

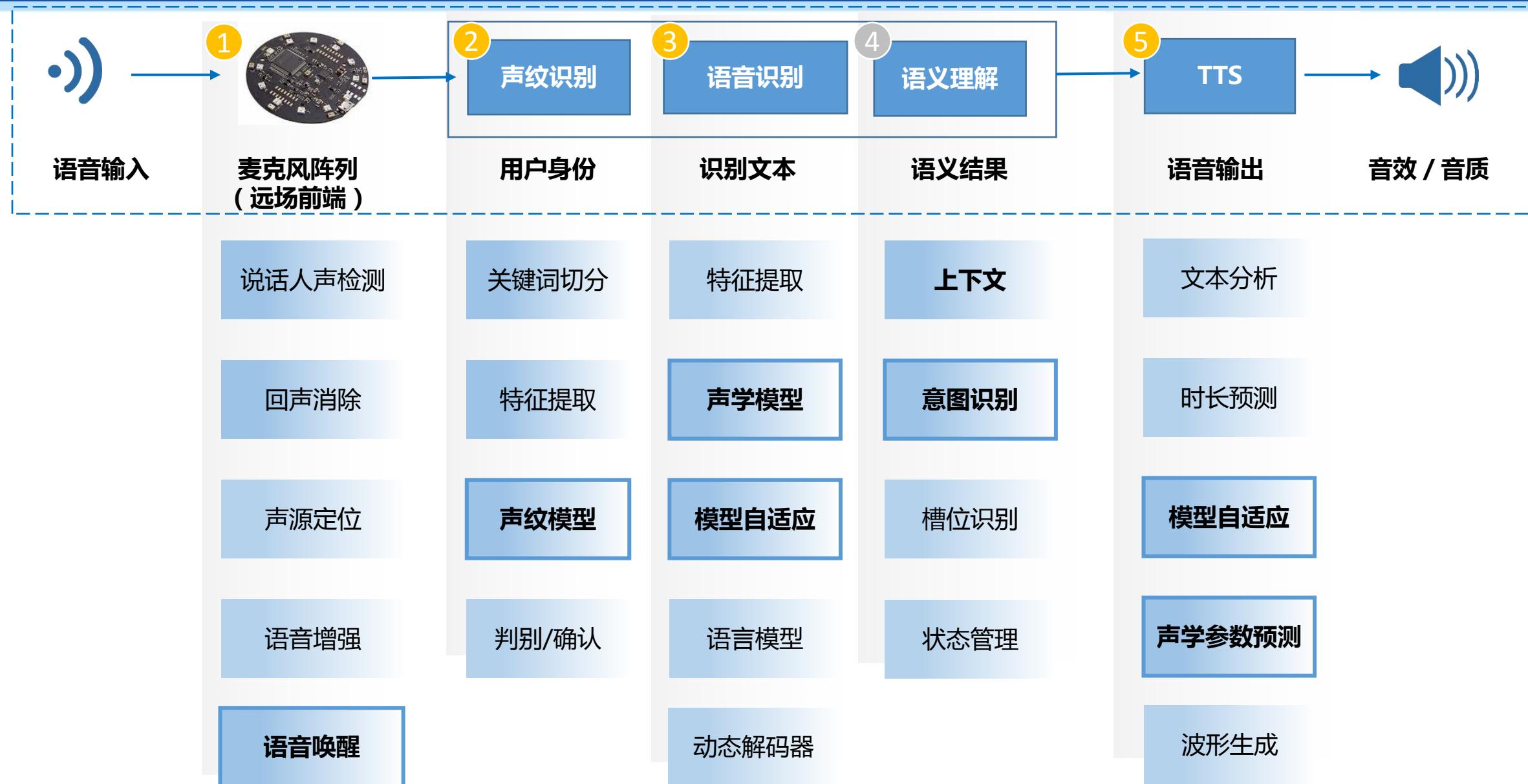


智能音箱语音技术分享

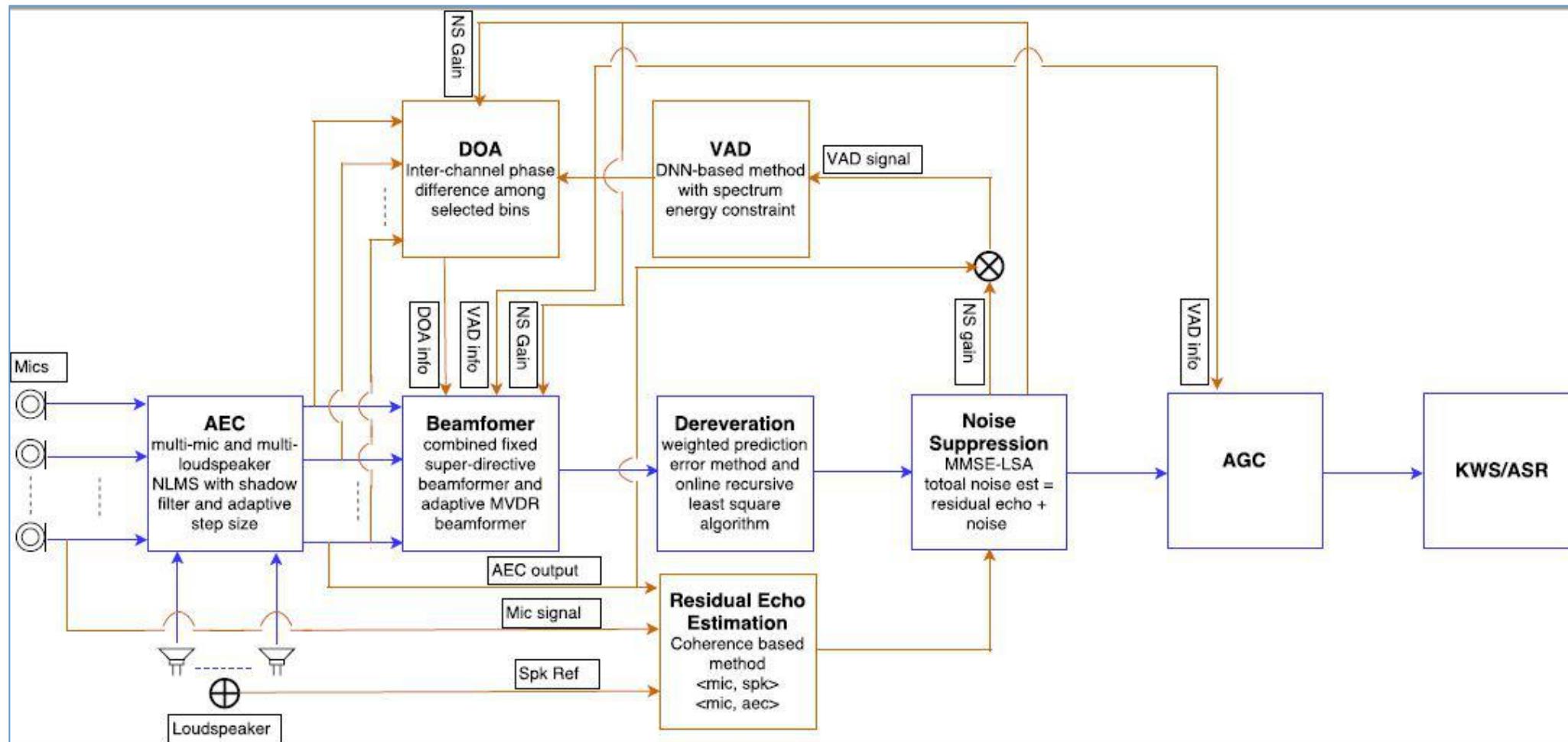
2018 / 09

- 智能音箱语音交互技术链条介绍
 - 麦克风阵列
 - 声纹识别
 - 语音识别
 - TTS
- 前沿研究和技术分享
 - 基于唤醒词信息的目标说话人语音提取
 - 端到端语音识别的Attention建模方法的关键技术点

