# 声纹识别开源工具ASV-Subtools

洪青阳

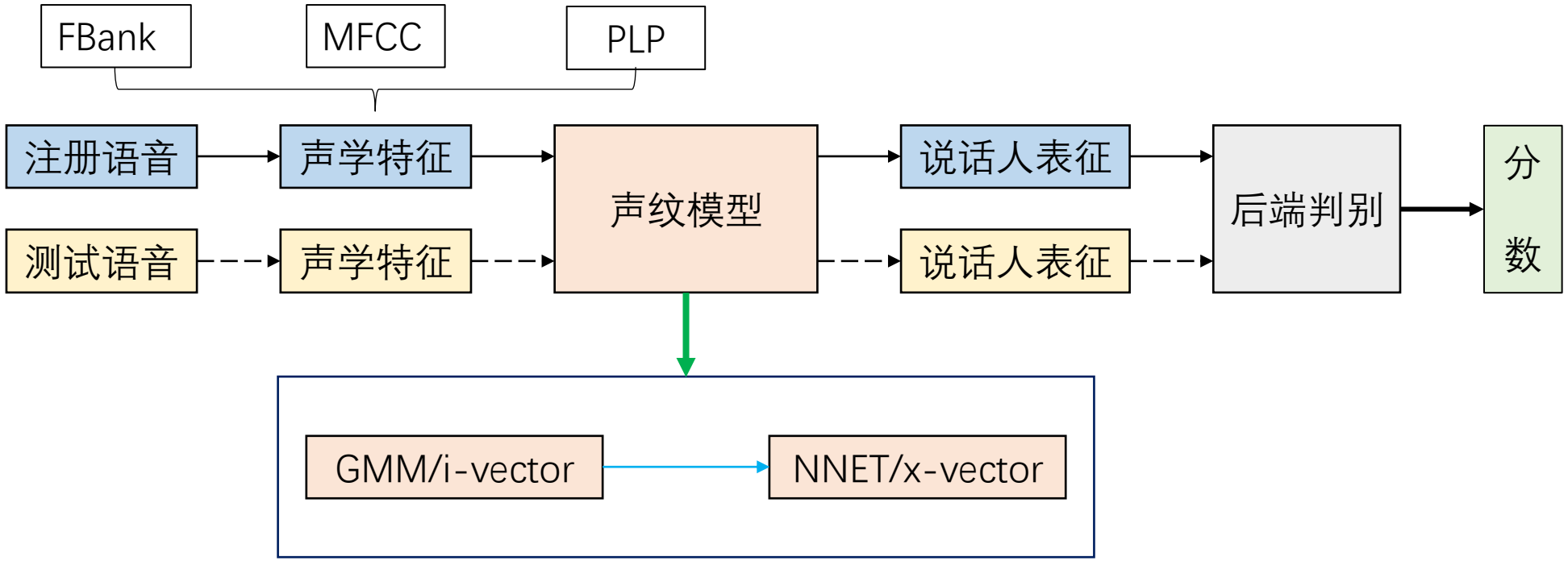
厦门大学智能语音实验室
http://speech.xmu.edu.cn

2020.11

# 纲 要

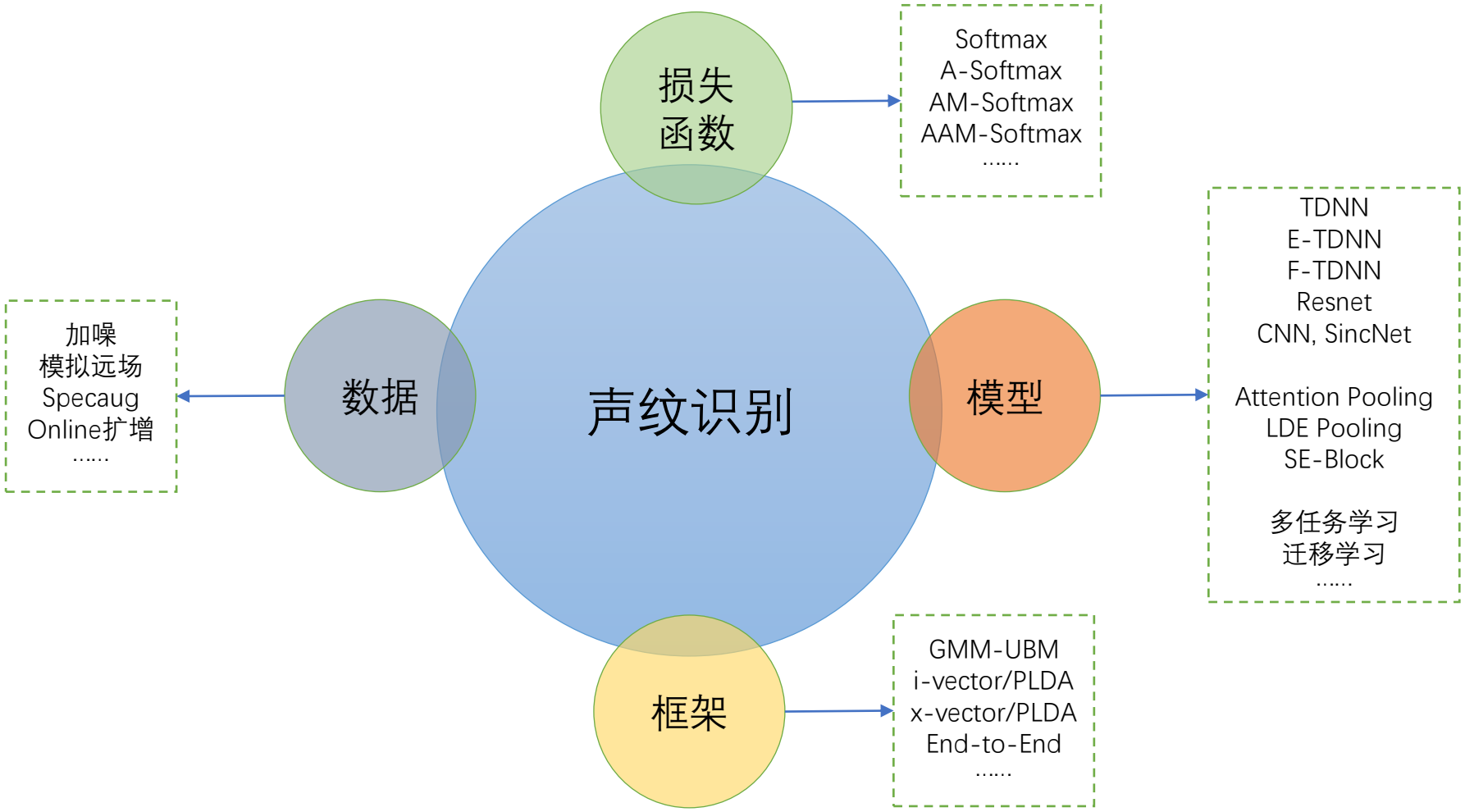
1. 背景介绍
2. 开源工具ASV-Subtools
3. 算法改进
4. 实验结果
5. 总结与展望

