腾讯说话人识别反欺骗技术进展与应用探索

Shanshan Zhang

参与人：Xinyue Ma, Ji Gao, Dan Su, Sheng Huang

2020/11/21

1 说话人AntiSpoofing 问题介绍

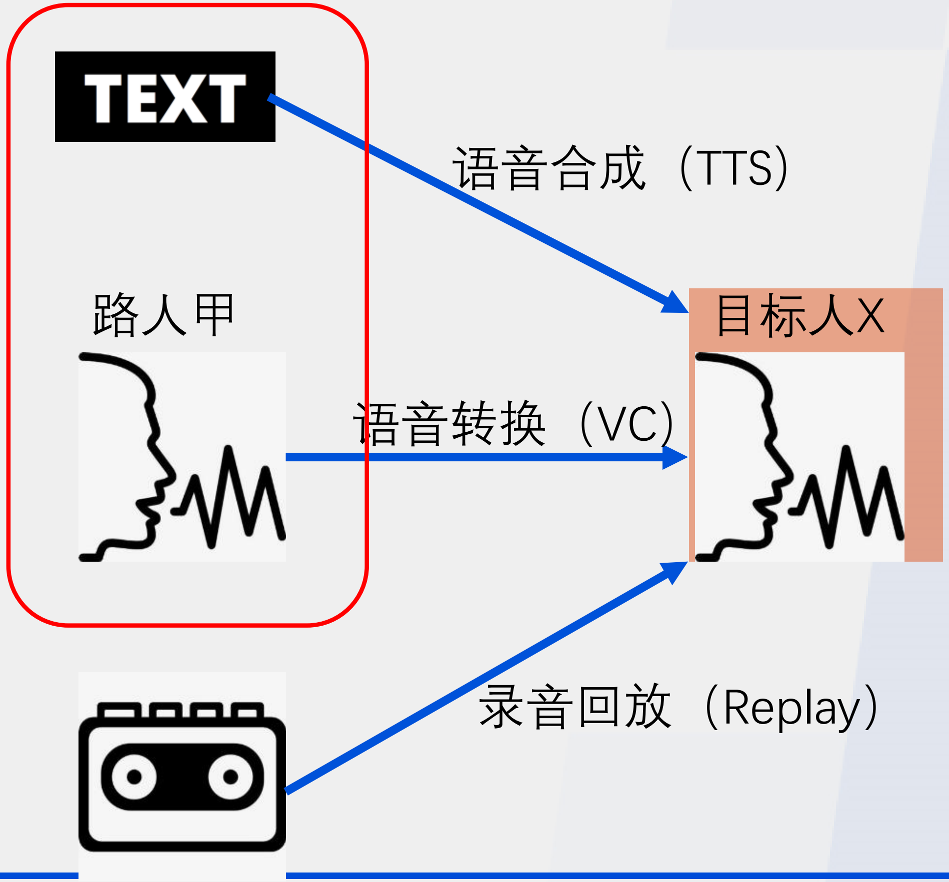
2 基于LCNN的识别系统

- Feature extraction
- Loss functions
- Attention modules

3 业务场景落地探索

说话人Anti-Spoofing问题

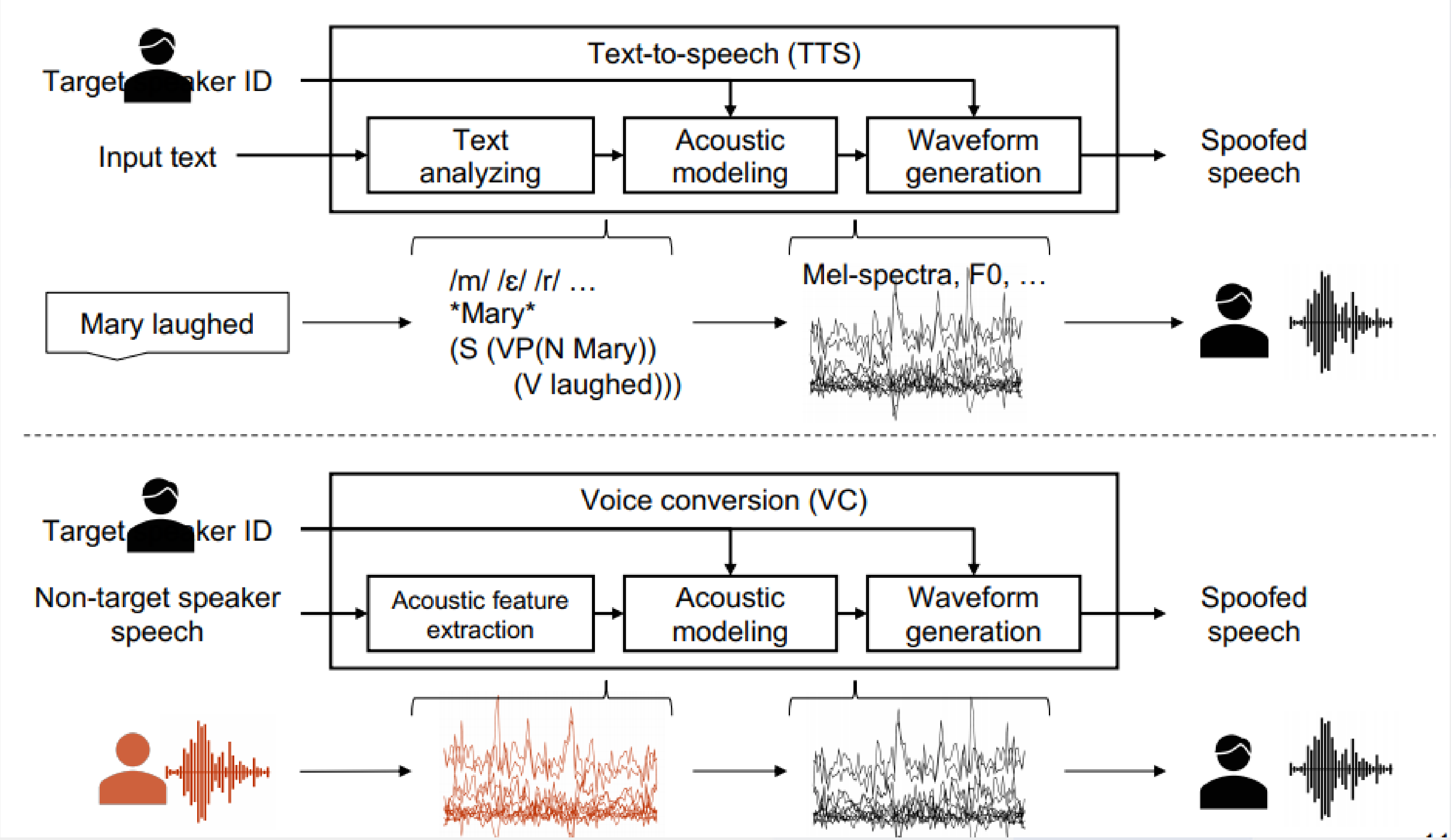
2019年《网络音视频信息管理规定》指出，明确禁止任何组织和个人滥用AI技术制作发布传播虚假新闻信息。



ASVspoof 比赛介绍

- logical access (LA)： TTS/VC
- physical access (PA): Replay

TTS vs VC


Text-to-speech (TTS)
Target speaker ID
Input text
Text analyzing
Acoustic modeling
Waveform generation
Spoofed speech
Mary laughed
/m/ /ɛ/ /r/ …
Mary
(S (VP(N Mary))
(V laughed)))
Mel-spectra, F0, …
Voice conversion (VC)
Target speaker ID
Non-target speaker speech
Acoustic feature extraction
Acoustic modeling
Waveform generation
Spoofed speech