智能音箱语音交互技术链条



- AI Lab Voice Processing (简称AIVP) 集成语音检测、声源测向、麦克风阵列波束形成、定向拾音、噪声抑制、混响消除、回声消除、自动增益等多种远场语音处理模块



- 语音唤醒难点:

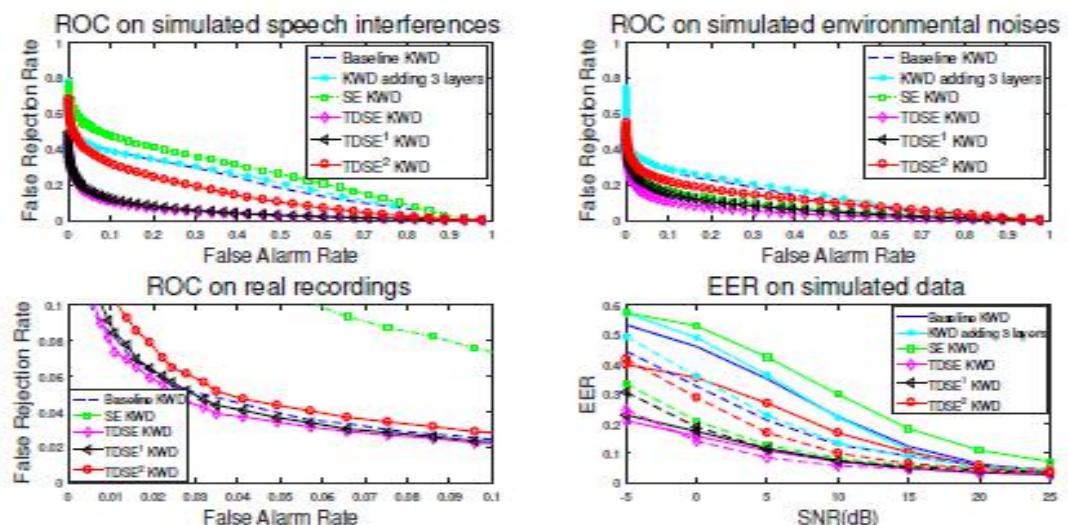
误唤醒;
噪声下唤醒率;
快语速、儿童唤醒率

• 语音唤醒改进

- 唤醒模型算法升级;
- 复杂度大幅提升后在保持性能的同时有效压缩到可用范围内;
- 误唤醒率降低60%以上~

• 唤醒词相关语音分离与增强

- 分离关键词与其它非关键词语音
- 分离关键词与其它环境噪声
- 说话人无关，文本相关
- 辅助KWS，提高单通道场景的远场噪声鲁棒性
- 降低对话和其它语音场景的误唤醒率
- 降低前端和KWS模块的功耗
- 大幅度提升噪声及背景人声下唤醒性能



- 目标

声纹认证

96%以上正确率确认身份

活体验证

结合声纹特征与内容，99.99%以上准确率

- 技术挑战

- 信道失配
- 环境噪声
- 短语音
- 远场

- 应用挑战

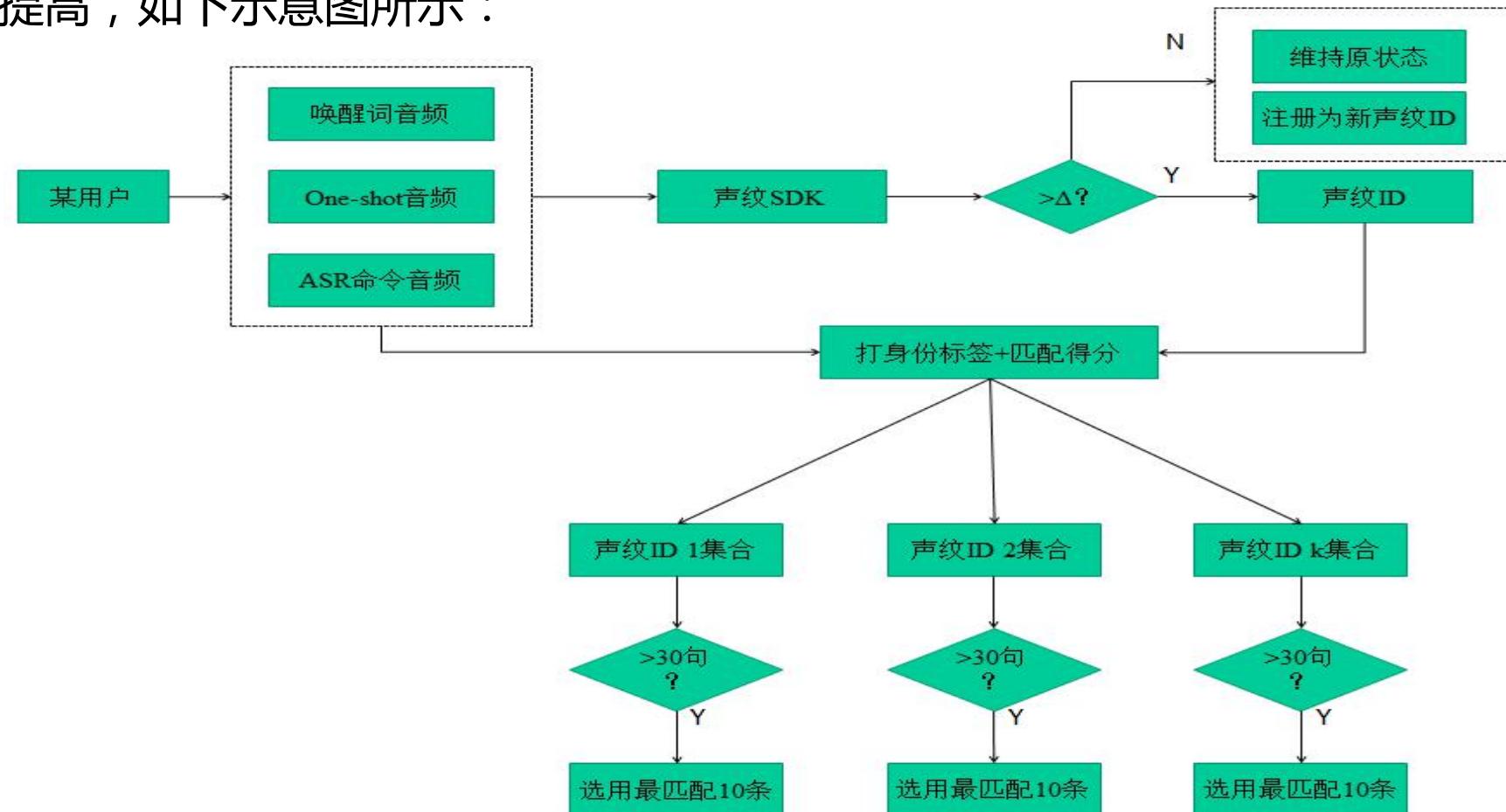
- 录音冒认
- 兼容能力
- 交互设计

- 当前实际问题

- 当前训练的500个人数据量偏少，持续采集中。
- 工程实现上，唤醒词截取不准确不干净问题；
- 音箱实际的使用场景比较复杂，有远场、噪声和多人同时说话的情况。

