深蓝学院
shenlanxueyuan.com

语音信号处理一些Topics

宋辉 & 李先刚

语音信号处理一些Topics

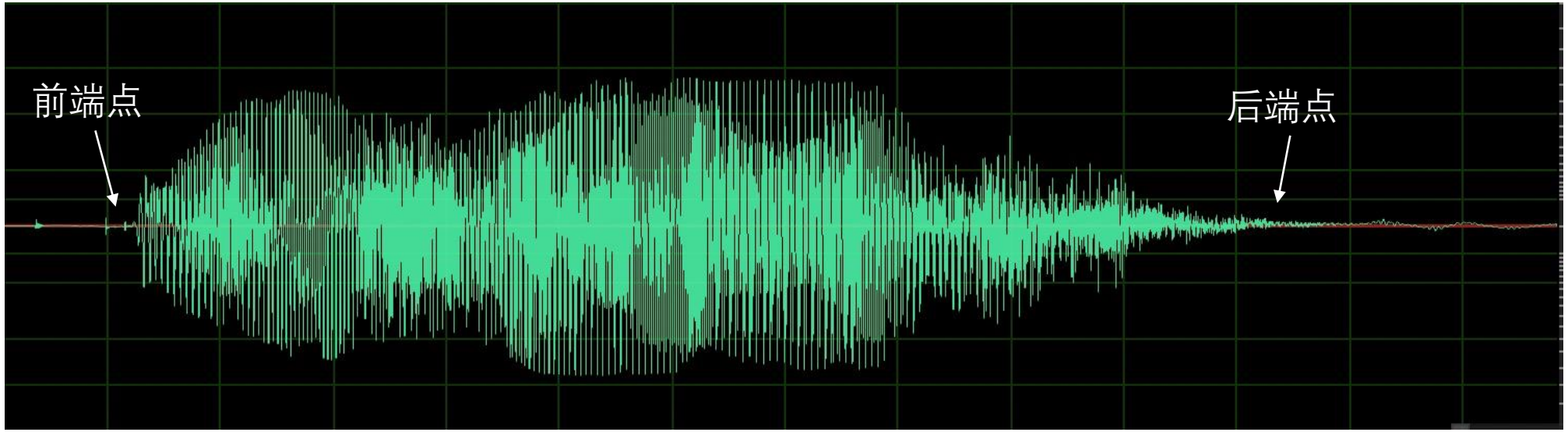
深蓝学院
shenlanxueyuan.com

- VAD
- 目标说话人语音分离

1 Voice Activity Detector (VAD)

深蓝学院 shenlanxueyuan.com

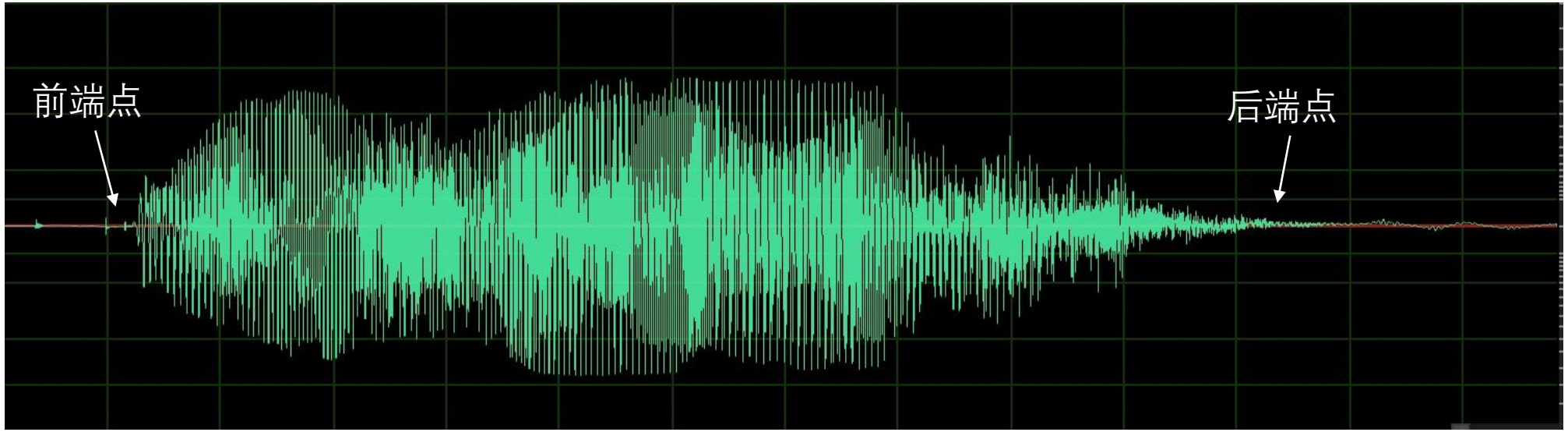
- VAD (端点检测)
  - 从语音信号中将语音（Speech）和非语音（Nonspeech）区分开，确定语音信号的端点，包括前端点和后端点