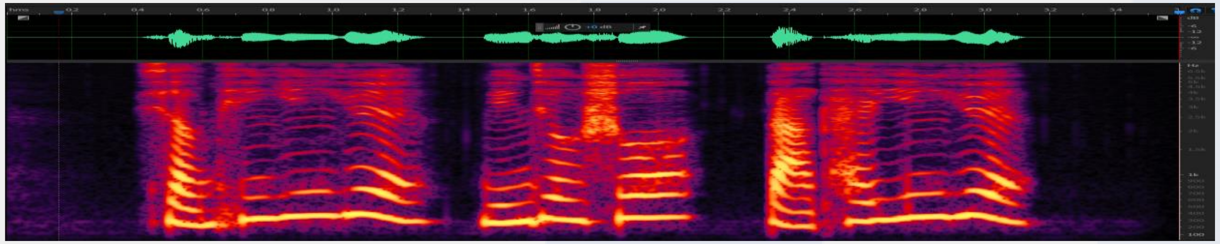
1 说话人AntiSpoofing 问题介绍

2 基于LCNN的识别系统

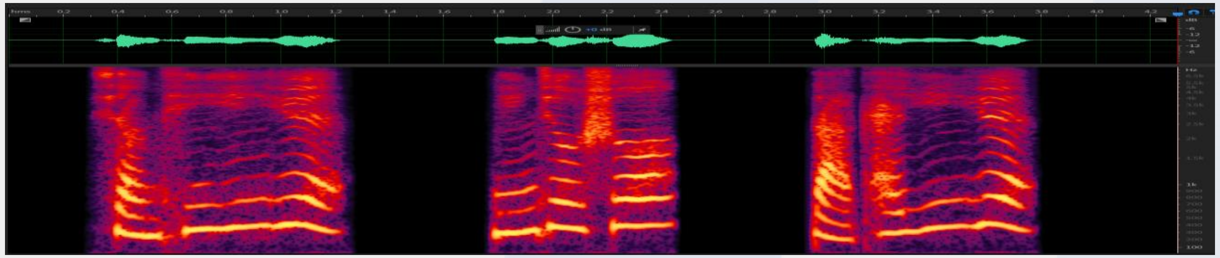
- Feature extraction
- Loss functions
- Attention modules

