语音库自动构建若干问题的研究

IHCIL

张巍

中国海洋大学 计算机科学与技术系

主要内容

一、研究背景及研究意义

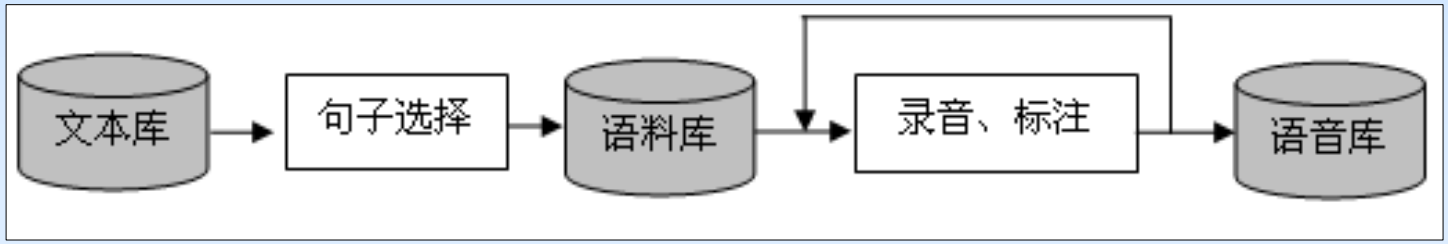
二、纯语音自动提取

三、文语自动对齐

四、无监督句子自动切分

五、无监督语种自动鉴别

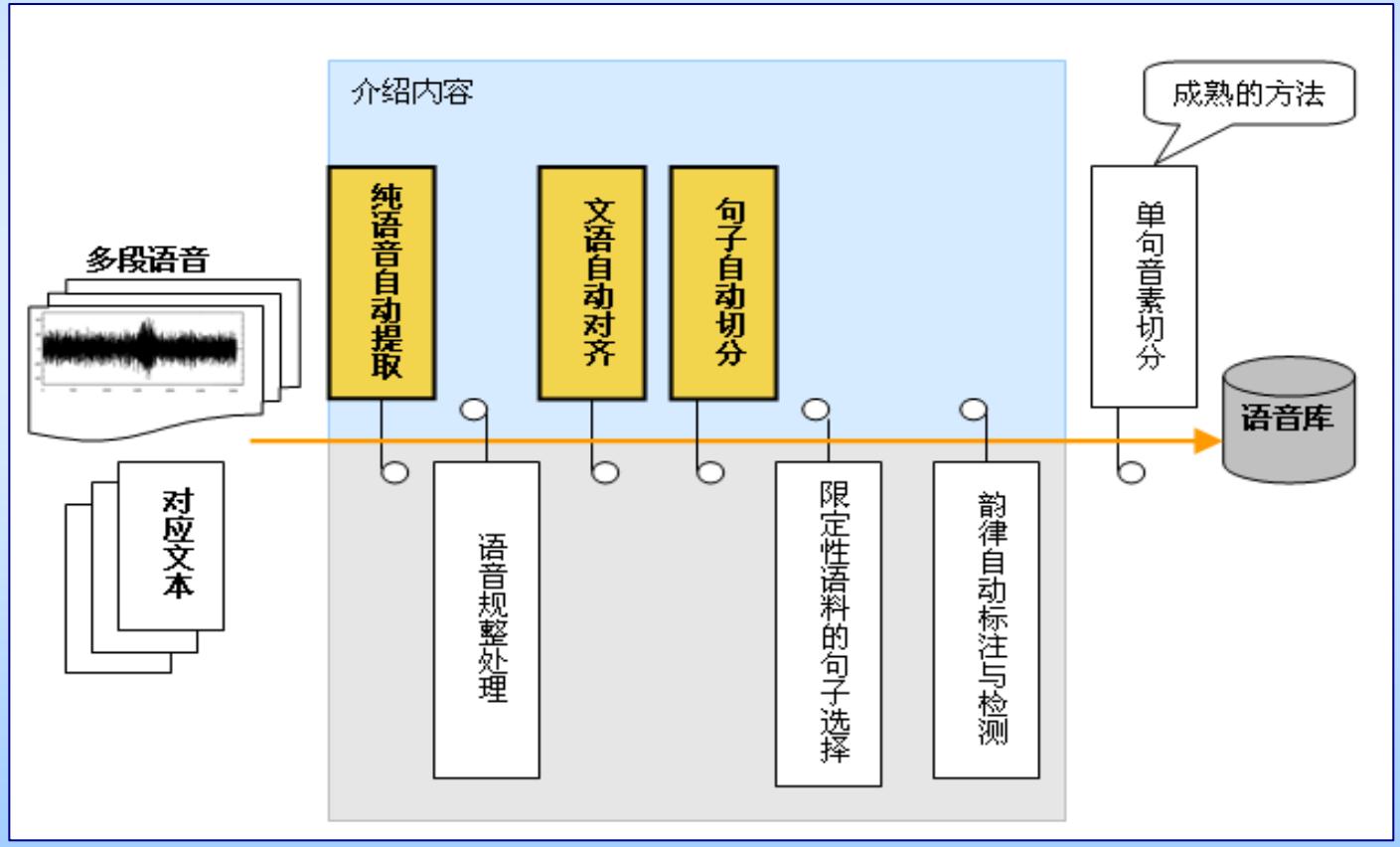
- 语音库(Corpus)是语音识别和语音合成的根基
- 传统语音Corpus的构建