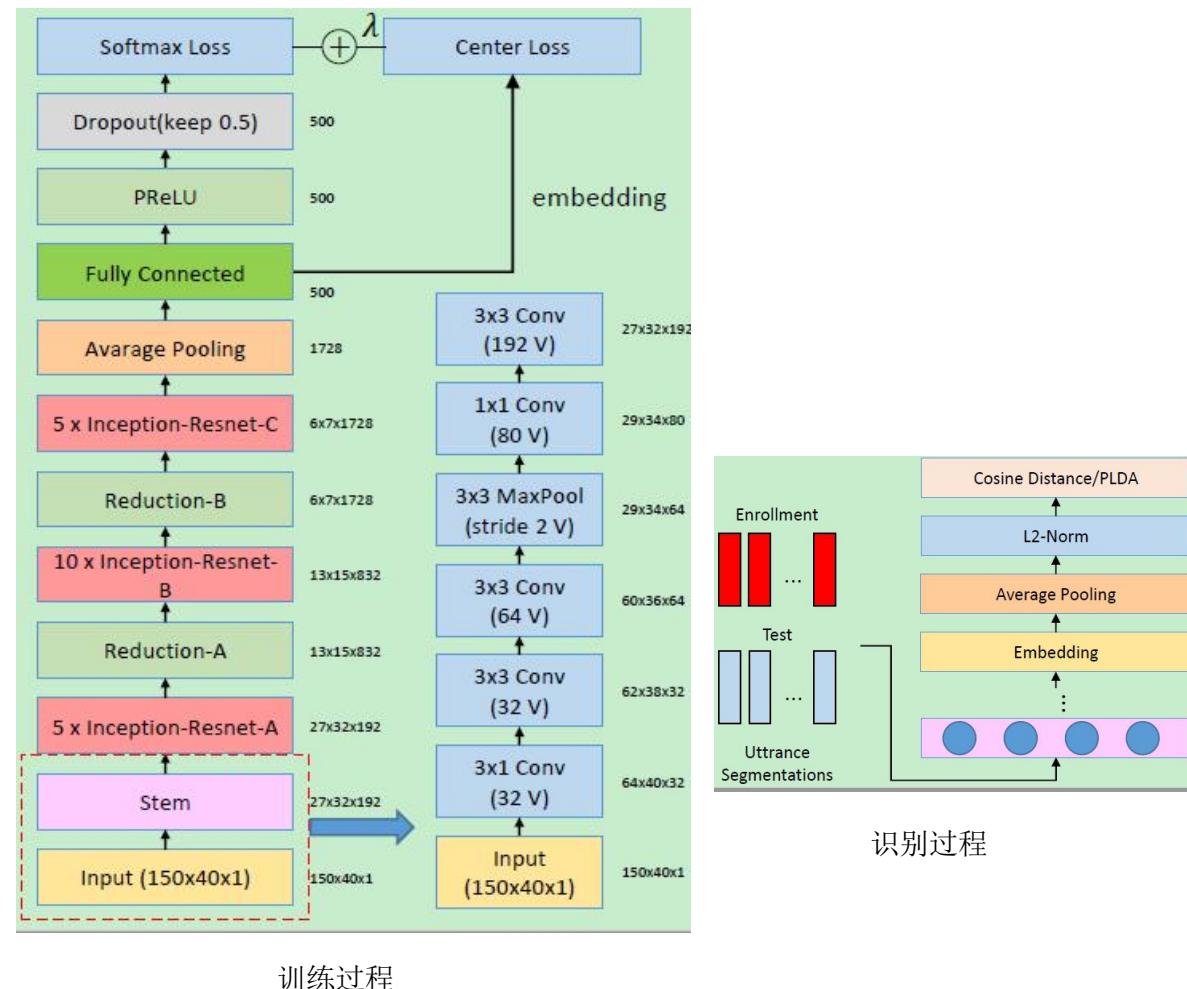
• 多识别任务融合

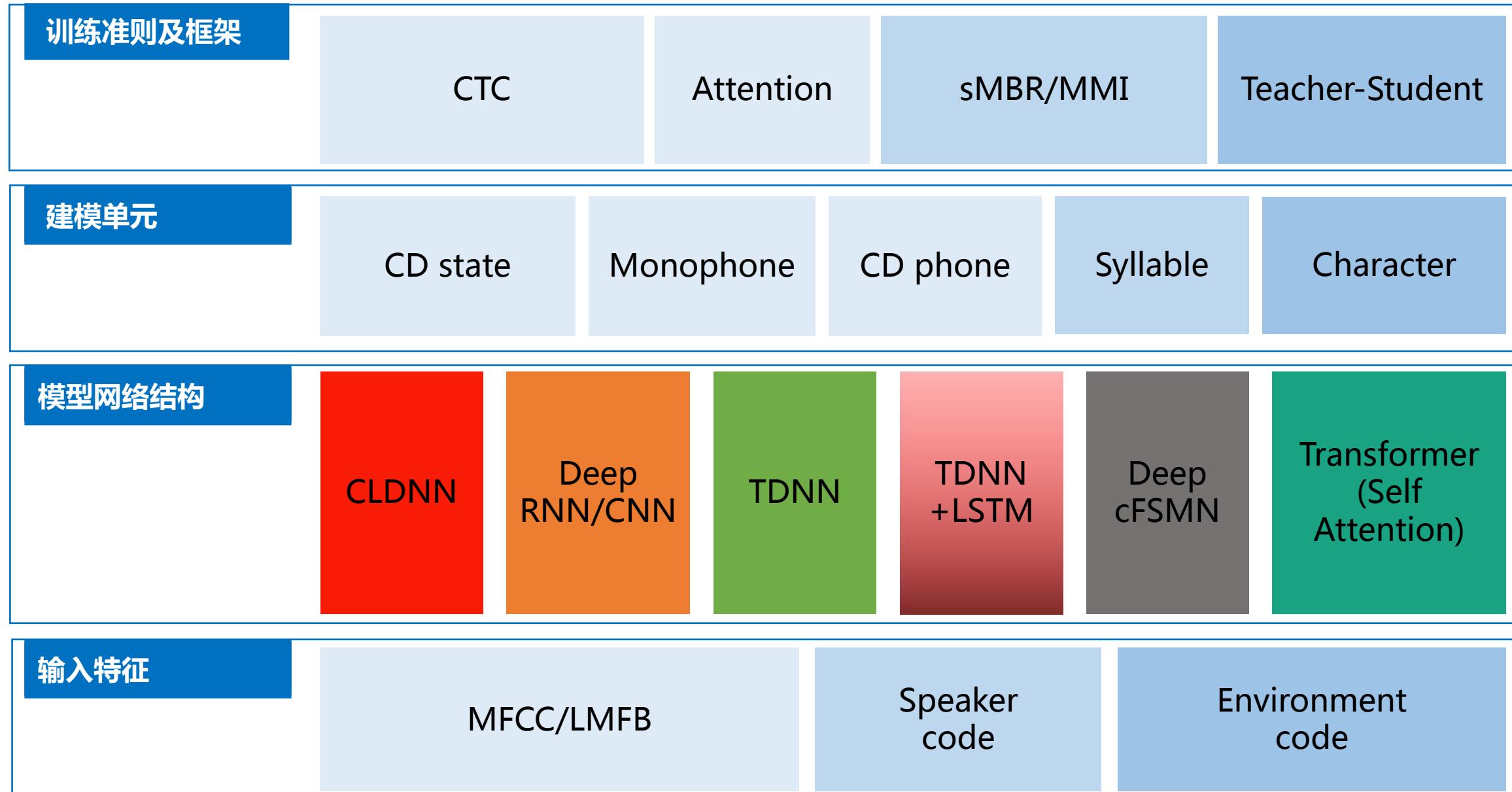
- 兼容确认和辨别功能，支持隐式更新和隐式注册，随着用户使用时间的增长，系统性能逐渐提高，如下示意图所示：



- 基于声纹特征，提供用户性别及年龄段属性
 - 无论用户是否注册，用户唤醒之后，声纹系统即会判断该用户的性别和年龄信息，便于在之后的互动中，根据用户属性进行相关推荐。
- 自研的多种类型声纹识别算法
 - 除了已实现的经典声纹识别算法外（GMM-UBM, GMM/Ivector, DNN/Ivector, GSV），探索和开发基于DNN embedding新方法，目前在短语音方面，具有比主流方法更精确的识别效果。部分核心自研算法及系统性能可参考语音顶级期刊。
 - 进行多系统融合的开发工作，合理布局全局框架，使具有较好互补性的声纹算法进行协同工作以进行更精准的识别。
 - 基于Inception-resnet的声纹识别框架