3 业务场景落地探索

真假语音对比

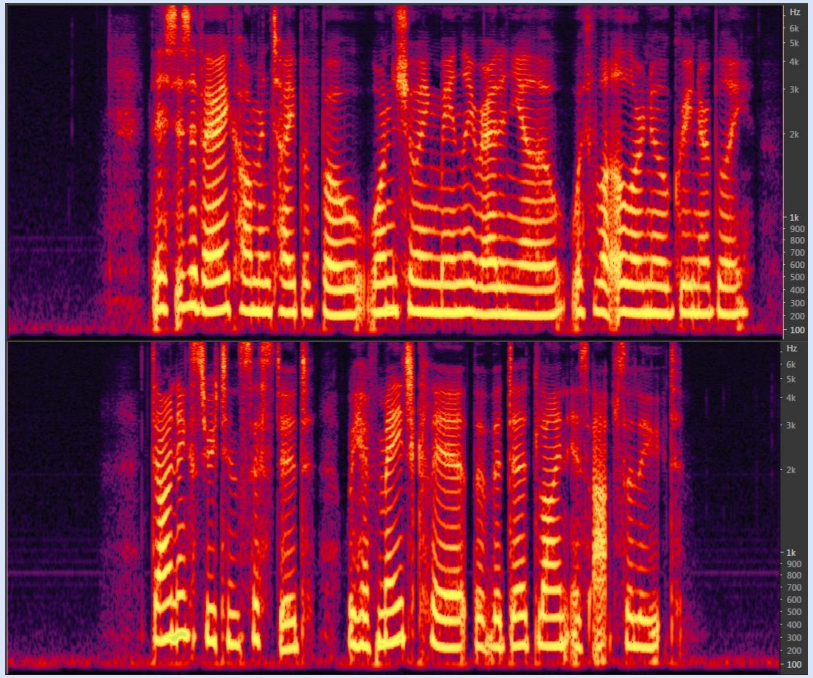


真实
语音



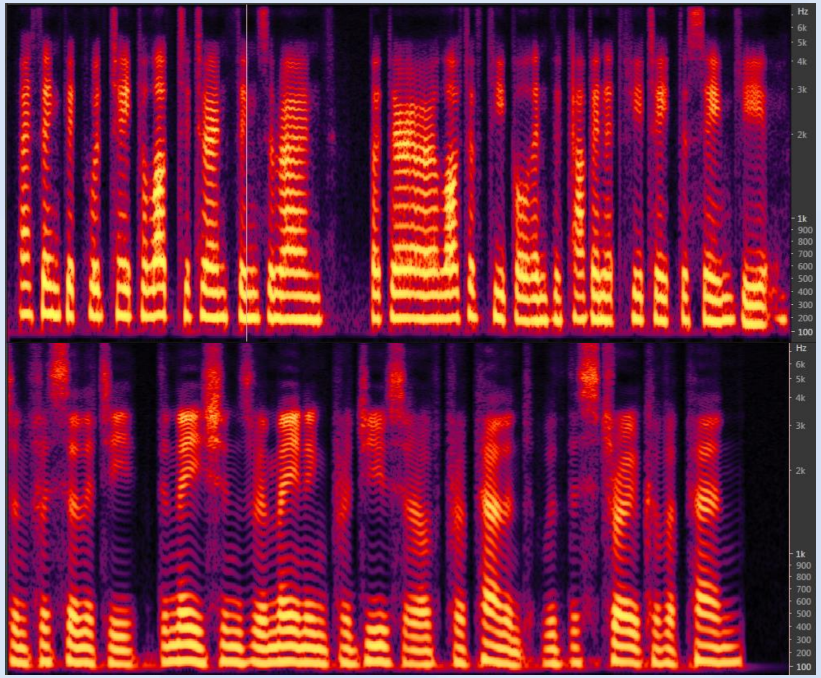
合成
语音

音频特点分析



真实语音

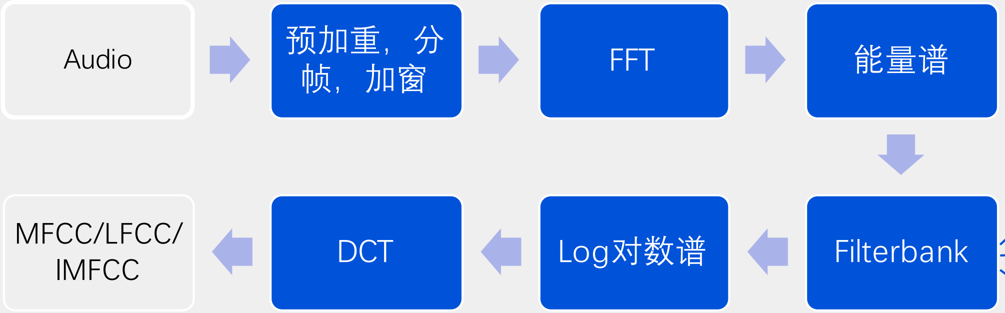
- 细节丰富
- 韵律自然



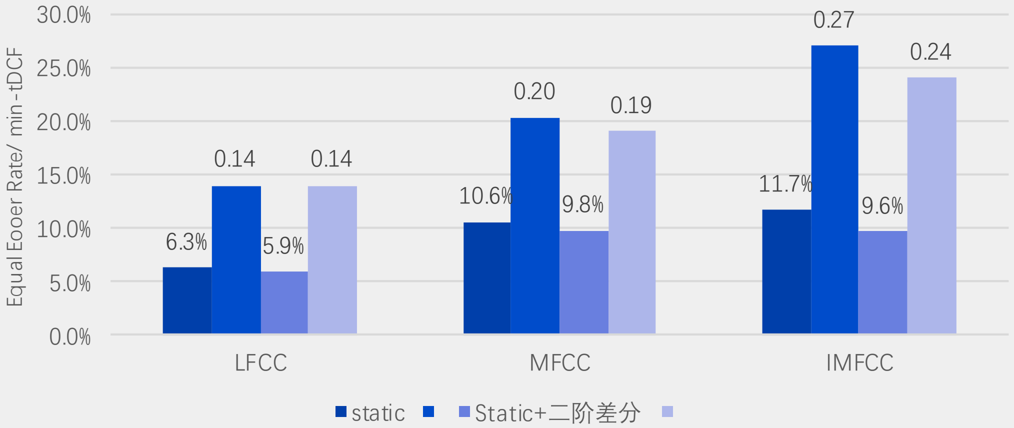
伪造语音

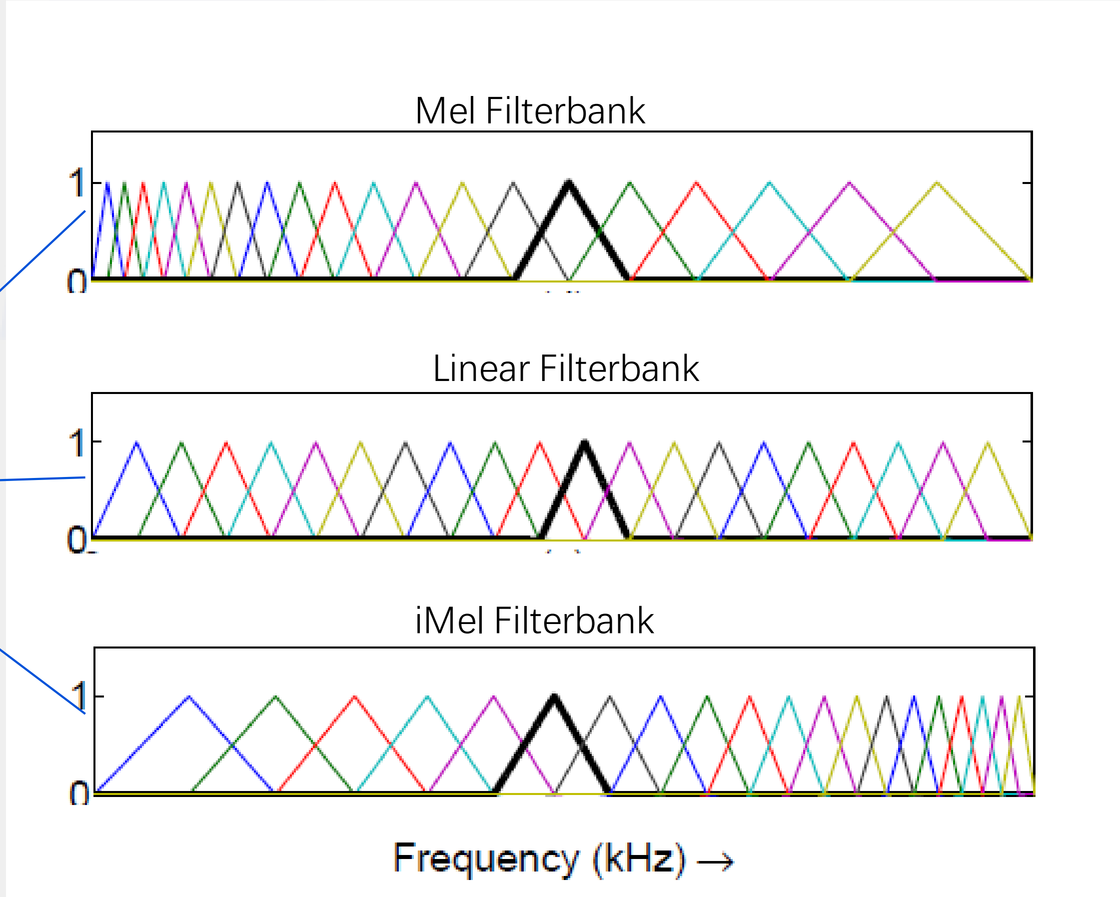
- 细节不自然—高频部分
- 韵律相对单一

Feature extraction



MFCC/LFCC/IMFCC特征在ASV2019 eval测试集上的性能对比





Modeling

- Light CNN with Max-Feature-Map(MFM) activation

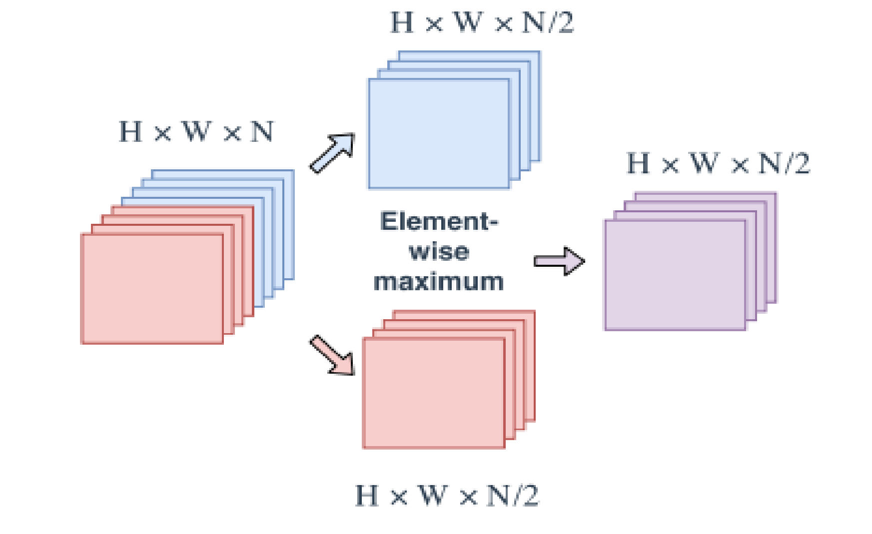


Table 1: *LCNN architecture*

Type	Filter / Stride	Output	Params
Conv_1	5 × 5 / 1 × 1	863 × 600 × 64	1.6K
MFM_2	–	864 × 600 × 32	–
MaxPool_3	2 × 2 / 2 × 2	431 × 300 × 32	–
Conv_4	1 × 1 / 1 × 1	431 × 300 × 64	2.1K
MFM_5	–	431 × 300 × 32	–
BatchNorm_6	–	431 × 300 × 32	–
Conv_7	3 × 3 / 1 × 1	431 × 300 × 96	27.7K
MFM_8	–	431 × 300 × 48	–
MaxPool_9	2 × 2 / 2 × 2	215 × 150 × 48	–
BatchNorm_10	–	215 × 150 × 48	–
Conv_11	1 × 1 / 1 × 1	215 × 150 × 96	4.7K
MFM_12	–	215 × 150 × 48	–
BatchNorm_13	–	215 × 150 × 48	–
Conv_14	3 × 3 / 1 × 1	215 × 150 × 128	55.4K
MFM_15	–	215 × 150 × 64	–
MaxPool_16	2 × 2 / 2 × 2	107 × 75 × 64	–
Conv_17	1 × 1 / 1 × 1	107 × 75 × 128	8.3K
MFM_18	–	107 × 75 × 64	–
BatchNorm_19	–	107 × 75 × 64	–
Conv_20	3 × 3 / 1 × 1	107 × 75 × 64	36.9K
MFM_21	–	107 × 75 × 32	–
BatchNorm_22	–	107 × 75 × 32	–
Conv_23	1 × 1 / 1 × 1	107 × 75 × 64	2.1K
MFM_24	–	107 × 75 × 32	–
BatchNorm_25	–	107 × 75 × 32	–
Conv_26	3 × 3 / 1 × 1	107 × 75 × 64	18.5K
MFM_27	–	107 × 75 × 32	–
MaxPool_28	2 × 2 / 2 × 2	53 × 37 × 32	–
FC_29	–	160	10.2 MM
MFM_30	–	80	–
BatchNorm_31	–	80	–
FC_32	–	2	64

Refer to: [1] Wu X, He R, Sun Z, et al. A light cnn for deep face representation with noisy labels[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2884-2896.
[2] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," Interspeech, 09 2019.