- 主要缺点
  - » 耗时，耗力，构建周期长
  - » 带有人为主观性和片面性
  - » 影响语音合成和识别的能力

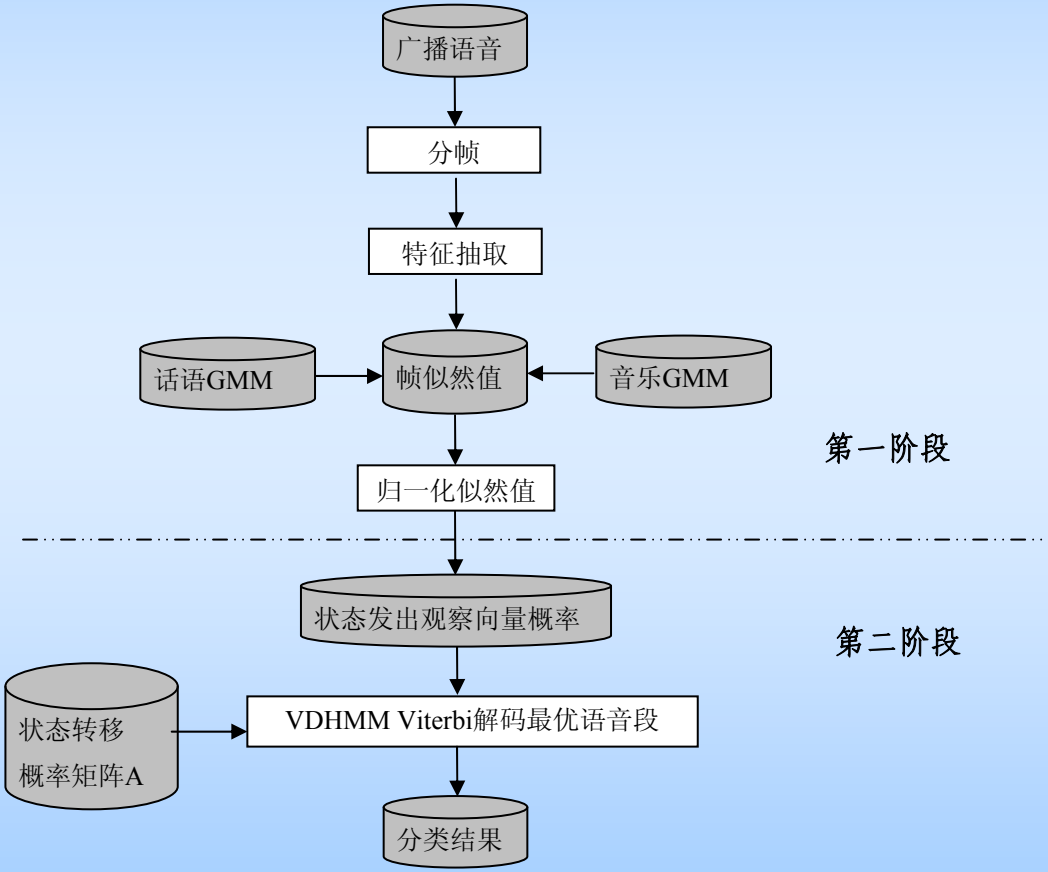
课题提出背景及研究意义

- 基于开放性资源快速构建语音Corpus
  - University of Edinburgh CSTR Simple4all
  - CMU TTS from Audio Book
  - Our Research



纯语音自动提取

GMM-VDHMM的音频分类



纯语音自动提取—GMM-VDHMM

- 分帧
  - 每25毫秒一帧，且不互相重叠,我们用{x(0), x(1),… , x(N −1)}
- 特征参数抽取(共42维)
  - Mel频率倒谱系数(MFCC)

$$c_m = \sum_{k=1}^{L} (\log \tilde{O}_k) \cos\left[m(k - \frac{1}{2}) \frac{\pi}{L} \right], \quad m = 1,..., L$$

  - 短时帧能量

$$E = \frac{1}{N} \sum_{n=0}^{N-1} x^2(n)$$

  - 过零率

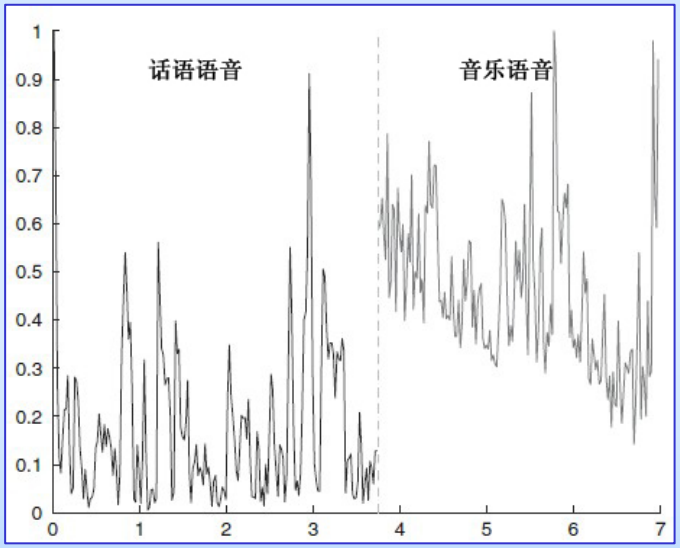
$$ZCR = \frac{1}{N} \sum_{n=1}^{N} \frac{|\operatorname{sgn}\{x(n)\} - \operatorname{sgn}\{x(n-1)\}|}{2}$$

  - 频谱熵

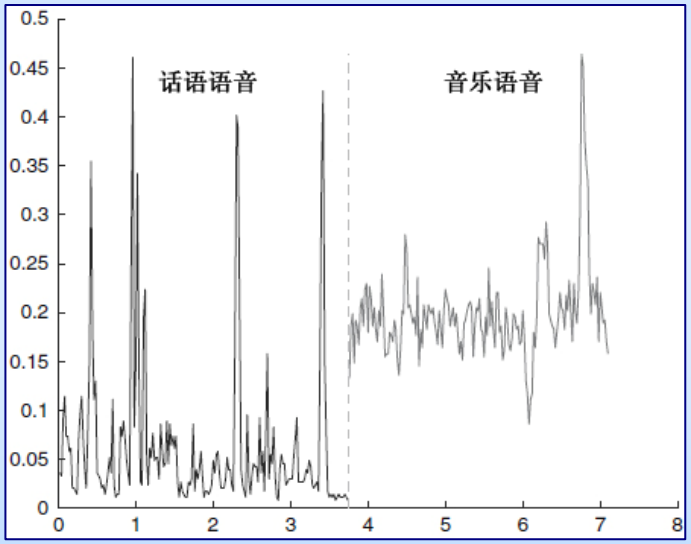
$$H = -\sum_{i=0}^{L-1} n_i \cdot \log_2(n_i), \ \mathrm{i} = 0, \ldots, \mathrm{L}-1$$