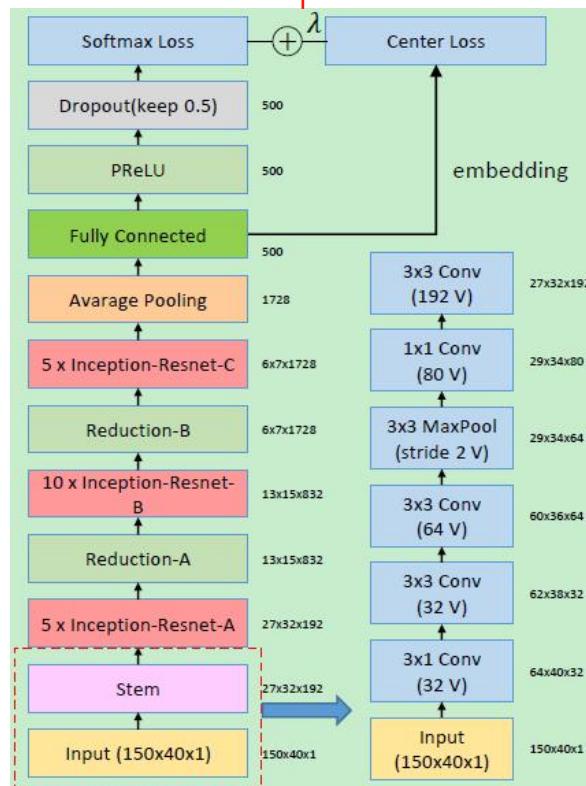
- 基于Inception-Resnet的声纹识别系统框架





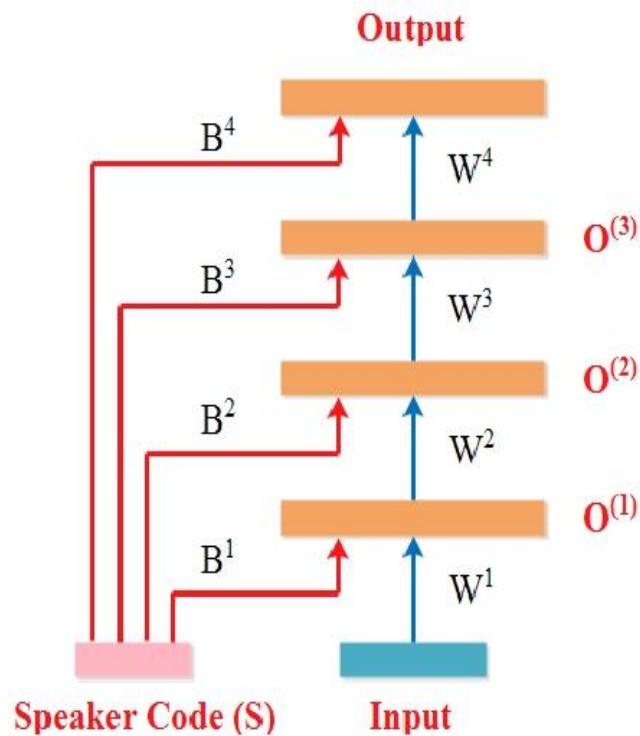
- 为每位用户提取并保存自己个性化声学信息特征；
- 随着用户数据积累，个性化特征自动更新，用户识别准确率可获得显著提升；

特有的声纹模型技术 用户说话人特征提取

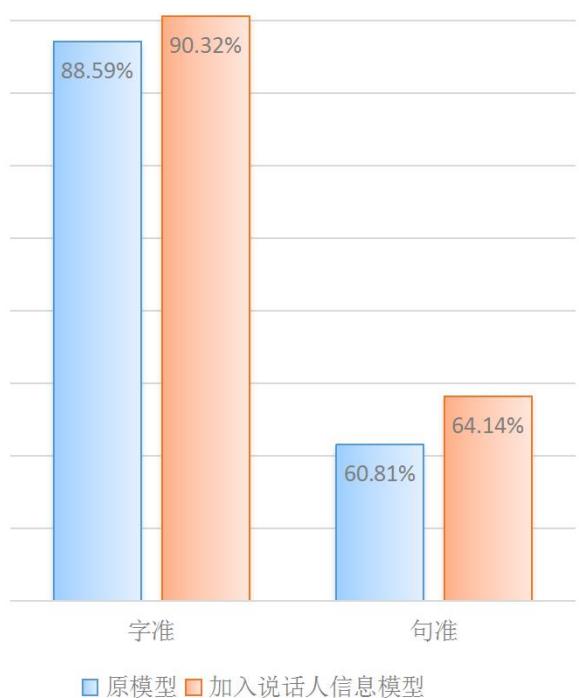


Speaker #1

结合说话人信息的声学模型训练 实验说话人特征引入方式及策略



目前采用用户一句话提取说话人特征而获得的准确率提升：



- 智能音箱语音交互技术链条介绍
 - 麦克风阵列
 - 声纹识别
 - 语音识别
 - TTS
- 前沿研究和技术分享
 - 基于唤醒词信息的目标说话人语音提取
 - 端到端语音识别的Attention建模方法的关键技术点

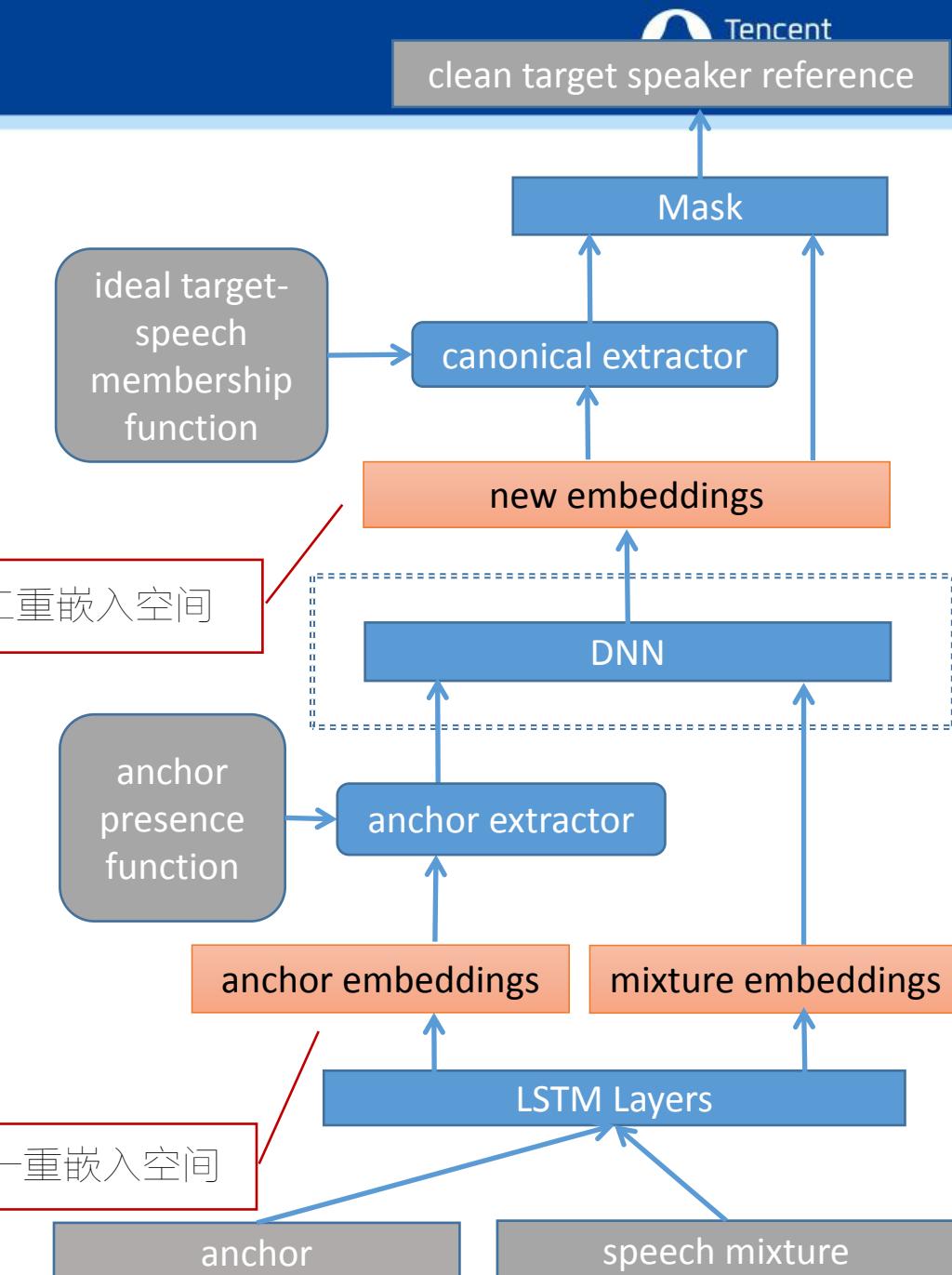
- 智能语音设备真实场景中经常出现其他说话人干扰的情况，使语音识别准确率急剧降低
- 现有最新技术有何缺陷？
 - 针对目标说话人的深度学习专用模型、专用隐层或专用偏置训练方法，只对闭集中的目标说话人有用，而无法用于未知目标说话人，**拓展性差**，且专用隐层或专用偏置训练的方法**不能有效捕获目标说话人特征**
 - 自适应到目标说话人的波束形成深度学习方法，目前最先进的技术都要求平均至少10s的自适应语音，远超真实应用场景可接受的自适应语音（例，唤醒词）长度，**可用性差，不易落地**
- 攻关目标
 - **性能最优。**多项指标评测，包括信号失真比（SDR）、主观语音质量评估（PESQ）、干扰说话人数、鲁棒性；
 - 从系统实时性、模型参数复杂度全方面评估，另还包括：
 - **拓展性：**能否用于开集？能否处理任意未知目标说话人？
 - **可用性：**能否落地——所需要的自适应语音时长能否大幅减少？能否满足真实应用场景？
 - **深度研究价值：**方法能否进一步拓展到处理无自适应语音的情况？