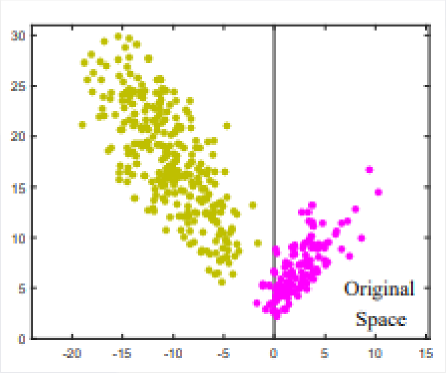
Loss function

- Two class classification

- Softmax

$$L_i = -\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\cos(\theta_{y_i})}}{\sum_j e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right)$$

Introduce margin m

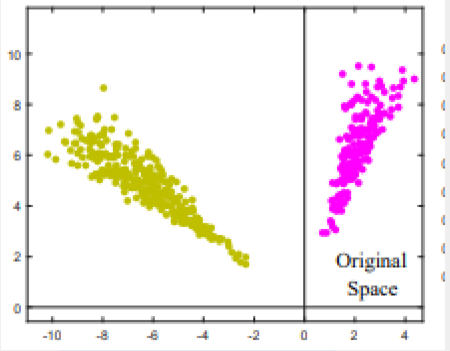


- Large margin softmax

$$L_i = -\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})}}{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})} + \sum_{j\neq y_i} e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right)$$

in which we generally require

$$\psi(\theta) = \begin{cases} \cos(m\theta), & 0 \leq \theta \leq \frac{\pi}{m} \\ \mathcal{D}(\theta), & \frac{\pi}{m} < \theta \leq \pi \end{cases}$$

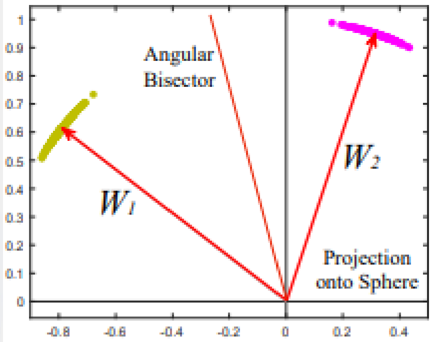


Normalize the weights and zero the biases

- Angular softmax

$$L_{\text{ang}} = \frac{1}{N}\sum_i -\log\left(\frac{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})}}{e^{\|\boldsymbol{x}_i\|\psi(\theta_{y_i,i})} + \sum_{j\neq y_i} e^{\|\boldsymbol{x}_i\|\cos(\theta_{j,i})}}\right)$$

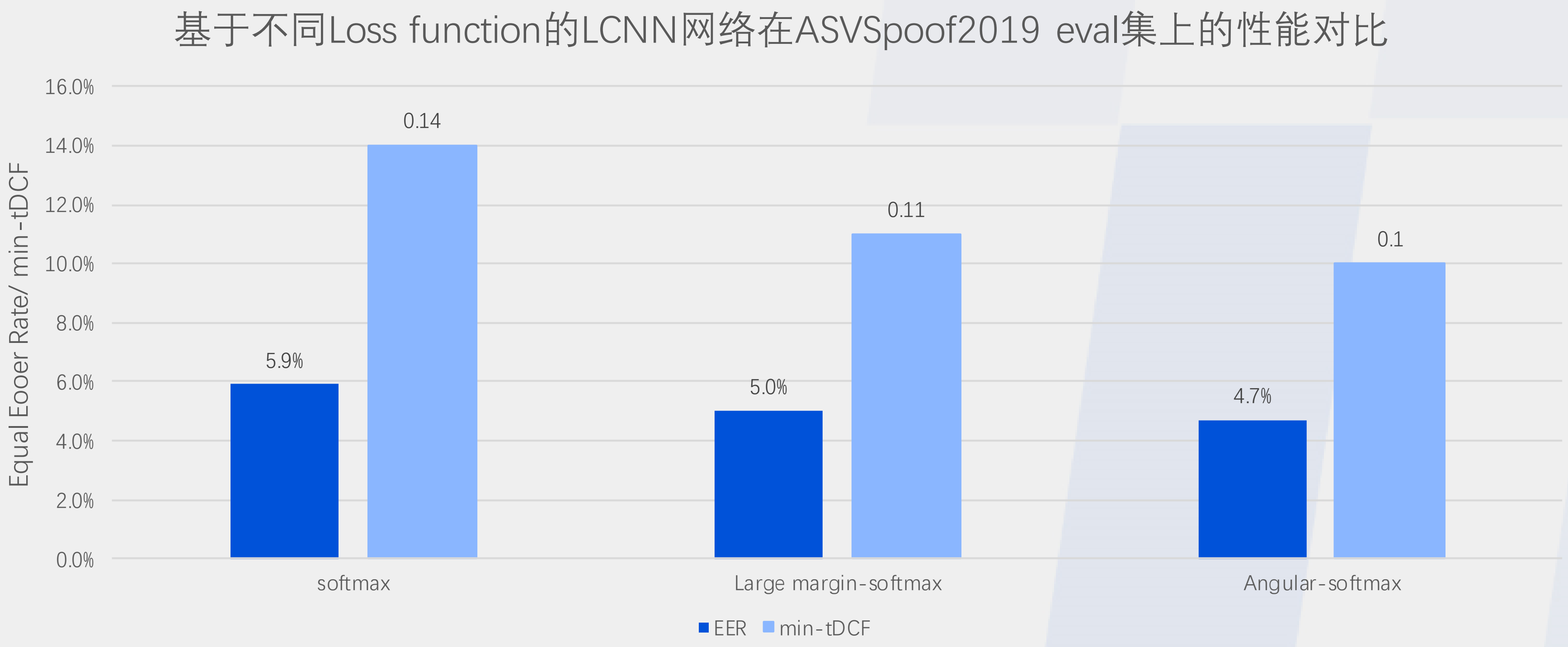
in which we define $\psi(\theta_{y_i,i}) = (-1)^k \cos(m\theta_{y_i,i}) - 2k$,



Refer to: [1] large-Margin Softmax Loss for Convolutional Neural Networks. Weiyang Liu, Yandong Wen, Zhiding Yu, Meng Yang. ICML 2016

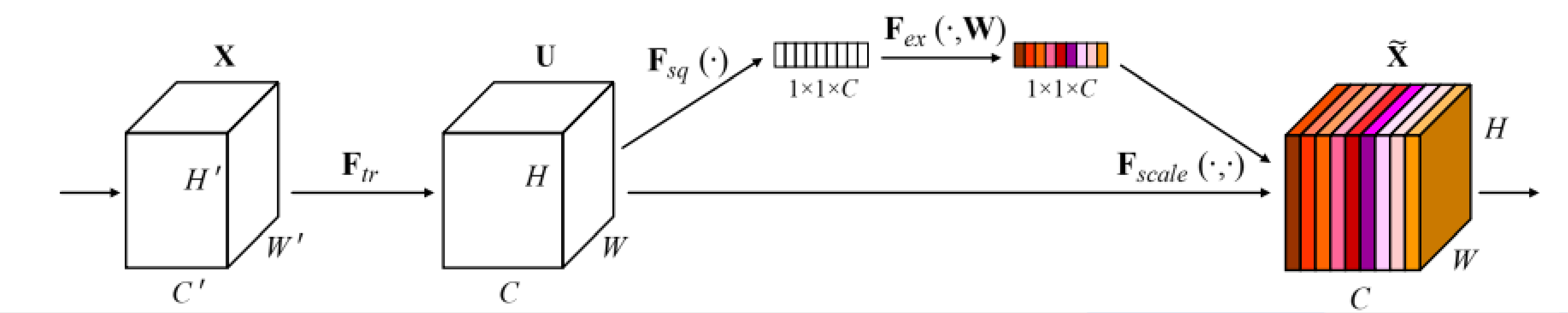
[2] SphereFace: Deep Hypersphere Embedding for Face Recognition. Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, Le Song. ICML 2017

Loss function



Attention modules

- Squeeze-and-Excitation (SE)
  - Squeeze: global average pooling in channel
  - Excitation: produces modulation weights



Refer to: Hu J, Shen L, and Sun G, “Squeeze-and-Excitation Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, 2017.