Voice Activity Detector (VAD)

深蓝学院 shenlanxueyuan.com

- 为什么VAD很重要
  - 太灵敏 or 太迟钝



能量VAD

深蓝学院 shenlanxueyuan.com

- 基于特征的方法
  - 采用能对语音和非语音（噪声）具有区分度的特征判断
  - 常用特征：能量，过零率，基频等

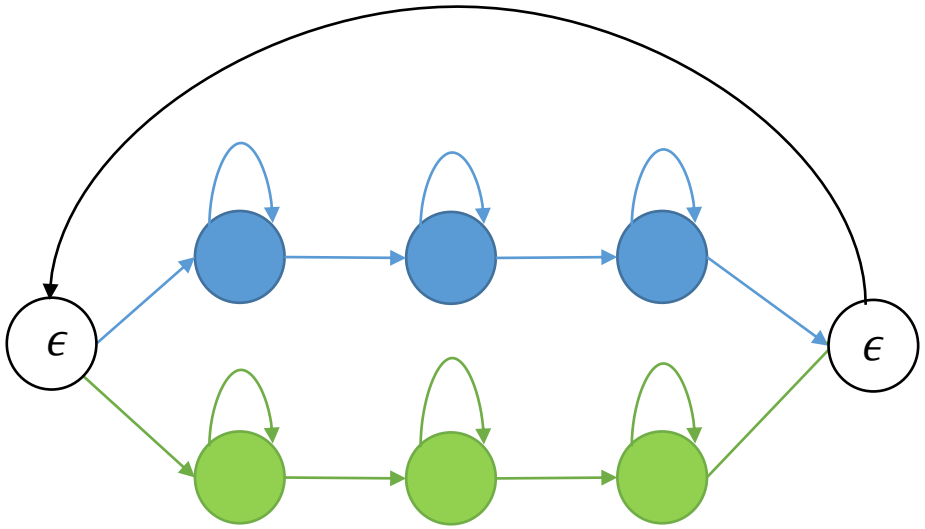
```
namespace kaldi {

void ComputeVadEnergy(const VadEnergyOptions &opts,
                      const MatrixBase<BaseFloat> &feats,
                      Vector<BaseFloat> *output_voiced) {
  int32 T = feats.NumRows();
  output_voiced->Resize(T);
  if (T == 0) {
    KALDI_WARN << "Empty features";
    return;
  }
  Vector<BaseFloat> log_energy(T);
  log_energy.CopyColFromMat(feats, 0); // column zero is log-energy.

  BaseFloat energy_threshold = opts.vad_energy_threshold;
  if (opts.vad_energy_mean_scale != 0.0) {
    KALDI_ASSERT(opts.vad_energy_mean_scale > 0.0);
    energy_threshold += opts.vad_energy_mean_scale * log_energy.Sum() / T;
  }
  KALDI_ASSERT(opts.vad_frames_context >= 0);
  KALDI_ASSERT(opts.vad_proportion_threshold > 0.0 &&
               opts.vad_proportion_threshold < 1.0);
  for (int32 t = 0; t < T; t++) {
    const BaseFloat *log_energy_data = log_energy.Data();
    int32 num_count = 0, den_count = 0, context = opts.vad_frames_context;
    for (int32 t2 = t - context; t2 <= t + context; t2++) {
      if (t2 >= 0 && t2 < T) {
        den_count++;
        if (log_energy_data[t2] > energy_threshold)
          num_count++;
      }
    }
    if (num_count >= den_count * opts.vad_proportion_threshold)
      (*output_voiced)(t) = 1.0;
    else
      (*output_voiced)(t) = 0.0;
  }
}
```