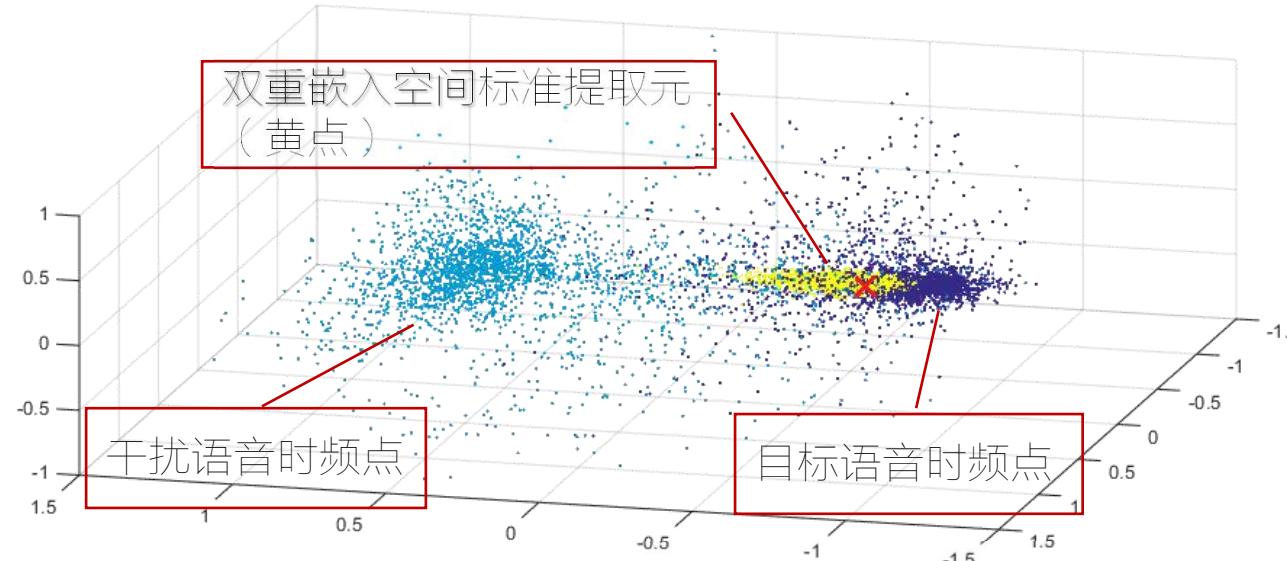
基于唤醒词信息的目标说话人语音提取



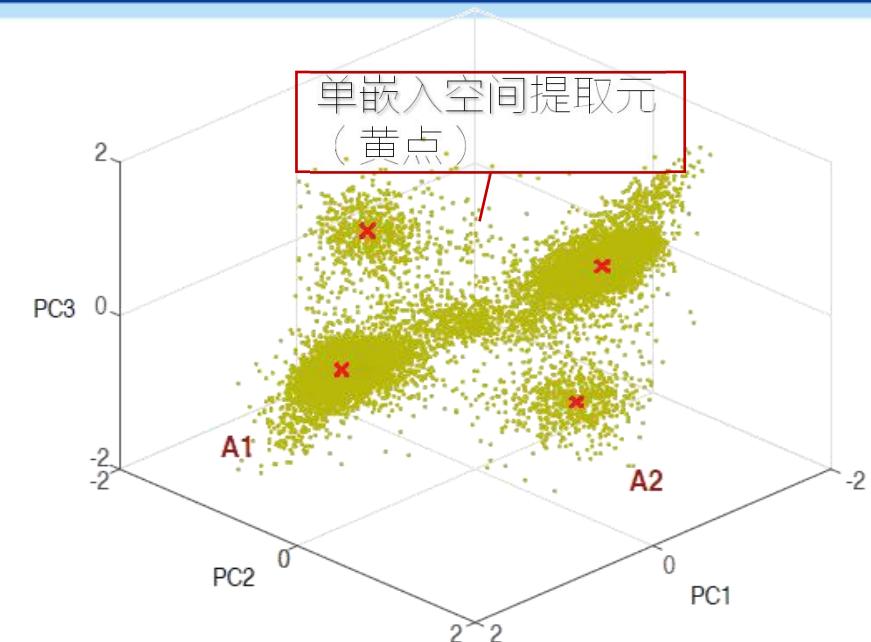
clean target speaker reference

- 心理声学研究发现人类听觉系统处理多说话人信号的注意力选择机制
 - 先采用某注意力选择机制初始化锚点
 - 然后基于该锚点回馈循环，对输入混合信号进行选择性增强
- 提出一种采用双重嵌入空间映射的“深度提取网络”（受上述发现启发）方法
 - 第一重嵌入空间：LSTM层提取目标说话人的锚参考语音（例，唤醒词）和混合语音的绝对特征
 - 第二重嵌入空间：基于上述绝对特征，通过前向网络计算混合语音以目标说话人语音特征为锚参考点的相对特征
 - 在第二重嵌入空间中，基于相对特征来计算标准提取元（canonical extractor），从而计算混合语音中时频点与该extractor点的距离来衡量此时频点归属于目标说话人的权重
 - 端到端训练右图统一网络，训练准则是最小化目标说话人频谱与干净的目标说话人频谱之间的误差
 - 测试时，采用训练数据所有标准提取元的质心点作为测试的标准提取元，无需在测试时重新估计





- 深度提取网络通过双重嵌入空间构建的标准提取元，分布十分稳定集中（左图黄色点）
 - 在端到端的训练中同时编码了有监督的标注信息和目标说话人语音锚参考点信息
 - 通过第二重空间，编码相对信息，较原始绝对位置信息更稳定
 - 即使锚参考信号极短（<1s）的条件下，亦能有效捕获目标说话人特征，如图：双重嵌入空间中构建的目标语音时频点与干扰语音明显区分开来
 - 由于标准提取元的稳定集中分布的特性，测试时可采用说话人无关的标准提取元（上图红×所示），从而可支持逐帧实时处理



- 其它基于深度学习的前沿方法[DANet](#)得到的提取元（右图）
 - 在单嵌入空间构建的提取元，仅利用了绝对位置信息
 - 对不同目标说话人，可能学到分布相对松散不稳定的提取元（下图黄点）
 - 测试时，若采用分散的提取元（下图红×所示）的质心来衡量距离，则结果次佳；若采用后期K-means聚类，则不能支持逐帧实时