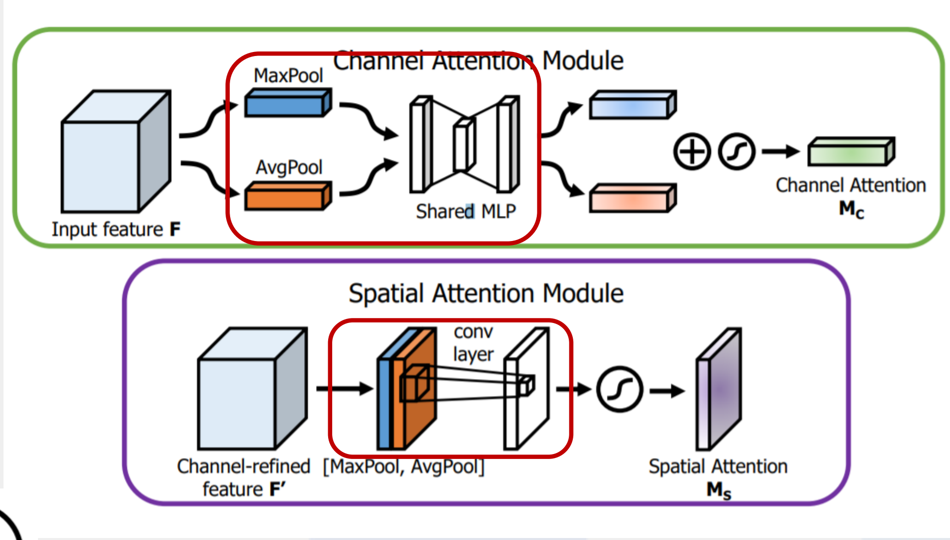
Attention modules

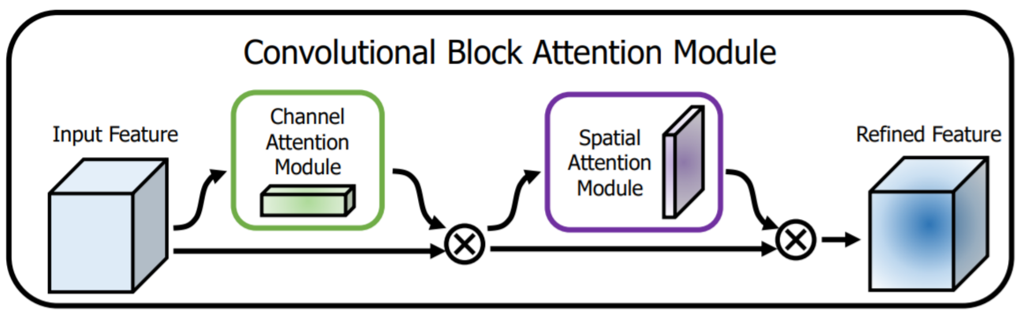
- Convolutional Block Attention Module (CBAM)
  - channel attention module

$$\begin{aligned}\mathbf{M_c}(\mathbf{F}) &= \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F})))\\ &= \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{avg}})) + \mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{max}}))),\end{aligned}$$

  - spatial attention module

$$\begin{aligned}\mathbf{M_s}(\mathbf{F}) &= \sigma(f^{7\times 7}([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})]))\\ &= \sigma(f^{7\times 7}([\mathbf{F^s_{avg}}; \mathbf{F^s_{max}}])),\end{aligned}$$

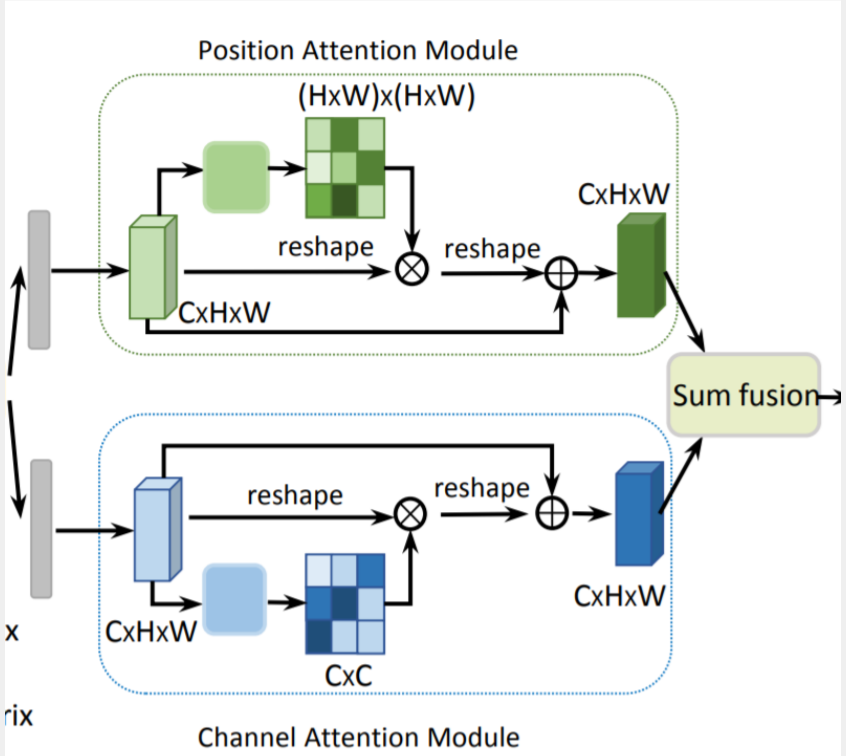


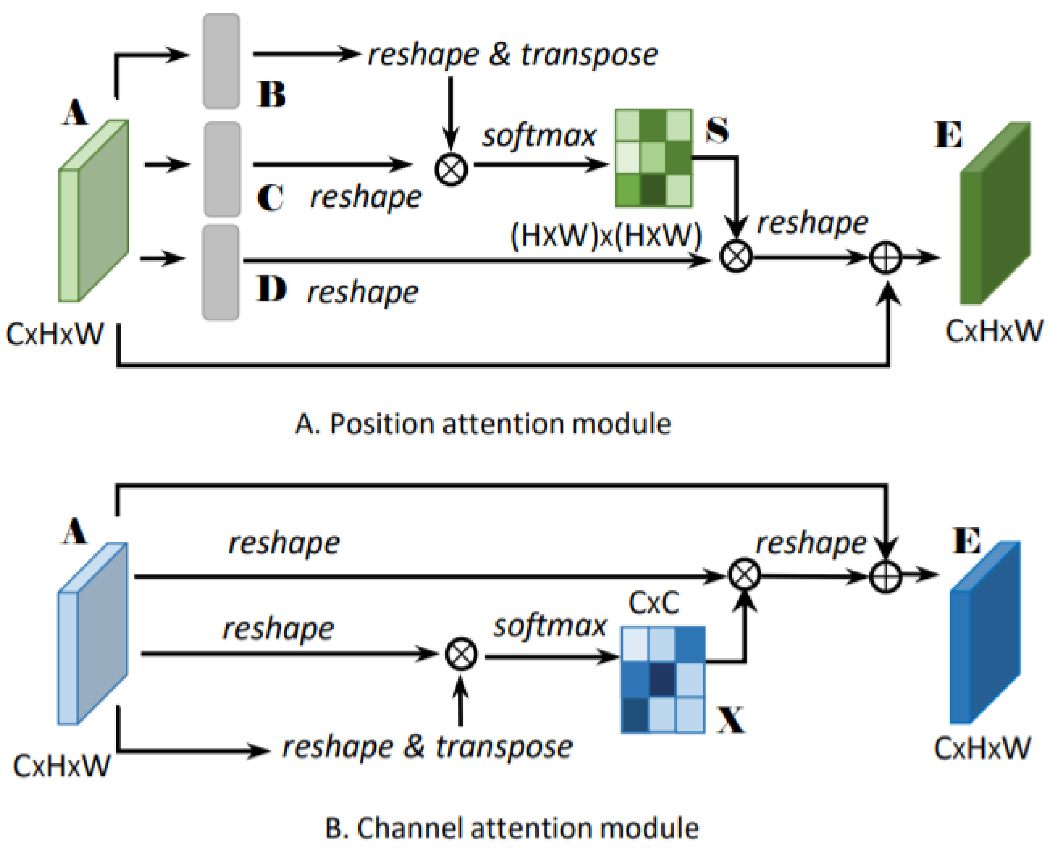


Refer to: Sanghyun Woo, JongChan Park, Joon-Young Lee, and Inso Kweon, "CBAM: Convolutional Block Attention Module," arXiv: Computer Vision and Pattern Recognition, pp. 3–19, 2018.