基于HMM的VAD

- 将VAD看做是一个特殊的语音识别任务
  - 其发音词典（或者声学模型）只有Silence和Speech
- 训练方法
  - 1）基于语音识别的Alignment得到每一帧特征对应的声学单元
  - 2）将非silence部分统一设置为speech
  - 3）基于EM算法训练GMM 或者DNN
- 一些小优化
  - 1）将Speech部分采用更多声学单元建模
- 基于DNN的VAD的两种框架
  - HMM框架、得分加窗平滑的方案



VAD的一些改进

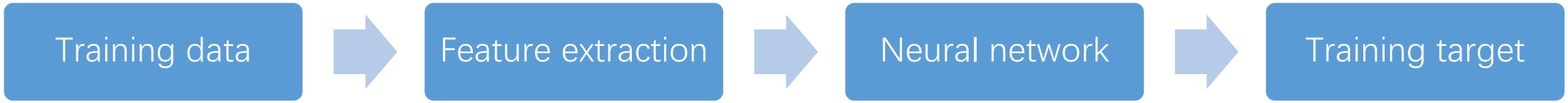
深蓝学院 shenlanxueyuan.com

- VAD与语音识别过程结合
  - 在语音识别声学建模过程中，将后端点（endpoint）的检测和声学模型一起联合建模
  - S. Chang, R. Prabhavalkar, Y. He, T. Sainath. Joint Endpointing and Decoding with End-to-End Models. ICASSP 2019
- VAD与语义理解结合
  - 基于文字内容判断一段语音识别说完整
  - 结合语音识别结果以及声学信号，训练分类模型
- 更小的模型
  - Binary Neural Network …

2 目标说话人语音分离

深蓝学院 shenlanxueyuan.com

- DNN based speech separation: typical solution



- Speaker separation
  - Speaker dependent: the underlying speakers are not allowed to change from training to testing
  - Target speaker dependent: interfering speakers are allowed to change, but the target speaker is fixed
  - Speaker independent: none of the speakers are required to be the same between the training and testing

鸡尾酒会问题

深蓝学院 shenlanxueyuan.com

- 鸡尾酒会问题
  - 多个说话人同时说话
  - 只希望听其中一个说话人
- 解决方案
  - 多通道盲源分离 Multi-channel blind separation
  - 单通道盲源分离 Single-channel blind separation
- 单通道语音分离算法
  - Deep clustering, Deep attractor network, Permutation invariant training

