基于电话信道的信息检测系统

(Design Spec)

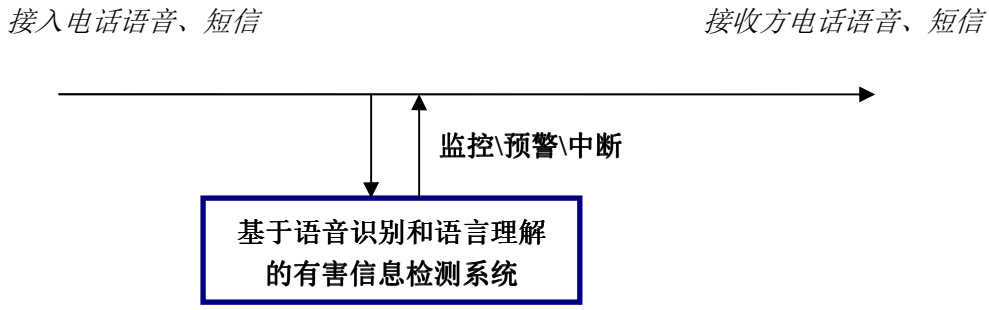
一. 研究背景

随着电子信息技术的发展，现代社会越来越依赖信息的高效流通和快速交互。然而，高速的信息流通在给人们带来便捷的同时，其负面影响也越来越显著。诸如色情、恐怖、暴力等信息的传播不仅违背公序良俗，也为社会安全带来巨大威胁和隐患。如何在保证信息自由流通的同时侦测和防止不良信息的传播，是摆在信息科学工作者面前的新课题、新挑战。

国家已经认识到不良信息传播带来的风险，并投入大量精力研究和开发智能化的检测监控系统\。然而，现有的信息安全系统绝大部分是针对文本信息的侦测。现实生活中，电话、手机等越来越成为信息交流的重要载体和工具，这些移动设备上的交换的信息数据量大，信息丰富，传播迅速；另一方面，移动设备上的交流具有高度随意性和模糊性。归因于此，到目前为止对移动设备上的语音和短信信息进行检测还缺乏一种成熟可靠的解决办法。本研究提出一种综合语音识别和语言理解技术的信息侦测系统来填补这方面的空白，为电话和短信检测提供一种系统、安全的解决方案。

二. 研究内容

本研究的目的是实现对电话信道（包括语音和短信短信）的实时检测应用，特别是对新疆维汉混杂语言环境下的应用。图 (1) 给出了一个电话信道检测系统的基本架构，其内核是语音识别和语言理解技术： 电话语音的内容经由语音识别系统识别成文本格式，由语言理解系统判断其中是否含有有害信息，进而完成监控、预警甚至中断。短信内容则直接进入语言理解系统进行判断和检测。



图(1) 电话信道有害信息检测系统框架

然而，要实现这样一个语音/短信检测系统非常困难，需要解决一系列研究课题和实现中的困难。

首先，电话语音信号具有数据量密集、噪声芜杂、随意性强等特点，这使得语音识别变得异常困难。到目前为止，现有的语音识别技术还远达不到准确识别语音内容的程度 (汉语的正确率只有 80%左右)。如何利用有限的识别精度取得可靠的检测效果，是本研究的中心内容。

第二，不同的说话人差异显著，这一方面对通用的语音识别技术带来挑战，另一方面也可以使我们通过语音来判断某些重点检测对象，并对其给以更高的检测权重。如何利用说话人信息提高检测质量，是我们要解决的另一个关键问题。