Attention modules

- Dual Attention (DA)
  - Channel 和position 并行关系





Refer to:Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and HanqingLu, "Dual Attention Network for Scene Segmentation," 2019 IEEE/CVF Conference on Computer Vision andPatternRecognition(CVPR),pp.3141–3149,2018.

Result on ASVSpoof 2019

- Attention modules
  - Squeeze-and-Excitation (SE)
  - Convolutional Block Attention Module (CBAM)
  - Dual Attention (DA)

	min-tDCF	EER（%）	min-tDCF	EER（%）
Baseline-GMM	0.066	2.71	0.021	8.09
LCNN	0.008	0.27	0.101	4.74
LCNN_SE	0.006	0.20	0.137	6.06
LCNN_CBAM	0.028	0.93	0.094	3.67
LCNN_DA	0.025	0.78	**0.078**	**2.76**

注：测试数据为 说话人Anti-Spoofing比赛 ASVSpoof2019 dev和eval测试集， 评价指标为EER和min-tDCF

Deepfake Detection Challenge

- 与图像团队参加由Facebook主办的DFDC百万美金比赛
  作为变声检测方案提供方

	EER(%)	min-tDCF
Dev集	1.16	0.03

	提交结果	
变声检测	0.67	
变脸检测	0.36	
两者融合	0.35	对总名次提升22名

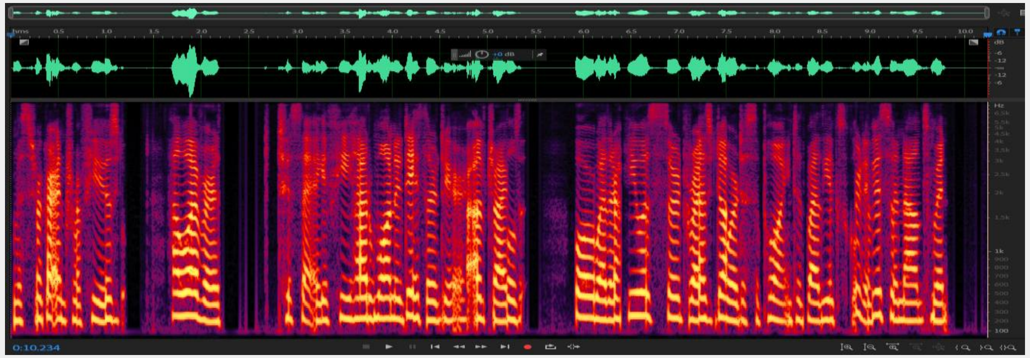
Kaggle-DFDC竞赛获得银牌

注：从提供的dev数据来看，变声数据比例大概占全部Fake数据10%。

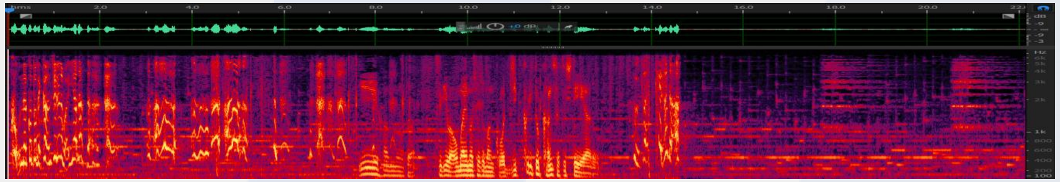
1. 说话人AntiSpoofing 问题介绍
2. 基于LCNN的识别系统
   - Feature extraction
   - Loss functions
   - Attention modules
3. **业务场景落地探索**

