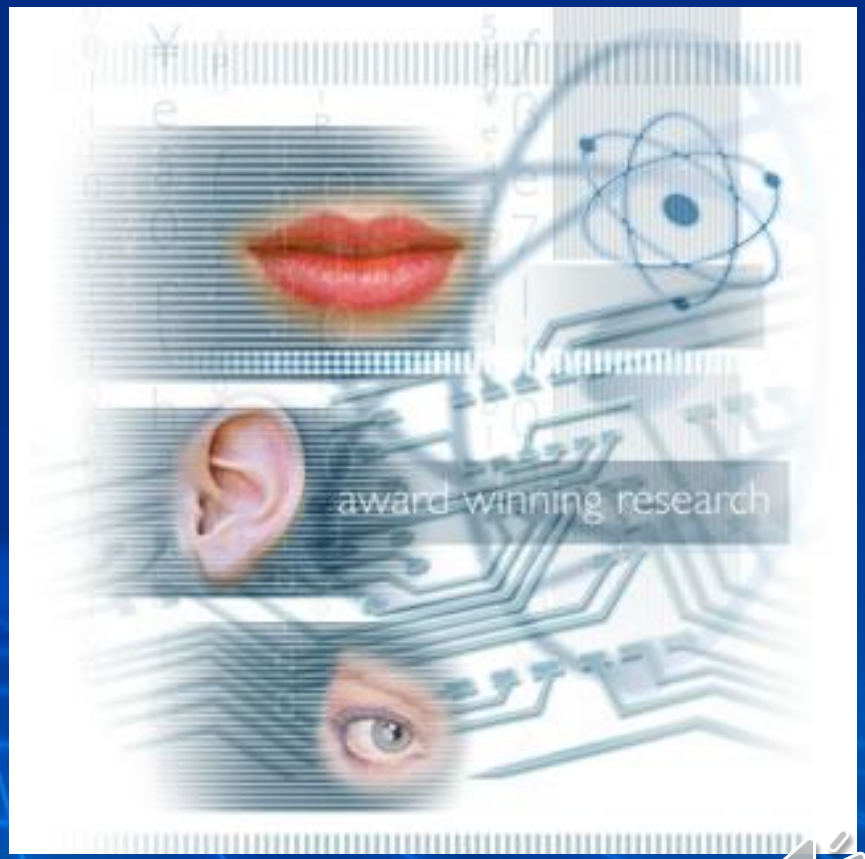
第六讲 唇读交互技术

Lip Reading

提纲

- 唇读和唇动身份识别概述
  - 唇读的目的和依据
  - 唇读的发展历史
  - 主要研究机构
- 传统算法介绍
  - 检测与定位
  - 视觉特征提取
  - 识别与音视频融合
- 最新发展态势



唇读和唇动身份识别概述

- 唇读的目的和依据
- 唇读的发展历史
- 主要研究机构

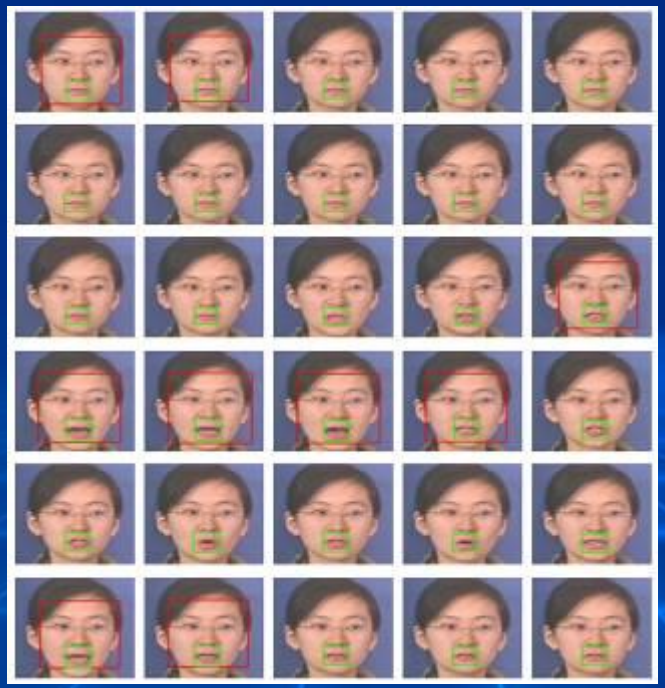
例子：齐达内头顶人事件

马特拉奇说了什么？

- 据英国一位唇语专家 杰西卡.里斯翻译，在齐达内头撞马特拉奇之前，马特拉奇曾说齐达内是“一个恐怖主义娼妓的儿子!”
- 巴西的一家环球电视台反复播放了当时的画面，通过读“唇语”和口形的方式，得出了一个答案，当时马特拉齐对齐达内说的是：“你姐姐是个娼妓！”

什么是唇读?

- 所谓唇读(lip reading/ speech reading)是指通过观察说话者的口型变化“,读出”或“部分读出”其所说的内容.



唇读的提出

- 80年代末90年代初，语音识别技术得到迅速发展,自动语音识别系统有了长足的进步,已有许多听写机之类的实用产品,较好的有
  - IBM开发的Viavoice语音系统。但是这类系统在噪声和干扰的条件下，识别率显著下降
- 为了解决语音识别系统在噪声条件下的精度问题，引入了视觉信息
  - 结合了视觉信息的语音识别系统(AVSR)
  - 目的: 结合视觉信息提高基于语音的语言识别系统(ASR)在噪声条件下的准确率

唇读的依据

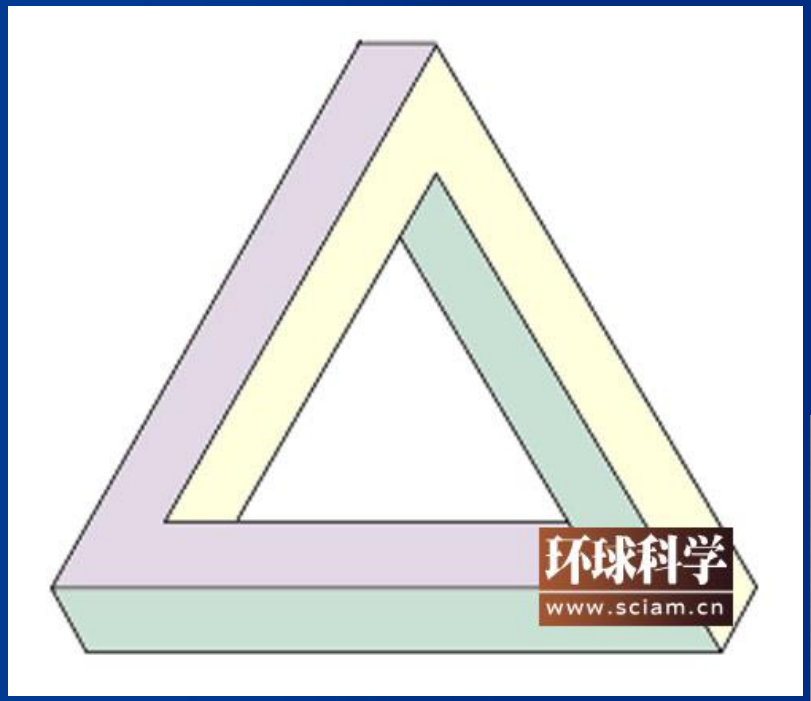
- 唇读是听力残障者获取语言信息的途径
  - 实际上，听力正常的人在语言交流的过程中，也会自然而然地使用视觉信息
- 人类语言交流过程是双模态的
  - 如果发生视觉通道获得的信息和听觉通道获得的信息不一致的情况，则会发生下列情况
    - 所获取的信息既不符合视觉信息，也不符合听觉信息——著名的McGurk实验(麦格克效应)

McGurk效应
(McGurk and MacDonald, 1976)

- *from Auditory Illusions*
  - “**What you hear is NOT what you receive.** ”
- More Example, See
  - http://v.youku.com/v_show/id_XODczMDU1MTY=.html

类似的情况

Audio +	Visual →	Perceived
ba	ga	da
pa	ga	ta
ma	ga	na


环球科学
www.sciam.cn


环球科学
www.sciam.cn

人类语音的双模态性

- McGurk效应表明：
  人类语音应该是双模态的
  - 大量研究结果表明，面部表情和声道形状以及发声之间存在着非常紧密的联系
  - 人类的Speech是由声带的振动引起的，声道与肌肉相适应的结构产生了嘴部动作和面部表情的变化

例子：身份辨认

嘴部区域，作为人脸面积最大的区域之一，在人身份辨认的过程中能起到什么作用呢?

