CNSRC 2022 Technical Report

Lantian Li, Tao Jiang, Qingyang Hong, Dong Wang

Tsinghua University & Xiamen University & AISHELL

2022.07.27

Odyssey-CNSRC 2022 Workshop
27 June 2022, Beijing, China

OUTLINE

- Data, Tasks and Baselines
- Technical Summary
- System Analysis
- The Next CNSRC

OUTLINE

- **Data, Tasks and Baselines**
- Technical Summary
- System Analysis
- The Next CNSRC

The Origin of CN-Celeb Datasets

Modern Challenge of speaker recognition technique

- Complex variation: recognizing speakers **in the wild**.
  - Intrinsic: speaking style (e.g., reading or spontaneous), speaking rate, emotion, and physical status ...
  - Extrinsic: recording device, ambient acoustics, background noise, and transmission channel ...

Multi-genre scenario

- Perhaps the **MOST** challenging scenario for speaker recognition.
  - Multi-genre involves nearly all the complex variations.





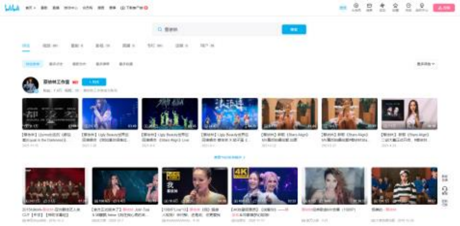
  - Multi-genre performance determines the practical success of speaker recognition research.

Data Collection Pipeline

1 Chinese POI list design

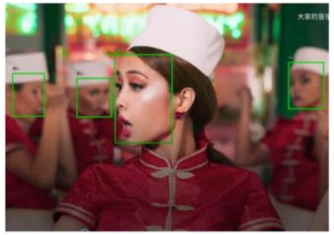
0001 周杰伦
0002 蔡依林
0003 刘德华
...

2 Pictures and videos download



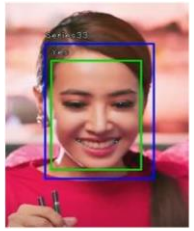
3 Face detection and tracking

Face detection
RetinaFace



Face verification
ArcFace

Face tracking
MOSSE

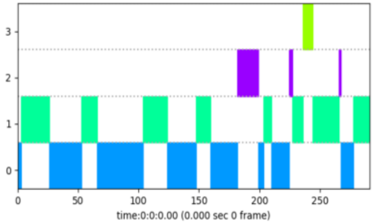


4 Mouth-Speech Sync

Synchronization
SyncNet

5 Speaker diarization

Relax & Recheck
UIS-RNN



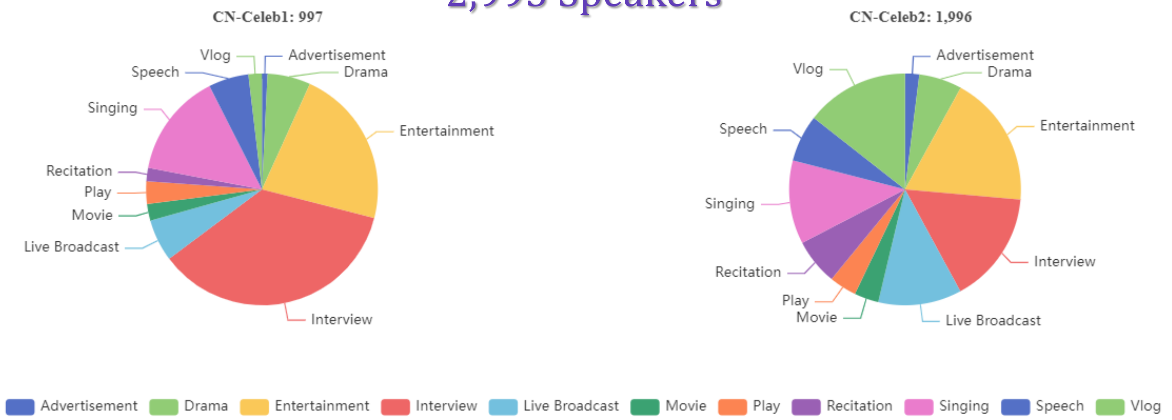
6 Human check

First-round **Full** check plus
Second-round **Spotting** check
Accuracy >= 90%

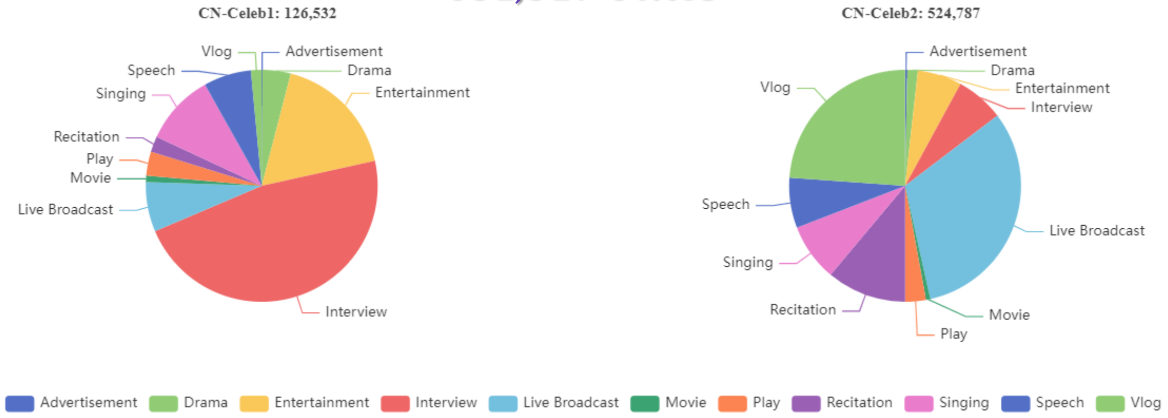
Data Profile

- Multi-genre data from multi-media sources

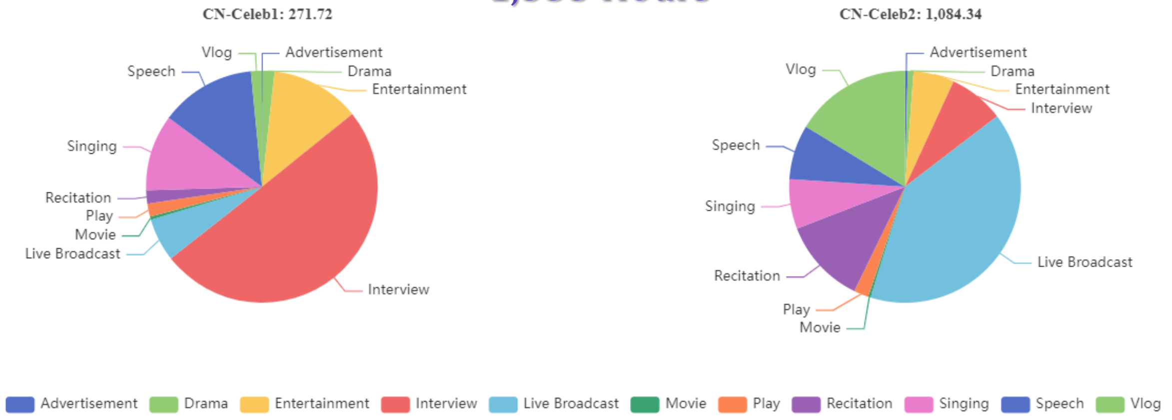
2,993 Speakers



651,319 Utters



~1,335 Hours



Task Description – Speaker Verification

Fixed Track

- **ONLY** CN-Celeb.T is allowed for training/tuning **ALL** the components of the system.
- This track is designed to compare different techniques under the **SAME** data resource.

CN-Celeb.T

CN-Celeb1/dev	# of Speakers	797
	# of Utters	107,953
CN-Celeb2	# of Speakers	1,996
	# of Utters	524,787
Overall	# of Speakers	2,793
	# of Utters	632,740

Open Track

- **ANY** data sources can be used for developing **ALL** the components of the system.
- This track is designed to examine the performance **Frontier** of the present technologies.