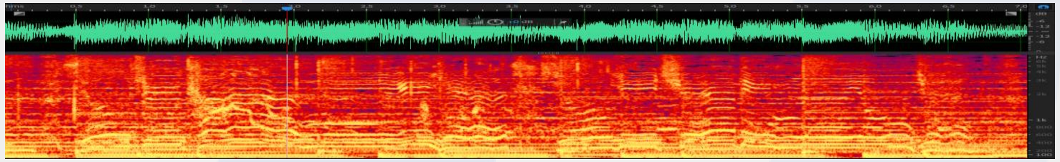
业务落地—数据分析

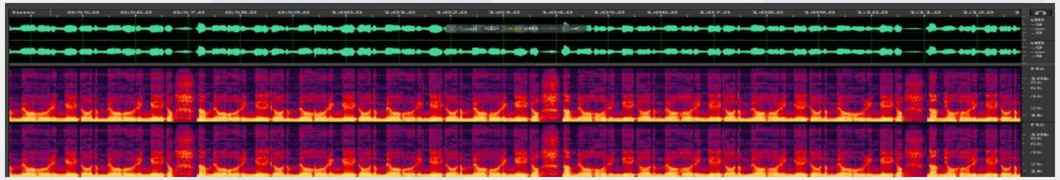
- 学术公开集数据
  - 单一纯净音频



- 业务场景数据
  - 多种音频内容
  - 信噪比低

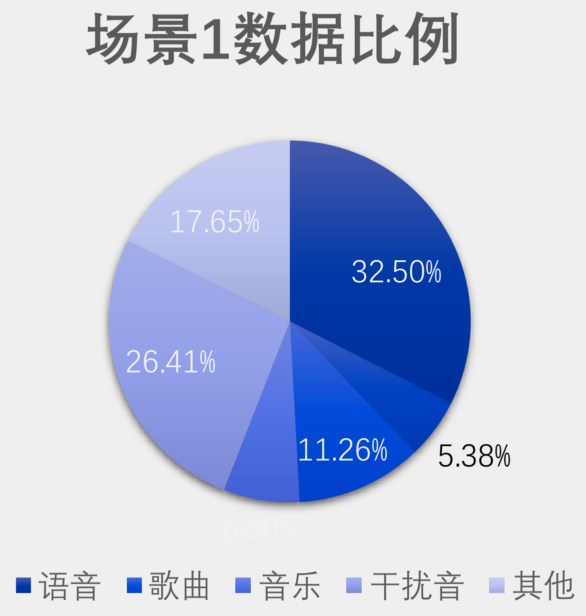


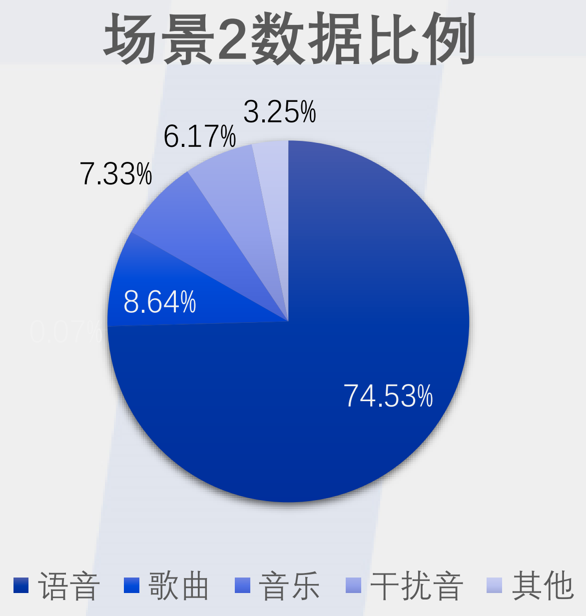




业务落地—数据分布

- 不同场景音频类型分布差异大



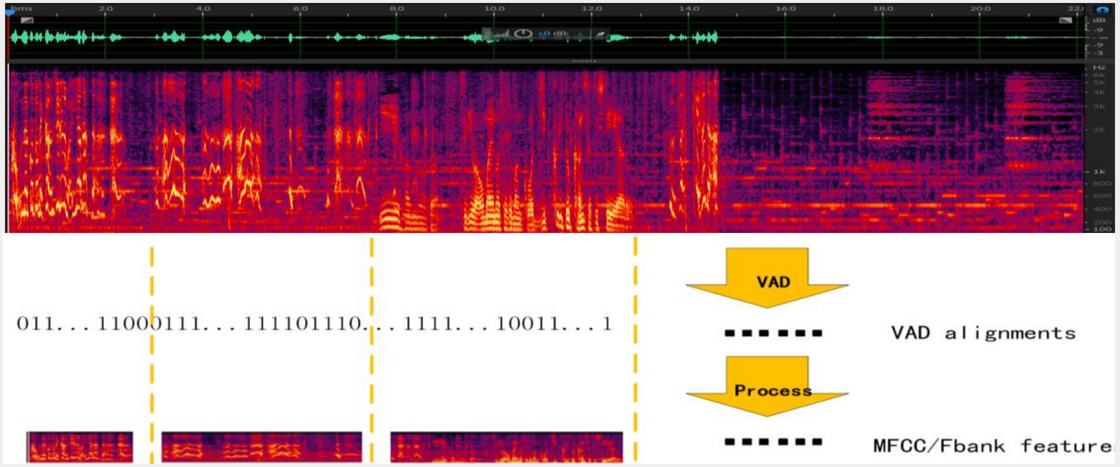


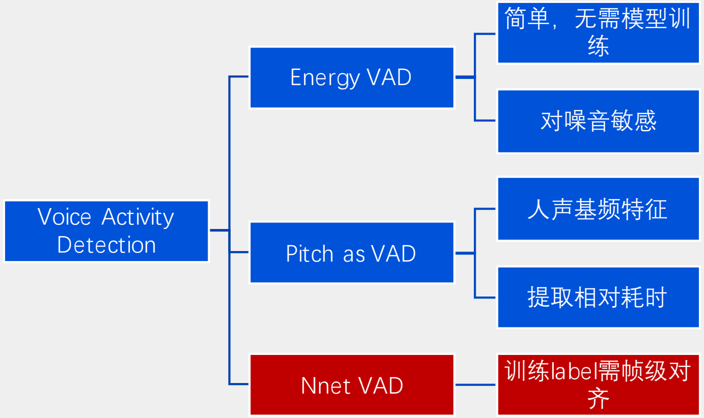
业务落地—问题定义

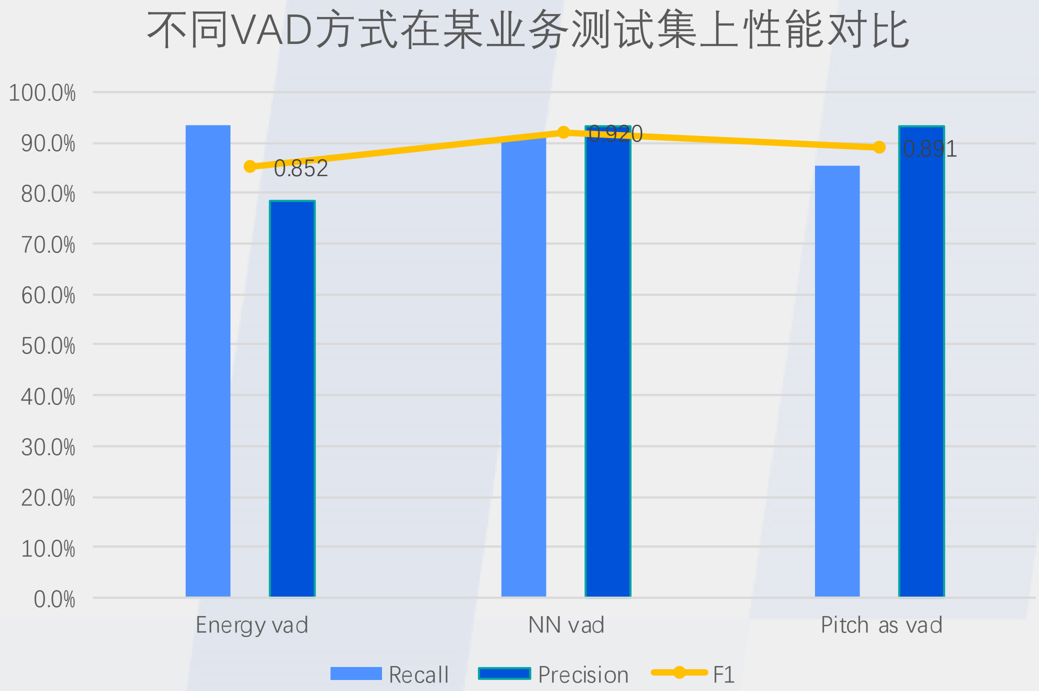
- **学术公开集任务**
  - 该音频中说话人是否为spoofed
- **学术公开集LA spoofing 范围**
  - TTS
  - VC

- **业务场景任务**
  - 该音频中是否出现spoofed说话人
- **业务场景中spoofing 范围**
  - TTS
  - VC
  - 鬼畜拼接

Voice Activity Detection (VAD)