# 1. 背景介绍—声纹识别框架




**声纹识别框架图**

# 1. 背景介绍—声纹识别技术



# 1. 背景介绍—研究工具

**研究工具**

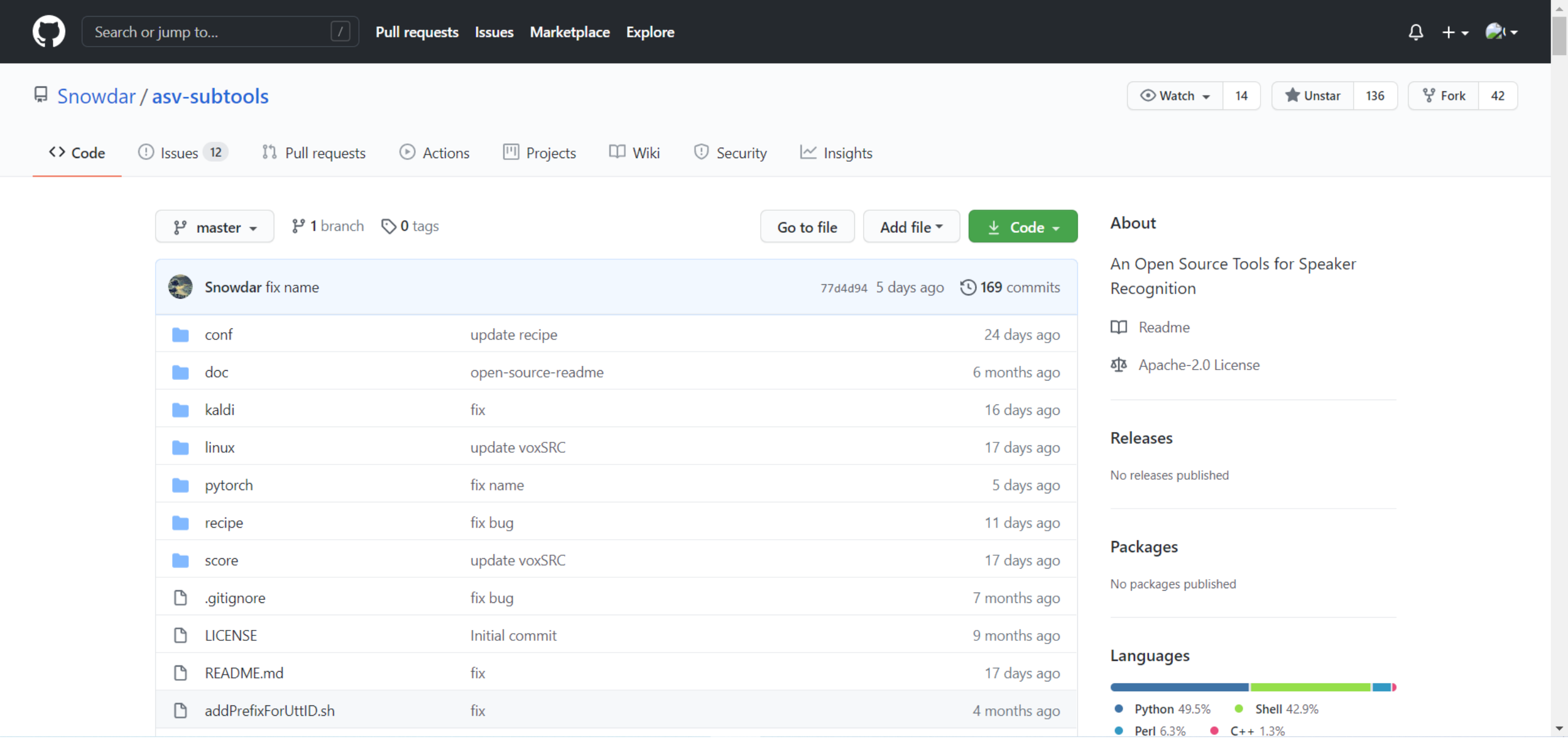
- Kaldi为语音领域主流研究工具，基于C++开发，但底层扩展开发困难，研究效率低。
- Pytorch/Tensorflow已成为深度学习热门工具，提供Python接口，使用灵活，研究效率高。
- 新算法层出不穷，需要快速验证，但很多基线系统性能不佳。