纯语音自动提取—GMM-VDHMM

- 特征参数抽取

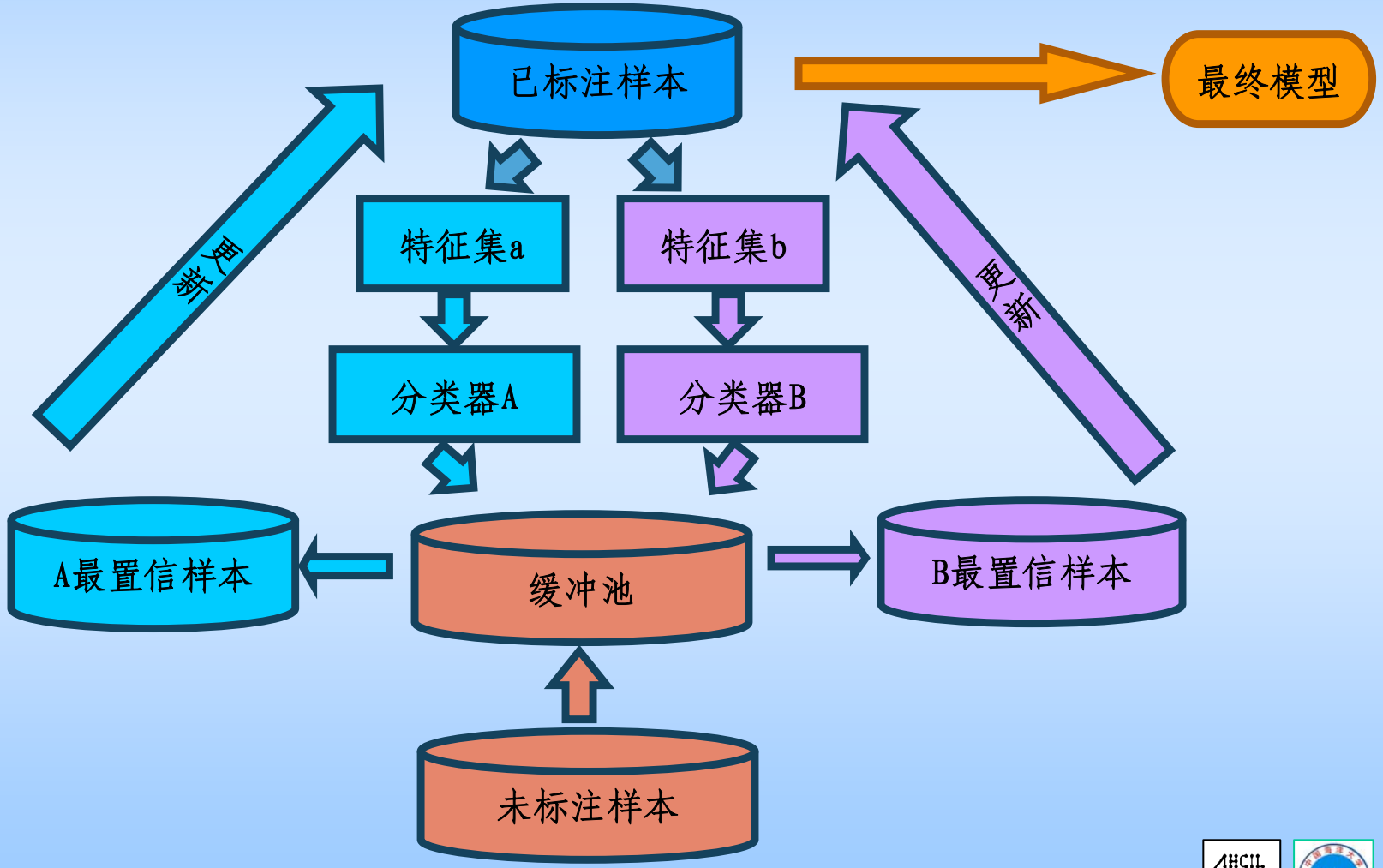


话语和音乐音频的短时能量

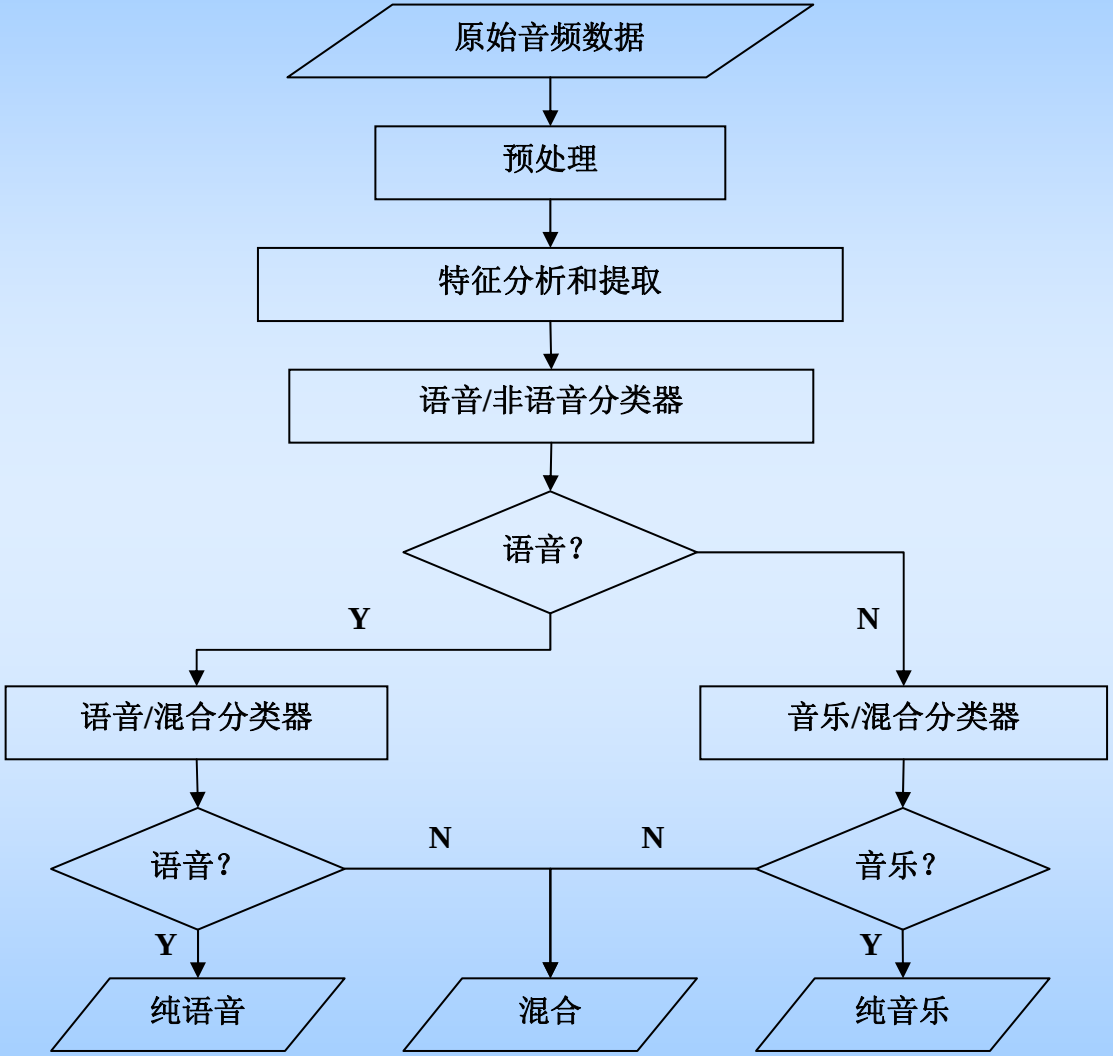


话语和音乐音频的过零率

❑ co-training协同训练



纯语音自动提取-co-training



实验分析

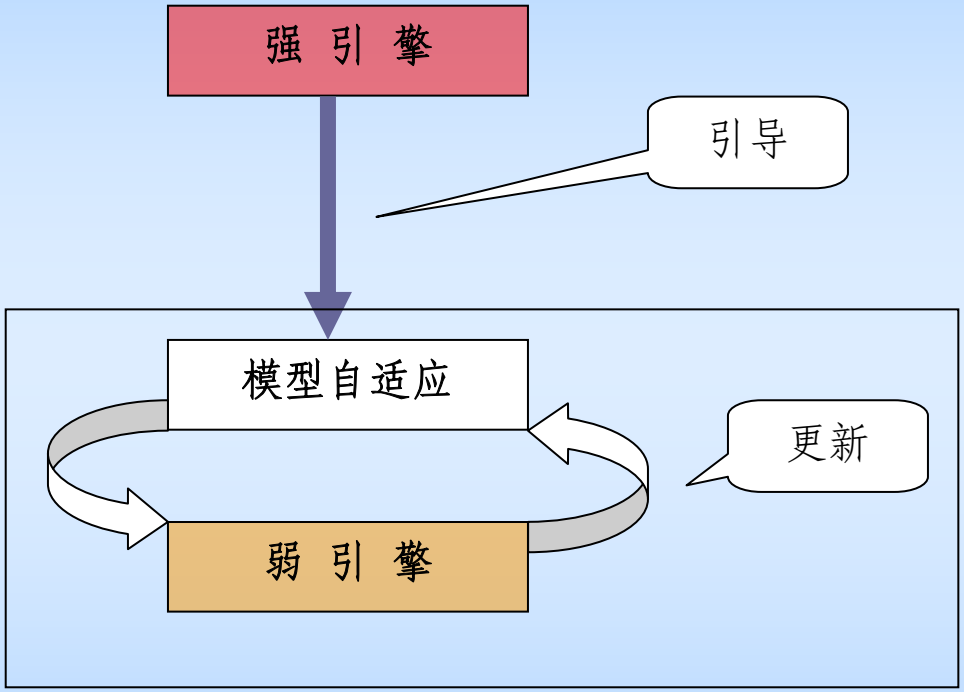
组号	话语准确率	混合准确率	音乐准确率	话语召回率	混合召回率	音乐召回率
1	0.96078	1	1	1	0.72079	0.99464
2	0.99244	0.7759	0.98475	1	0.96854	0.75341
3	0.99255	0.95037	0.99889	1	0.98101	0.92985
4	0.98278	0.98587	0.7483	1	0.80823	0.9814
5	0.95469	0.90588	1	0.99709	0.6428	0.99227
6	0.94394	0.96753	0.96264	1	0.72292	0.97386
7	0.97039	0.99424	0.97207	1	0.82792	0.99607
8	0.98027	0.9919	0.99608	1	0.84296	0.99694
9	0.96594	0.90236	1	1	0.73258	0.96492
10	0.99059	0.81357	0.96506	1	0.93401	0.82382
11	0.97514	0.9415	0.99949	1	0.7984	0.98338
12	0.97525	0.9956	0.99567	1	0.80321	0.99835
13	0.99407	0.97716	1	1	0.94082	0.99643
平均	0.975295	0.938606	0.970996	0.999776	0.824938	0.952718

文语自动对齐

•语音识别器的识别结果		•原始的文本
•全国	8710000 9210000	•全国
•人大常委会	9210000 1100000	•人大常委会
•副委员	1100000 1310000	•副委员长
•陈成之	1310000 1420000	•陈昌智、
•国务委员	1420000 1600000	•国务委员
•杨杰	1600000 1710000	•杨洁篪、
•出席	1710000 1810000	•出席
•欢迎	18100000 1920000	•欢迎
•仪式	1920000 2100000	•仪式。

Google voice recognition (GVR)