- 针对背景人声问题---目标说话人提取分离

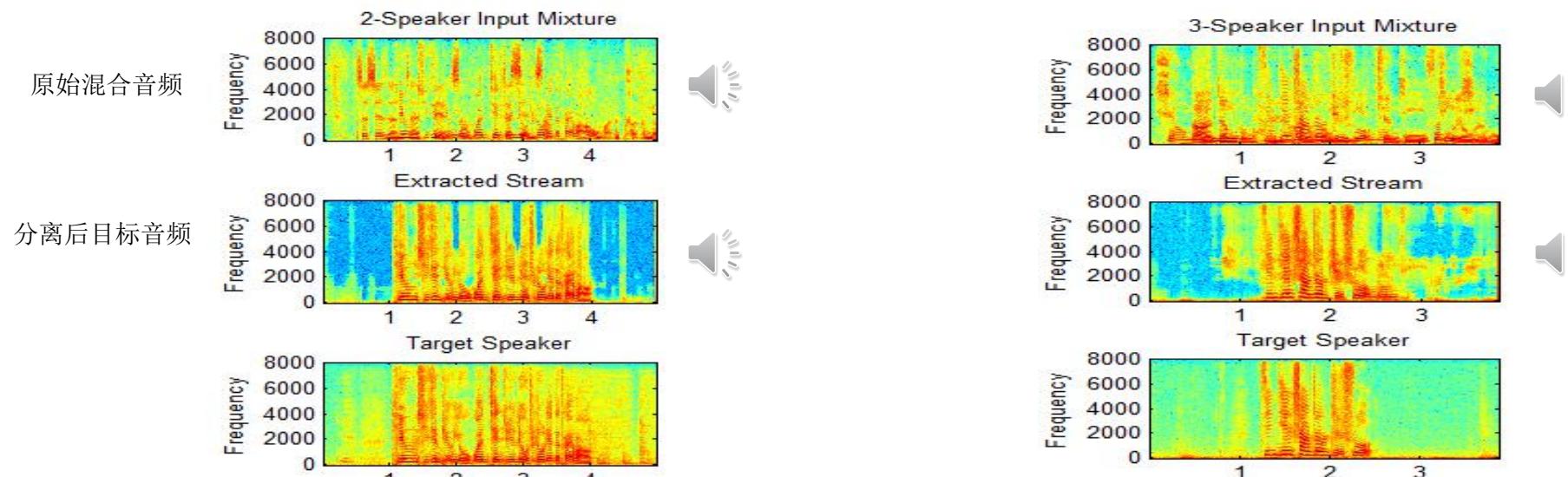
- 排列不变性训练 (Permutation Invariant Training)

D. Yu, K. Morten, T. Zheng-Hua, and J. Jesper, “Permutation invariant training of deep models for speaker-independent multitalker speech separation,” ICASSP’ 17, pp. 31 - 35, 2017.

D. Yu, X. Chang, and Y. M. Qian, “Recognizing multi-talker speech with permutation invariant training,” INTERSPEECH’ 17, 2017.

M Kolbæk, D Yu, ZH Tan and J Jensen. “Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks”, ACM Transactions on Audio, Speech and Language Processing, 2018

- 基于关键词的单通道目标说话人语音分离



J Wang, J Chen, D Su, LW Chen, M Yu, D Yu “Deep Extractor Network for Target Speaker Recovery From Single Channel Speech Mixtures,” INTERSPEECH 2018

模型	单干扰说话人		多干扰说话人&中等干扰程度		多干扰说话人&严重干扰程度	
	信号失真比	主观语音质量评估	信号失真比	主观语音质量评估	信号失真比	主观语音质量评估
原始混合语音	0.99	1.96	0.44	1.87	-2.05	1.46
DANet Nearest	15.71	2.51	11.99	2.30	8.84 (26%)	2.13
DANet Oracle	16.67	2.58	11.98	2.27	8.85 (26%)	2.10
DANet Anchor	17.14	2.72	13.32	2.48	10.39 (22%)	2.12
深度提取网络	17.53	2.75	13.44	2.52	10.67 (20%)	2.14

- 性能最优。**从“单个干扰说话人”到“多干扰说话人”，从“中等”到“严重”干扰程度的全部测试条件下，信号失真比（SDR）和主观语音质量评估（PESQ）指标结果显示，我们提出的“深度提取网络”一致最优，且提升显著
 - 系统的鲁棒性亦最优：表中括号里面的百分比是从“中等干扰”到“严重干扰”的信号失真比的相对恶化程度，该数值越小反映系统鲁棒性越强
- 实时性：**系统可实时运行；模型参数复杂度279,797k，与参考系统的279,425k基本相当；
- 拓展性：**开集测试，能处理任意未知目标说话人；并且，上述所有结果都是用“单干扰说话人”训练语料训练的系统，直接应用到“多干扰说话人”场景；可以预见，加入匹配场景的训练语料后，系统性能还将获得进一步提升
- 可用性：**自适应语音时长 < 1s，可落地到绝大多数智能语音设备的真实应用场景
- 深度研究价值：**支持通过模拟人类听觉系统处理多说话人时的注意力选择机制来初始化锚点，本端到端系统研究方向是支持无自适应语音。

从Hybrid系统升级到端到端系统是近年语音识别研究热点，最前沿的序列到序列技术各有优劣点：

基于注意机制（Attention）的方法

优势：无需假设当前帧的输出和之前的输出标注独立；对下一单元的预测同时用到了声学模型和语言模型的信息

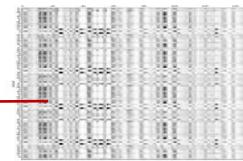
缺陷：Attention缺乏从左到右的对齐限制；为缩小与Hybrid系统的差距，一些前沿研究采取引入外部语言模型等技巧，系统越来越复杂，逐渐背离“端到端”的设计初衷

CTC方法

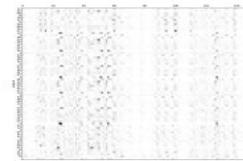
优势：从左到右的序列到序列模型，简单灵活，无需帧级别标注，解码快

缺陷：模型建立的前提是假设当前帧的输出和之前的输出标注独立

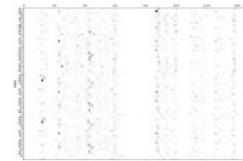
Attention Only



(a) Attention 1 epoch



(b) Attention 3 epoch



(c) Attention 5 epoch



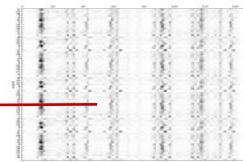
(d) Attention 7 epoch



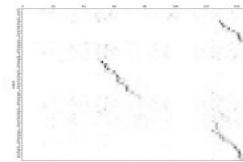
(e) Attention 9 epoch

错误对齐：堆集到尾部

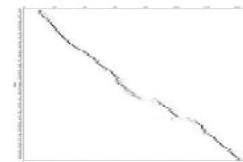
Attention + CTC



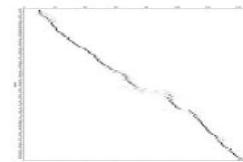
(f) MTL 1 epoch



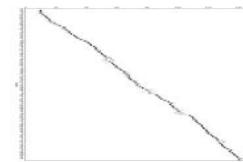
(g) MTL 3 epoch



(h) MTL 5 epoch



(i) MTL 7 epoch



(j) MTL 9 epoch

错误对齐：非左右次序

端到端语音识别的Attention建模方法的关键技术点

- 引入最小风险贝叶斯决策 (MBR) 损失，并结合交叉熵损失作为初始化：

$$L'_{MBR} = L_{MBR} + \lambda \sum_{u=1}^U \sum_{y_u} \gamma(y_u) L_{xent}(y_u, y'_u)$$