→ 基于Pytorch自主开发一套声纹识别工具

# 开源工具ASV-Subtools

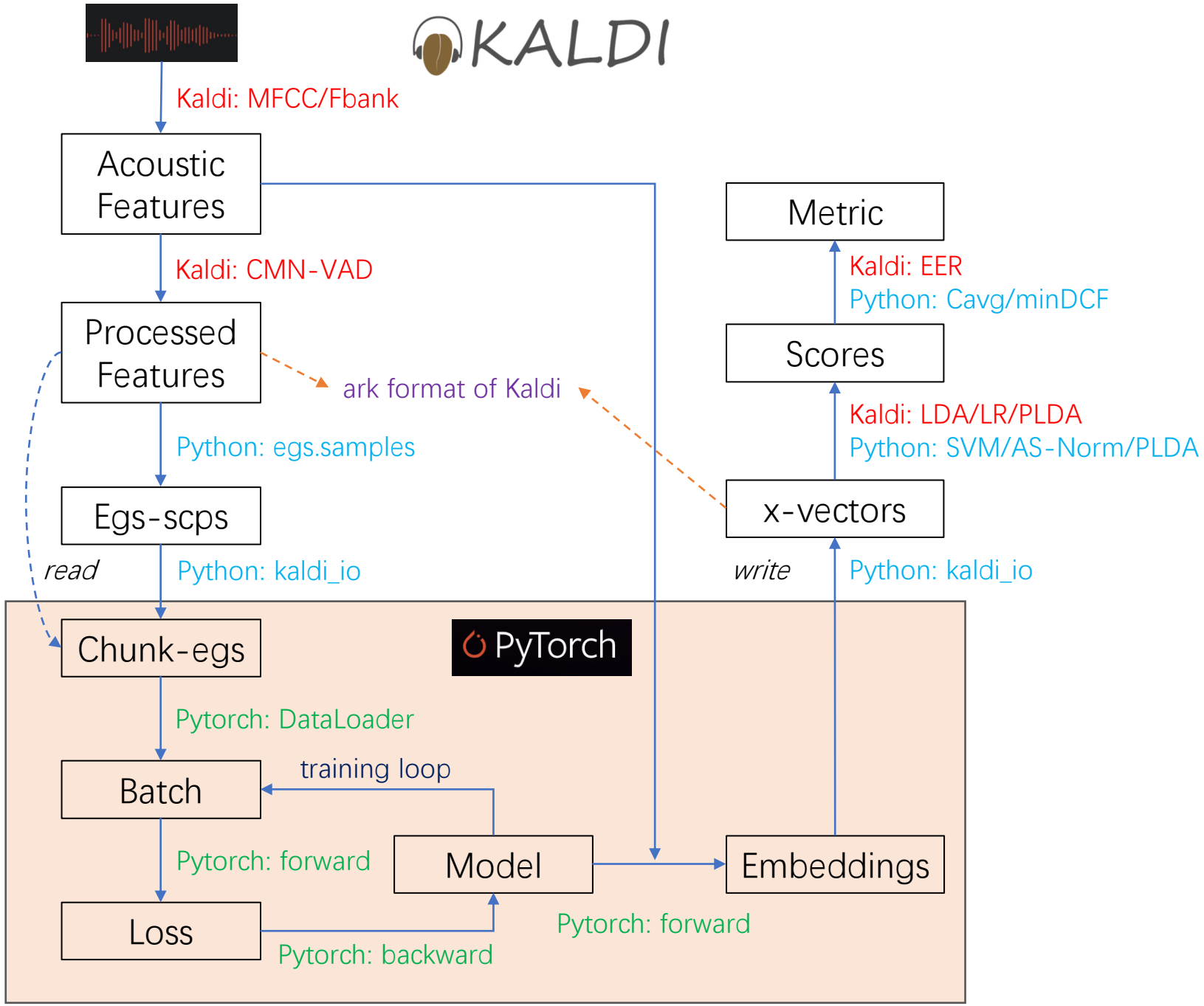
2020年5月 正式发布

https://github.com/Snowdar/asv-subtools

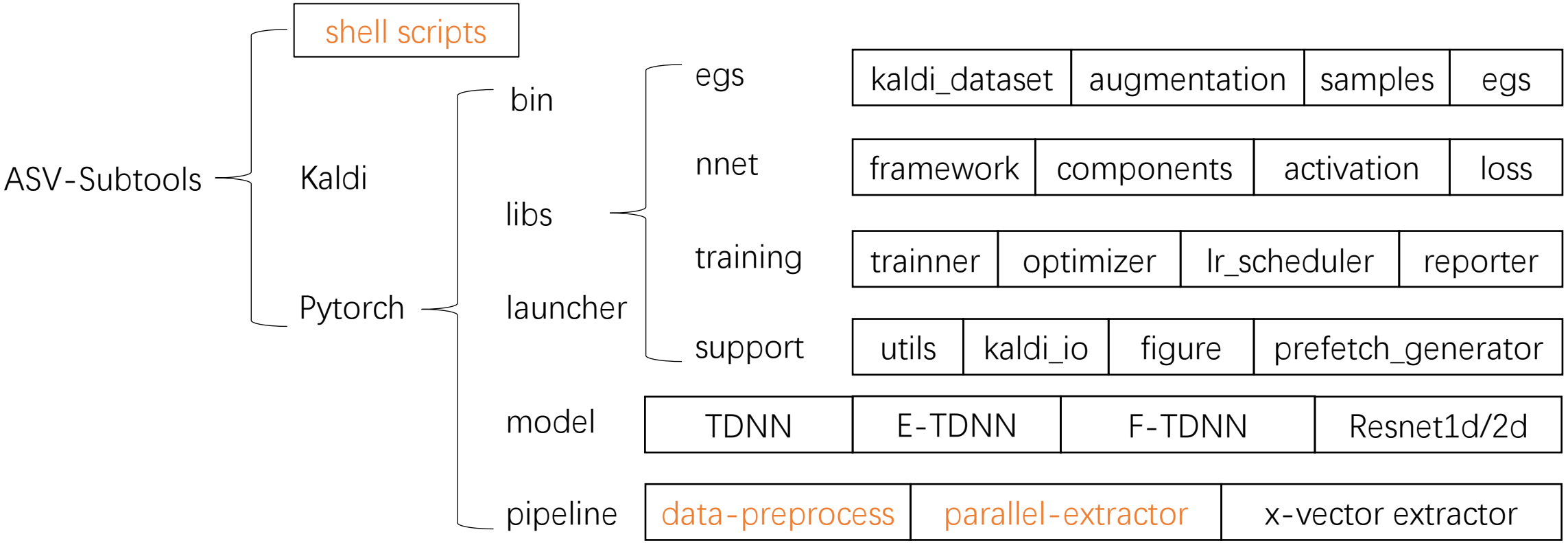



Author: Miao Zhao, Jianfeng Zhou, Zheng Li, Hao Lu
Co-author: Lin Li, Qingyang Hong