CN-Celeb.E

Enroll Data	# of Speakers	196
	# of Utters	196
	Avg. Duration	28s
Test Data	# of Speakers	200
	# of Utters	17,777
	Avg. Duration	8s
Trials	# of Target	17,755
	# of Non-target	3,466,537

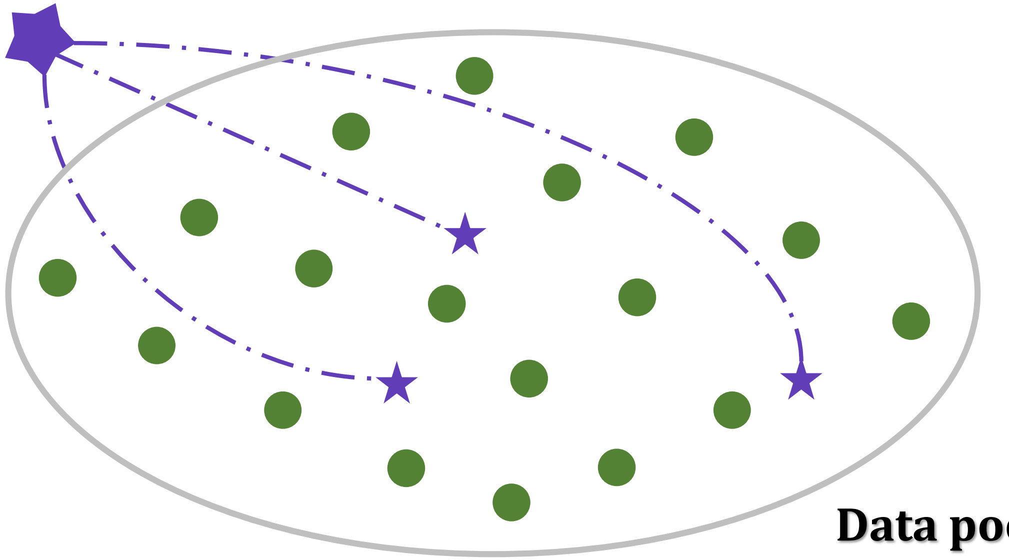
Task Description – Speaker Retrieval

Definition

- SR task is to find out the utterances spoken by a *target* speaker from a large data pool, given an enrollment data of the target speaker.
- The dataset contains **TWO** parts:
  - Target speakers' enrollment data ★
  - Data pool involves utters of the target speakers ★ as well as a large amount of non-target utters ●

Data profile

- **ANY** data sources except CN-Celeb.E are allowed for system building.
- SR.dev/SR.eval as development/evaluation sets.
  - 5/25 target speakers ★
  - 50/250 target utterances ★
  - 20,000/500,000 non-target utterances ●



Performance measurement

Task 1: Speaker Verification

- Primary: Minimum Detection Cost Function ($minDCF$)

$$C_{Det}(\theta) = \overset{1.0}{C_{Miss}} \times \overset{0.01}{P_{Target}} \times P_{Miss}(\theta) + \overset{1.0}{C_{FalseAlarm}} \times (1 - \overset{0.99}{P_{Target})} \times P_{FalseAlarm}(\theta)$$

$$minDCF = \arg\min_{\theta} \{0.01 \times P_{Miss}(\theta) + 0.99 \times P_{FalseAlarm}(\theta)\}$$

- Secondary: Equal Error Rate (EER)

$$P_{Miss}(\theta^*) = P_{FalseAlarm}(\theta^*)$$

Task 2: Speaker Retrieval

- Mean Average Precision (mAP)

➢ *Precision* of top-***k*** for a speaker ***i***

$$Precision(i,k) = \frac{\sum_{j=1}^{k} \delta(utt_j \ is \ from \ speaker \ i)}{k}$$

➢ *Average Precision* of top-***N***

$$AP(i) = \frac{1}{N} \sum_{k=1}^{N} Precision(i,k)$$

➢ *Mean AP over all **S***

$$mAP = \frac{1}{S} \sum_{i=1}^{S} AP(i)$$

Baselines – ASV-Subtools for Seniors

ABOUT *ASV-Subtools*

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

ASV-Subtools

ASV-Subtools: An Open Source Tools for Speaker Recognition

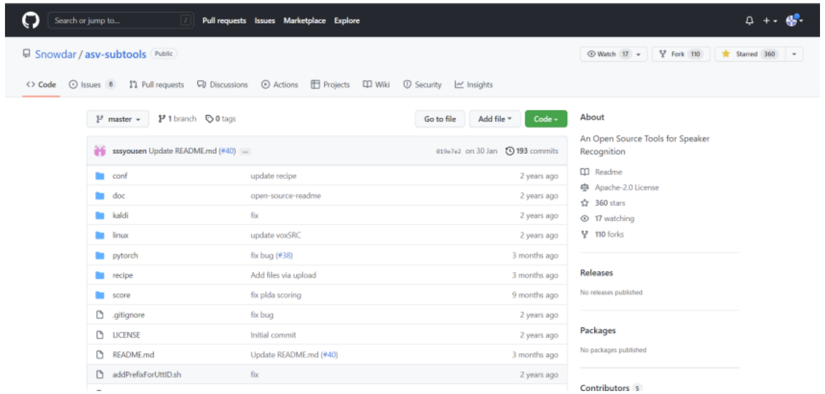
ASV-Subtools is developed based on Pytorch and Kaldi for the task of speaker recognition, language identification, etc.
The 'sub' of 'subtools' means that there are many modular tools and the parts constitute the whole.

> Copyright: XMU Speech Lab (Xiamen University, China)
> Apache 2.0
>
> Author : Miao Zhao (Email: snowdar@stu.xmu.edu.cn), Jianfeng Zhou, Zheng Li, Hao Lu, Fuchuan Tong, Tao Jiang
> Current Maintainer: Fuchuan Tong (Email: 1017549629@qq.com)
> Co-author: Lin Li, Qingyang Hong

Citation:

```
@inproceedings{tong2021asv,
  title={{ASV-Subtools}: {Open} Source Toolkit for Automatic Speaker Verification},
  author={Tong, Fuchuan and Zhao, Miao and Zhou, Jianfeng and Lu, Hao and Li, Zheng and Li, Lin and Hong, Qin
  booktitle={ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASS
  pages={6184--6188},
  year={2021},
  organization={IEEE}
}
```

TalentedSoft 天聪智能软件



Prepared Baselines

- Data processing
  - Data/Spec Augment
  - VAD
- Feature selection
  - Fbanks (80)
- Backbone
  - ResNet34SE
- Backend
  - Cosine similarity

Neural Backbone

Layer	Module	Output
Input	–	F×T×1
Conv2D	3×3×32, Stride 1	F×T×32
ResNet Block-1	$\begin{bmatrix}3\times3\times32\\3\times3\times32\\\text{SE Layer}\end{bmatrix}\times3$, Stride 1	F×T×32
ResNet Block-2	$\begin{bmatrix}3\times3\times64\\3\times3\times64\\\text{SE Layer}\end{bmatrix}\times4$, Stride 2	$[\frac{F}{2}]\times[\frac{T}{2}]\times64$
ResNet Block-3	$\begin{bmatrix}3\times3\times128\\3\times3\times128\\\text{SE Layer}\end{bmatrix}\times6$, Stride 2	$[\frac{F}{4}]\times[\frac{T}{4}]\times128$
ResNet Block-4	$\begin{bmatrix}3\times3\times256\\3\times3\times256\\\text{SE Layer}\end{bmatrix}\times3$, Stride 2	$[\frac{F}{8}]\times[\frac{T}{8}]\times256$
Pooling	TSP	$2\times[\frac{F}{8}]\times256$
Dense	–	256
Dense	AM-Softmax	2793

Tasks	Training	Evaluation	Metrics
Task 1 SV	CN-Celeb.T	CN-Celeb.E	minDCF: 0.463, EER: 9.141%
Task 2 SR		SR.eval	mAP: 0.242

Baselines – Sunine for Juniors

ABOUT *Sunine*

- *https://gitlab.com/csltstu/sunine*

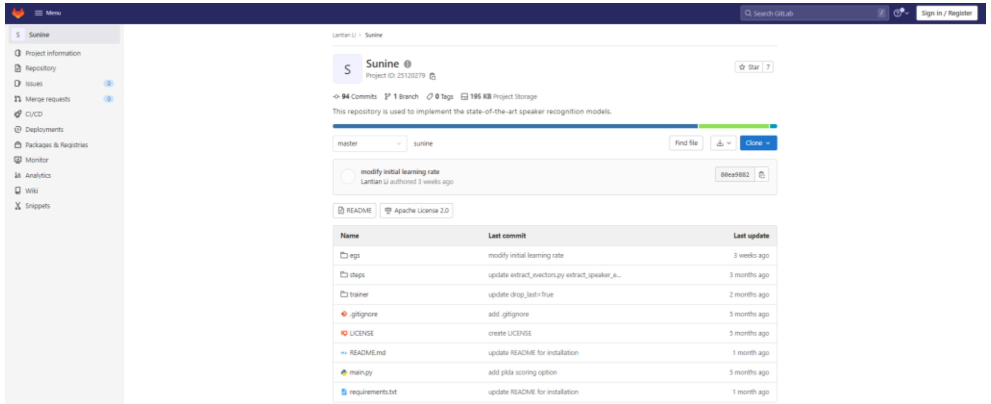
Sunine: THU-CSLT Speaker Recognition Toolkit

Center for Speech and Language Technologies GROUPING

Sunine is an **open-source** speaker recognition toolkit based on PyTorch.