用于初始化的
交叉熵损失

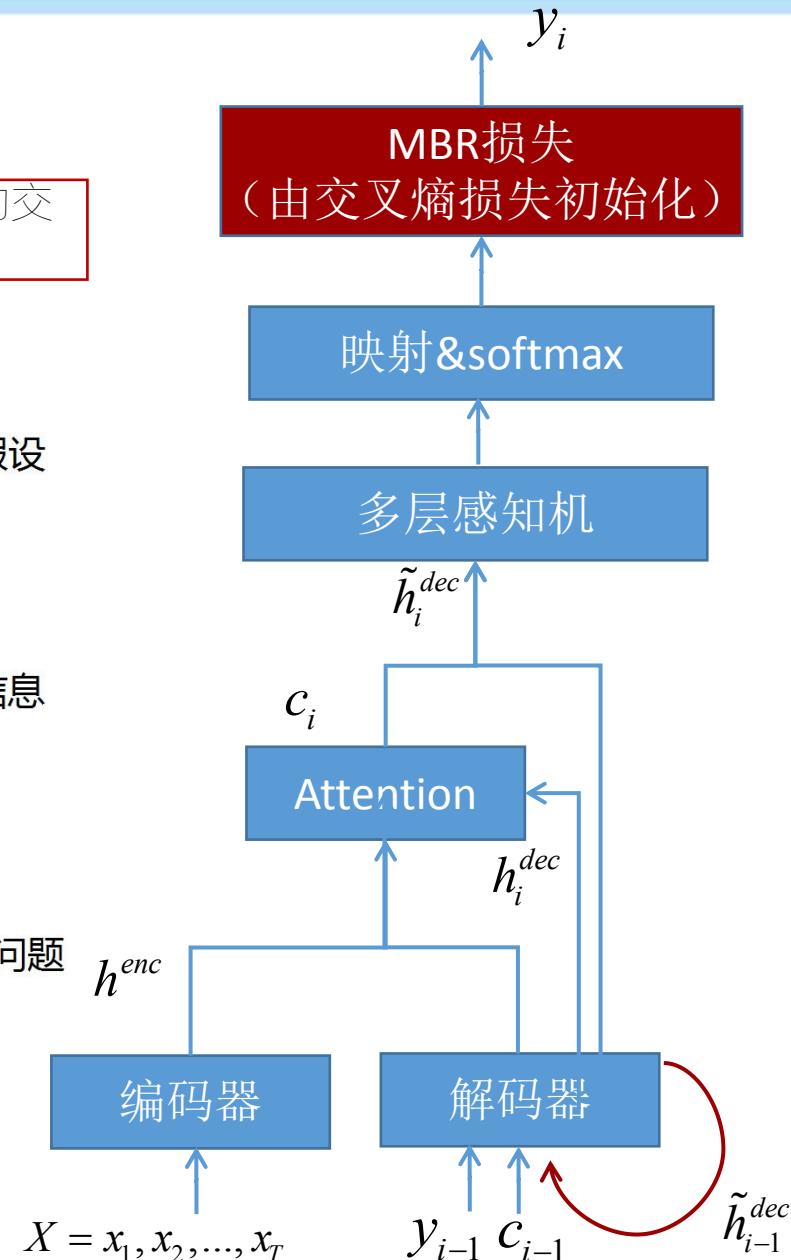
- 交叉熵作初始化：对于成功的MBR训练非常关键
- Softmax smoothing用于N-Best：能生成更好的序列到序列假设，解决了序列到序列生成假设过程中易于过强预测的问题

- 提出解码器反馈输入的结构：

- 原Attention：解码器输入仅包括前上下文向量 c_{i-1} ，该向量编码了Attention隐状态所有信息
- 改进Attention：1) 解码器输入增加了之前解码输出反馈回来的隐状态向量 \tilde{h}_{i-1}^{dec} ，该向量关注上一步中概率得分最高的标注对应的隐状态信息；
2) 采用teacher-forcing + 预定采样的方法，解决训练过程使用有监督的标注作输入而测试过程只有预测反馈作输入所导致的不匹配的问题

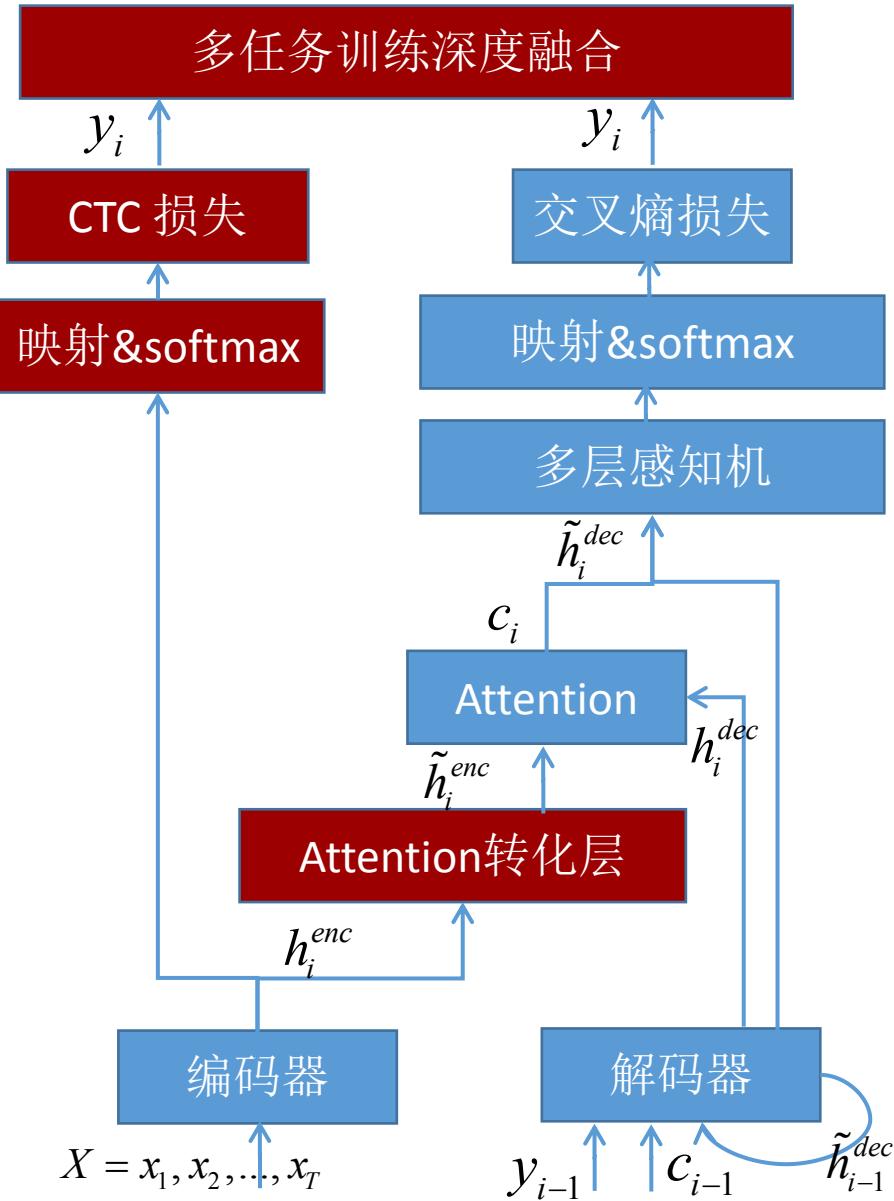
- 减轻序列到序列模型容易过拟合的问题：

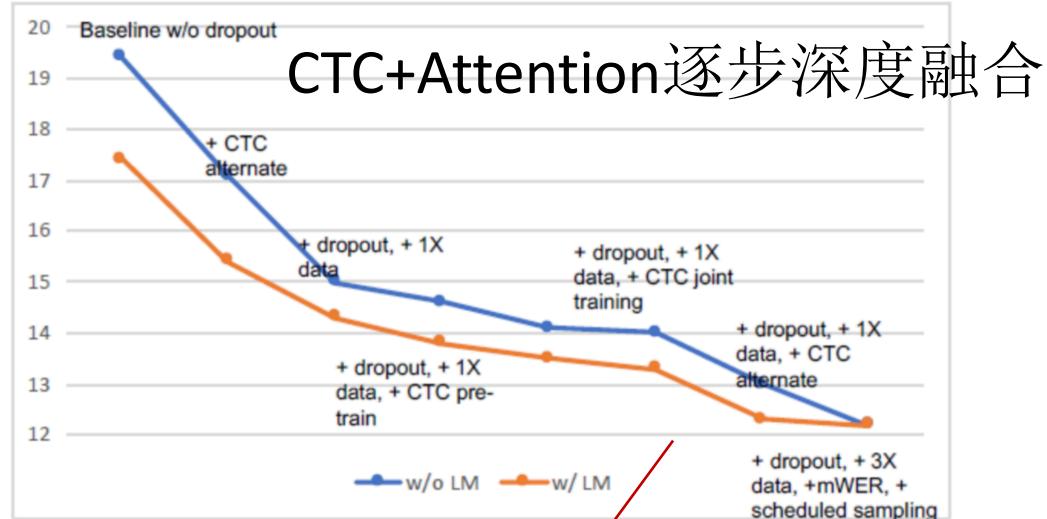
- 通过Speed permutation扩张训练数据 (3X data)
- 采用dropout并调整超参