# 工程结构

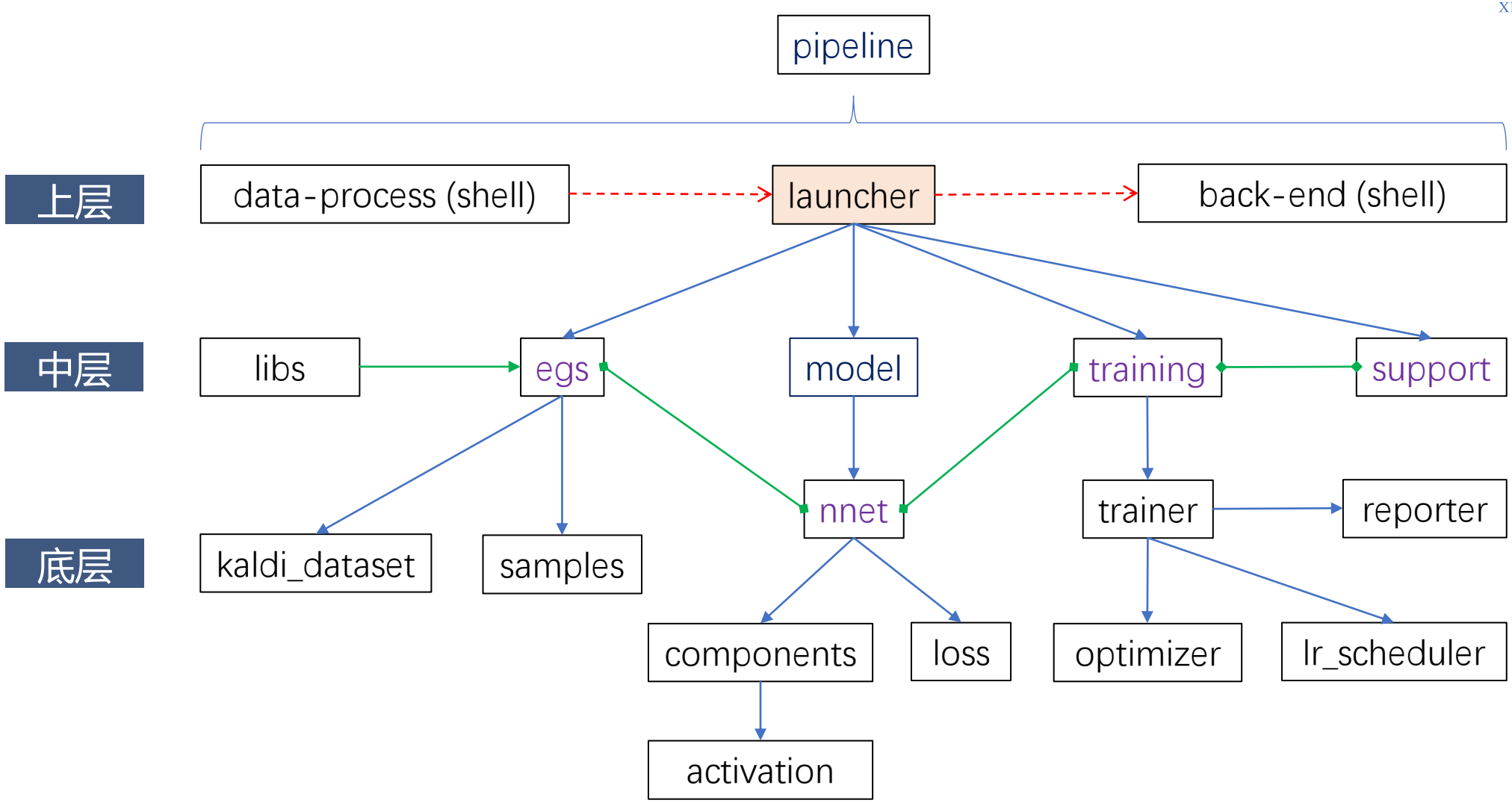

Kaldi: MFCC/Fbank
Acoustic Features
Kaldi: CMN-VAD
Processed Features
ark format of Kaldi
Python: egs.samples
Egs-scps
read
Python: kaldi_io
Chunk-egs
PyTorch
Pytorch: DataLoader
Batch
training loop
Pytorch: forward
Model
Loss
Pytorch: backward
Embeddings
Pytorch: forward
write
Python: kaldi_io
x-vectors
Kaldi: LDA/LR/PLDA
Python: SVM/AS-Norm/PLDA
Scores
Kaldi: EER
Python: Cavg/minDCF
Metric