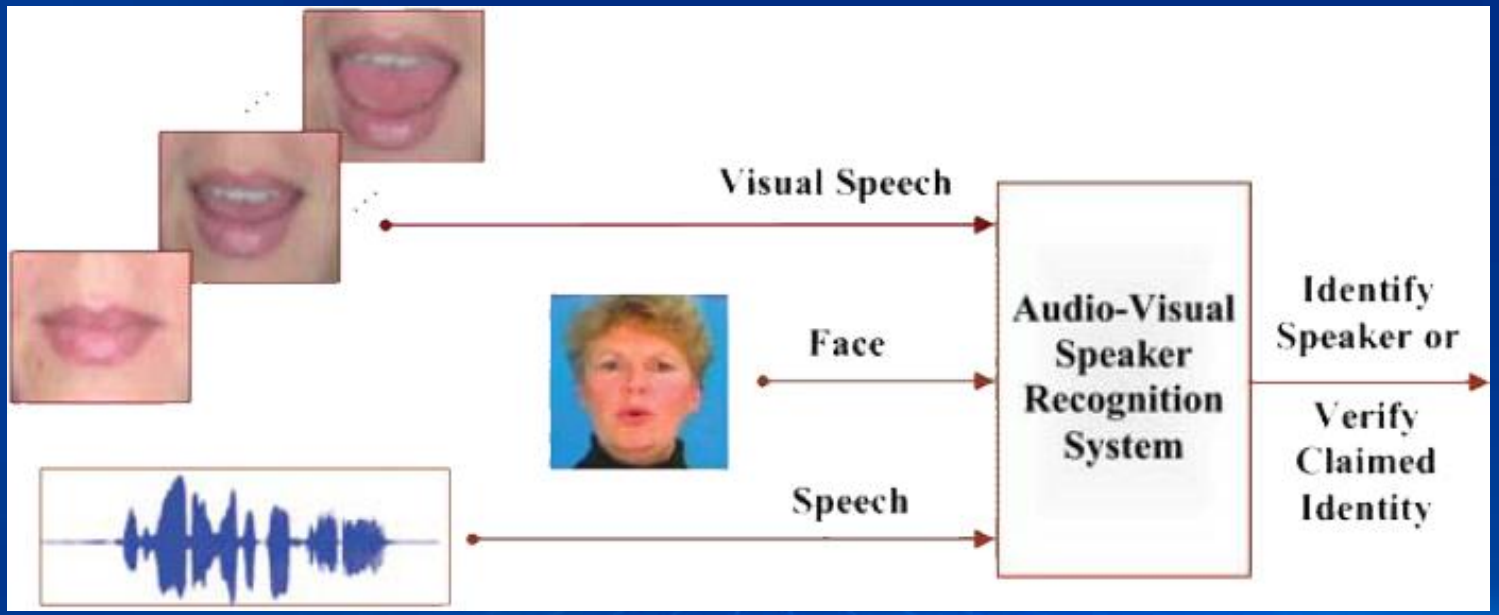
例子：身份辨认

?

视觉语言信息对身份识别的可能贡献

- 仅仅利用声音信号来进行话者识别，容易受噪声的干扰
- 声音+人脸识别的方式
  - 相对容易被冒充：一张图片+一个录音机
- 声音+动态视觉信息
  - 相对鲁棒，同时具有简单易用，容易被用户接受的优点
  - 在对系统安全性要求不是特别苛刻的场合，这种应用是有效的

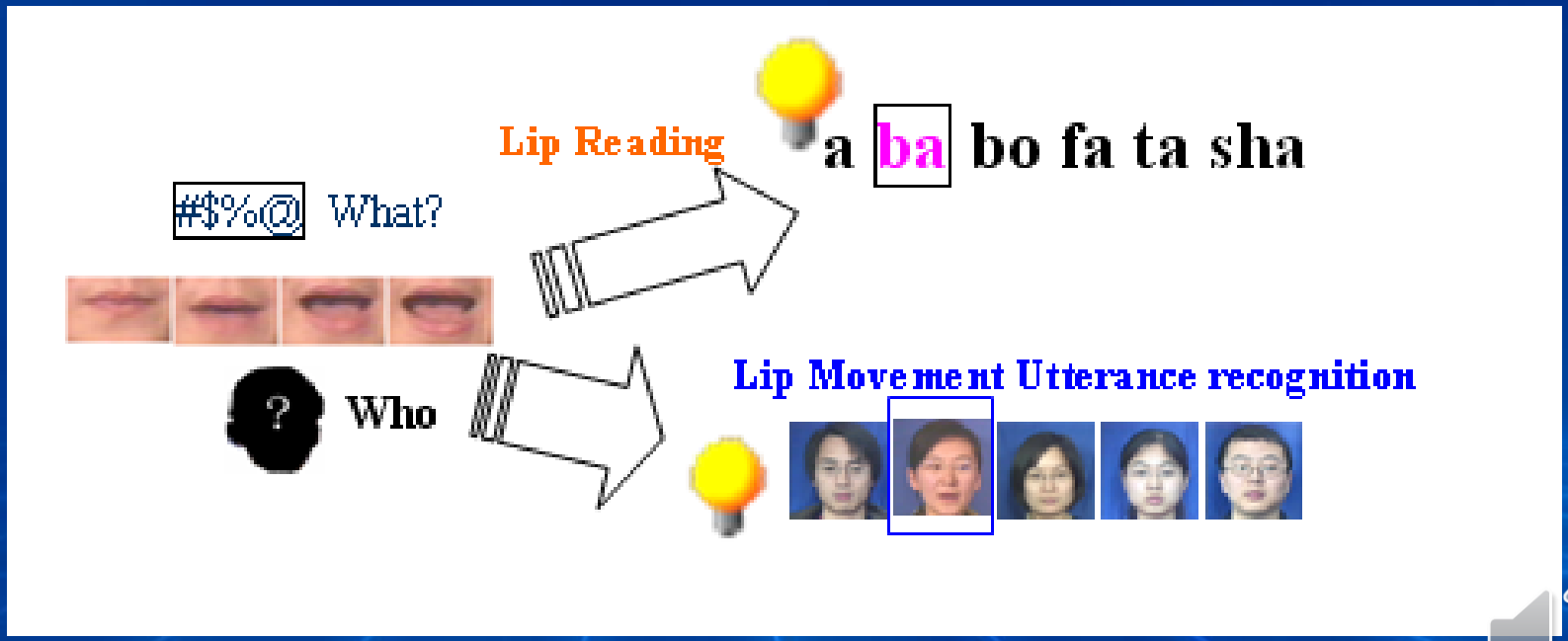
一个多通道的话者识别系统示意图



From Aleksic and Katsaggelos 06

关于唇动的两种研究方向

- 从上世纪90年中期开始，唇读逐渐成为一个研究的方向 (Audio-Visual Speech Recognition / Lip Reading)
- 而从1997年开始，视觉语言信息被引入身份识别领域 (Audio-Visual Speaker Identification / AV Biometrics)



唇读大致的发展阶段

- 启蒙阶段1984 - 1994
- 积累阶段1994 - 2004
- 寻求突破阶段2004至现在

启蒙阶段

- 1984年，伊利诺斯大学的Petajan将唇动作为听觉通道的信息补充，用以解决噪声环境或多话者环境下语音识别率下降的问题，使用的是简单二值图像和几何距离度量
- 1989年，Yuhas将人工神经网络ANN引入唇读，并且第一次将整幅唇区灰度图像作为特征向量来使用
- 1989年，MIT 媒体实验的Pentland则使用了帧间差分的特征提取方法，利用线性时间归整(LTW)进行识别.
- 1992年， Stork第一次引入时延神经网络TDNN；
- 1994年，Rao第一次使用动态时间归整(DTW)；

积累阶段-1

- 1994年, 受HMM在语音识别领域取得巨大成功的影响，Goldschen第一次在唇读领域使用HMM
- 此后10年间，逐渐形成了以离散余弦变换DCT为特征，以HMM为主流识别器，视觉信息和听觉信息相结合的音视频语音识别框架(AVSR)
- 这期间，部分研究者对视觉信息对于语言理解的贡献进行单独研究，形成单独的唇读框架