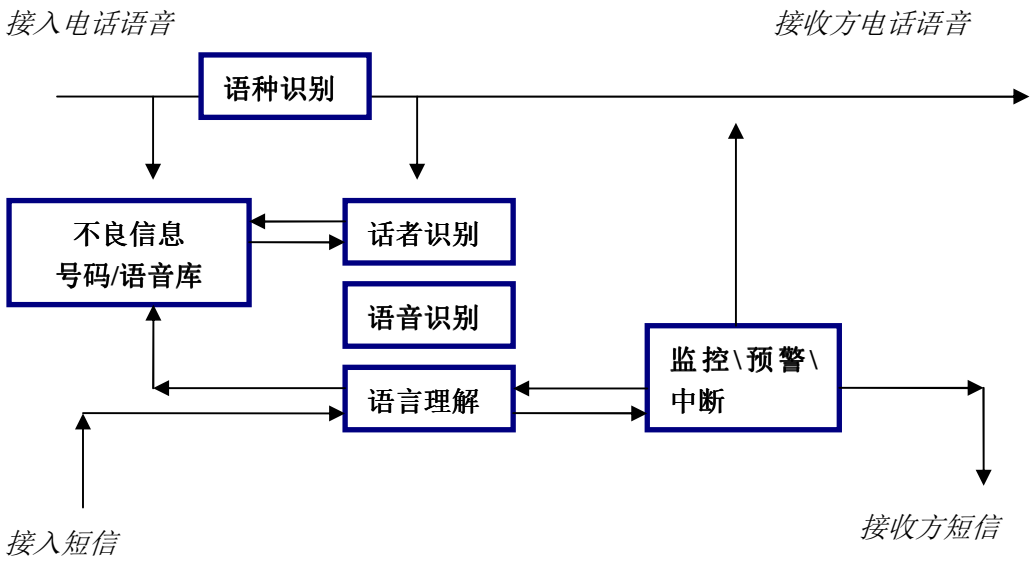
第三，多民族地区通常会有多种方言甚至多种语言的混杂，如新疆地区就有汉/维/哈等多种语言并存。 对语种和方言进行判断以提高识别和检测的精确度，是我们要研究的另一个课题。

第四，不论是语音识别的结果还是用户的短信内容，一个共同的特点是含有大量文法和语法错误，传统的语言理解技术必须进行改进以增加容错性。

第五, 有害信息的定义缺乏正则性，很难在系统设计时明确界定。如何利用各种机器学习方法使系统可以自主学习有害信息的判断准则， 是我们研究的另一个重点。

最后，语音数据量要远远大于文本数据量，因此需要多的多的计算资源，这对系统的实时性提出严峻挑战。如何提高系统的响应速度，达到实时检测，是我们必须攻克的一个重要课题。

为了解决上诉困难，在现有技术条件下构造一个快速、可靠的检出系统，我们提出一种综合语种识别(LID)、说话人识别(SID)、语音识别(ASR) 以及语言理解(NLU)技术，以概率模型为基础的，基于可信度累积的智能型检测方法。 该方法的基本思路是充分利用当前语音处理和信息检索技术所能达到的准确率，集成多模块的可信度信息，并利用机器学习方法对可信度进行累积和建模，进而构造一种可靠的、可以自主学习的判决器， 以适应动态的、高复杂的检测环境和实时性、智能化的检测要求。



图(2) 基于语音和短信的电话检测系统设计原型

图(2)给出了这一系统的设计原型。在该设计中，电话语音首先进入语种识别模块，用以判断说话人使用的语言，并调用相应的语音识别模块识别出讲话内容，并以文本方式送入语言理解模块。同时，短信系统可以直接接入语言理解模块进行检测。语言理解模块对输入内容是否存在有害信息进行判断，并给出危害评分。依照有害信息评分结果，系统会提出预警，触发检测和录音，甚至断开连接。另一方面，含有大量有害通信信息的电话号码会被自动记录，语音片断会被提取并自动生成说话人模型，形成敏感号码和语音库。这些敏感号码和语音库用于对比接入的电话和声音， 如果接入号码或声音（通过说话人识别） 存在于敏感号码和语音库中， 则识别过程会被赋予更高的权重， 以加强检测力度。最后，监控和预警结果通过人工确认的方式对系统进行反馈，该反馈被用以更新语言理解模块中的判决模型。

三. 研究方法

图（2）所示的解决方案综合了 LID, SID, ASR, NLU 等语音和语言处理技术，每一项技术都是当前研究的热点并远未完善。 在努力提高每一模块自身性能的同时，我们需要设计一套适合于电话语音实时识别和判决的研究方案。

（一） 语种识别 (Language identification , LID)

LID 技术对于距离较远的语系可以有比较好的识别效果，例如英语/汉语, 法语/英语，而对较近距离的语系则比较困难，如汉语/日语， 维语/哈萨克斯坦语。

在我们研究中，需要识别的是汉语、维语、哈语、柯语等四种语言。依据识别任务难度和训练数据量大小，我们采取层次化识别策略，即先区分汉语和其它三种语言，再进一步对三种少数民族语言进行细分。

从方法上，我们研究如下两种方法:

1) 基于 Ivector 的声学方法

这一方法首先对训练数据进行聚类，得到 M 个高斯分布模型。测试语音在 M 个高斯分布上的后验概率形成 Ivector。以 Ivector 为特征向量，构造 GMM 模型，LID 实现为基于 GMM 的最大概率。