# 工程结构



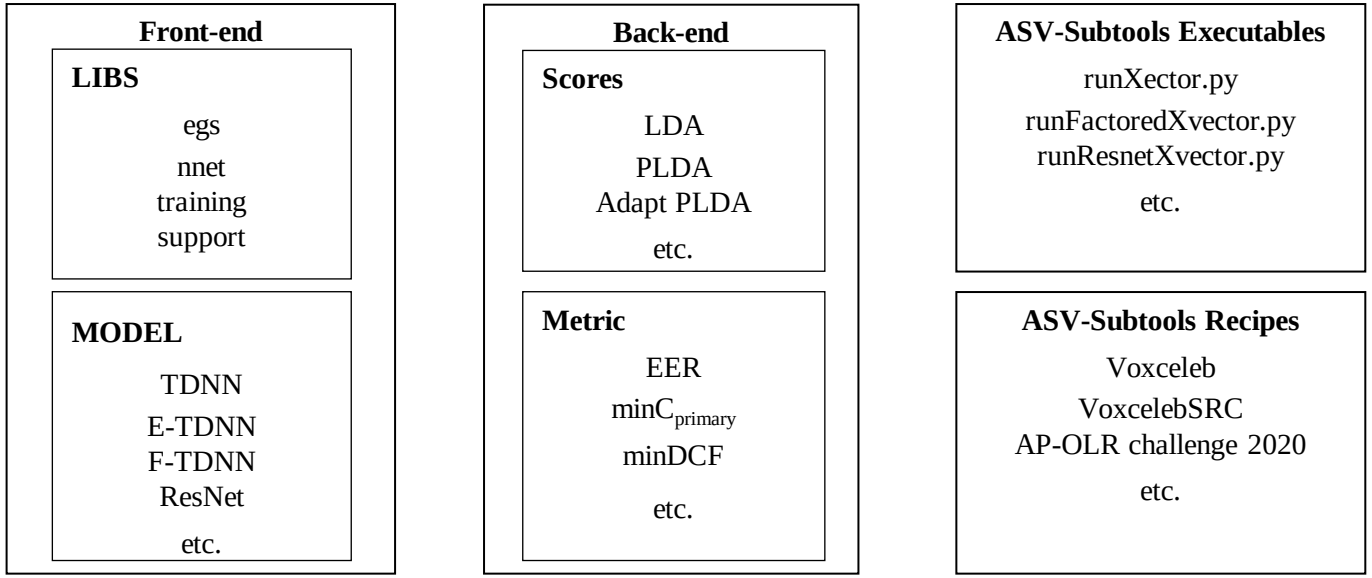

**ASV-Subtools工程结构**

# Pytorch训练架构




**Pytorch训练框架**

# ASV-Subtools系统组件



# 3. 算法改进

# 数据扩增：Inverted Specaugment

| **Specaugment Algorithm：** |
| --- |
| Input feature matrix $M$, drop probability $p_f$,$p_t$, multi-mask times $n_f$,$n_t$ |
| $M$ <- TimeWarping($M$) |
| Get total $Num_f$, $Num_t$ from $M$ |
| for i in 1,2,3,…,$n_f$: |
| $f$ <- Uniform(0, $Num_f * p_f$), $f_0$ <- Uniform(0, $Num_f - f$) |
| $M$ <- Drop **frequency** dimension of $M$ from $f_0$ to $f_0 + f$ |
| for i in 1,2,3,…,$n_t$: |
| $t$ <- Uniform(0, $Num_t * p_t$), $t_0$ <- Uniform(0, $Num_t - t$) |
| $M$ <- Drop **time** dimension of $M$ from $t_0$ to $t_0 + t$ |
| Return $M$ |

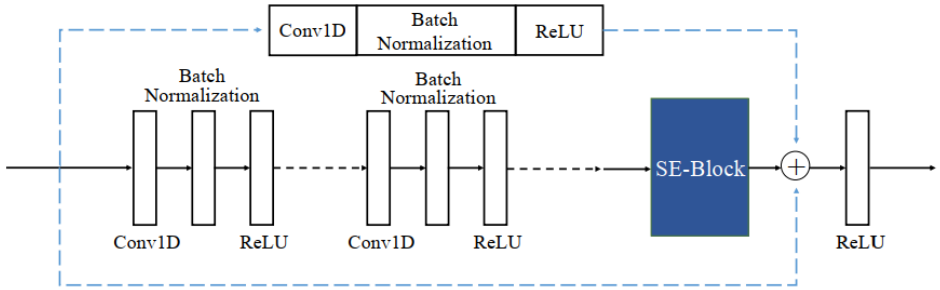
| **Inverted Specaugment Algorithm：** |
| --- |
| Input feature matrix $M$, drop probability $p_f$,$p_t$, multi-mask times $n_f$,$n_t$ |
| Get total $Num_f$, $Num_t$ from $M$ |
| $r_f$ <- Uniform(1, $n_f$) |
| for i in 1,2,3,…,$r_f$: |
| $f$ <- Uniform(0, $Num_f * p_f$), $f_0$ <- Uniform(0, $Num_f - f$) |
| $M$ <- Drop **frequency** dimension of $M$ from $f_0$ to $f_0 + f$ |
| $M$ <- $M * Num_f/(Num_f - f)$ |
| $r_t$ <- Uniform(1, $n_t$) |
| for i in 1,2,3,…,$r_t$: |
| $t$ <- Uniform(0, $Num_t * p_t$), $t_0$ <- Uniform(0, $Num_t - t$) |
| $M$ <- Drop **time** dimension of $M$ from $t_0$ to $t_0 + t$ |
| Return $M$ |