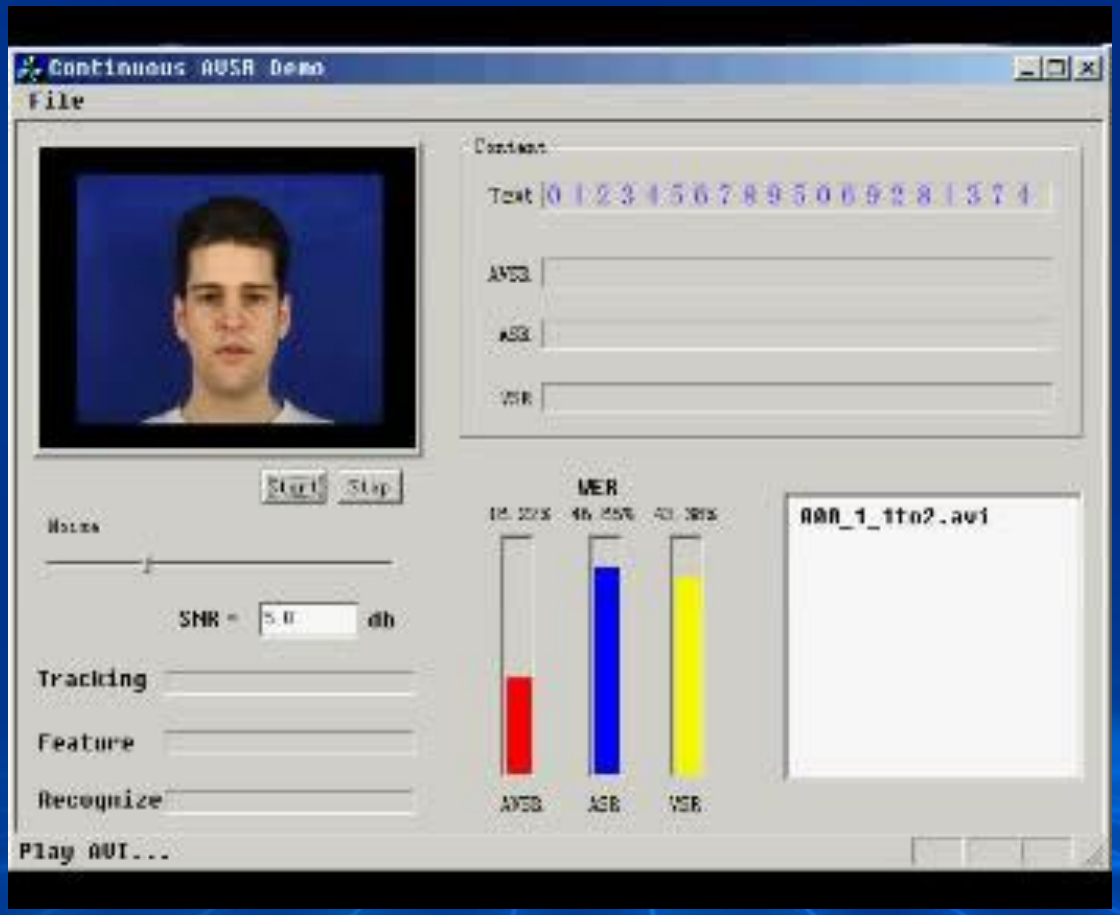
积累阶段-2

- 2000年夏天WorkShop2000: Audio-Visual Speech Recognition
  - IBM华沙实验室的Neti和Potamianos倡议，瑞士AI感知研究所的Luettin，CMU的Matthews和约翰霍普金斯大学的Vergyri等AVSR和LR专家在约翰霍普金斯大学举办该Workshop
  - 与会者在IBM的ViaVioce视音频数据库的统一平台下对当时视觉语言领域的主流技术，包括系统框架，特征提取和融合技术进行了全面的介绍比较，并对日后发展进行了展望
  - 会议形成了最终报告Audio-Visual Speech Recognition

积累阶段-3

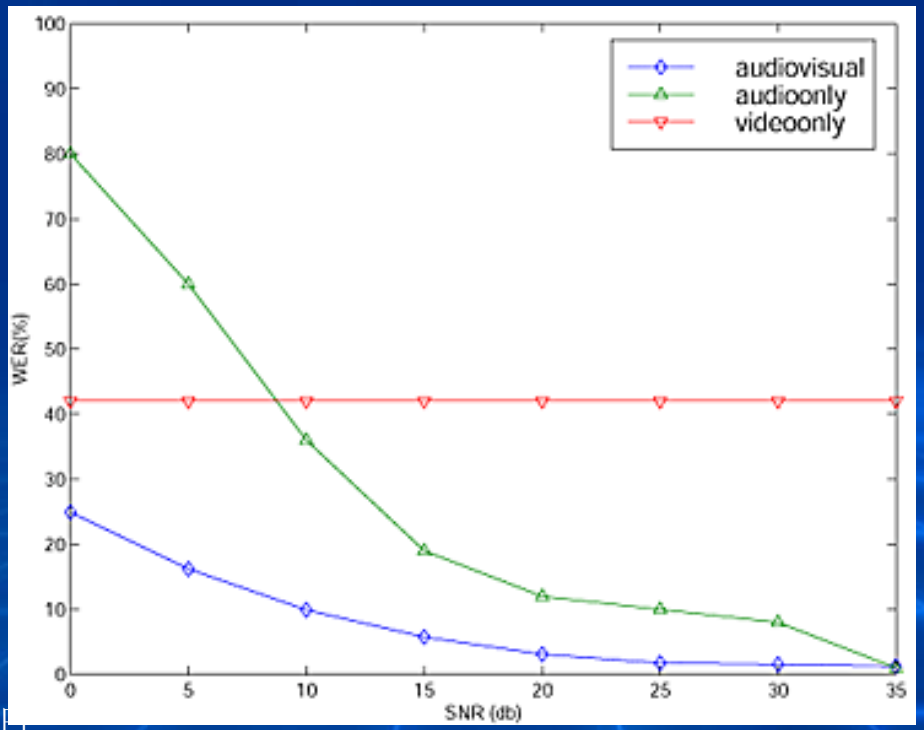
- 从01-04年间，Scanlon，Potamianos和Luettin，Matthews和Cootes等人相继在PAMI等期刊会议上发表唇读综述性文章
  - 这些文章成为日后唇读研究者的经典文献之一
- 2003年，Intel发布了开源唇读原型系统AVCSR

Intel 2003 AVCSR



Intel 2003 AVCSR

- Intel的唇读演示系统表明
  - 结合了视觉信息的语音识别系统，其Word Error Rate有明显的下降



寻求突破阶段2004 – 现在

- 以DCT+LDA+HMM的唇读系统，在中等规模的数据库上实验，准确率在70%左右
  - 远远不能达到实用的要求
- 因此进入2004年之后，科学家们开始尝试寻求突破这个框架的自动唇读识别技术，包括以下方面：
  - 代表语言的最好的视觉特征——(突破DCT)
  - 与人类语言感知功能最匹配的识别模式——(突破HMM)
  - 从语言的角度，音素Phoneme和视素viseme新的映射算法.
  - 多姿态的唇读问题.
  - 一些崭新的应用
    - 如在Car Environment和Embeded Device上的尝试

Audio-Visual Biometrics的发展

- Audio-Visual Biometrics最早是由瑞士IDIAP人工智能研究所的Luettin 等人在1996 年提出的
  - 他们发现，唇动不仅可以提供用于交流的信息，还可以被用作身份鉴别和识别
- 随后，一些学者纷纷加入到这项研究中来，如
  - 爱尔兰国立都柏林大学的Niall Fox
  - Intel 公司微处理器研究中心的Nefian
  - CMU 的Simon Lucey
  - 澳大利亚昆士兰大学
  - 美国西北大学和IBM华生实验室的研究者
  - 英国Surrey大学的J.Kittler教授

AVSR的研究机构(国内)

- 哈尔滨工业大学
  - 姚鸿勋，王晓龙
- 西北工业大学
  - 赵荣椿，蒋冬梅，蒋晓悦等
- 江苏大学
  - 张建明，王良民， 詹永照
- 上海交通大学
- 浙江大学
- 中科院声学所
  - 李彦君
- 。 。 。