The goal is to create a **user-friendly** toolkit that can be used to easily develop **state-of-the-art speaker recognition technologies**.

> Copyright: THU-CSLT (Tsinghua University, China)
> Apache License, Version 2.0 LICENSE
>
> Authors : Lantian Li (lilt@cslt.org), Yang Zhang (zhangy20@mails.tsinghua.edu.cn)
> Co-author: Dong Wang (wangdong99@mails.tsinghua.edu.cn)



Prepared Baselines

- Data processing
  - NULL
- Feature selection
  - Fbanks (80)
- Backbone
  - ResNet34SE
  - Attentive pooling
- Backend
  - Cosine similarity

Neural Backbone

Layer	Module	Output
Input	–	F×T×1
Conv2D	3×3×32, Stride 1	F×T×32
ResNet Block-1	$\begin{bmatrix} 3\times3\times32 \\ 3\times3\times32 \\ \text{SE Layer} \end{bmatrix}\times3$, Stride 1	F×T×32
ResNet Block-2	$\begin{bmatrix} 3\times3\times64 \\ 3\times3\times64 \\ \text{SE Layer} \end{bmatrix}\times4$, Stride 2	$[\frac{F}{2}]\times[\frac{T}{2}]\times64$
ResNet Block-3	$\begin{bmatrix} 3\times3\times128 \\ 3\times3\times128 \\ \text{SE Layer} \end{bmatrix}\times6$, Stride 2	$[\frac{F}{4}]\times[\frac{T}{4}]\times128$
ResNet Block-4	$\begin{bmatrix} 3\times3\times256 \\ 3\times3\times256 \\ \text{SE Layer} \end{bmatrix}\times3$, Stride 2	$[\frac{F}{8}]\times[\frac{T}{8}]\times256$
Pooling	ASP	$2\times[\frac{F}{8}]\times256$
Dense	–	256
Dense	AM-Softmax	2793

Tasks	Training	Evaluation	Metrics
Task 1 SV	CN-Celeb.T	CN-Celeb.E	minDCF: 0.549, EER: 10.611%
Task 2 SR		SR.eval	mAP: 0.152

OUTLINE

- Data, Tasks and Baselines
- **Technical Summary**
- System Analysis
- The Next CNSRC

Representative Techniques (16 Teams)

Components	Methods
Data processing	SpecAugment (Time/Frequency masking), Speed perturbation, Noise & Music & Reverberation & Babble augmentation, Short-clip concatenation/filtering
Feature selection	FBank, MFCC, PCEN, Energy, Spectrogram plus CMN, pre-trained WavLM as feature extractor
Neural backbone	ECAPA-TDNN variants (dynamic/multi-scale convolution, multi-scale attention, various ResBlocks), ResNet family (34/74/101/152/221/293) with SE, Split-attention, Gated attention RepVGG, Hybrid NN (CNN/TDNN/LSTM), Transformer
Pooling strategy	Multi-query Multi-head attention pooling, Global-local statistic pooling, SPoC pooling
Auxiliary design	Gradient reversal layer, Genre embedding, Mutual information, Data uncertainty learning
Loss function	Margin-based loss (AM, AAM, AdaFace) with Subcenter constraint, Inter-TopK penalty, Circle loss
Training strategy	Multi-stage training (e.g., Chunk size increasing, Large margin finetuning), LR schedulers (ReduceLROnPlateau/CyclicLR)
Backend scoring	Cosine (αQE), PLDA-(diag), Attention back-end + AS-Norm + QMF/music calibration
System fusion	Score-level average, Embedding-level ensemble

Technical Highlights

Components	Methods
Data processing	SpecAugment (Time/Frequency masking), Speed perturbation, Noise & Music & Reverberation & Babble augmentation, Short-clip concatenation/filtering
Feature selection	FBank, MFCC, PCEN, Energy, Spectrogram plus CMN, pre-trained WavLM as feature extractor
Neural backbone	ECAPA-TDNN variants (dynamic/multi-scale convolution, multi-scale attention, various ResBlocks), ResNet family (34/74/101/152/221/293) with SE, Split-attention, Gated attention RepVGG, Hybrid NN (CNN/TDNN/LSTM), Transformer
Pooling strategy	Multi-query Multi-head attention pooling, Global-local statistic pooling, SPoC pooling
Auxiliary design	Gradient reversal layer, Genre embedding, Mutual information, Data uncertainty learning
Loss function	Margin-based loss (AM, AAM, AdaFace) with Subcenter constraint, Inter-TopK penalty, Circle loss
Training strategy	Multi-stage training (e.g., Chunk size increasing, Large margin finetuning), LR schedulers (ReduceLROnPlateau/CyclicLR)
Backend scoring	Cosine (αQE), PLDA-(diag), Attention back-end + AS-Norm + QMF/music calibration
System fusion	Score-level average, Embedding-level ensemble

Speed perturbation

Components	Methods
Data processing	SpecAugment (Time/Frequency masking), Speed perturbation, Noise & Music & Reverberation & Babble augmentation, Short-clip concatenation/filtering
Feature	n p
Neural b	multi-sca 21/293) LSTM), T
Pooling strategy	Multi-query Multi-head attention pooling, (
Auxiliary design	Gradient reversal layer, Genre embedding,
Loss function	Margin-based loss (AM, AAM, AdaFace) wi
Training strategy	Multi-stage training (e.g., Chunk size increasing, Large margin finetuning), LR schedulers (ReduceLROnPlateau/CyclicLR)
Backend scoring	Cosine (αQE), PLDA-(diag), Attention back-end + AS-Norm + QMF/music calibration
System fusion	Score-level average, Embedding-level ensemble

The utterance with a new speed will be considered from a new speaker.

T022, T082, T102, T106, T121

Results from T121

System Index	*Features*	*EER*	*MinDCF(0.01)*
1	MFCC-AM	9.930	0.5045
2	Fbank-AM	10.18	0.4894
3	Fbank-Sub-centers	10.54	0.4672
4	Fbank(Speed Perturbation)-AM	8.696	0.4607
5	Spectrogram-AM	10.90	0.4717

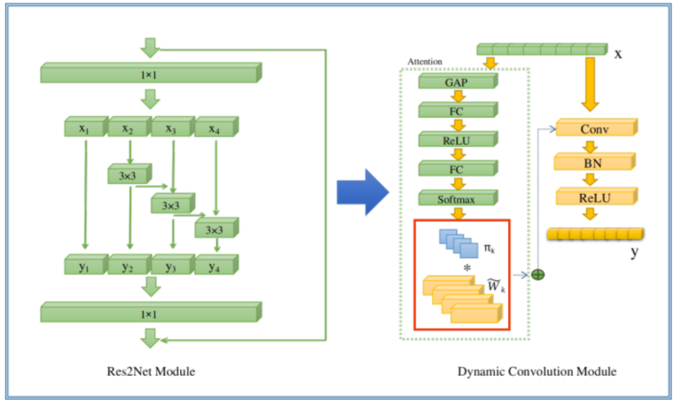
[a] Scoring with cosine distance.

ECAPA-TDNN variants

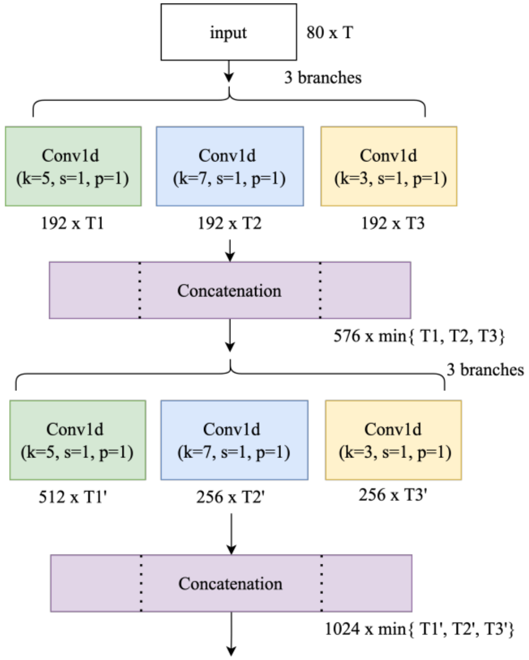
Components	Methods
Data processing	SpecAugment (Time/Fre… …erberation & Babble augmentation, Sh…
Feature selection	FBank, MFCC, PCE… Ene… …xtractor
	ECAPA-TDNN variants (dynamic/multi-scale convolution, multi-scale attention, various ResBlocks),