盲源分离 -> 目标说话人语音分离

深蓝学院 shenlanxueyuan.com

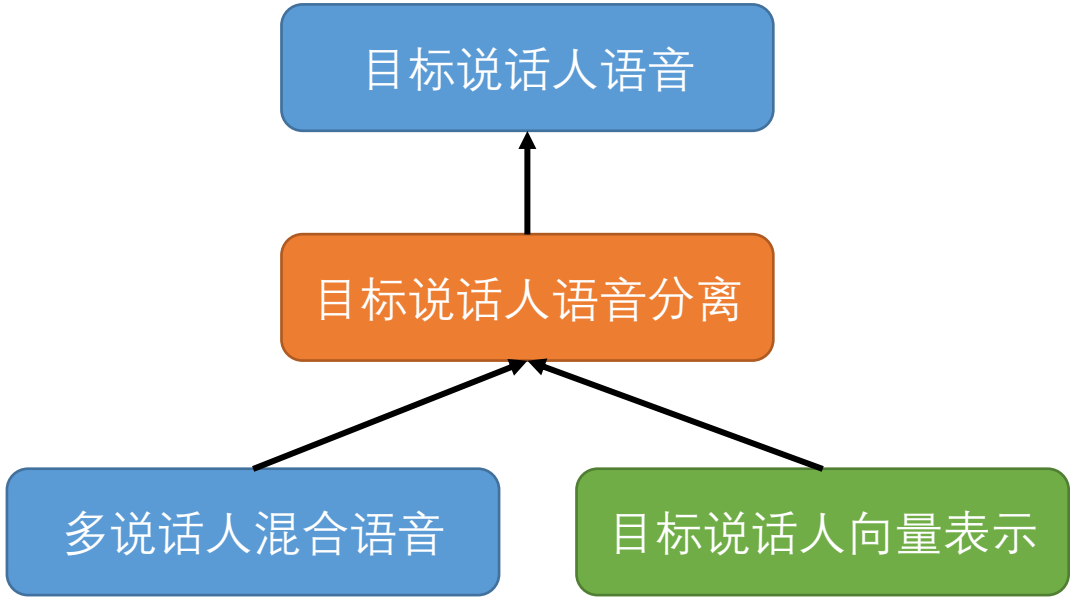
- 盲源分离面临的挑战
  - 声源个数不确定
  - 分离后多个输出，仍不知道目标用户所对应的语音
  - 计算代价较高（PIT）
- 在很多现实应用场景中，知道“whom to listen to”