一个典型的AVSR系统框架

人脸检测

唇区检测

视觉特征提取

听觉特征提取

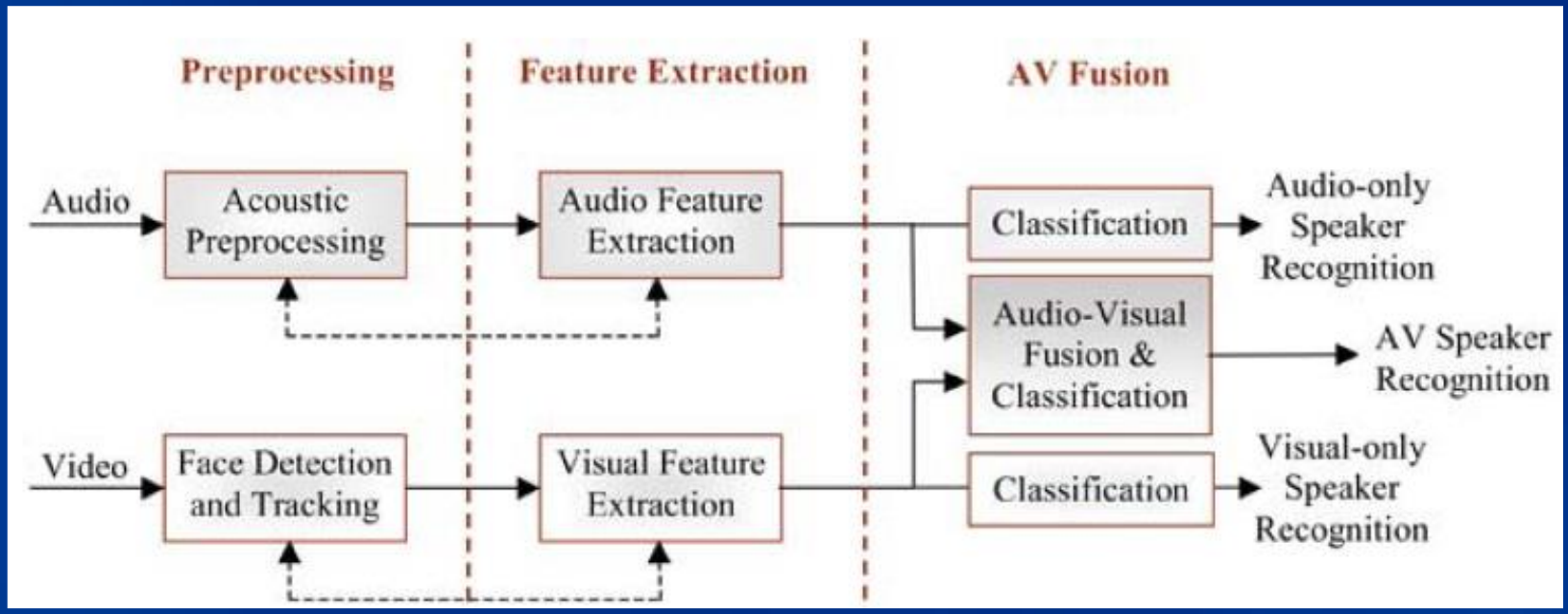
MFCC

识别与融合

传统算法介绍

- 自动唇读系统，主要包含以下部分
  - 唇部检测与定位
  - 视觉特征提取
  - 唇读识别
  - AV融合

一个典型的AV Biometrics框架



From Aleksic and Katsaggelos 06

通过上述示意图可以看出，AVSR和AV Biometrics系统框架有很多相似之处，因此后面将做统一的介绍

嘴部检测

- 早期的唇读研究者，主要把精力集中在特征和识别器上，往往采用人工标注的办法获得嘴部区域
- 自动的AVSR要求实现嘴部区域的获取，一般的，嘴部区域获取过程如下

输入图像

人脸检测

嘴部检测

嘴部检测

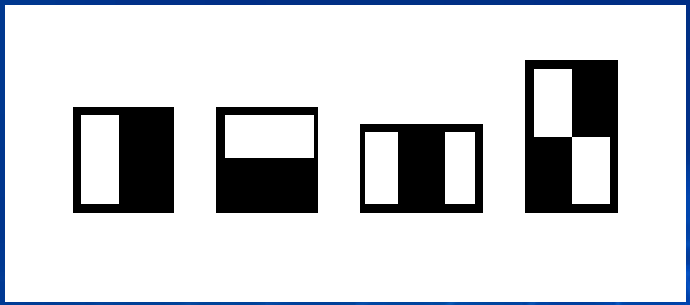
- 根据特征提取模块的需要，嘴部检测算法结果层次不一
  - 如果需要提取总体灰度特征，只需要提取出嘴部框就足够了
  - 如果需要提取唇型轮廓信息，则还需进行变形模板匹配
- 两个比较有代表性的唇区ROI检测算法
  - Haar + Adaboost (Intel OpenCV) ——获得嘴框
  - Real time Lip Contour Extraction System (HK CityU) ——获得轮廓
- 仍然是唇读的一个难点问题
  - 虽然陆陆续续提出了一些方法，但是始终缺乏统一的测试标准和测试平台，各种方法之间仍然缺乏横向的比较

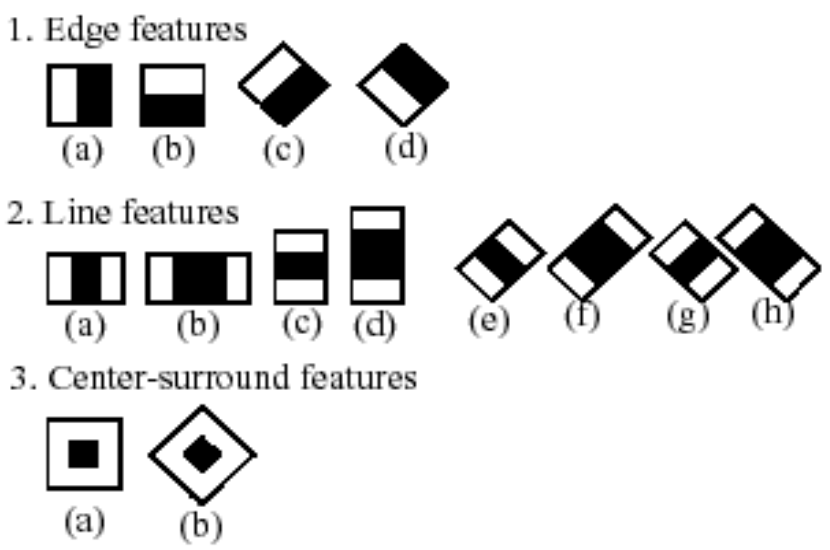
Haar + AdaBoost 嘴部检测

- OpenCV 提供了该算法的主要模块
- 该检测算法主要包含三个部分
  - 大量的、过完备的Haar-like矩形特征
  - Cascade级联结构
  - 基于AdaBoost建立每层分类器的集成

Haar + AdaBoost 嘴部检测

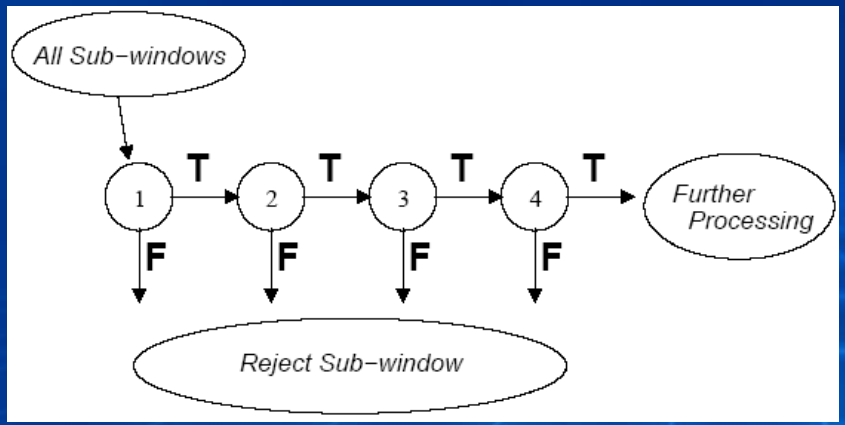
- 大量的过完备的矩形特征
  - 能够反映绝大部分的细节
  - 整个检测器力量的来源

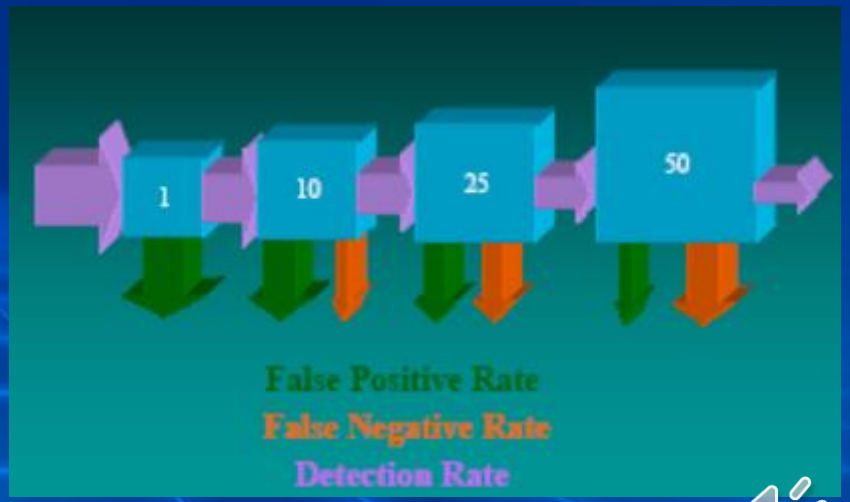




Haar + AdaBoost 嘴部检测

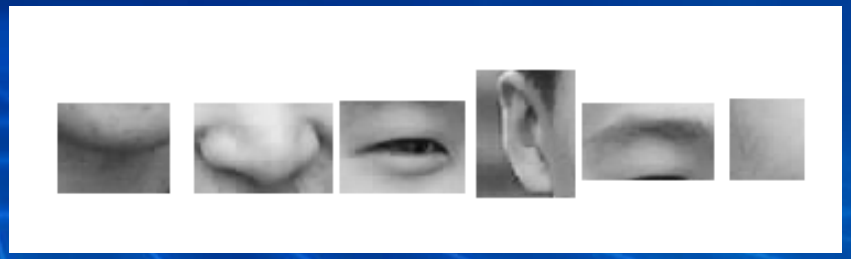
- Cascade级联结构
  - 降低单个识别器的设计难度，提高检测速度
  - “最难的”样本才用“最复杂”的识别器去检测
    - 杀鸡焉用牛刀?