ECAPA-TDNN: one of the most favorite backbones.
- Res2Net (multi-scale receptive fields)
- Squeeze-and-Excitation (channel interaction)

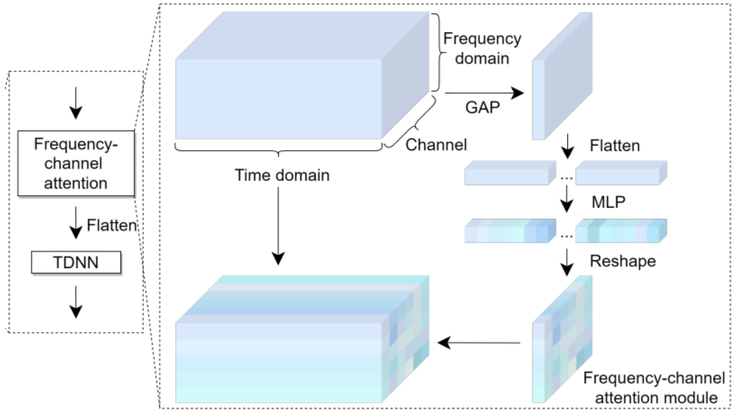
T009: DyNet replaces Res2Net



T070: Add Multi-scale convolution



T106: Adopt Multi-scale MFA



ResNet family

Components	Methods
Data processing	SpecAugment (Time/Frequency masking), Speed perturbation, Noise & Music & Reverberation & Babble augmentation, Short-clip concatenation/filtering
Feature selection	, pre-trained WavLM as feature extractor
Neural backbone	ECAPA- N variants (dynamic/multi-scale convolution, multi-scale attention, various ResBlocks), ResNet family (34/74/101/152/221/293) with SE, Split-attention, Gated attention

ResNet: one of the most favorite backbones.

T022/T082: Growing deeper

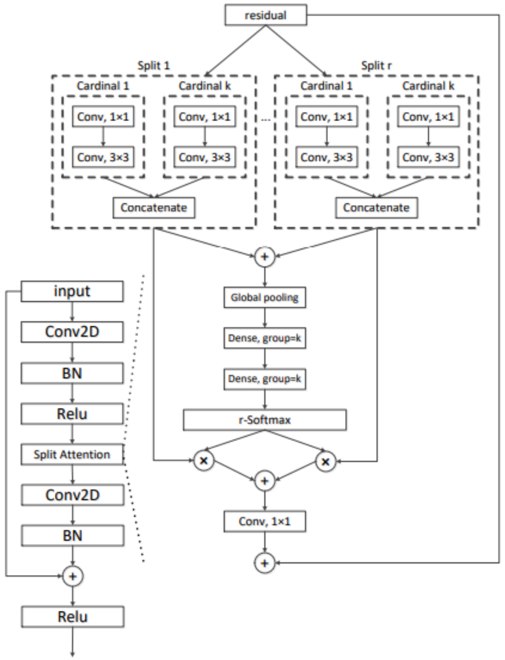
T022

5*System	CN-Celeb.E		CN-Celeb.E (submean asnorm)	
	eer	minc	eer	minc
ResNet34	7.8231	0.3755	7.4571	0.3518
Ecapa-tdnnL	8.8257	0.4086	8.6623	0.3914
ResNet74	7.7274	0.3785	7.3162	0.3475
RepVGG_A2	7.7387	0.3681	7.4233	0.3460
ResNet101	6.2518	0.3619	6.1335	0.3358
fused	-	-	5.9530	0.3185

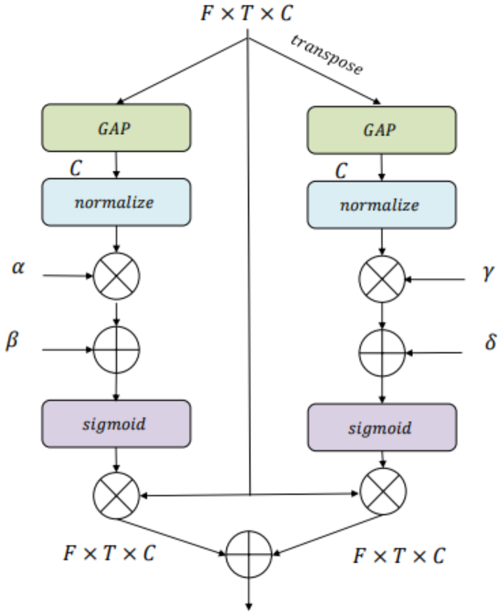
T082

System	Params #	minDCF (0.01)	EER (%)
ResNet34 *	6.63M	0.3958	7.981
ResNet34	6.63M	0.3707	6.590
ResNet152	19.8M	0.3386	5.762
ResNet221	23.8M	0.3270	**5.543**
ResNet293	28.6M	**0.3202**	5.553
DF-ResNet	14.8M	0.3361	6.279

T053: Split-attention block



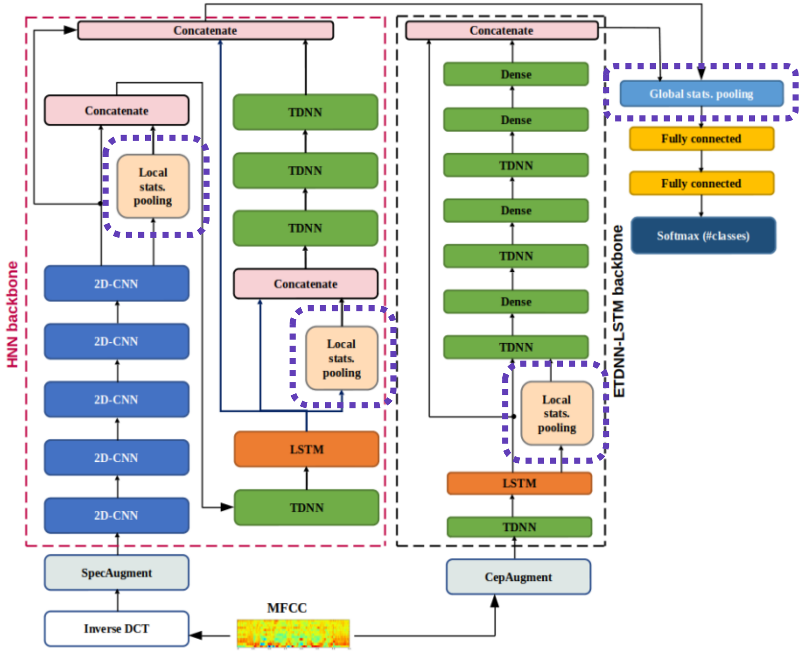
T106: Gated F-C attention module



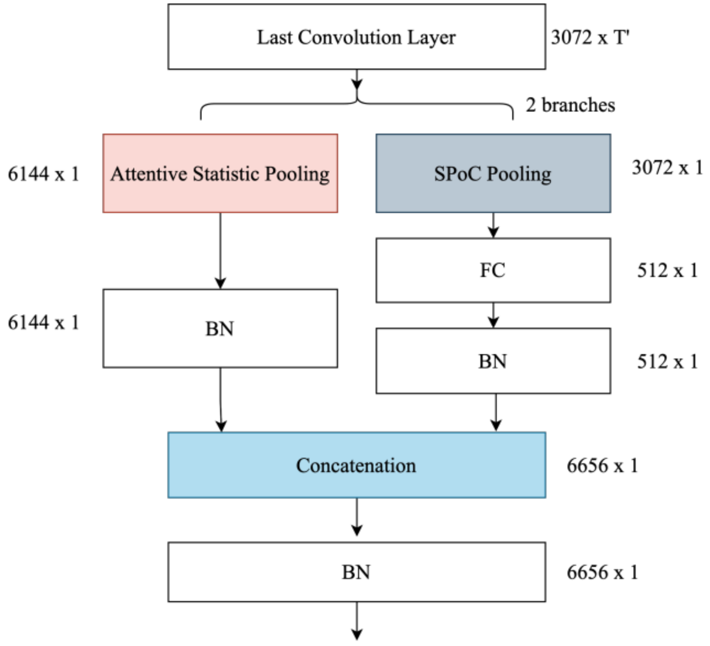
Pooling strategy

Components	Methods
Data processing	SpecAugment (Time/Frequency masking), Speed perturbation, Noise & Music & Reverberation & Babble augmentation, Short-clip concatenation/filtering
Feature selection	ctor
Neural backbone	ResBlocks),
Pooling strategy	
Auxiliary design	
Loss function	rcle loss
Training strategy	
Backend scoring	
System fusion	Score-level average, Embedding-level ensemble

T042: Global-Local statistic pooling in Hybrid backbone



T070: Integrate Attentive statistic pooling and SPoC pooling



Auxiliary design

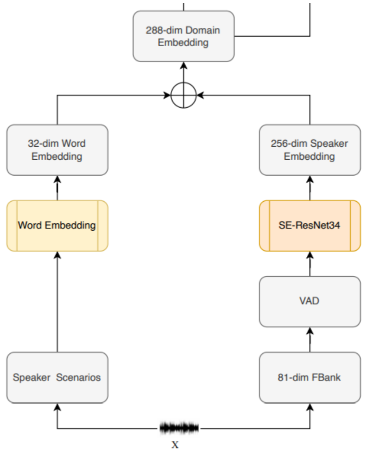
Components	Methods
Data processing	…tion &
Feature selection	…or
Neural backbone	…sBlocks),
Pooling strategy	
Auxiliary design	
Loss function	…le loss
Training strategy	
Backend scoring	
System fusion	

Goal: to eliminate genre interference and learn genre-invariant speaker representation.