多引擎多次重识别

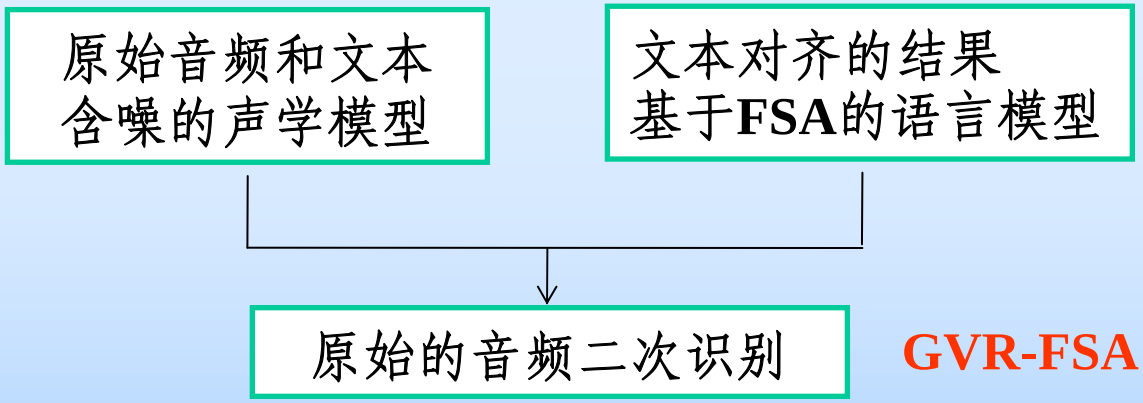


➢GVR识别的问题

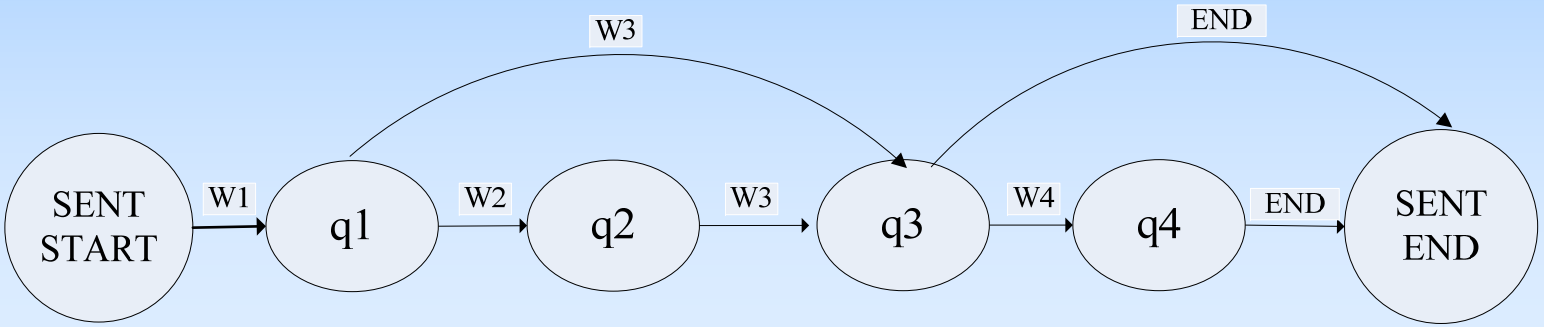
缺少了识别文本的时间信息

➢解决方法

有限状态自动机(**Finite State Automaton**，简记为**FSA**)的语言模型



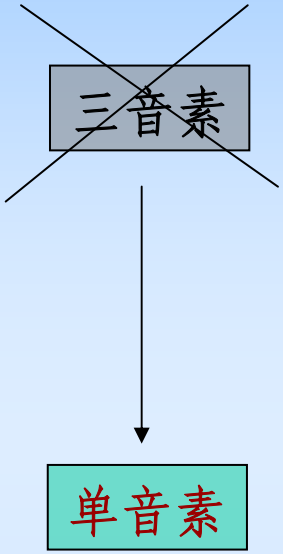
➢GVR-FSA 的语言模型



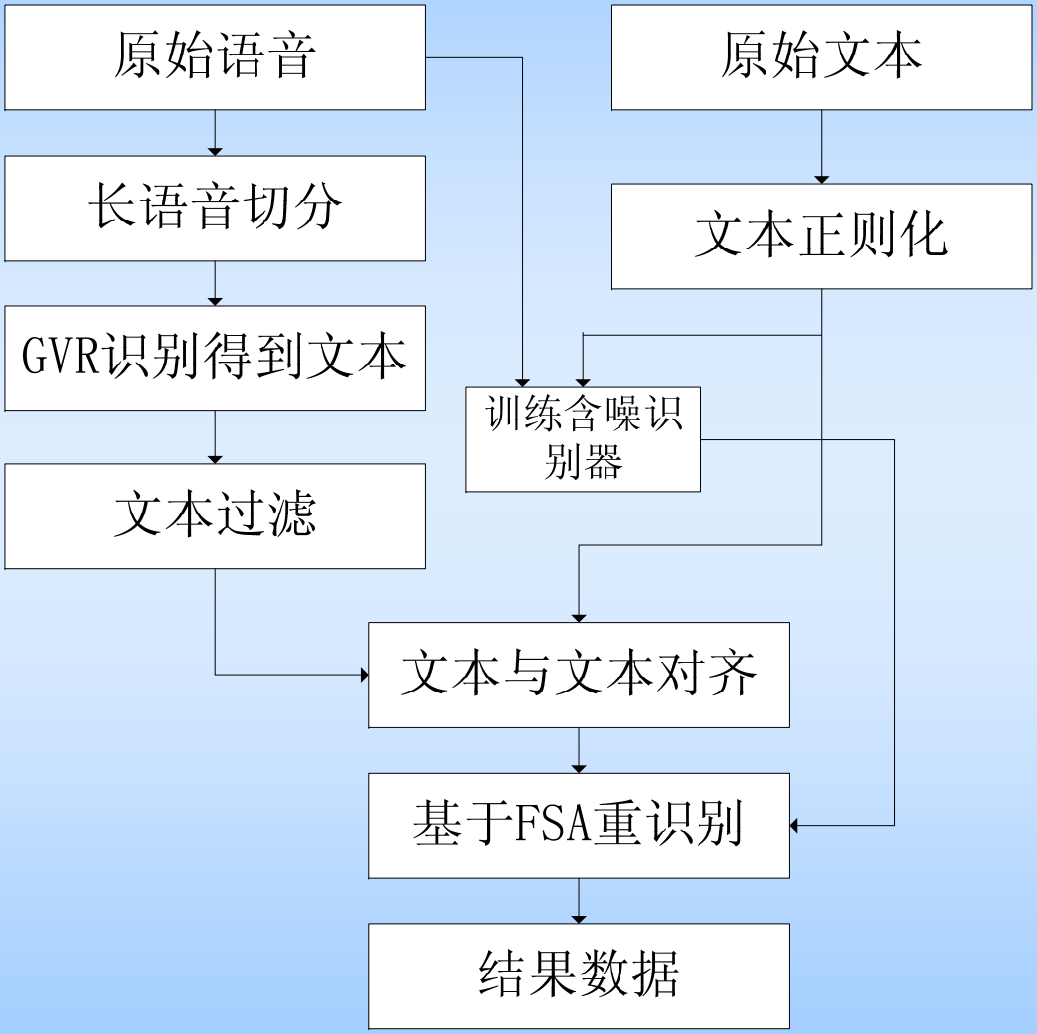
Skip Network: **W1 W2 W3 W4**

W1 W3

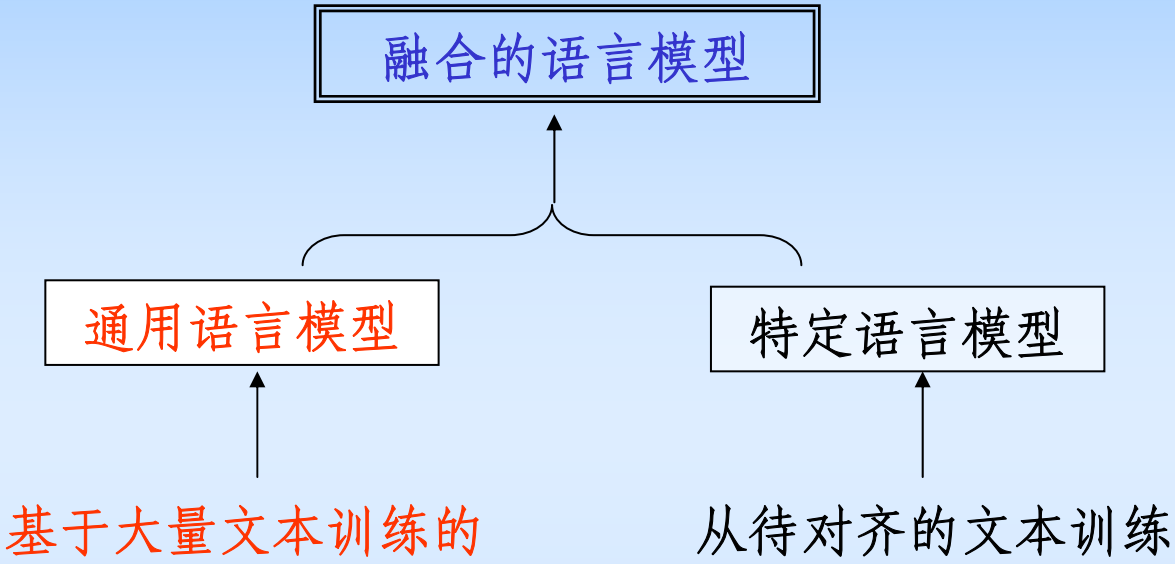
➢GVR-FSA 的声学模型



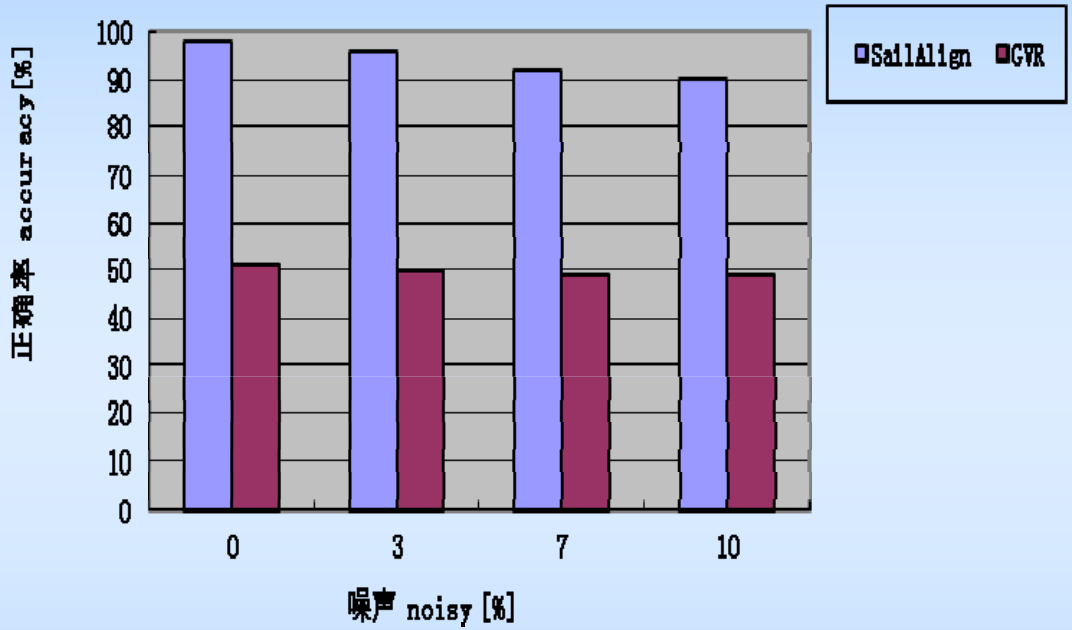