Haar + AdaBoost 嘴部检测

- 基于AdaBoost建立每层分类器的集成
  - 合理高效挑选反例样本
    - 利用前一层错误分类的反例样本，对下一层识别器进行训练
  - 在每一层中，对训练出来的识别器进行加权

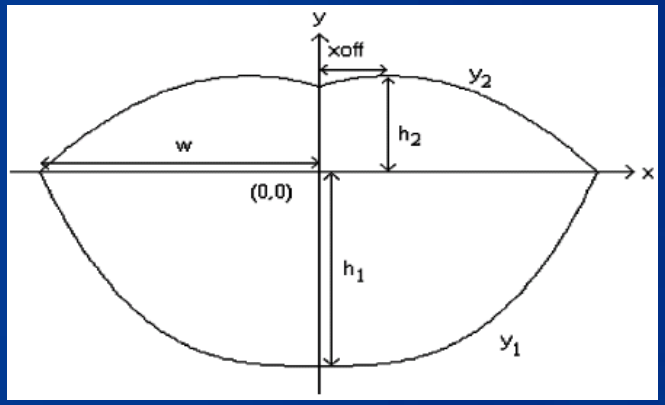


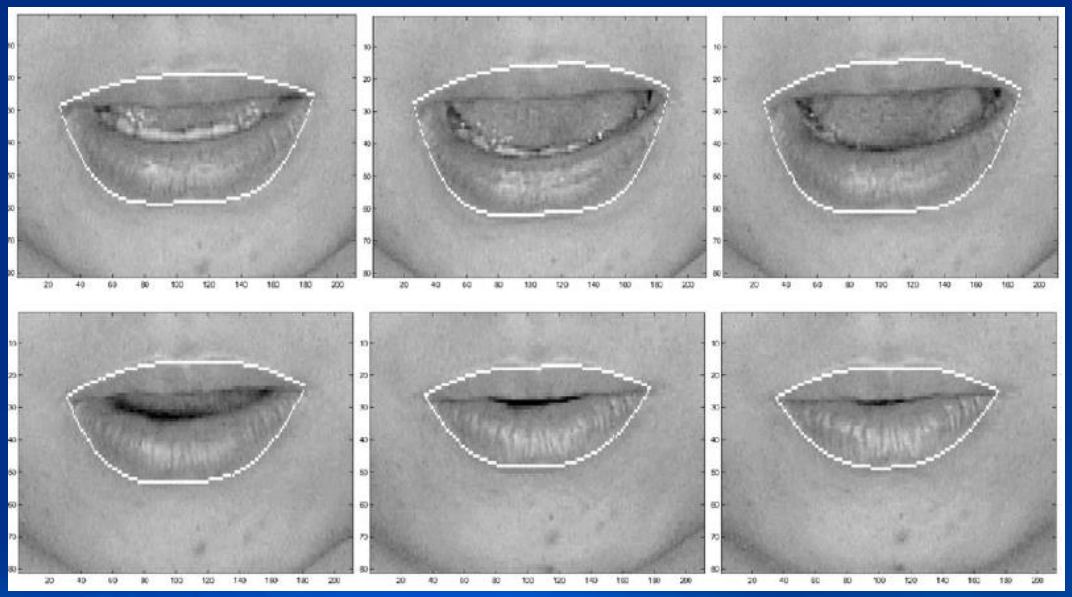
实时的唇部轮廓提取系统

- Work from HK CityU
- 大致包括2个主要步骤
  - 唇像素分割
  - 模板匹配

唇像素分割

模板匹配





传统算法介绍

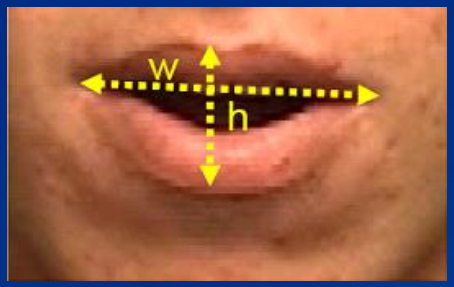
- 自动唇读系统，主要包含以下部分：
  - 检测与定位
  - 视觉特征提取
  - 唇读识别方法
  - 音视频融合方法

特征提取

- 主要有
  - 形状特征
  - 像素特征
  - 上述两者的结合
- 唇区检测与特征提取方式紧密相关
  - 嘴框的检测适用于基于像素的特征
  - 轮廓提取则适用于基于形状的特征

基于轮廓的特征

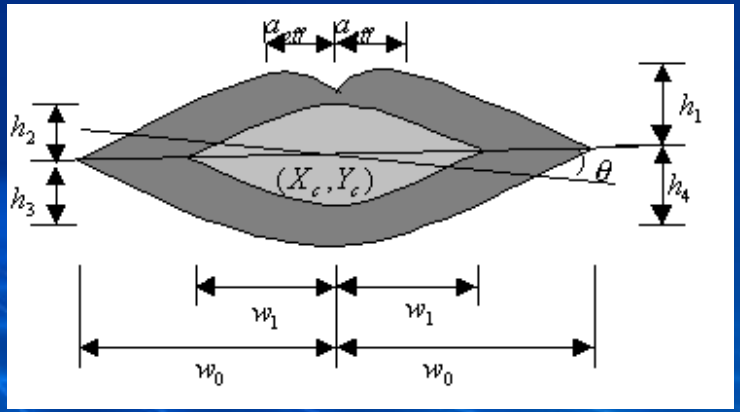
- 基于轮廓的特征:
  - *Height, Width, Area (Benoit)*
  - *Moments, Fourier descriptor(Potamianos)*
  - *Deformable Template(Yuille，Silsbee),*
  - *B-spline (Blake, Sanchez),*
  - *quartic curve (Hennecke) ,*
  - *parabola (Coianiz/Tian)，*
  - *Snake (Chiou)*
  - *Active Shape Mode (Luettin)*
  - *…*



- 特征提取
  - 对唇模型进行假设，将模型与唇轮廓匹配的过程
  - 本质是一个代价函数最小化的过程（优化迭代过程）

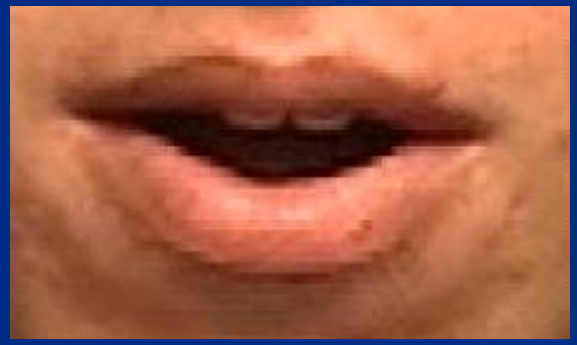
基于轮廓特征

- 优点
  - 特征维数少
    - 刻画右侧的唇型模版只需要13个参数
  - 对形状和旋转变化不敏感
- 缺点
  - 丢失了很大的信息
  - 对噪声点十分敏感
  - 涉及的匹配过程消耗大
  - 在目前识别率不佳



基于像素的特征

- 基于像素的特征
  - *PCA (Bregler),*
  - *whole ROI (Waibel),*
  - *DCT (Duchnowski) ,*
  - *DWT (Potamianos),*
  - *LDA (Duchnowski)*



- 已有的正交变化方法，均可被用来提取像素特征：DCT, KLT, DWT…

$$O = P \times I \qquad O = P^T \times I \times P$$

基于像素的特征

- 优点
  - 保留了全部的信息
  - 运算效率高
  - 目前识别率最高
- 缺点
  - 对形状旋转位置变化敏感
  - 特征维数过高，因此，往往需要降维
- 降维方法
  - DCT和唇读特性 ([Potamianos 98])
  - 特征空间本身特性(Correlation, 统计特性[Heckman02], Mutual Information[Scanlon04])
  - 从对识别的贡献来考虑(Genetic, Adaboost)
  - 线形变换（LDA[Potamianos 98], PCA ）
  - 。。。