将目标说话人的声纹信息作为输入

深蓝学院 shenlanxueyuan.com

- 说话人识别
  - 输入：语音特征
  - 输出：说话人向量表示（Speaker Embedding）

- 目标说话人语音分离
  - 输入1：语音特征
  - 输入2：说话人向量表示
  - 输出：mask/spectrum…

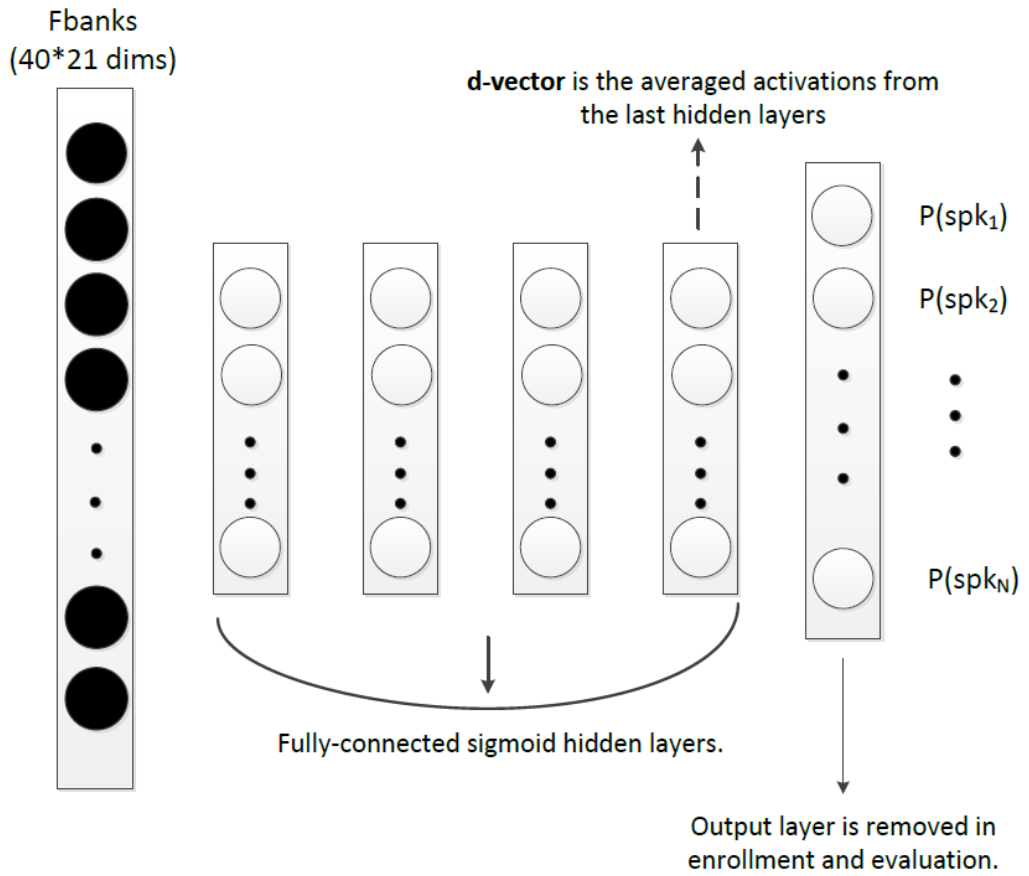


将目标说话人的声纹信息作为输入

深蓝学院 shenlanxueyuan.com

- 说话人识别系统d-vector

1. Ehsan Variani et al. “Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification”. ICASSP. 2014.
2. Lantian Li et al. “Deep speaker vectors for semi text-independent speaker verification”. arXiv: 1505.06427(2015).
3. Yuan Liu et al. “Deep feature for text-dependent speaker verification”. Speech Communication 2015.