➢GVR-FSA


原始语音
原始文本
长语音切分
文本正则化
GVR识别得到文本
训练含噪识别器
文本过滤
文本与文本对齐
基于FSA重识别
结果数据

➢Improved_SailAlign

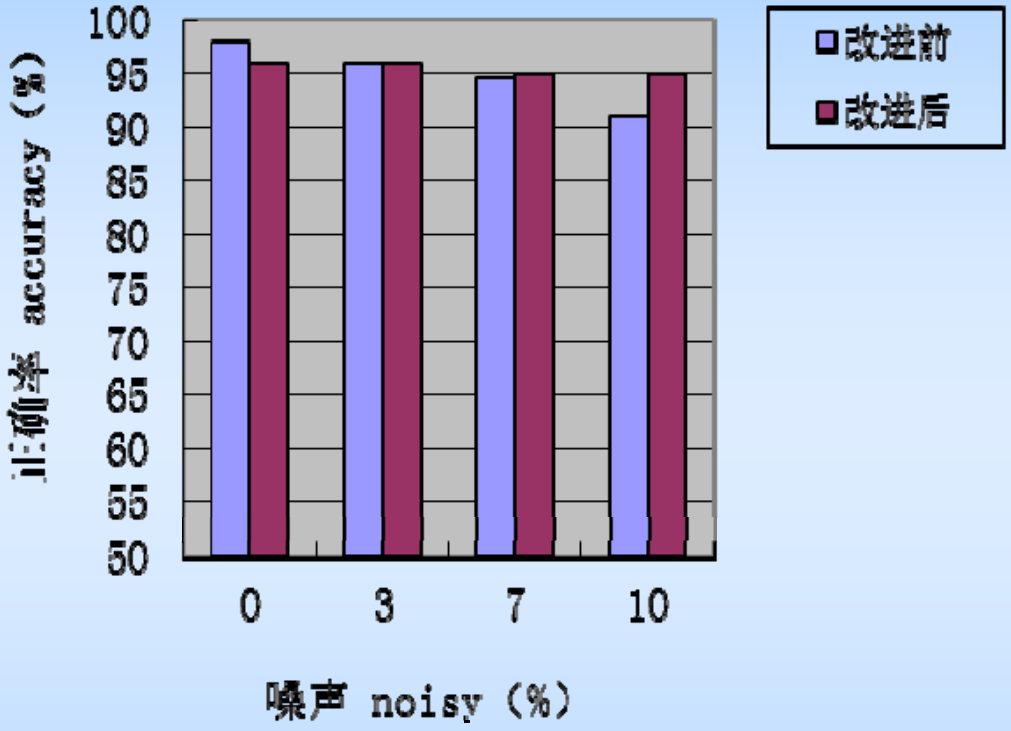


实验分析



ISA和GVR的性能比较

文语自动对齐-实验分析



改进前后的SailAlign性能比较

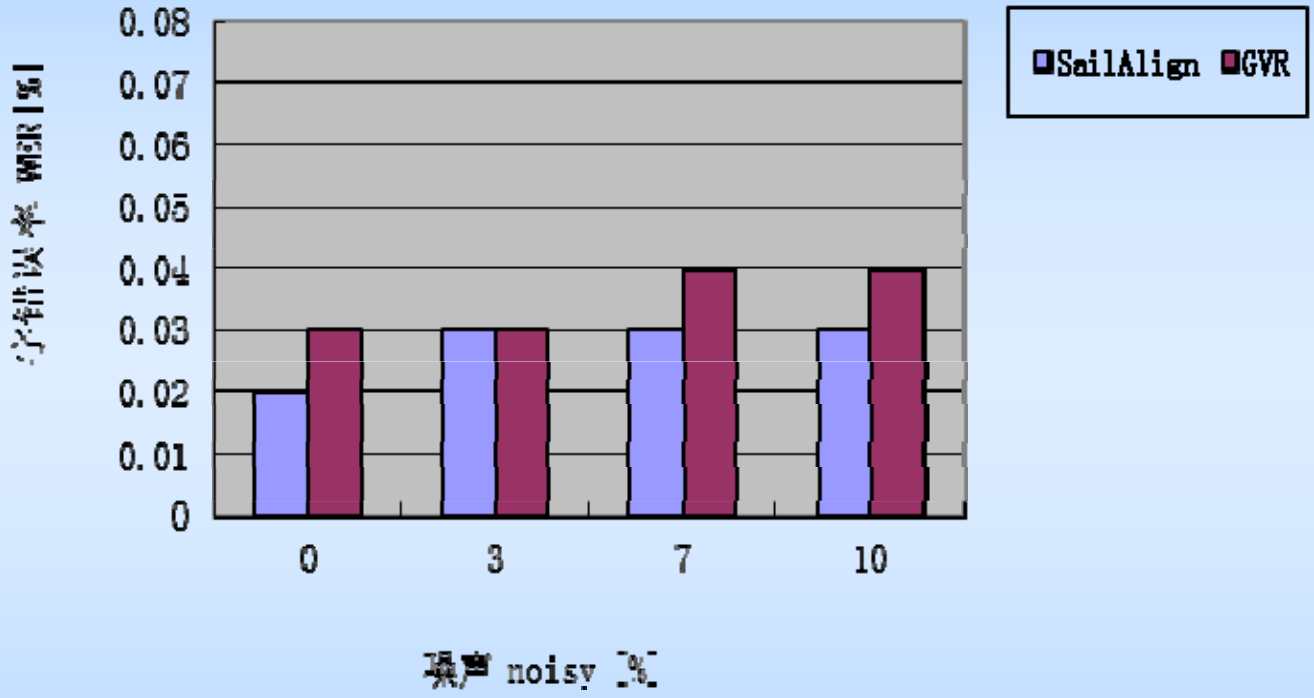
error margin

例子:
算法得到的时间信息
中共中央　12000-13000（ms）

实际该词所对应的时间信息
中共中央　12500-13500

本文的允许**error margin** 为**50ms**，超过**50ms**，认为对齐错误这里我们用**WER** 来表示不匹配的概率，即对齐错误字的总数占文本中总字数的百分比