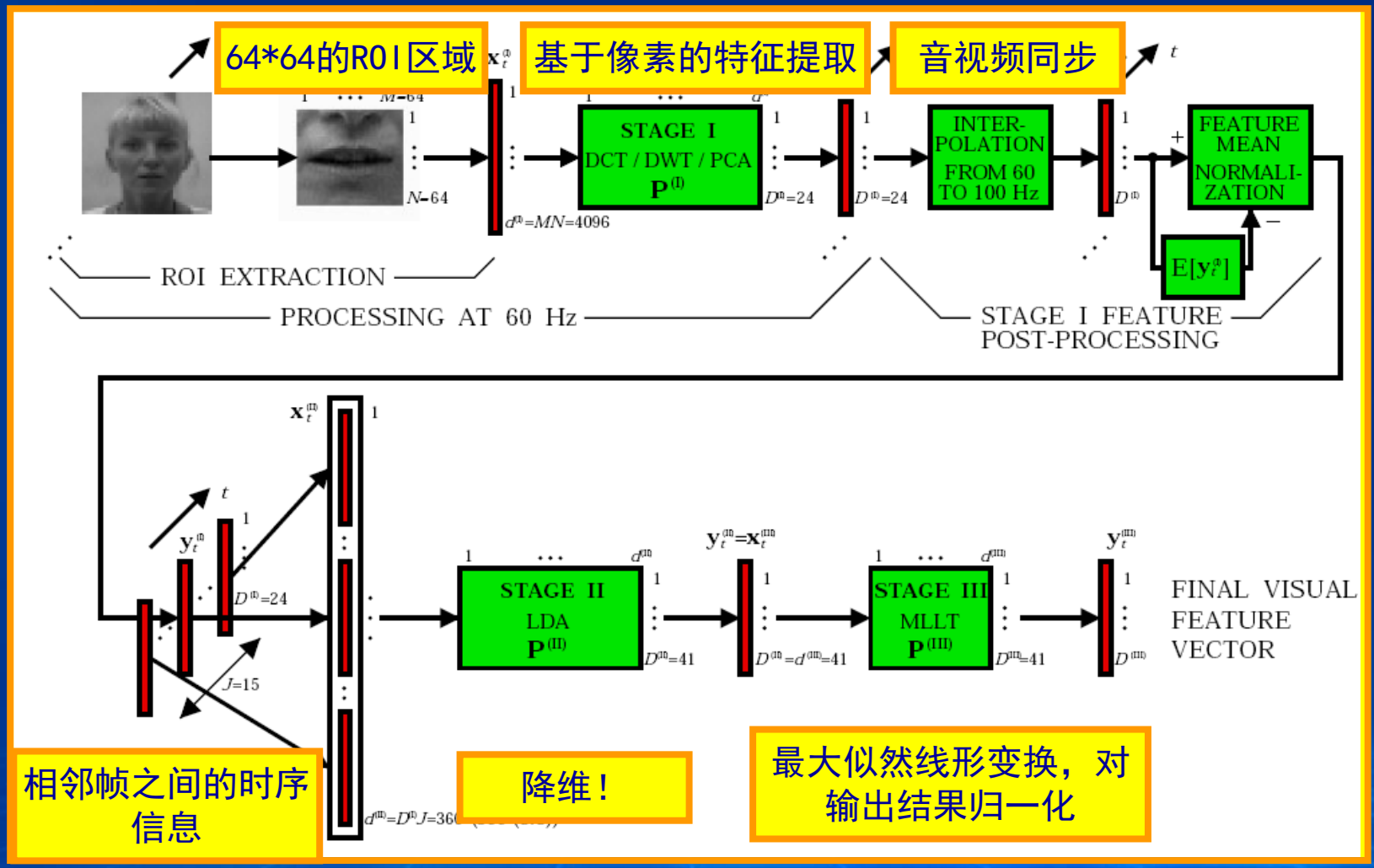
混合特征

- 结合形状特征和像素特征的特征
  - *Active appearance mode (Matthews)*
  - *Direct Combination (Yao)*
- 一般认为，结合形状特征和像素特征的结合特征，识别率最高，代价最大

基本结论

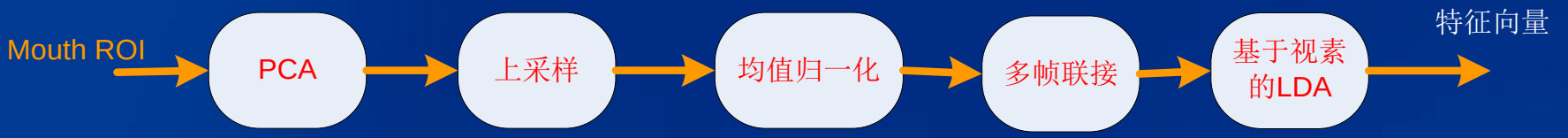
- 面积特征在AVSR应用中是比较鲁棒的特征
  - Potamianos et al. 2004
- 一般认为，基于像素特征的方法是最有效的
  - (不是准确率最高) [Matthews PAMI02]
  - 其中，DCT特征perform equally well or better than others *[Potamianos 01]*
- 举例

IBM的AVSR特征提取Cascade结构



Potamianos， IBM Watson Research Center

Intel AVSR OpenSource



- 与IBM的AVSR特征提取框架非常相似

传统算法介绍

- 自动唇读系统，主要包含以下部分：
  - 检测与定位
  - 视觉特征提取
  - 唇读识别方法
  - 音视频融合方法

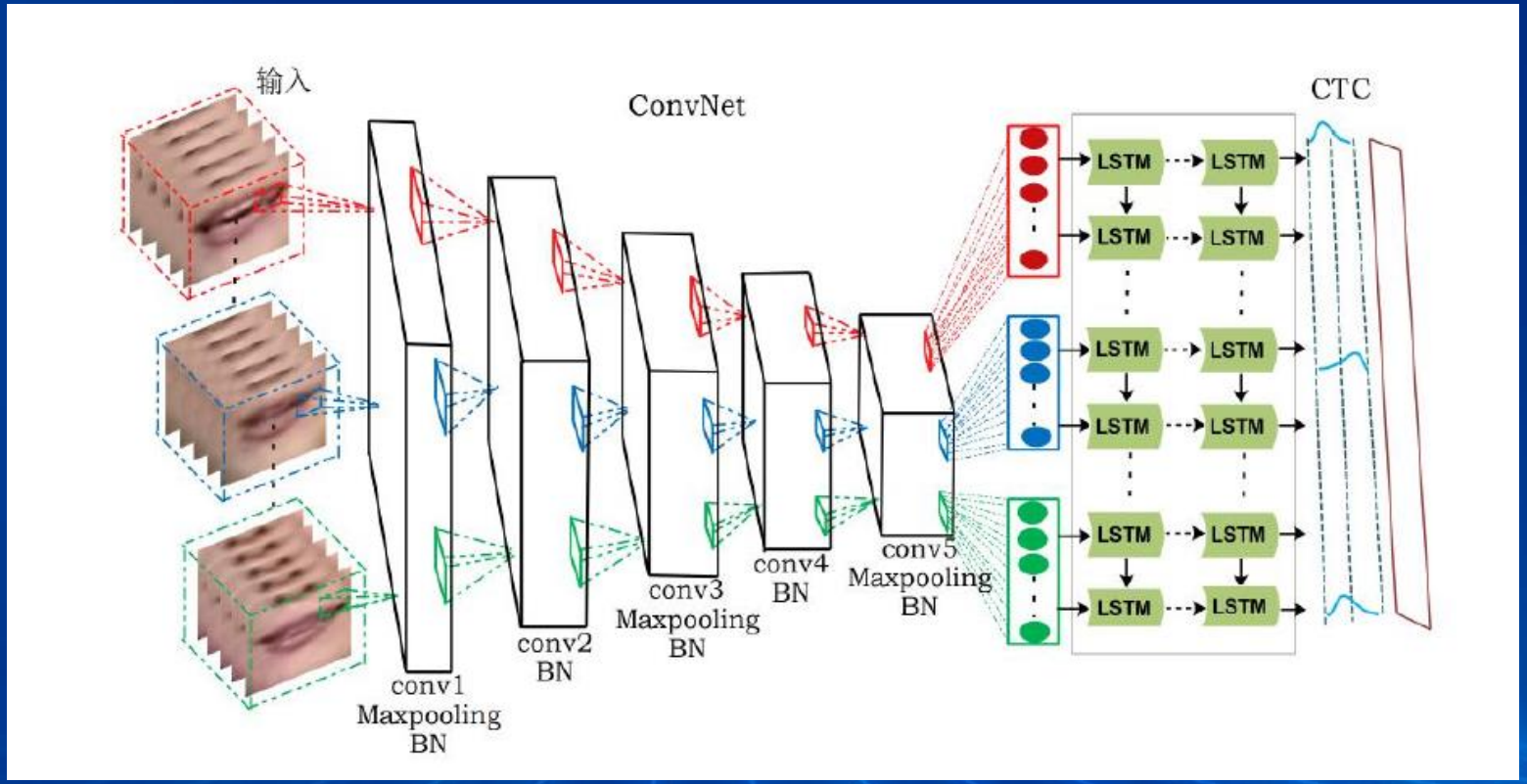
识别算法

- 隐马尔科夫模型HMM是当前音视频语音识别技术主流识别器
  - 如果一个过程的“将来”仅依赖“过去和现在”而不依赖“将来”，则此过程具有马尔可夫性，或称此过程为马尔可夫过程
  - 如果仅与之前的n个状态有关，则称为n阶马尔科夫模型