端到端语音识别的Attention建模方法的关键技术点

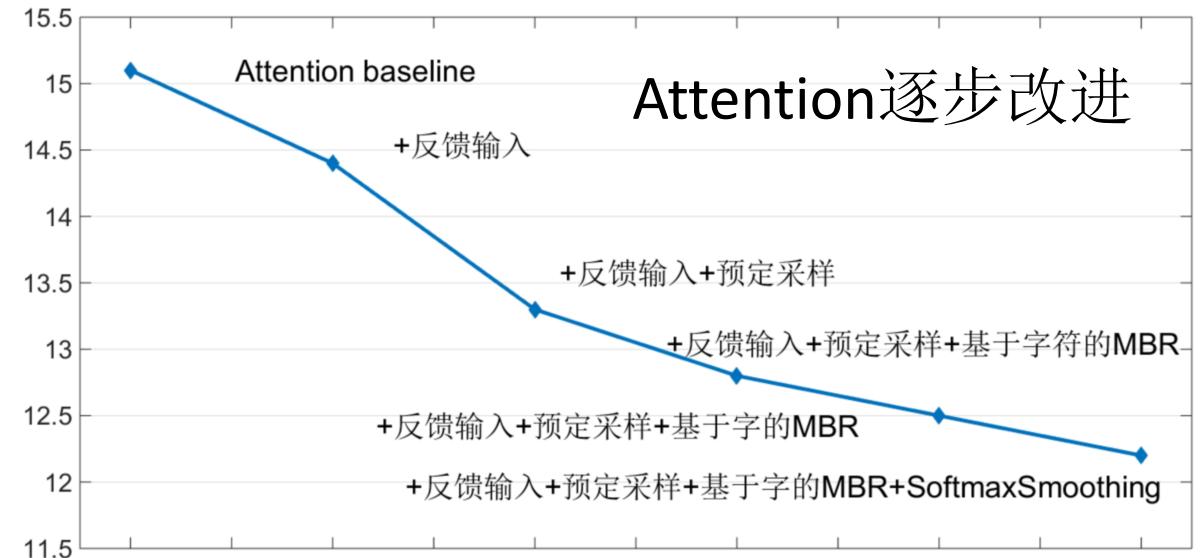
- 提出“Attention转化层”：直接融合的效果较差，我们分析本质原因是CTC架构和Attention的架构要求不同的隐层特性，故提出增加“Attention转化层”（具体实现中我们采用两层BLSTM）来实现它们之间的转化
- 研究用于深度融合Attention和CTC的多任务训练方法：
 - 预训练：用CTC训练初始化编码器参数，然后作Attention模型的训练
 - 联合训练：用CTC损失和Attention的交叉熵损失插值得到联合训练的损失
 - 替换训练：每个epoch中，先用CTC训练编码器参数，再用Attention交叉熵训练整个网络





随着深度融合方法的逐步优化，语言模型的作用变得越来越小，直至可忽略，满足“端到端”简洁之美。

在不使用外部语言模型的条件下，我们的系统达到了比其它使用外部模型的最新端到端系统显著低的字错误率。



系统 (注，相应参考文献见本页备注)	是否用外部 语言模型？	SwitchBoard 数据集	CallHome 数据集
Attention Seq2Seq + Trigram [1]	是	25.8	46.0
BRNN Grapheme CTC + Ngram [2]	是	20.0	31.8
BLSTM Phoneme CTC + Fisher LM [3]	是	14.8	NA
Acoustic-to-Word + noLM [4]	否	14.5	25.1
Iterated CTC + RNN WLM [5]	是	14.0	25.3
我们的CTC + Attention 深度融合方法	是	12.3	23.3
我们改进的Attention方法	否	12.2	17.8

智能音箱语音交互技术链条相关的研究



	会议/期刊名称	论文题目	姓名
1 前端	Symmetry	An Improved Set-membership Proportionate Adaptive Algorithm For A Block-sparse System	Zhan Jin , Yingsong Li , And Jianming Liu
	Interspeech 2018	Text-dependent Speech Enhancement For Small-footprint Robust Keyword Detection	Meng Yu, Xuan Ji, Yi Gao, Lianwu Chen, Jie Chen, Jimeng Zheng, Dan Su, Dong Yu
2 声纹	Interspeech 2018	Deep Discriminative Embeddings For Duration Robust Speaker Verification	Na Li, Deyi Tuo, Dan Su, Zhifeng Li, And Dong Yu
	Interspeech 2018	Deep Extractor Network For Target Speaker Recovery From Single Channel Speech Mixtures	Jun Wang, Jie Chen, Dan Su, Lianwu Chen, Meng Yu, Dong Yu
3 ASR	Interspeech 2018	Permutation Invariant Training Of Generative Adversarial Network For Monaural Speech Separation	Lianwu Chen, Meng Yu, Dan Su, Dong Yu
	ICASSP 2018	Adaptive Permutation Invariant Training With Auxiliary Information For Monaural Multi-talker Speech Recognition	Xuankai Chang, Yanmin Qian, Dong Yu
	ICASSP 2018	Knowledge Transfer In Permutation Invariant Training For Single-channel Multi-talker Speech Recognition	Tian Tan, Yanmin Qian, Dong Yu
	FITEE 2018	Past Review, Current Progress, And Challenges Ahead On The Cocktail Party Problem	Yan-min Qian, Chaoweng, Xuan-kai Chang, Shuai Wang, Dong Yu
	Interspeech 2018	Monaural Multi-Talker Speech Recognition with Attention Mechanism and Gated Convolutional Networks	Xuankai Chang, Yanmin Qian, Dong Yu
	Interspeech 2018	Improving Attention Based Sequence-to-sequence Models For End-to-end English Conversational Speech Recognition	Chao Weng, Jia Cui, Guangsen Wang, Jun Wang, Chengzhu Yu, Dan Su, Dong Yu
	Interspeech 2018	A Multistage Training Framework For Acoustic-to-Word Model	Chengzhu Yu, Chunlei Zhang, Chao Weng, Jia Cui, Dong Yu
	ICASSP 2018	Neural Network Language Modeling With Letter-based Features And Importance Sampling	Hainan Xu , Ke Li , Yiming Wang , Jian Wang , Shiycin Kang , Xie Chen, Daniel Povey , Sanjeev Khudanpur
4 NLU	ACL 2018 5偏、IJCAI 2018 4偏 (ai lab 共11偏) 、NAACL 4偏		
5 合成	ICASSP 2018	Feature Based Adaptation For Speaking Style Synthesis	Xixin Wu, Lifa Sun, Shiycin Kang, Songxiang Liu, Zhiyong Wu, Xunying Liu, Helen Meng
	Interspeech 2018	Rapid Style Adaptation Using Residual Error Embedding For Expressive Speech Synthesis	Xixin Wu, Yuewen Cao, Mu Wang, Songxiang Liu, Shiycin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, Helen Meng

Thank You



Tencent
AI Lab