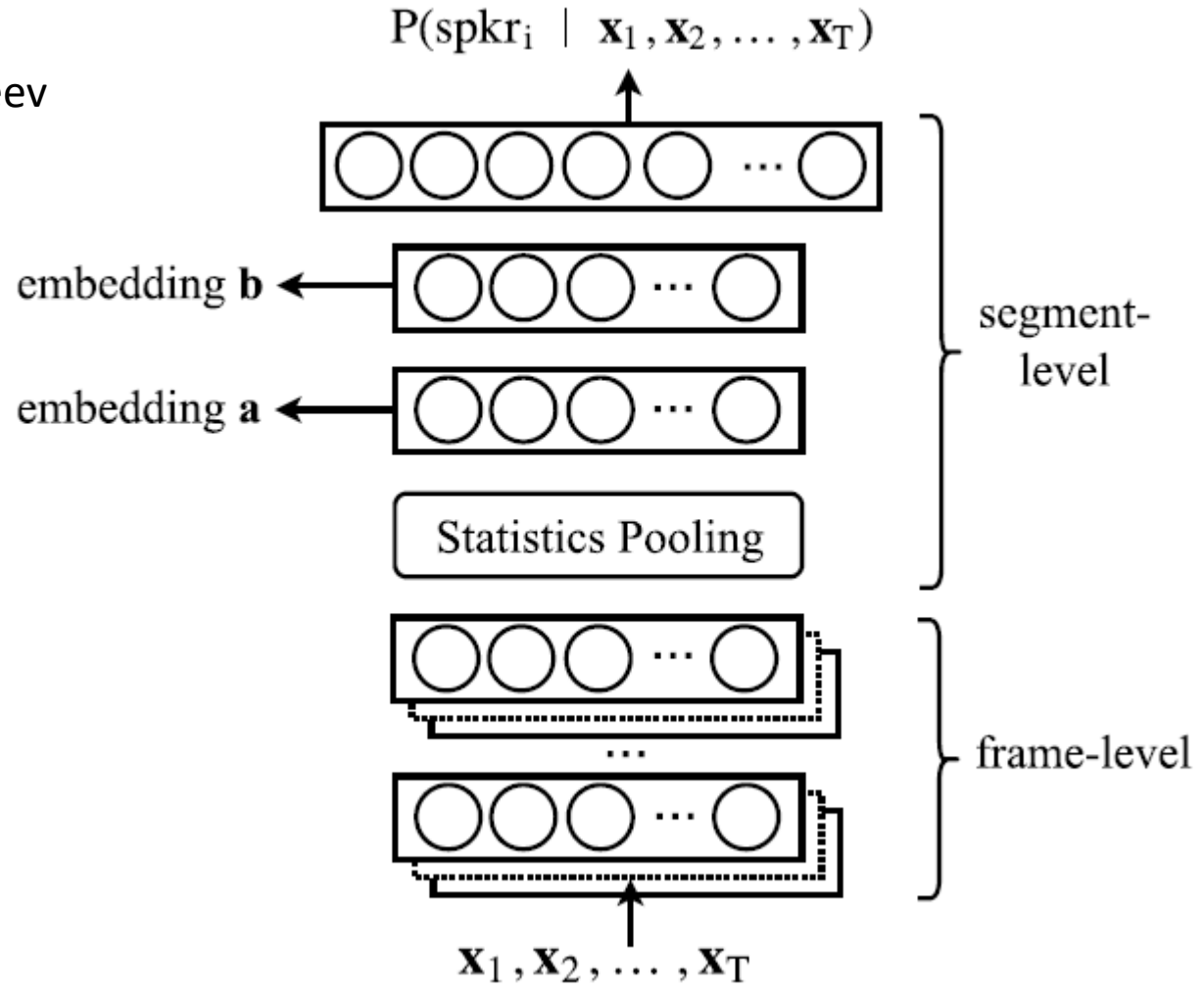


将目标说话人的声纹信息作为输入

深蓝学院 shenlanxueyuan.com

- 说话人识别系统x-vector

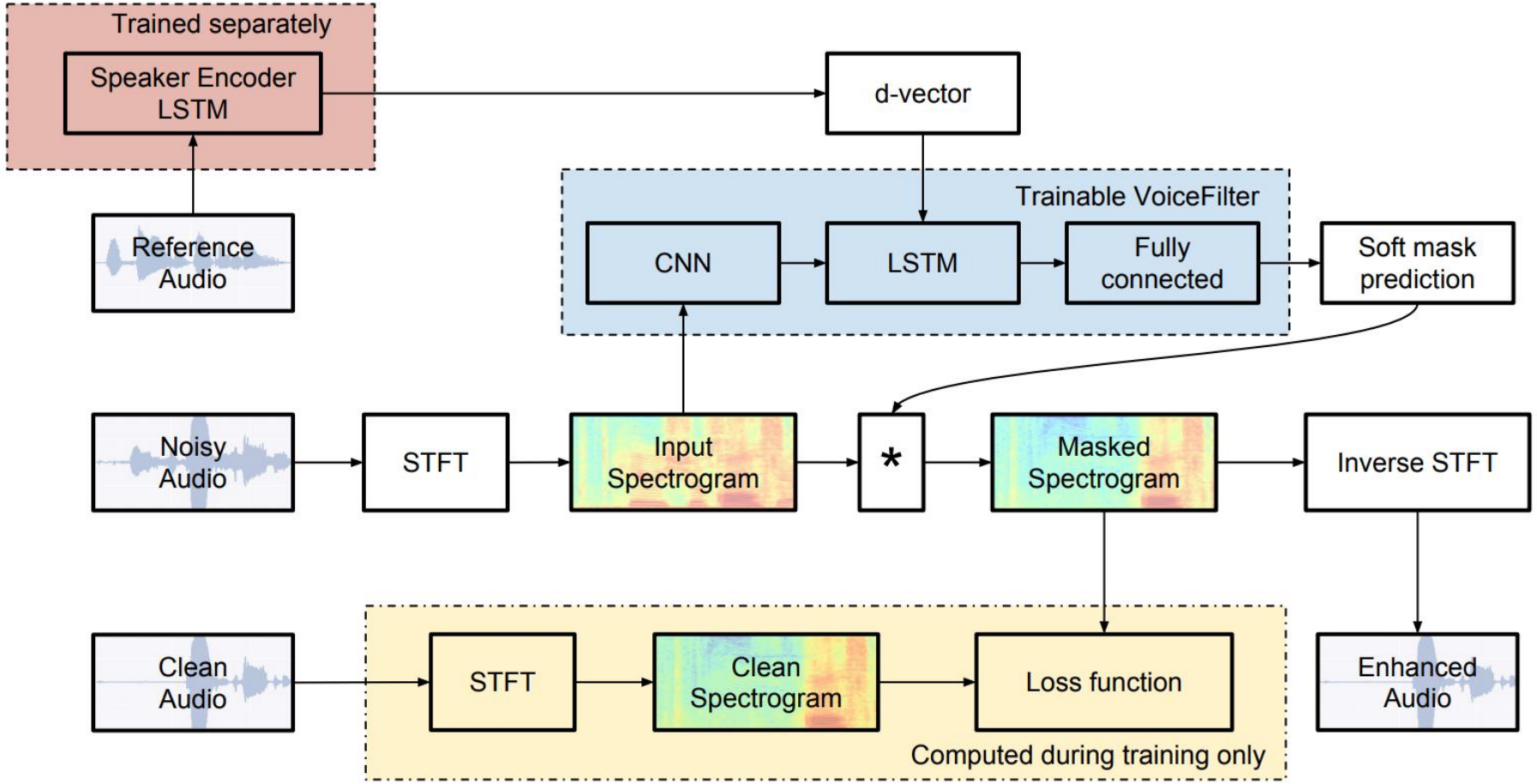
1. David Snyder, Daniel Garcia-Romero, Daniel Povey and Sanjeev Khudanpur. “Deep Neural Network Embeddings for Text-Independent Speaker Verification”. Interspeech 2017.
2. David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur. “X-vectors: Robust DNN Embeddings for Speaker Recognition”. ICASSP 2018.
3. Yingke Zhu, Tom Ko, David Snyder, Brian Mak, Daniel Povey. “Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification”. Interspeech 2018.