基于HMM的唇读方法

- 唇读过程本身是一个HMM过程
  - 能观察到唇动过程，可以拆解成一帧帧的图片，即“观察值的序列”
  - 这些序列或者序列的片段，背后对应着一些特定语义信息，这是直接观察不到的”状态”
- 如何用HMM做唇读？
  - 通过可观察的序列（一帧一帧唇动图片），去猜测其背后的隐状态（特定的语义信息）
    - 这就是HMM的“评估问题”，可以用前向后向算法解决(实际中多用Viterbi算法近似)
  - 而HMM的建立过程，是“学习问题”，用Baum-Welch算法解决

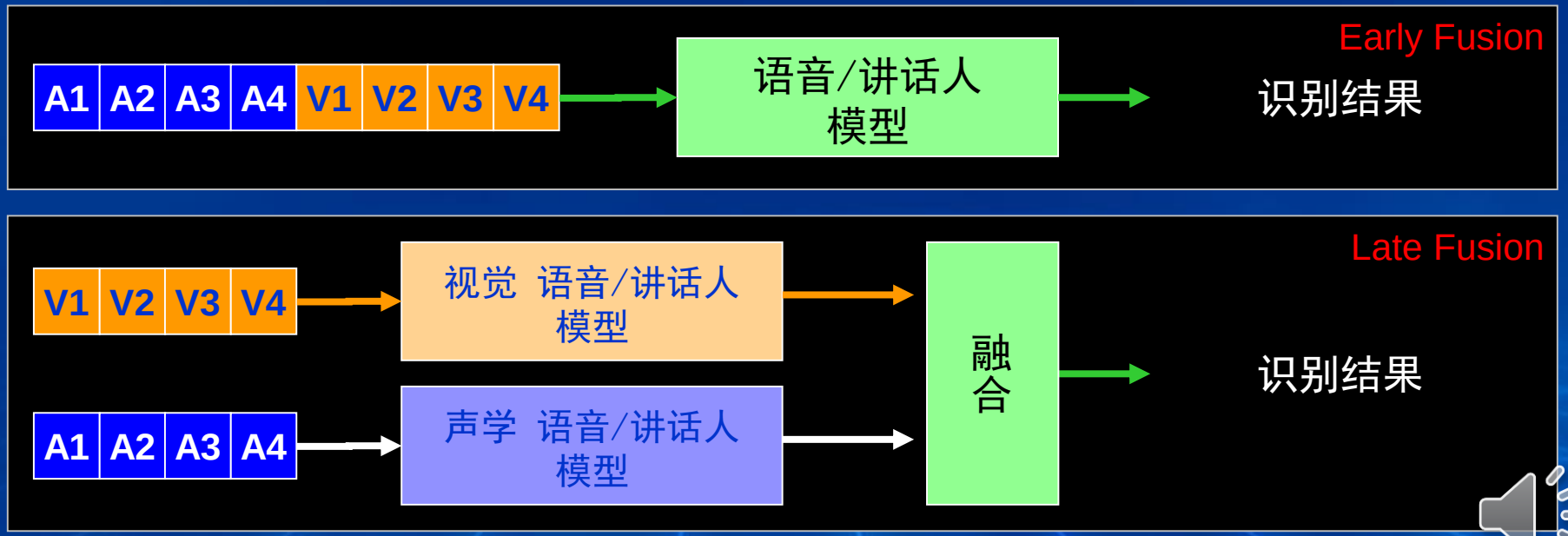
基于深度学习的唇语识别



P2P网络

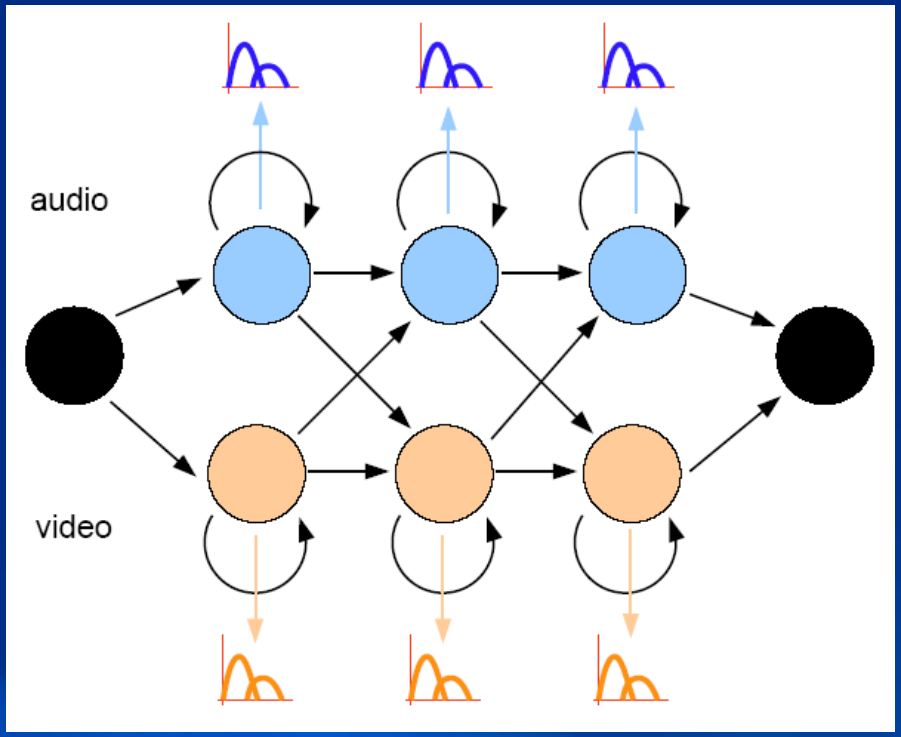
音视频融合算法

- 音视频的融合
  - 多个通道信息如何作出统一决策？
- 融合技术主要有
  - 前融合(又叫特征融合)和后融合(决策融合)



中融合

- 问题
  - 决策融合无法刻画音视频之间的依赖关系
  - 特征融合饱受噪声，以及音视频信号之间的不同步的影响
- 为解决上述问题，中融合——也就是识别器内部融合被提出
- coupled HMMs
  - 典型的中融合算法