这一方法的特点是只判断语音特征，不需要了解语言特性，对训练数据要求不高。 因为不考虑发音的连续性，判决速度快。这一方法适用于对差别较大的语言进行判断。

2) 基于音素的语言学方法

这一方法是对每种语言训练基于音素的声学模型和语言模型。测试语音在声学模型上进行解码得到音素串，所得音素串再由语言模型来计算生成概率，概率最大的即为测试语音的语种。为进一步提高准确度，基于区分性的建模方法 (discriminative modeling)，如 Logistic regression, MLP 或 SVM 等经常用来做判决模型。

这一方法是当前 LID 的主流，需要比较大的语音和文本训练语料，速度较慢，但因为集成了语音和语言信息，特别适合对发音相近而语法差异较大的语言对进行区分。

（二） 说话人识别和确认 (speaker identification and verification, SID/SVR)

SID 技术要求定位到说话人，而 SVR 技术则要求判断是否为指认的说话人。SID 技术受限于判断集的大小，而 SVR 技术则受限于说话人模型与背景模型的可区分度。我们研究中所用的说话人识别技术既不同于 SID，也不同于 SVR， 而是集中说话人确认(Speaker Set Verification SSV) 和开集说话人识别(open set SID, OSSID)。SSV 的目的是对是否为敏感说话人进行判断从而采取不同的判决准则，OSSID 的目的是进一步对说话人进行识别。

1) 基于 SID 的 SSV/OSSID 方法

这一方法将说话人集设定为闭集进行 SID，找到最匹配的说话人后再与背景模型进行比较，从而得到 SSV 的可信度。这一方法的好处是 SSV 和 OSSID 可以同时得到，缺点在于计算量大， 每个说话人需要较多的训练语料。各种说话人聚类方法可以克服这些困难。

2) 基于多候选的 SSV 方法

基于 SID 的 SSV 方法的另一个问题是， SID 只选择最大似然的说话人与背景模型进行比较评估，而不考虑其它说话人的影响。一个合理的假设是当候选集中有多人与测试语音接近时，

SSI 结果的可信度应该做相应修正。如何利用多候选进行 SSV，是有待解决的问题。

3) 基于区分性模型的方法

基于 SID 的方法，不论是单候选还是多候选，都是以生成性模型(generative model)为基础。这种方式具有很好的扩展性和自适应性，但不利用生成判决可信度 (Likelihood Ratio 从 0 到正无穷)。简单的方法是用 Logistic function 对生成评分进行正规化，更合适的办法是用各种区分性模型直接得到 SSI 的后验概率。

总之，我们需要扩展说话人识别的传统方法，使其适合当前的 SSV/OSSID 任务。

(三) 语音识别 （Automatic speech recognition, ASR）

语音识别技术应用到有害信息检测系统时可以采用如下三种模式：

1. 全文本模式

这种模式应用大规模连续语音识别技术将所有输入语音转化成文本，供后续语言理解模块进行有害性判断。这一模式的特点是识别准确率高，后续处理方便，但计算量大，识别速度慢，且不利于词表外词的检出。

2. 音素串模式

这种模式应用基于音素的连续语音识别系统将语音转化成音素串，之后由语言理解模块在串中搜索出需要关注的有害信息。这一模式的特点是不受词表限制，可以自由加入搜索词，且识别速度快。但由于缺少语言信息，识别率较低，语言理解模块工作量大。

3. 关键词模式

这一模式介于前两种模式之间，将关注的标识性词表和音素模型结合起来，生成混合有关键词和音素的串用于后续有害信息检测。这一模式可以兼顾识别精度和速度，既可以集中到关键词，又可以不受词表限制，是我们研究的主要方向。

不论哪种模式，语音识别系统都可以生成识别网格(lattice)，实现对识别错误的补偿。这需要语言理解模块做更多工作，对多元识别候选进行甄别和利用。

(四) 语言理解 (Natural language understanding, NLU)

对于短信输入，传统的 NLU 技术经过相应改进可以直接用于有害信息的判断；对于语音输入，依 ASR 系统的输出不同，NLU 系统需要完成不同的任务。总体而言，我们的研究思路是从 ASR 结果中查找到需要关注的关键词，并依此建立主题模型(topic model)，据此判断测试串中是否含有有害信息。其中关键的研究方向包括：