目标说话人语音分离

深蓝学院 shenlanxueyuan.com

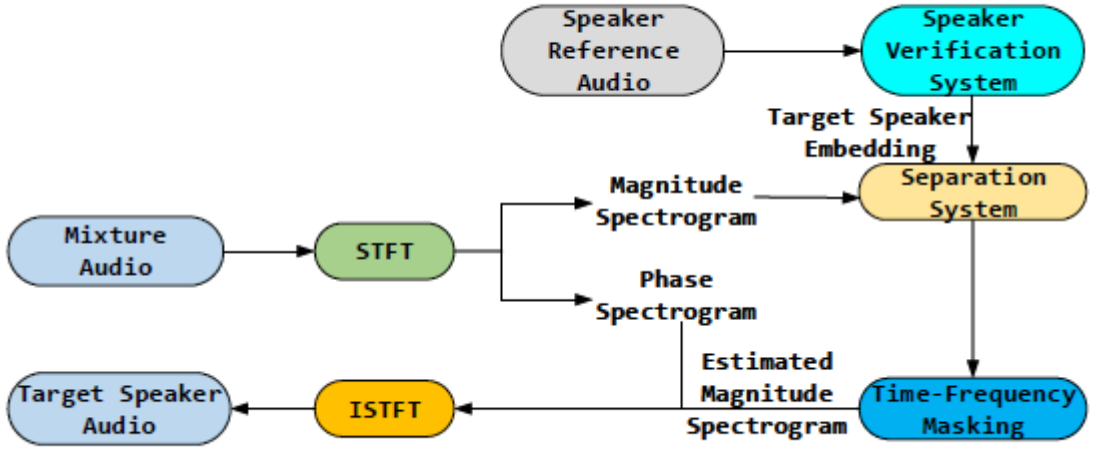
- VoiceFilter
  - Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, et. al. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. Interspeech, 2019