T009: Gradient reversal layer

(b) ECAPA-TDNN with Auxiliary Task

T028: Auxiliary genre embedding



T042: Minimize the *MI* between speaker embedding and genre label

$$L_{nuisance} = E_{p(\omega,c)}[\log p_\omega(\omega|c)] - E_{p(\omega)p(c)}[\log p_\omega(\omega|c)]$$

$$L_{MIM-DG} = -L_{speaker} + \beta L_{nuisance}$$

T106: Data uncertainty learning

$$e_i = \mu_i + \epsilon\delta_i, \epsilon \in N(0, I)$$

$$L_{DUL} = L_{class}(e_i) + \lambda KL(N(z_i|\mu_i, \delta_i^2)||N(\epsilon|0, I))$$

Loss function

Subcenter: to alleviate the effect of noisy and low-quality samples

T022/T102/T106/T121

$$\cos\left(\theta_{i,j}\right) = \max_{1 \le k \le K} \left(\|\boldsymbol{z}_i\| \cdot \|\boldsymbol{W}_{j,k}\|\right)$$

Inter-TopK or AdaFace: to choose sample / margin based on data quality.

T022/T102: Inter-TopK penalty

$$\phi(\theta_j) = \begin{cases} cos(\theta_j + m), & j \in \underset{1 \le n \le N}{\arg topK}(cos(\theta_j)) \\ cos(\theta_j). & Others \end{cases}$$

T106: AdaFace-Softmax

$$f\left(\theta_{i,j}, m\right)_{\text{AdaFace}} = \begin{cases} s\cos\left(\theta_{i,j} + g_{\text{angle}}\right) - g_{\text{add}} & j = y_i \\ s\cos\theta_{i,j} & j \neq y_i \end{cases}$$

Components	Methods
Data processing	
Feature selection	
Neural backbone	
Pooling strategy	
Auxiliary design	
Loss function	Margin-based loss (AM, AAM, AdaFace) with Subcenter constraint, Inter-TopK penalty, Circle loss
Training strategy	Multi-stage training (e.g., Chunk size increasing, Large margin finetuning), LR schedulers (ReduceLROnPlateau/CyclicLR)
Backend scoring	Cosine (αQE), PLDA-(diag), Attention back-end + AS-Norm + QMF/music calibration
System fusion	Score-level average, Embedding-level ensemble

Backend scoring

Components	Methods
Data processing	
Feature selection	
Neural backbone	
Pooling strategy	
Auxiliary design	
Loss function	
Training strategy	
Backend scoring	Cosine (αQE), PLDA-(diag), Attention back-end + AS-Norm + QMF/music calibration
System fusion	Score-level average, Embedding-level ensemble

αQE and attention back-end: to leverage multiple enrollment utterances.

T057: αQE (alpha query expansion) to enrollment aggregation at test time

Ablation Study on ResNet34(Fbank) System

	EER(%)	MinDCF
ResNet34(Fbank)	9.94	0.5261
+αQE	8.80	0.4323
++AS-Norm	**8.31**	**0.4041**

QMF: quality-aware calibration based on quality metric function

T053: QMF of duration

$$\hat{S} = S + C \cdot f(d_t) \qquad f(d_t) = \frac{1}{d_t}$$

T106: self-attention to learn the intra-relationship of enrollments

Sub-Systems	CN-Celeb.E(TTA) EER(%)	CN-Celeb.E(TTA) minDCF
Speaker-ViT	9.012	0.4318
S1: Speaker-ViT + back-end	7.282	0.4240
ResNet-Att	7.502	0.4234
S2: ResNet-Att + back-end	7.187	0.4039
CNN-ECAPA-TDNN	8.037	0.4137
S3: CNN-ECAPA-TDNN + back-end	6.928	0.4048
MFA-ECAPA	8.237	0.4162
S4: MFA-ECAPA + back-end	**6.860**	0.3974
CNN-Speaker-ViT	7.851	0.3957
S5: CNN-Speaker-ViT + back-end	6.888	**0.3832**

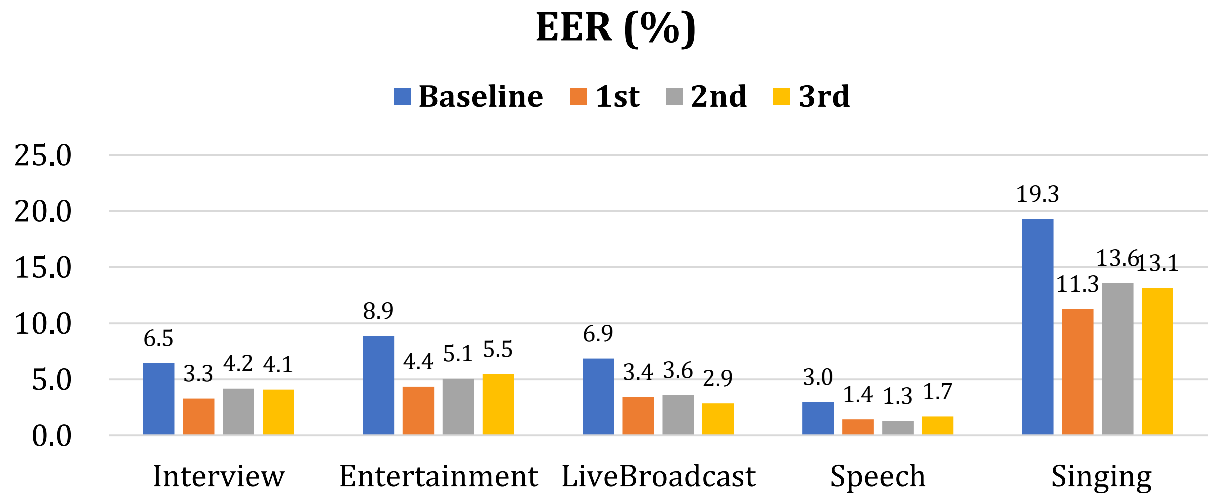
OUTLINE

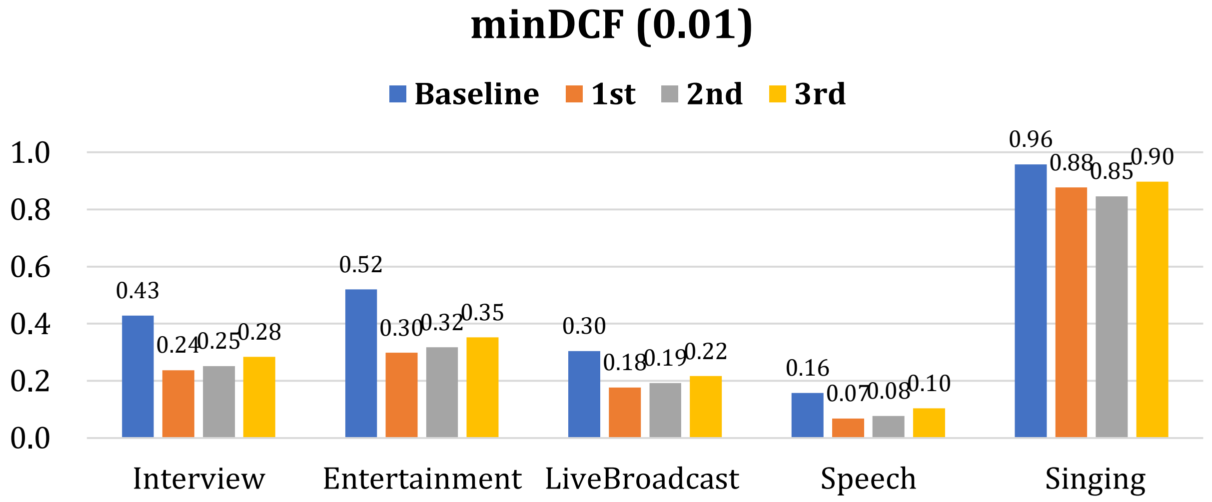
- Data, Tasks and Baselines
- Technical Summary
- **System Analysis**
- The Next CNSRC

Task 1 SV: Fixed Track

Multi-genre analysis

- Select 5 genres which have the most number of test trials.
- Make comparison amongst Baseline and Top-3 systems.