1) 关键词提取

对本文本模式的 ASR 输出，关键词提取相对直观，而对音素串模式，NLU 需要从音素串中提取关键词，故而必须考虑模糊匹配问题。对关键词模式，虽然关键词已经在 ASR 输出中表示出来，但对于非词表中的关键词和漏标的关键词，NLU 系统通常需要对他们在音素层进行追回。另一方面，如果 ASR 输出为识别网格，NLU 系统还必须解决在网格上查找关键词的问题。

2) 背景信息的提取

除了关键词，语音中的其它成份对有害信息的判断也起至关重要作用。如何提取以音素或全文本表达的背景信息，比如常用词词频，词间关系，文本量大小，这些都对建立主题模型和判决至关重要。

3) 主题模型构造

现有主题模型绝大部分基于词的全文本，而对于音素串模式和关键词模式，主题模型的构造还缺乏有效的手段。另一方面，基于系统的开放性和动态性，静态的主题模型很难适应系统长期运行的要求。利用 Non Parametric Bayesian 模型 可以部分解决相关问题，但如何解决模型自适应，实现有效的自主学习，依然是个未解决的问题。

4) 信任度积累

我们系统的一个基本特征是以概率为基础的模糊判决。具体思路是，系统在侦测过程中发现关键字后并不会立即采取动作，而是记录这一关键字并进行累积。当关键字和背景信息累积到足以通过主题模型进行判决的时候才触发相应的侦测动作。这意味着有害信息的判断并不是刚性的，即时的，而是一段时间的观测结果，这可以有效防止过度敏感性误判。实现这一模糊判决的基础是信任度积累算法：不论是 SSI, ASR, 还是 NLU, 所有处理结果都以信任度为基础，整个处理管道是信任度的累乘，而在时间上记录是信任度的累加。这一积累过程和现实中检索和侦测理念是相对应的，也是我们研究的核心内容，即如何对各模块的信任度进行积累，何时进行重置，如何进行剪枝，如何利用冗余候选…这些都是需要解决的问题。

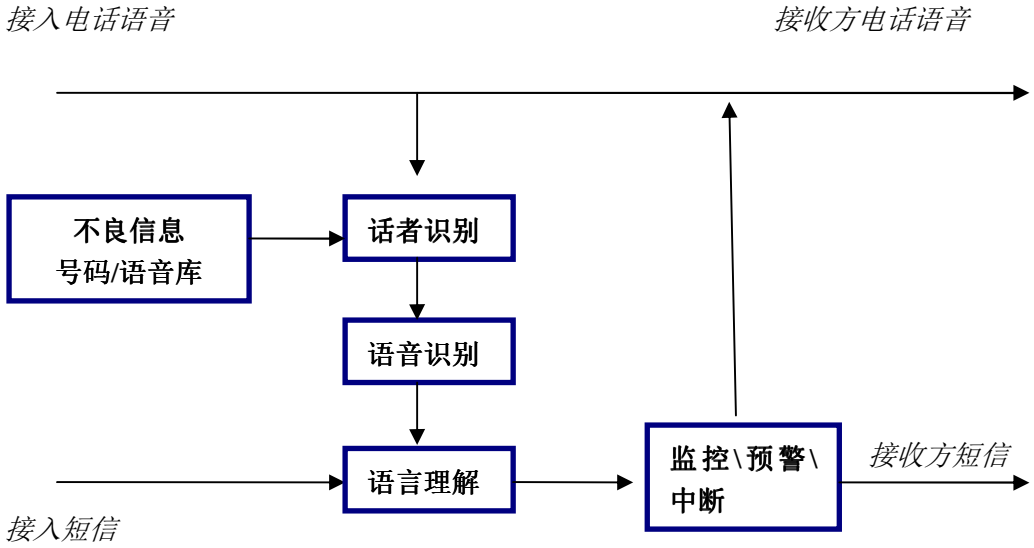
5) 基于反馈的模型自适应

NLU 模块的初始模型可以通过标注文本训练得到， 但这一模型可能无法完全适应实际需要。一个重要的研究方向是如何利用用户的反馈信息进行模型自适应，比如利用 MAP 或 model interpolation 等方法，通过学习用户反馈，自动更新判决模型以达到最佳判决效果。

四. 研究步骤

4.1 系统框架

我们计划在五个月内实现一个维语电话检测的系统原型，如图（3）所示。这是图（2）的简化版，实现在线关键词检出和有害信息判断功能， 并进行简单的敏感说话集确认，其中不包括语种识别，模型号自适应和说话人模型更新。