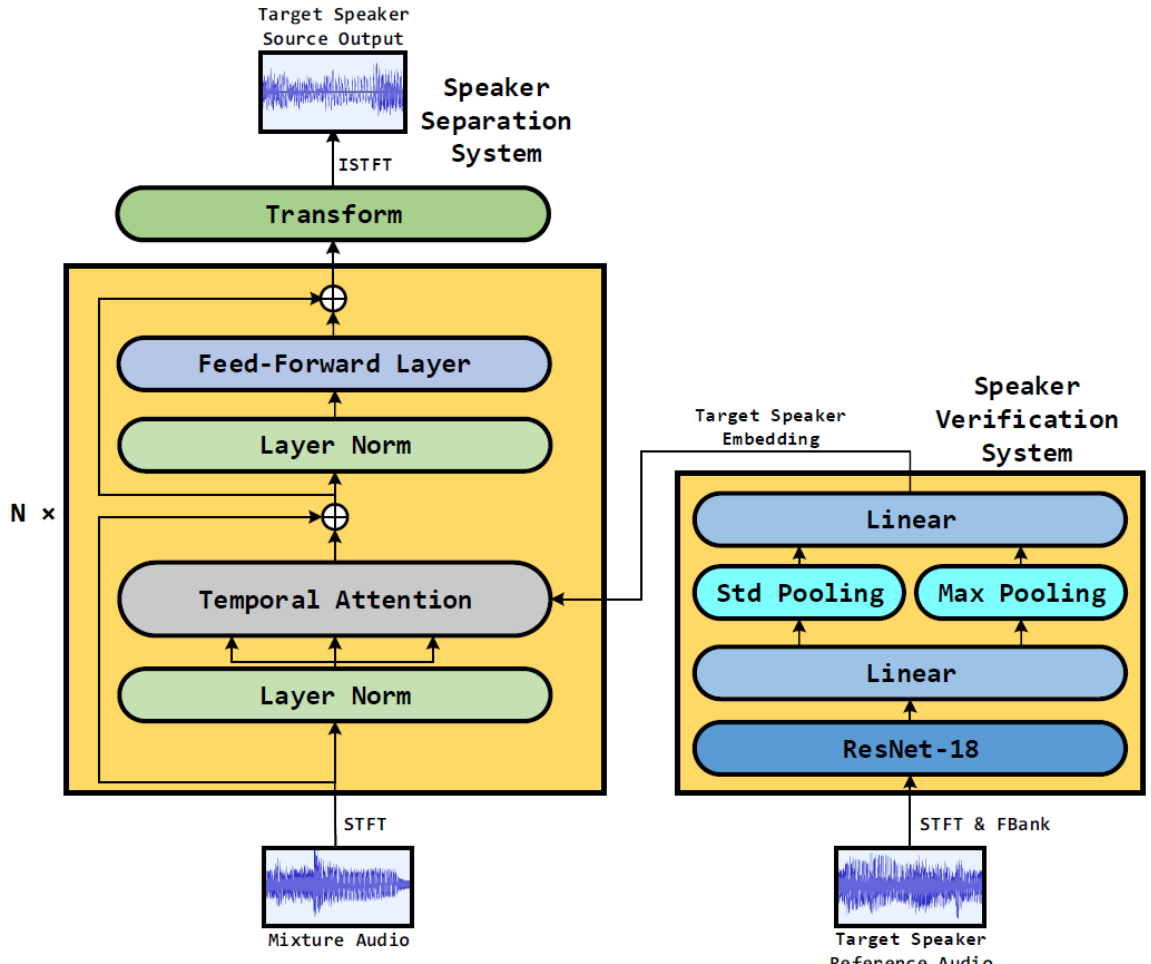


目标说话人语音分离

深蓝学院 shenlanxueyuan.com

- Atss-Net
  - T. Li, Q. Lin, Y. Bao, M. Li, “Atss-Net: Target Speaker Separation via Attention-based Neural Network”, arXiv:2005.09200





目标说话人语音分离：一些应用举例

深蓝学院 shenlanxueyuan.com

- 嘈杂场景语音交互应用
- 替代AEC
- 车载场景录音分析：分离出司机的声音