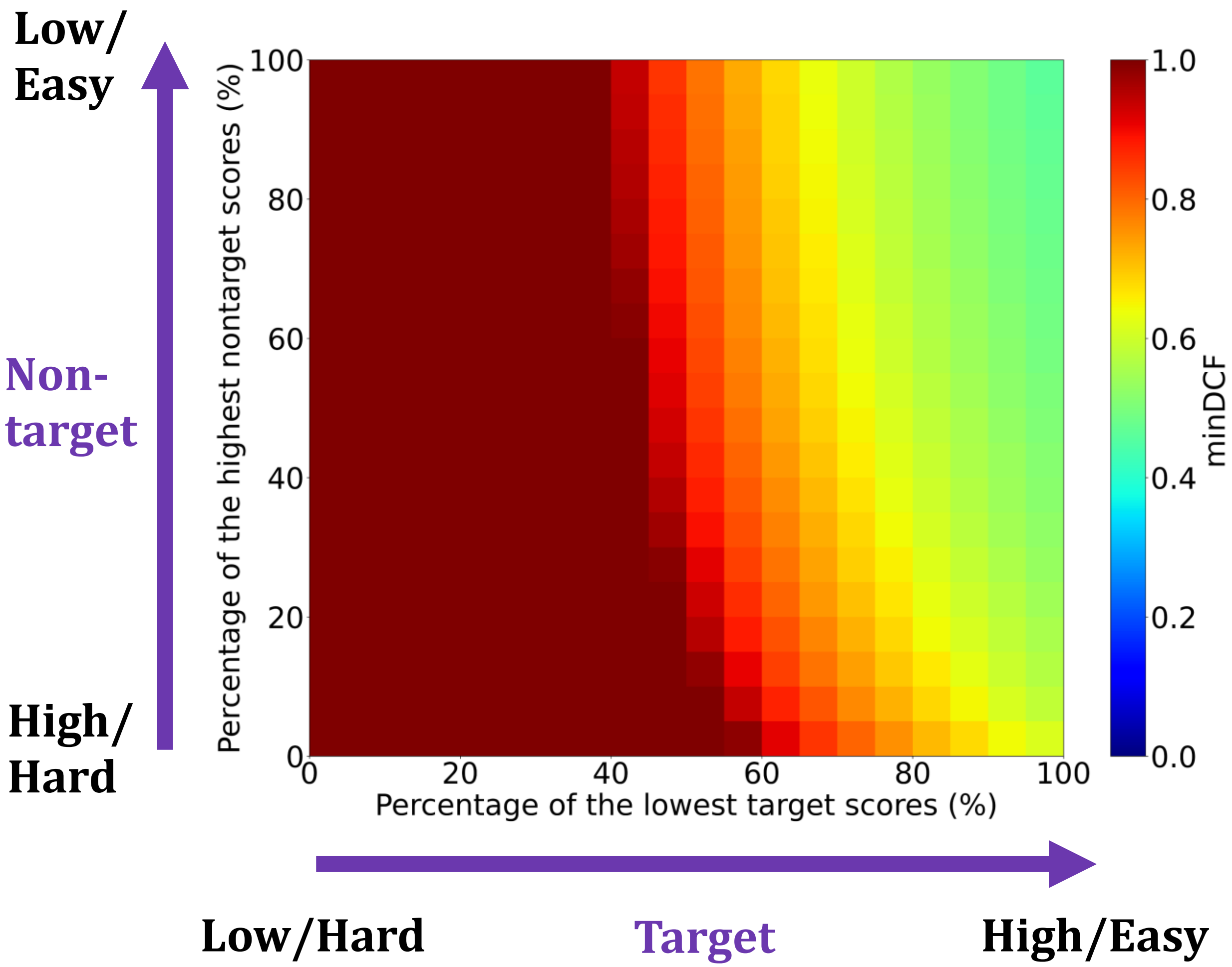




- Different genres present obviously different performance.
- Under different genres, the Top-3 systems show different advantage.

Task 1 SV: Fixed Track

- More fine-grained comparison with minDCF (0.01) via *C-P map*

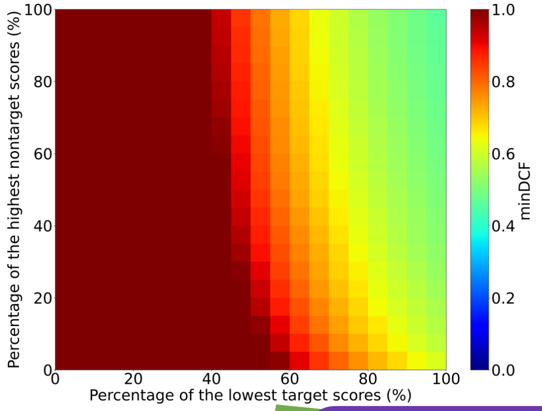


- Each point is a subset of the full trials, defined as **a trial config**.
- **x-axis**: scores of **Target** trials are *increased* from left to right. [from hard to easy]
- **y-axis**: scores of **Non-target** trials are *decreased* from bottom to up. [from hard to easy]
- The **color** in the map represents the metric values corresponding to each trial config.