- 时间扭曲(TimeWarping)，避免破坏基频信息。

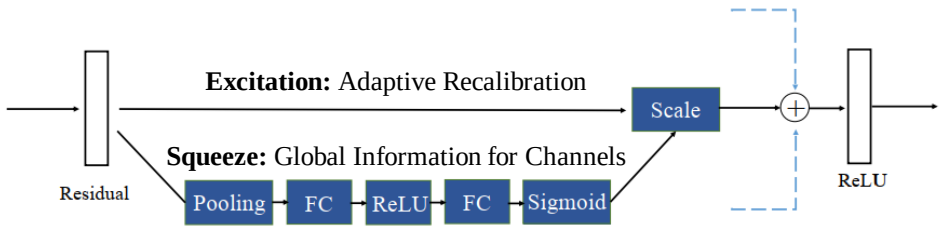
+ 频率均值修正（绿色）

+ 随机多次丢弃（蓝色）

# 通道/维度增强模块(SE-Block)




(a) ResNet incorporated SE-Block




(b) Details of SE-Block

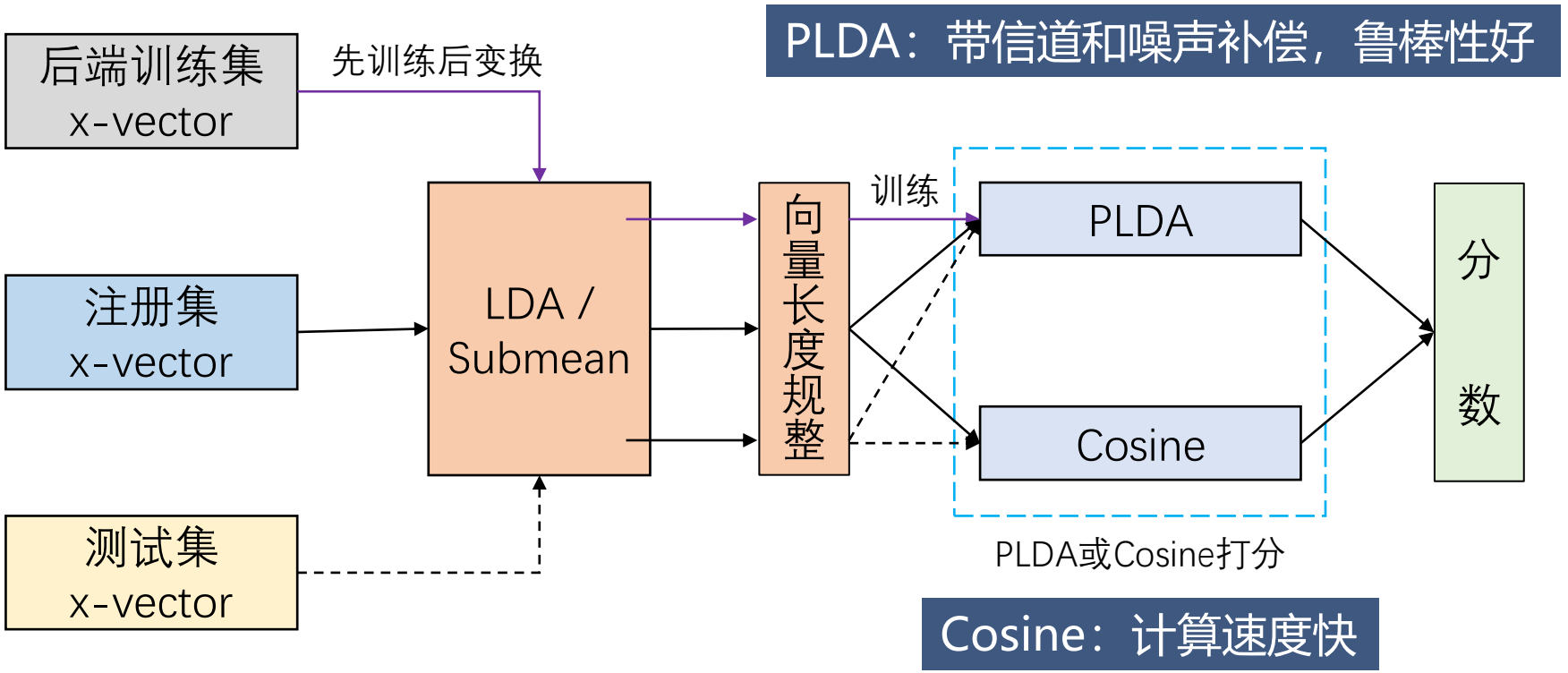
借助SE-Block学习全局信息以突出有用的特征信息而抑制无用的特征信息。

网络结构，可在标准tdnn-xvector的前3层帧级别层后都添加SE-Block层。

1. **Squeeze**：学习不同通道(特征维度)对于区分性的贡献度
2. **Excitation**：将原始数据乘上贡献度增强有用信息、削弱无用信息


Jianfeng Zhou, Tao Jiang, Zheng Li, Lin Li, Qingyang Hong, “Deep Speaker Embedding Extraction with Channel-Wise Feature Responses and Additive Supervision Softmax Loss Function”, *INTERSPEECH 2019*.

# 后端打分




**后端打分流程图**

# ASV-Subtools实现的PLDA自适应(Python代码)

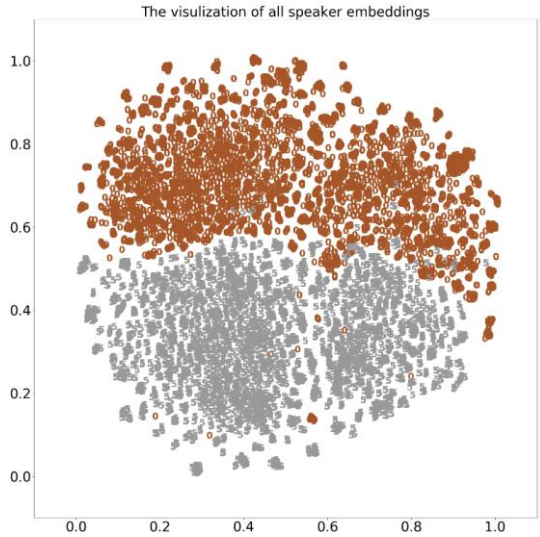
XIAMEN UNIVERSITY

**无监督方法：**

- CORAL
  - feature-based domain adaptation
  - model-based domain adaptation
- CORAL+

**有监督方法：**

- 监督线性差值法(LIP)
- 相关对齐差值法(CIP)
  - 基于CORAL的监督参数调整法
  - 基于CORAL+的监督参数调整法




Different Domains

Qiongqiong Wang, Koji Okabe, Kong Aik Lee, and Takafumi Koshinaka, "A Generalized Framework for Domain Adaptation of PLDA in Speaker Recognition", *ICASSP2020*.