by David Dean

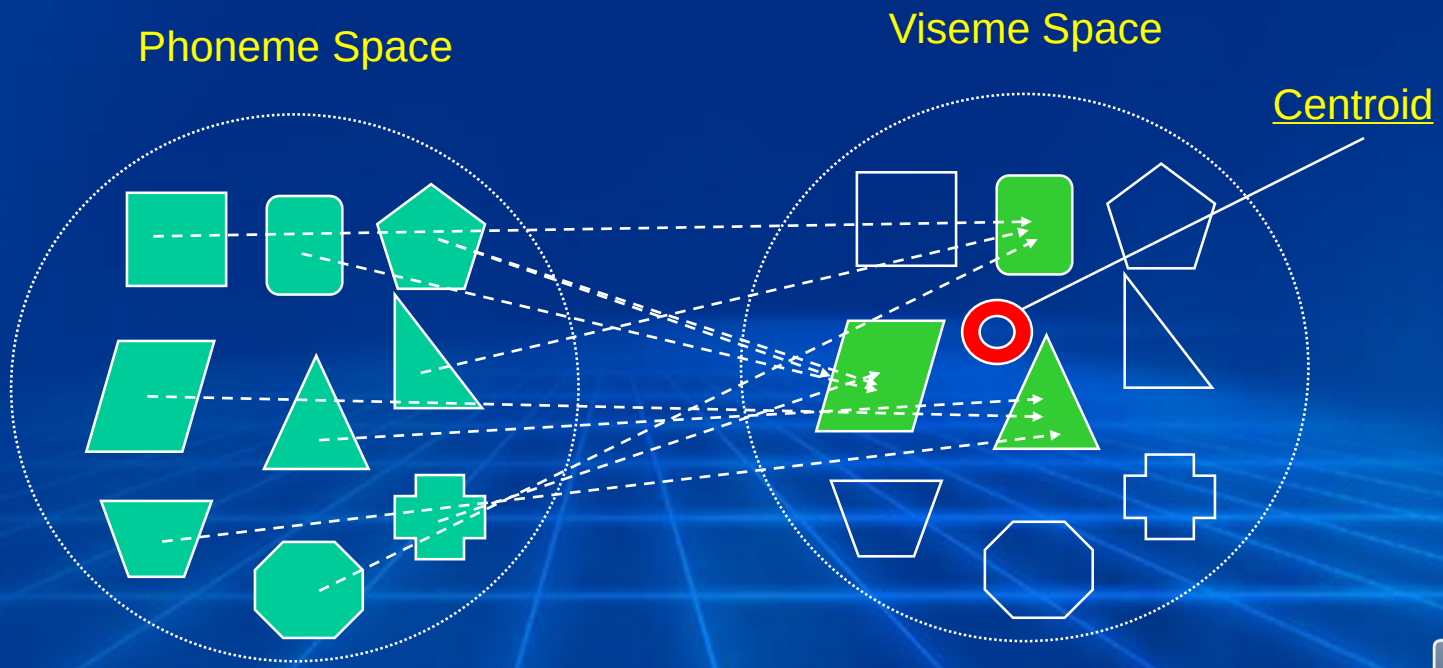
面向唇读的Audio-to-Visual Mapping

- 音素：声学上可区分的最小语义单元。
- 视素：视觉上可见的最小语义单元。
- 音素和视素之间是多对一的关系
  - 不同的音素，可能在视觉上并不可区分
  - 从视觉的角度，很多时候只能看到一部分发音器官，如嘴唇和下巴，对舌头和声门的变化，是看不到的
- 音素→视素对应关系的确定方法
  - 聚类的办法

近期有学者质疑视素存在的必要性，并且建议取消这个Mapping过程(Hazen06)

基于聚类的映射

- 步骤1：声学信号分类
- 步骤2：将这些声学类映射到相应的视觉输出，然后计算它们的映射的“中心”



from Courtesy, Hung Nguyen

唇读的最新发展动态

- 进入2004年之后，唇读的发展出现了一些新的动态，集中表现在
  - 鲁棒有效代表语言的视觉特征
    - ——开始尝试突破DCT等像素特征
  - 与人类语言感知功能最匹配的识别模式
    - ——开始尝试突破HMM识别方法
  - 新的融合算法
  - 多姿态的唇读问题
  - 一些崭新的应用
    - Car Environment
    - Embeded Device上的尝试

一些崭新的应用

- 车载环境
- 嵌入式设备
- 多姿态唇读

车载环境

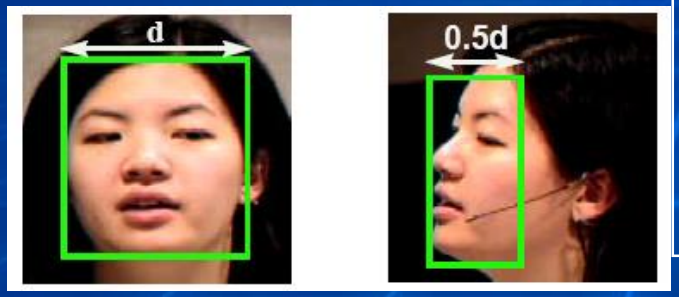
- UIUC的Tomas Huang研究小组
  - 建立了一个在轿车环境下采集的音视频语音库，开始进行真实环境下的AVSR研究
- 该数据库立足于：
  - 真实移动环境下数据的获取；
  - 开发和应用鲁棒的音视频特征提取算法；
  - 在提取的特征上测试利用小词汇库建立的模型的有效性；

开始多姿态的研究

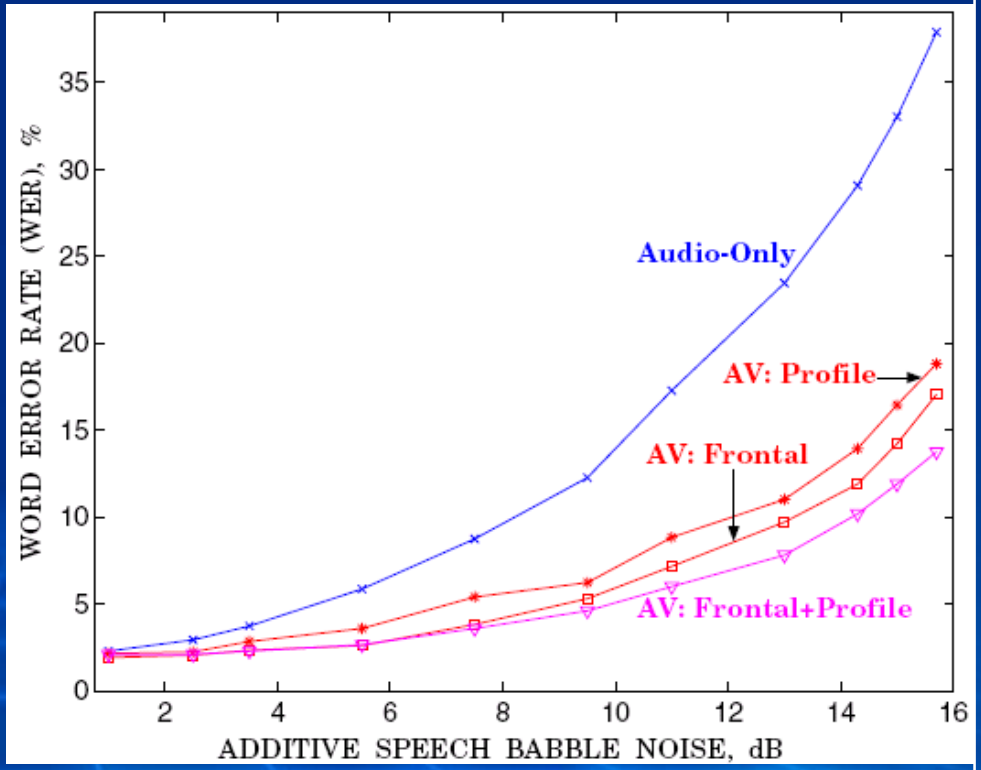
- 从06年以来，IBM和CMU的科学家们开始研究多姿态的唇读问题
- 力求回答
  - 侧面(Profile View)的视觉语言信息是否与正面(Frontal Views)的视觉语言信息有所不同？

Profile Versus Frontal!!

- 作为基准，正面唇读系统的word error rate (WER)为25.4%。
- 侧面系统的为39.9%，相对有60%的退化。
- 两者进行特征级融合WER为23.7%
- 类似的情况见右图



P.Lucey, Potamianos 2006



IBM CHIL Project

- CHIL: “Computer in the Human Interaction Loop”,
  - 是由IBM牵头的，来自９个国家15个机构合作的项目
  - 主要面向研讨会和交互式会议应用场合
- 有以下几个主要目标：
  - “connector” → connect humans using the right medium at the right time. 恰当的互联？
  - “memory jog” → helps humans remember facts at the right time. 助推记忆？
  - “Socially supportive work spaces” → helps humans collaborate 推进协作
  - “Attention cockpit” → helps facilitate human interaction
  - Others → meeting browser, translator, etc.
- 其中，视觉信息部分，主要由安装在会议室各个角落５个固定摄像机和４个pan-tilt-zoom (PTZ) 摄像机提供。

主要涉及的技术

- 视觉跟踪技术
  - 3D Head Tracking
  - 2D face detection
- 语音技术
  - 声音检测
  - Speaker Diarization
  - 语音识别；
- 音视频技术
  - 双模态语音识别

其他资料

- 目前对多姿态唇读的研究，就是在这种情形下提出来的
- 对CHIL项目感兴趣的同学，可以访问
  - http://chil.server.de/
- 对唇读最新研究成果感兴趣的同学，可以
  - 访问*HCSNet Workshop on the Use of Vision in HCI - VisHCI 2006*主页
  - http://users.rsise.anu.edu.au/~vishci/program.html

其他资料

- 牛津大学人工智能实验室/谷歌 DeepMind 团队 和加拿大高等研究院（CIFAR）联合发布**结合深度学习技术的唇读程序 LipNet**。
- 在 GRID 语料库上，LipNet 实现了 93.4% 准确度，超过了经验丰富的人类唇读者和之前的 79.6% 的最佳准确度。研究人员还将 LipNet 的表现和听觉受损会读唇的人的表现进行比较。平均来看，他们可以达到 52.3% 的准确度，LipNet 在相同句子上的表现是这个成绩的 1.78 倍
- GRID 语料库包含 34 个志愿者录的短视频(每个3 秒)
- https://www.leiphone.com/news/201611/lmrRpn2DdOUoex3E.html

总结与展望

- 在过去的20年间，取得了很大的进展
- 但是仍然离实际应用较为遥远
  - 打一个比方：现在的唇读技术发展水平，与80年代中期的语音识别技术水平非常相似
  - 已经具备了相当积累，亟待突破的时刻