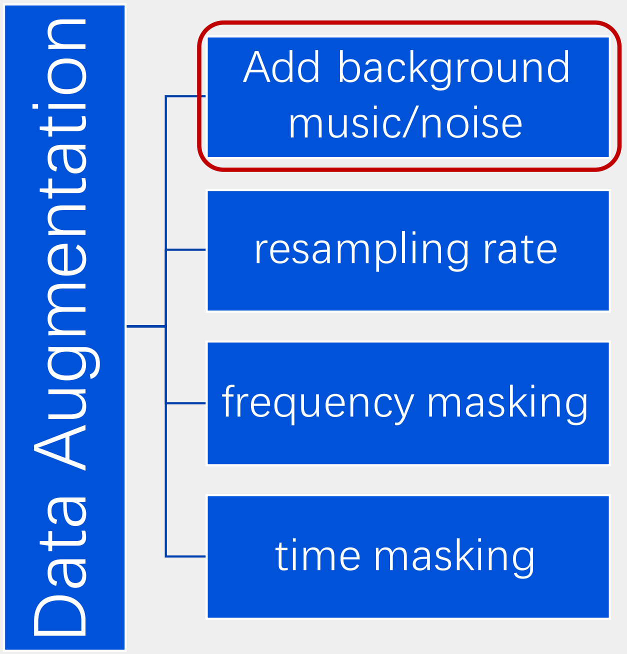


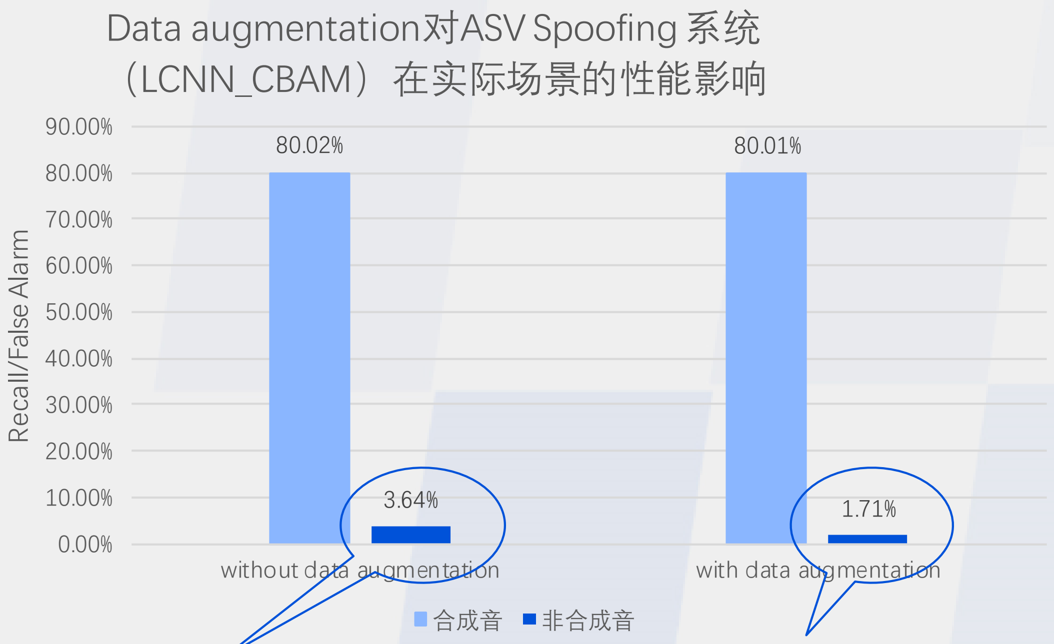


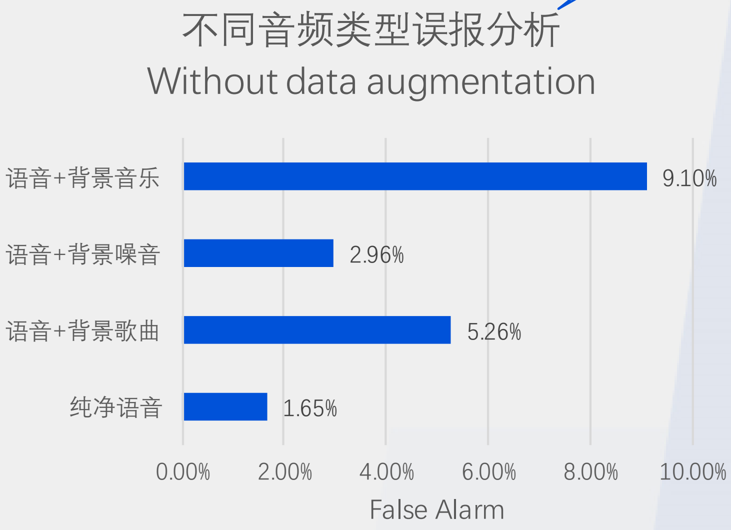


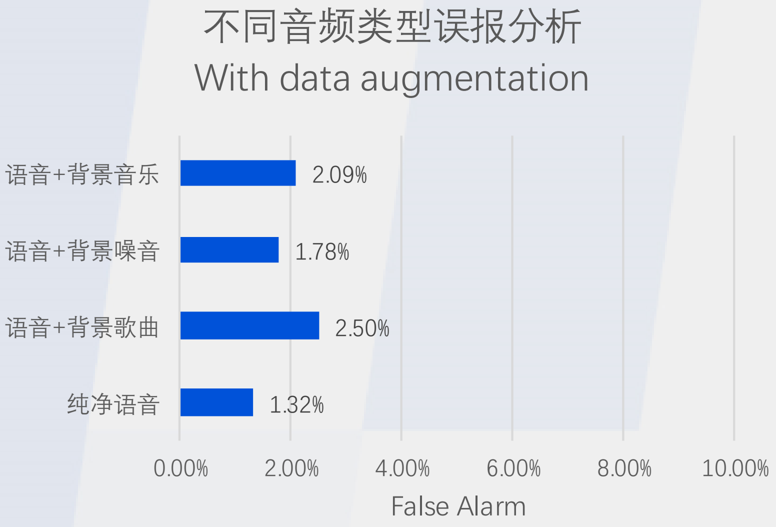
Data Augmentation

- 加入真实背景音乐、噪声数据

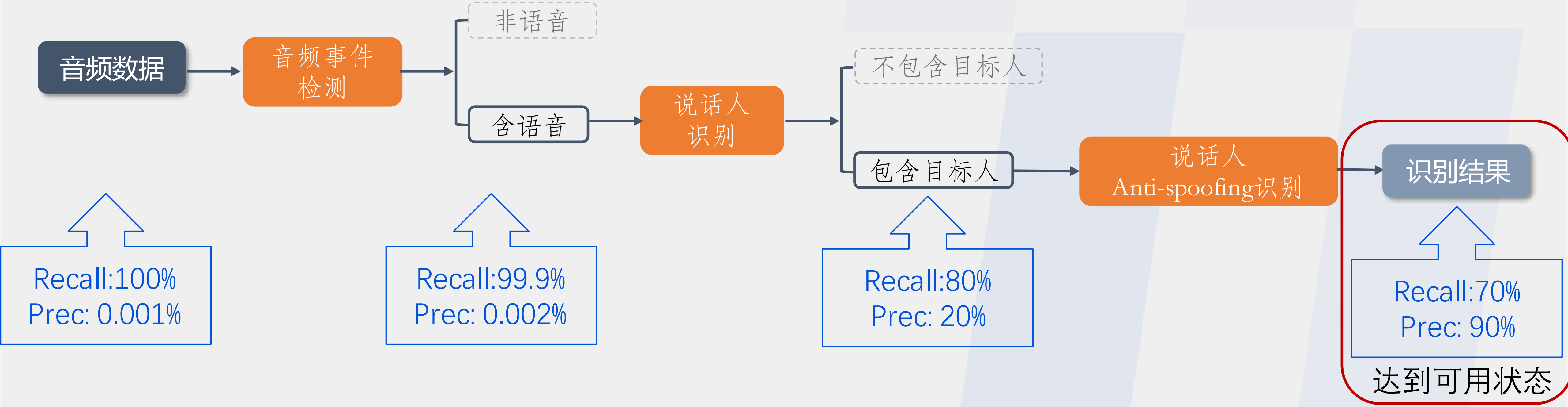




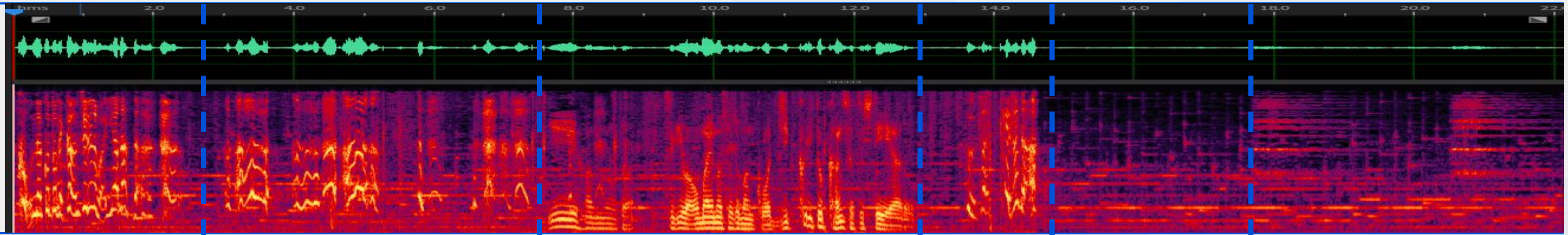




特定人的antispoofing 识别



音频内容理解

音频						
分类	语音	音乐	语音	其他	音乐	音乐
子类标签	语种	歌曲	语种 说话人，真假语音	干扰音	纯音乐	歌曲
内容	“今天天气真好。”		“ Hello, welcome to Tencent！ ”			
业务应用						

Thanks