文语自动对齐-实验分析

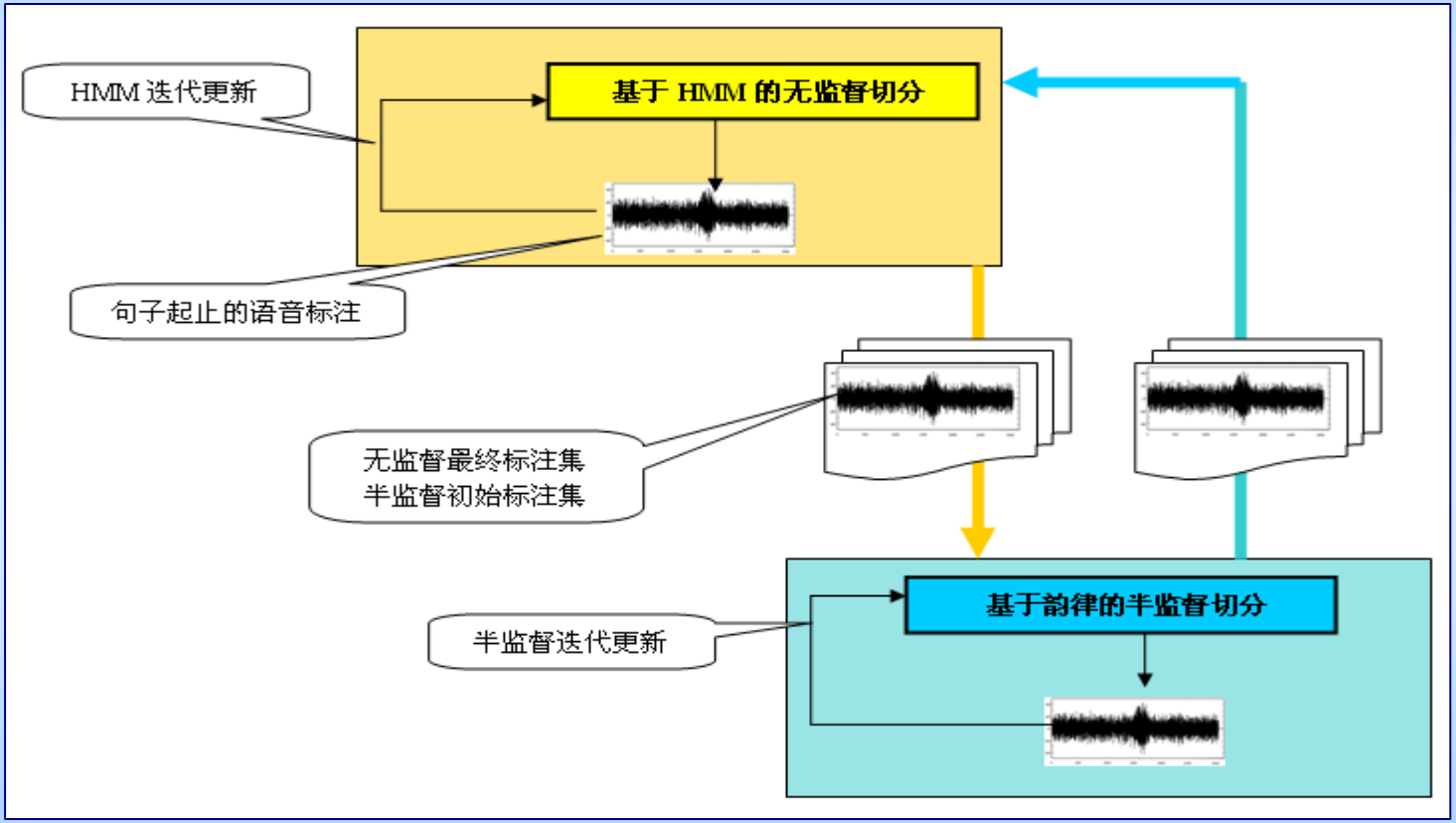


ISA和GVR的字错误率能比较

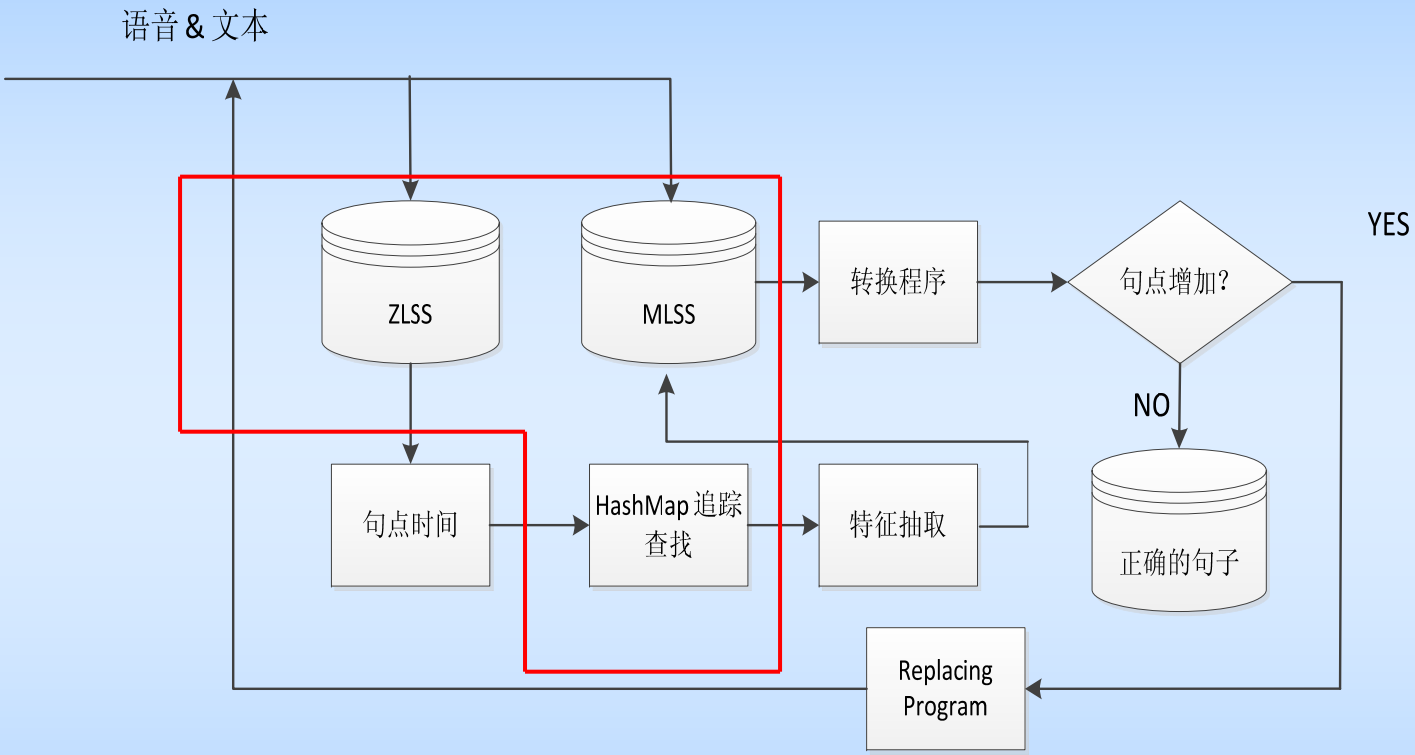
无监督句子自动切分

无监督句子自动切分

整体框架



整体框架-续


语音 & 文本
ZLSS
MLSS
转换程序
句点增加?
YES
NO
句点时间
HashMap 追踪
查找
特征抽取
正确的句子
Replacing
Program