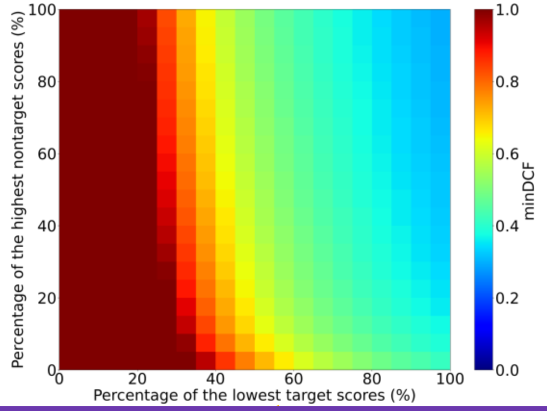
Task 1 SV: Fixed Track

- More fine-grained comparison with minDCF (0.01) via *C-P map*

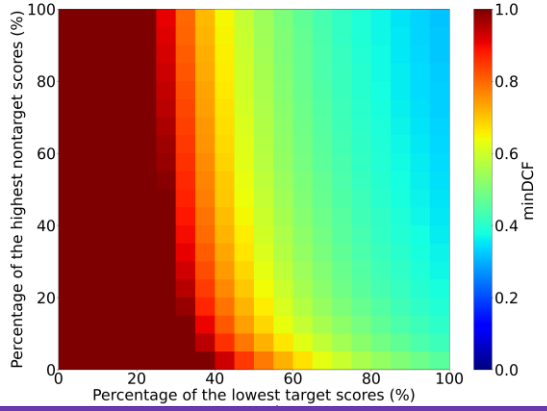
Baseline



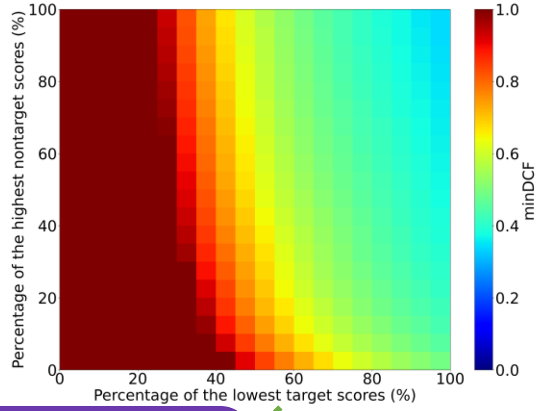
1st



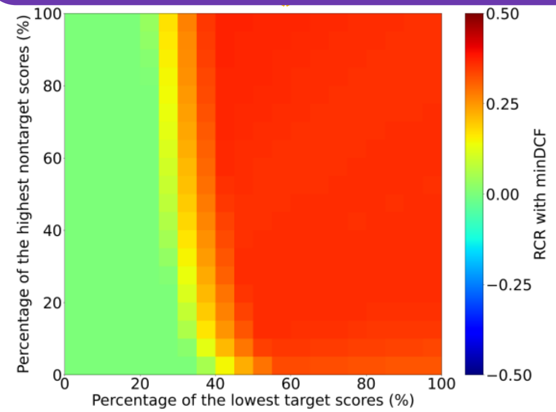
2nd

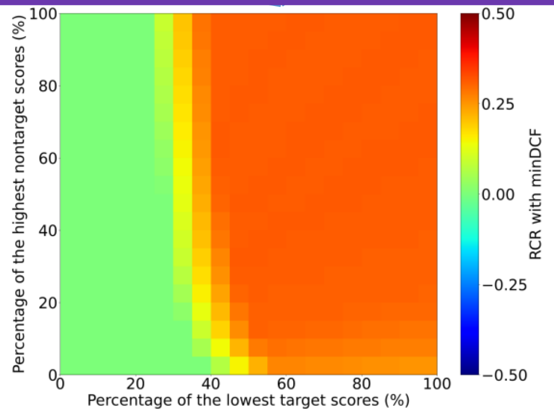


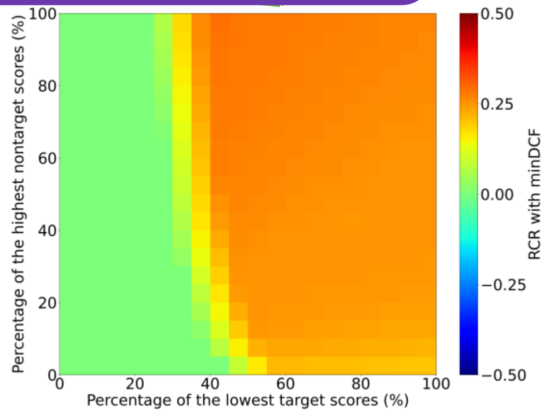
3rd



The Top-3 systems clearly outperform the Baseline system under *Easy/Normal* trial configs. However, this advantage becomes marginal under *Hard* trial configs.



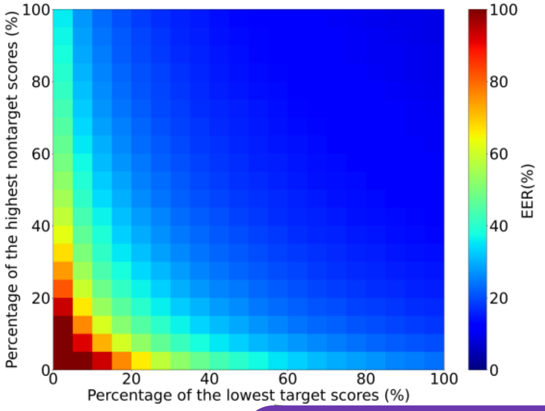




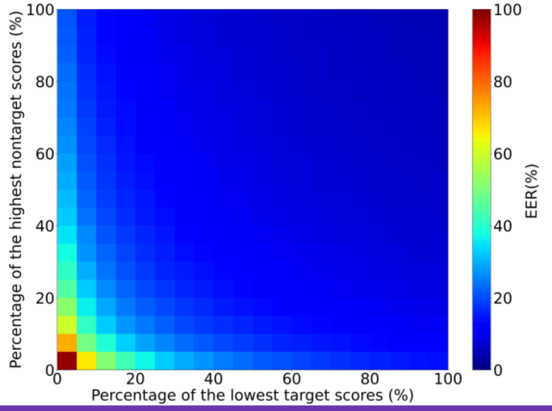
Task 1 SV: Fixed Track

- More fine-grained comparison with EER (%) via *C-P map*

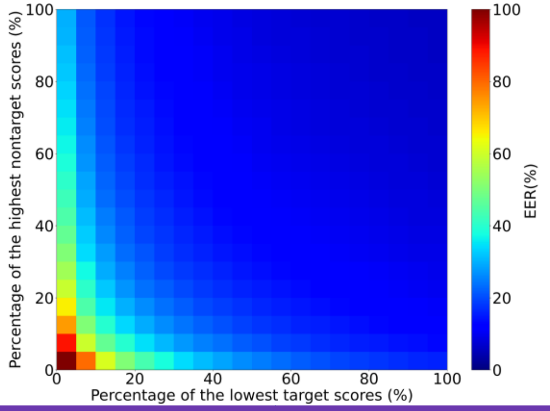
Baseline



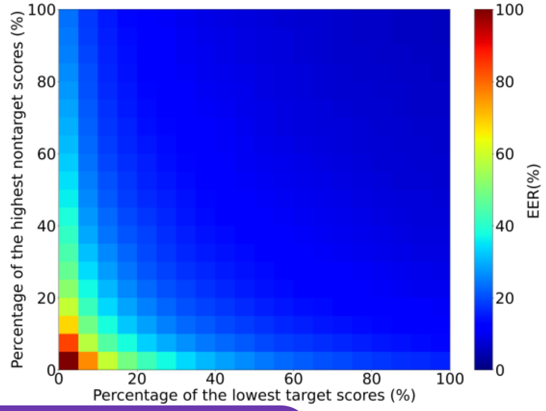
1^{st}



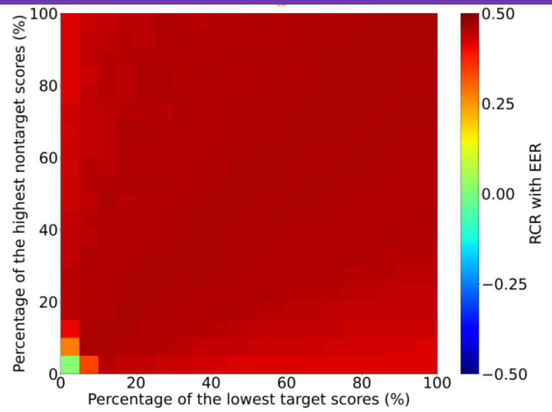
2^{nd}

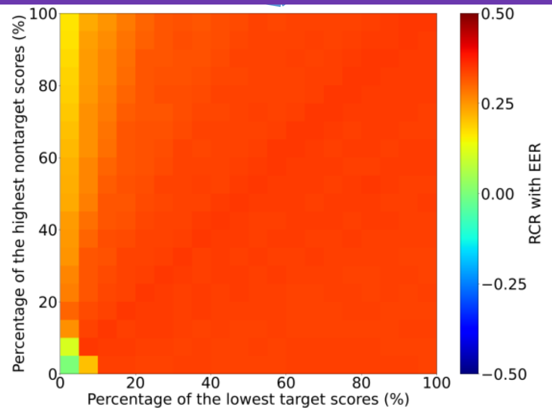


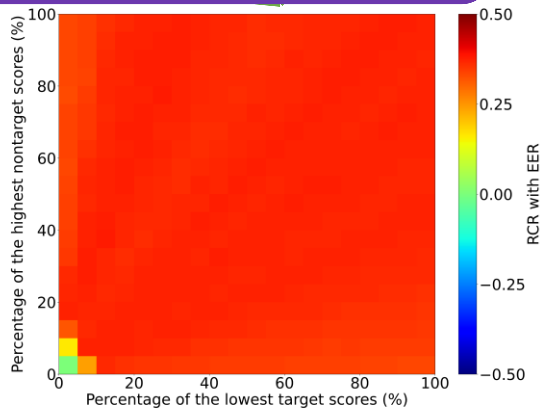
3^{rd}



The Top-3 systems outperform the Baseline system under a majority of trial configs.
The large proportion of high-performance area reveals that there are larger amount of *Easy* trials.