图(3) 计划完成的第一阶段系统原型

4.1 进度表

表（1） 研究进度 Gant Chart

	Aug.	Sept.	Oct.	Nov.	Dec.
Plan	Discussion				
SSV		SID Migration	SSV research	SSV deployment	
ASR		AM training	Keyword decoding		
NLU		Data collection and labeling	Decision model training	Simulation	
Integration			framework Dev.	deployment	deployment

表(1)给出了项目进度 Gant-Chart.，其中八月份用于技术讨论和方案设定，并准备相关数据和资源，九至十一月份进行模块设计和实现，同时开始设计原型系统的程序框架，十一和十二月份进行系统集成。

4.2 研发步骤

4.2.1. 说话人识别

(1) 移植 CSLT 文本无关的 SVR/SID 系统到维语，测试 100 人集上 500 个测试语料的 SSV FOM 和 OSSID FOM。
(2) 利用多元候选/说话人聚类/区分性模型 等技术 增加 SSV FOM.
(3) 比较维语和汉语 SVR/SID/SSV/OSSID 的性能区别

4.2.2 语音识别

(1) 结合 XJU 现有 HTK AM/HTK 关键词表 (200 词) 和 CSLT 的 KWS 系统，实现维语上的 KWS 系统。测试 10 小时语音测试集上的平均关键词 FOM 和总体 FOM。

(2) 利用 XJU 的 AM 语料，训练 Kaldi 格式的 AM， 加入 LDA+MLLT+MMI+CMLLR， 实现增强的维语 KWS 系统。测试 10 小时测试集上的平均关键词 FOM 和总体 FOM。

(3) 利用 Kaldi AM, 实现维语的 KW+phone 系统， 利用模糊查找技术进一步增强 KWS 系统。测试 10 小时测试集上的平均关键词 FOM 和总体 FOM。

4.2.3 语言理解

(1) 利用 XJU 提供的标注语料训练基于全文本词向量的初始维语有害信息分类系统，测试 10MW 文本的 F-measure.
(2) 利用 XJU 提供的标注语料，训练基于关键词词向量的初始维语有害信息分类系统，测试 10MW 文本的 F-measure.
(3) 利用 LDA 模型，构造基于主题模型的分类系统，在全文本和关键词模式下分别测试 10MW 文本的 F-measure.
(4) 扩展关键词模式为基于 KWS+phone-bigram 的词向量和主题向量模型，测试 10MW ASR 文本的 F-measure.

4.2.3 系统集成

(1) 构造语音/短信输入平台
(2) 利用模拟检测模块构造系统控制台界面
(3) 加入/测试初始 SVR/SID 系统
(4) 加入/测试初始 NLU 检测系统，测试短信检测
(5) 加入/测试初始 KWS 系统，测试语音检测
(6) 加入/测试求精后的 KWS 和 NLU 系统

4.3 资源整合

为整合研究力量和数据资源，我们列出 CSLT 和 XJU 需要提供的共享内容。 这些内容需要在八月底前完成。

表（2） CSLT 与 XJU 合作资源列表

	CSLT	XJU
说话人识别	1. 现有非特定内容说话人识别和确认系统 (WXJ,10/01)	1. 100 人 500 段维语测试语料 (Carl, 10/01)
语音识别	1. 现有汉语关键词提取系统 (WD,10/01)	1. 现有基于 HTK 的语音模型 (Carl, 10/01) 2. 维语发音词典(Carl, 10/01) 3. 不少于 50 小时的维语训练语音库 (Carl, 11/01)(4. 不少于 100M 的维语训练文本库(Carl, 11/01) 5. 10 个小时含敏感词的测试语音(Carl, 10/20)
语言理解	1. 现有汉语主题模型训练工具 (XYQ,ZQ, 10/01) 2. 汉语基于关键字的敏感信息分类工具(XYQ, ZQ, 10/15) 3. 汉语基于主题模型的敏感信息分类工具(XYQ, ZQ, 11/01)	1. 敏感词列表(Carl, 10/01) 2. 带有敏感分类的正反训练文本，敏感文本不少于 50M(Carl, 11/15) 3. 带有标注的 10M 测试文本数据 (Carl, 10/15)
系统集成	1. 演示机器 (ZF, 10/01) 2. 系统设计 (WD,WXJ, 09/20) 3. 系统集成实现(TBD, 11/01)	