无监督句子自动切分

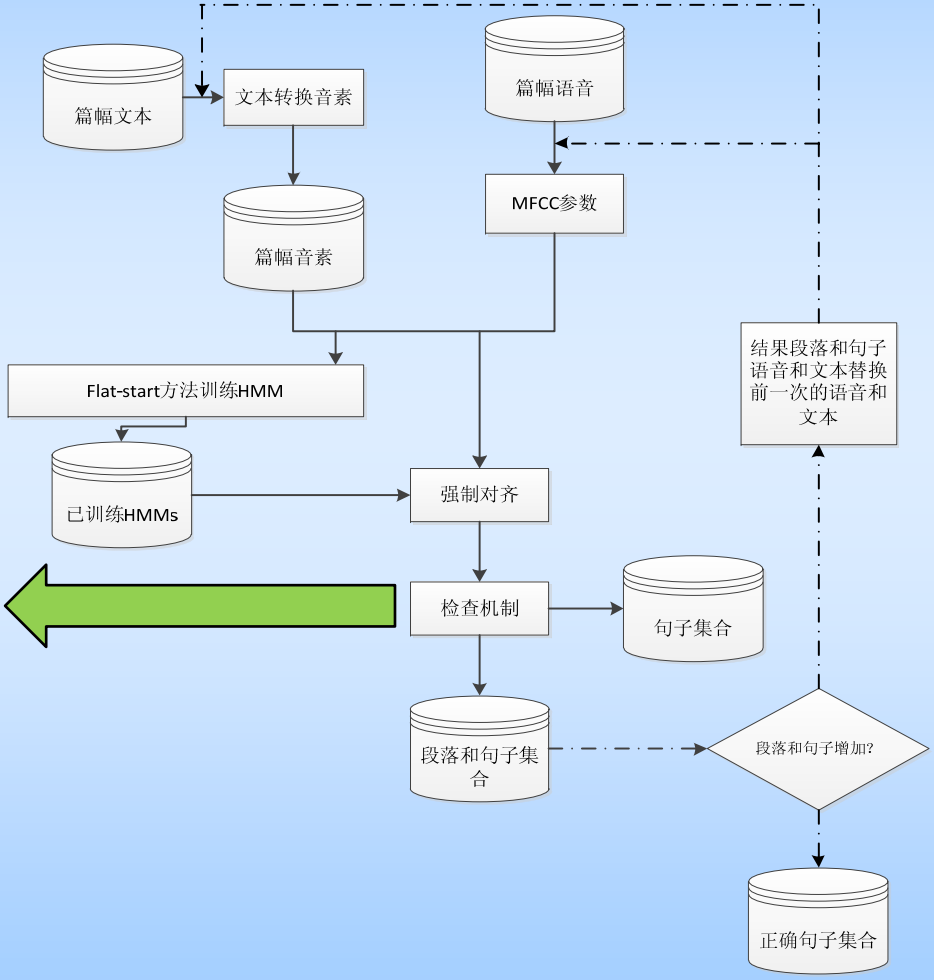
- 迭代算法
  - ZLSS和MLSS
  - 相同的时间标度，基于同一时间轴互为输入输出,相互翻滚迭代

- 迭代算法关键步骤
  - ZLSS算法(Zero-labeling Sentence Segmentation)
    - 无标注的句子切分算法
    - 得到初始的精标数据集
  - MLSS算法(Minimum Labeling Sentence Segmentation)
    - 半监督学习算法（Co_training）[10]
    - 二元分类器-基于最大熵
  - HashMap追踪查找机制
    - 利用哈希表追踪定位每一次迭代相对原始篇幅语音数据的位置

无监督句子自动切分

ZLSS算法的迭代过程

切分准确率为
98.93%[8]



MLSS算法的迭代过程

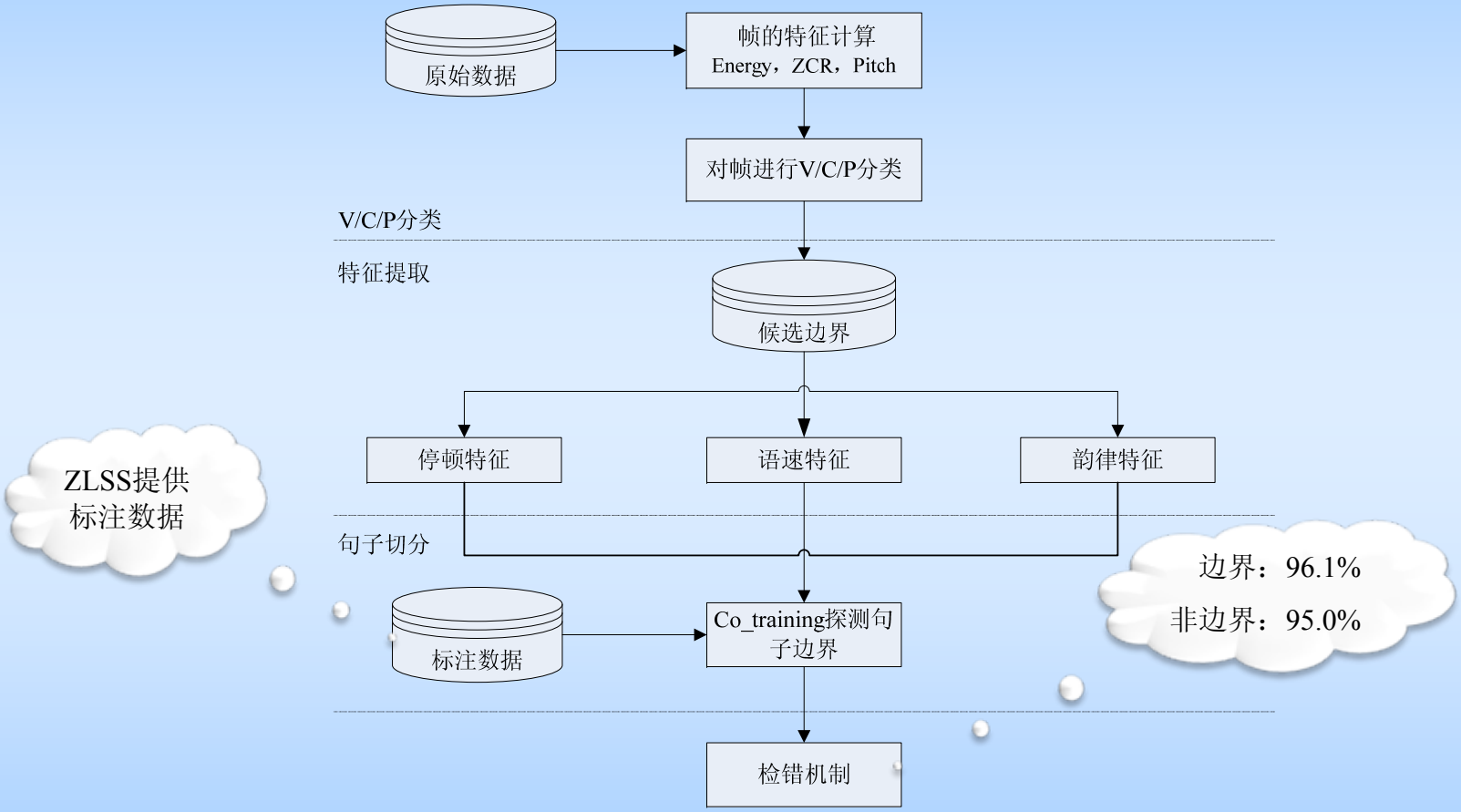


图4 MLSS算法迭代过程

- 实验分析
  - 切分系统经过完整的四次迭代过程（中文结果）
  - 标注系统用于提供少量的精标数据集
  - 分类系统在此基础上，继续训练探测出更多的句子边界

迭代次数	标注系统	分类系统	句子总数	正确比率
1	411	218	629	64.7%
2	720	108	828	89.3%
3	875	35	910	93.6%
4	920	15	935	96.2%

无监督语种鉴别 Slides

无监督语种鉴别 Slides

Thank you

Q&A