# 自适应效果—SRE18

| | |
|---|---|
| 网络框架 | x-vector |
| 网络训练数据 | Switchboard, and SREs04–06,08,10,and12 |
| Out-of-Domain PLDA训练数据 | SREs04–06,08,10,and12 (English) |
| In-Domain PLDA训练数据 | SRE18-Eval (Tunisian Arabic) |
| 测试数据 | SRE18-Dev |
| 后处理 | 无 |

**Table 3**. Comparison results of PLDA domain adaptations on SRE'18 development set.

| System | EER(%) | minDCF08 | minDCF10 | $\text{minC}_{primary}$ |
|---|---|---|---|---|
| OOD PLDA | 12.96 | 0.701 | 0.877 | 0.728 |
| InD PLDA | 8.17 | 0.538 | 0.715 | 0.569 |
| Kaldi | 9.56 | 0.664 | 0.844 | 0.698 |
| CORAL+ | 7.62 | 0.604 | 0.807 | 0.637 |
| LIP | 6.85 | 0.522 | 0.731 | 0.555 |
| LIP reg | 7.03 | 0.518 | **0.706** | **0.546** |
| CIP | **6.26** | **0.515** | 0.737 | 0.548 |
| CIP reg | 6.34 | 0.524 | 0.736 | 0.559 |

reg - covariance regularization

Fuchuan Tong, Miao Zhao, Jianfeng Zhou, Hao Lu, Zheng Li, Lin Li, Qingyang Hong, "ASV-SUBTOOLS: OPEN SOURCE TOOLKIT FOR AUTOMATIC SPEAKER VERIFICATION", *Submitted to ICASSP2021*.

# 实验结果

- 文本无关：Voxceleb1&2
- 文本相关：SdSV2020
- 文本相关：RSR2015
- 语种识别：OLR2020

# 实验结果—Voxceleb

厦门大学 XIAMEN UNIVERSITY

**Table 1**. Results for verification on the VoxCeleb1 *test*, "Specaug" represents SpecAugment, "InSpecaug" represents Inverted SpecAugment.

| Reference | Model | EER(%) | minDCF08 |
|---|---|---|---|
| | *Trainset: VoxCeleb1&2 dev* | | |
| Kaldi recipe | TDNN | 3.13 | 0.326 |
| Liu et al. [17] | TDNN | 2.34 | - |
| Liu et al. [17] | TDNN + AM-softmax | 2.00 | - |
| ASV-Subtools | TDNN | 2.02 | 0.224 |
| | TDNN + AM-softmax | **1.97** | **0.215** |
| | *Trainset: VoxCeleb1 dev* | | |
| Okabe et al. [12] | TDNN + Attentive Pooling | 3.85 | 0.406 |
| Cai et al. [18] | ResNet-34 + A-softmax | 3.46 | - |
| Insoo et al. [19] | OrthResCNN-OVP | 2.85 | - |
| ASV-Subtools | TDNN | 3.36 | 0.370 |
| | TDNN + Specaug | 3.11 | 0.348 |
| | TDNN + InSpecaug | 2.78 | 0.322 |
| | TDNN + InSpecaug + AM-softmax | **2.64** | **0.289** |

**Table 2**. Results for verification on three tasks, all using the VoxCeleb2 *dev* for training, "InSpecaug" represents Inverted SpecAugment.

| Model | InSpecaug | EER(%) | | |
|---|---|---|---|---|
| | | VoxCeleb1-test | VoxCeleb1-E | VoxCeleb1-H |
| E-TDNN + PLDA | No | 2.08 | 2.22 | 3.84 |
| | Yes | 2.03 | 2.18 | 3.78 |
| F-TDNN + PLDA | No | 2.01 | 2.08 | 3.68 |
| | Yes | 1.98 | 2.08 | 3.66 |
| ResNet34 + PLDA | No | 1.89 | 1.96 | 3.56 |
| | Yes | 1.81 | 1.92 | 3.46 |
| ResNet34 + Cosine | No | 1.71 | **1.81** | 3.26 |
| | Yes | **1.67** | **1.81** | **3.23** |
| **Training ResNet34 with SGD (fine-tuning super-parameters)** | | | | |
| Cosine | - | **1.26** | **1.34** | **2.36** |
| + AS-Norm | - | **1.16** | - | - |

Fuchuan Tong, Miao Zhao, Jianfeng Zhou, Hao Lu, Zheng Li, Lin Li, Qingyang Hong, "ASV-SUBTOOLS: OPEN SOURCE TOOLKIT FOR AUTOMATIC SPEAKER VERIFICATION", *Submitted to ICASSP2021*.

# 实验结果—SdSV2020

XIAMEN UNIVERSITY

Table 3: *The EER% and minDCF results of our single systems on the Task 1 of SdSV Challenge 2020 on the Progress subset and Evalution subset*

| # | Feature | Systems | Progress subset | | Evaluation subset | |
|---|---|---|---|---|---|---|
| | | | minDCF | EER% | minDCF | EER% |
| 0 | MFCC | official i-vector/HMM baseline [23] | 0.1472 | 3.47 | 0.1464 | 3.49 |
| 1 | MFCC | TDNN-Xvector | 0.1210 | 3.05 | 0.1219 | 3.04 |
| 2 | MFCC | TDNN-Xvector-SpecAug | 0.1149 | 2.90 | 0.1156 | 2.92 |
| 3 | MFCC | ETDNN-Xvector-SpecAug | 0.1091 | 2.61 | 0.1089 | 2.64 |
| 4 | PLP | TDNN-Xvector-SpecAug | 0.1171 | 2.98 | 0.1176 | 3.02 |
| 5 | PLP | ETDNN-Xvector-SpecAug | 0.1078 | 2.66 | 0.1082 | 2.68 |
| 6 | FBank | TDNN-Xvector-SpecAug | 0.1066 | 2.71 | 0.1077 | 2.72 |
| 7 | FBank | ETDNN-Xvector-SpecAug | 0.1056 | 2.66 | 0.1059 | 2.67 |
| 8 | MFCC | Transfer-Xvector-SpecAug | 0.1067 | 2.63 | 0.1069 | 2.63 |
| 9 | MFCC | Transfer-Xvector-SpecAug-Attentive | **0.1016** | **2.48** | **0.1024** | **2.51** |

Tao Jiang, Miao Zhao, Lin Li, Qingyang Hong, "The XMUSPEECH System for Short-Duration Speaker Verification Challenge 2020", *INTERSPEECH 2020*.

