CASIA. He received his Ph. D. degree from CASIA in 2007. His research interest covers speech signal processing, CASA, and speech recognition.)

刘文举 副研究员. 主要研究方向为语音识别与合成, 说话人识别和计算听觉场景分析. E-mail: lwj@nlpr.ia.ac.cn
(**LIU Wen-Ju** Associate professor at CASIA. His research interest covers speech recognition, speech synthesis, speaker recognition, and CASA.)



徐 波 研究员. 主要研究方向为多媒体内容管理, 语音信号处理, 语音识别与合成和统计机器翻译.
E-mail: xubo@hitic.ia.ac.cn
(**XU Bo** Professor at CASIA. His research interest covers multiple media content management, speech signal processing, speech recognition and synthesis, and statistical machine translation.)