# 实验结果—RSR2015

| System | Development set | | | | | |
|---|---|---|---|---|---|---|
| | Male | | | Female | | |
| | TW | IC | IW | TW | IC | IW |
| x-vector | 0.672 | 1.713 | 0.045 | 0.333 | 1.223 | 0.024 |
| xv-SE | 0.627 | 1.713 | 0.045 | 0.214 | 1.123 | 0.012 |
| xv-AP | **0.482** | 1.556 | 0.045 | 0.238 | 1.081 | 0.012 |
| xv-SE-AP | 0.493 | **1.478** | 0.045 | **0.119** | **0.998** | **0.012** |

| System | Male | | | | | | Female | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Development | | | Evaluation | | | Development | | | Evaluation | | |
| | TW | IC | IW | TW | IC | IW | TW | IC | IW | TW | IC | IW |
| i-vector [20] | 2.870 | 5.950 | 0.740 | 1.950 | 4.030 | 0.320 | 3.050 | 7.870 | 0.940 | 1.910 | 6.610 | 0.750 |
| HiLAM [20] | 1.660 | 3.690 | 0.490 | 0.820 | 2.470 | 0.190 | 1.770 | 3.240 | 0.450 | 0.610 | 2.960 | 0.140 |
| Joint-spk-utt [23] | 5.565 | 1.981 | 1.792 | 5.125 | 2.079 | 0.888 | 5.179 | 1.699 | 0.831 | 3.110 | 1.453 | 0.499 |
| SUDA [24] | 0.202 | **0.728** | **0.022** | 0.068 | 0.722 | **0.010** | 0.297 | 1.449 | 0.024 | 0.125 | **0.863** | 0.023 |
| xv-SE-AP | 0.493 | 1.478 | 0.045 | 0.303 | 1.005 | **0.010** | 0.119 | 0.998 | **0.012** | 0.114 | 1.044 | 0.023 |
| MT-SE-AP | 0.112 | 3.493 | 0.034 | 0.039 | 2.362 | **0.010** | 0.047 | 2.768 | **0.012** | 0.057 | 2.679 | 0.034 |
| **GRL-MT-SE1-AP** | **0.056** | 1.601 | **0.022** | **0.020** | **0.703** | **0.010** | **0.012** | **0.938** | **0.012** | **0.023** | 0.931 | **0.011** |

AP-Attention Pooling, MT-Multi-task, GRL-Gradient Reversal Layer

Yan Liu, Miao Zhao, Jianfeng Zhou, Zheng Li, Lin Li, Qingyang Hong, "TEMPORAL AND CHANNEL-WISE ATTENTIVE LEARNING FOR PHONETIC-LEVEL BASED TEXT-DEPENDENT SPEAKER VERIFICATION", *Submitted to ICASSP2021*.

# 实验结果—OLR2020 Baseline

$C_{avg}$ AND EER RESULTS ON THE REFERENCED DEVELOPMENT SETS

| Task | Cross-channel LID | | Dialect Identification | |
|---|---|---|---|---|
| Enrollment Set | AP20-ref-dev-task1 | | AP20-OLR-dialect | |
| Test Set | AP19-OLR-channel | | AP19-OLR-dev&eval-task3-test | |
| | *Cavg* | EER% | *Cavg* | EER% |
| [Kaldi] i-vector | 0.2965 | 29.12 | 0.0703 | 9.33 |
| [Kaldi] x-vector | 0.3583 | 36.37 | 0.0807 | 14.67 |
| [Pytorch] x-vector | 0.2696 | 26.94 | 0.0849 | 12.40 |

$C_{avg}$ AND EER RESULTS ON THE AP20 EVALUATION SETS

| Task | Cross-channel LID | | Dialect Identification | | Noisy LID | |
|---|---|---|---|---|---|---|
| Enrollment Set | AP20-ref-enroll-task1 | | AP20-OLR-dialect | | AP20-ref-enroll-task3 | |
| Test Set | AP20-OLR-channel-test | | AP20-OLR-dialect-test | | AP20-OLR-noisy-test | |
| | *Cavg* | EER% | *Cavg* | EER% | *Cavg* | EER% |
| [Kaldi] i-vector | 0.1542 | 19.40 | 0.2214 | 23.94 | 0.0967 | 9.77 |
| [Kaldi] x-vector | 0.2098 | 22.49 | 0.2117 | 22.25 | 0.1079 | 11.12 |
| [Pytorch] x-vector | 0.1321 | 14.58 | 0.1752 | 19.74 | 0.0715 | 7.14 |

Zheng Li, Miao Zhao, Qingyang Hong, Lin Li, Zhiyuan Tang, Dong Wang, Liming Song and Cheng Yang, "AP20-OLR Challenge: Three Tasks and Their Baselines", *APSIPA ASC 2020*.

# 总结与展望

- **总结**
  - 已发布ASV-Subtools，支持TDNN/E-TDNN/F-TDNN/Resnet、Multi-head Attention/LDE Pooling、AM/AAM Softmax，并在Voxceleb等数据库取得很好的性能，研究人员可用做基线系统；
  - 集成Inverted Specaugment、SE、PLDA自适应等改进算法；
  - 工具提供了Voxceleb、VoxcelebSRC、AP-OLR challenge2020实验脚本。

- **展望**
  - 发布多任务学习版本；
  - 发布GAN训练版本；
  - 集成更多新方法和新网络，如online扩增、ECAPA等；
  - 研究工具能对接产品落地。

# 致谢


- 团队成员：
  - 主体工程：赵淼（就职国音智能）
  - PLDA自适应：周健峰（就职海康威视）
  - 语种识别、多任务学习：李铮
  - F-TDNN：陆昊
  - GAN训练：童福川
  - 文本相关：江涛、刘妍
  - 测试调优：郅艺铭、李静
- 感谢厦门大学智能语音实验室其他同学的贡献。

- 感谢使用和支持ASV-Subtools的研究人员，欢迎更多参与者，共同推动声纹识别开源工具。

# 汇报完毕，敬请指正！

厦门大学智能语音实验室

speech.xmu.edu.cn