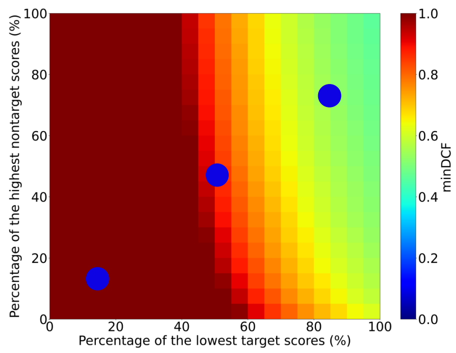






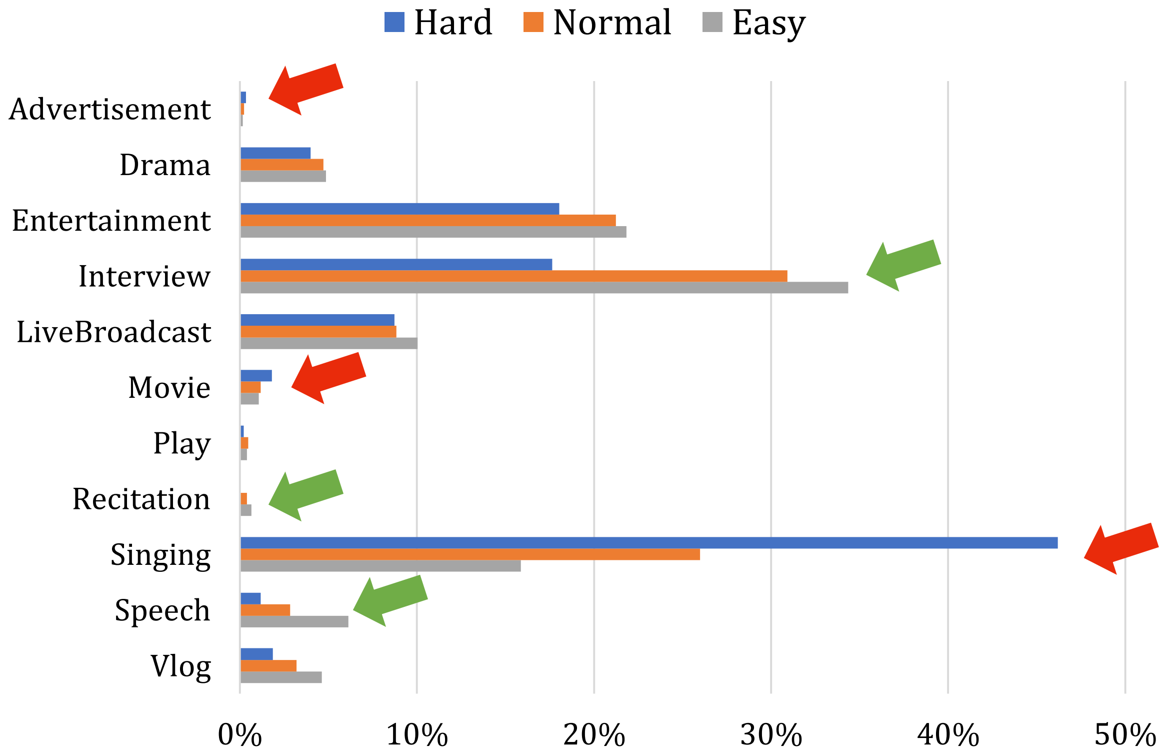
Task 1 SV: Fixed Track

- Genre distributions of Easy/Normal/Hard trial configs.
  - The genre distribution of target trials plays a key role.
  - Different genres show different levels of difficulty.

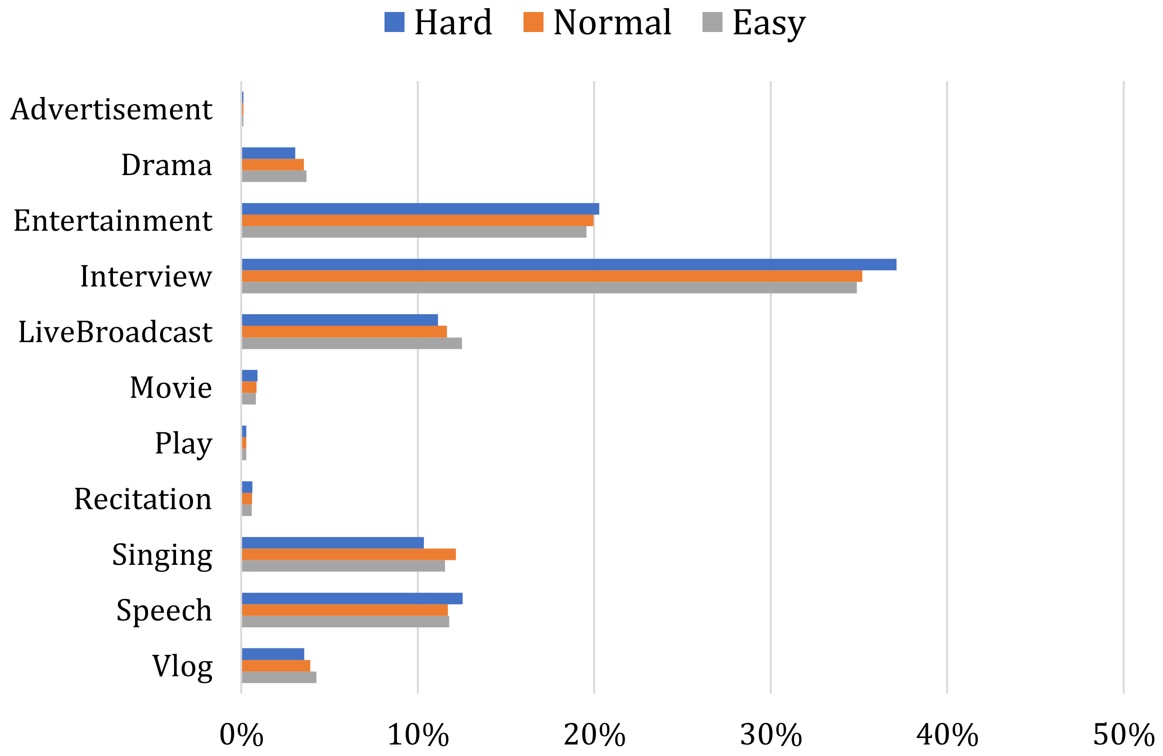


Baseline C-P Map

Genre distributions of three trial configs (Target trials)

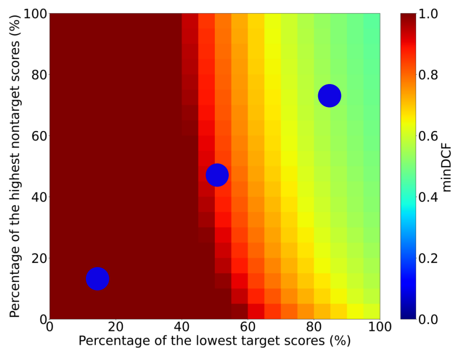


Genre distributions of three trial configs (Non-target trials)



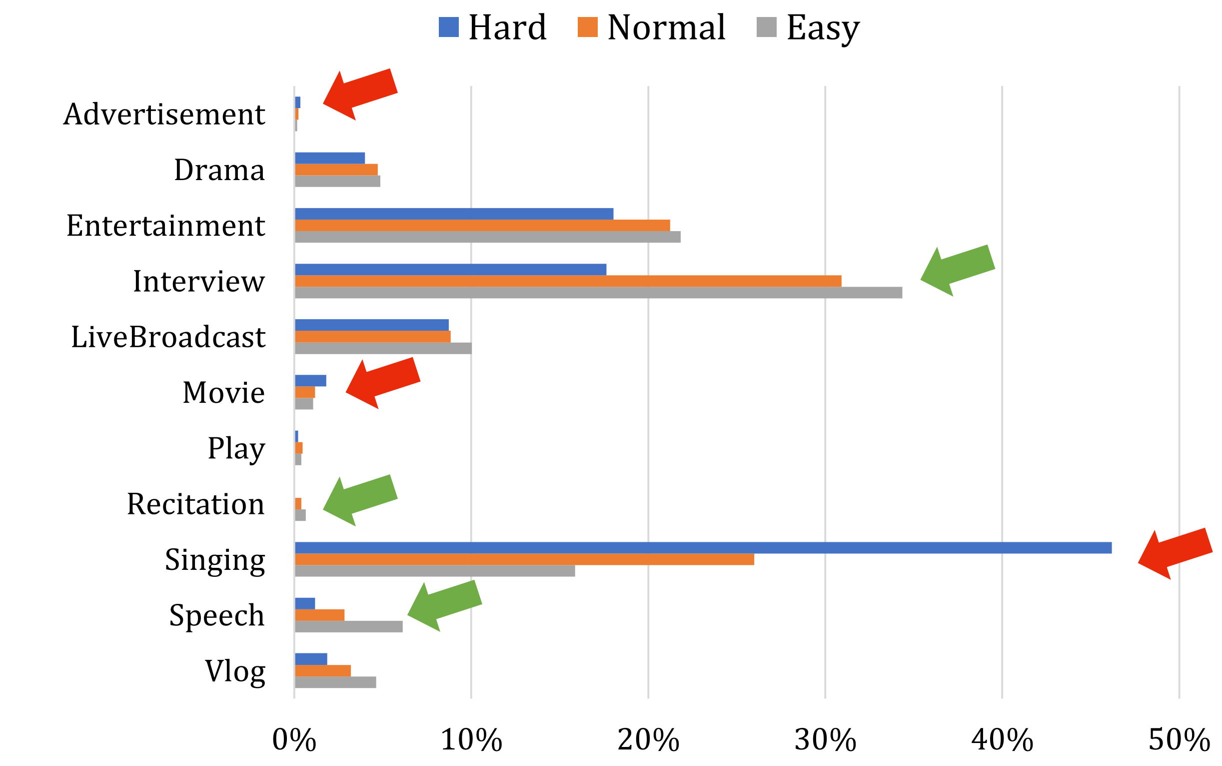
Task 1 SV: Fixed Track

- Genre distributions of Easy/Normal/Hard trial configs.
  - The genre distribution of target trials plays a key role.
  - Different genres show different levels of difficulty.

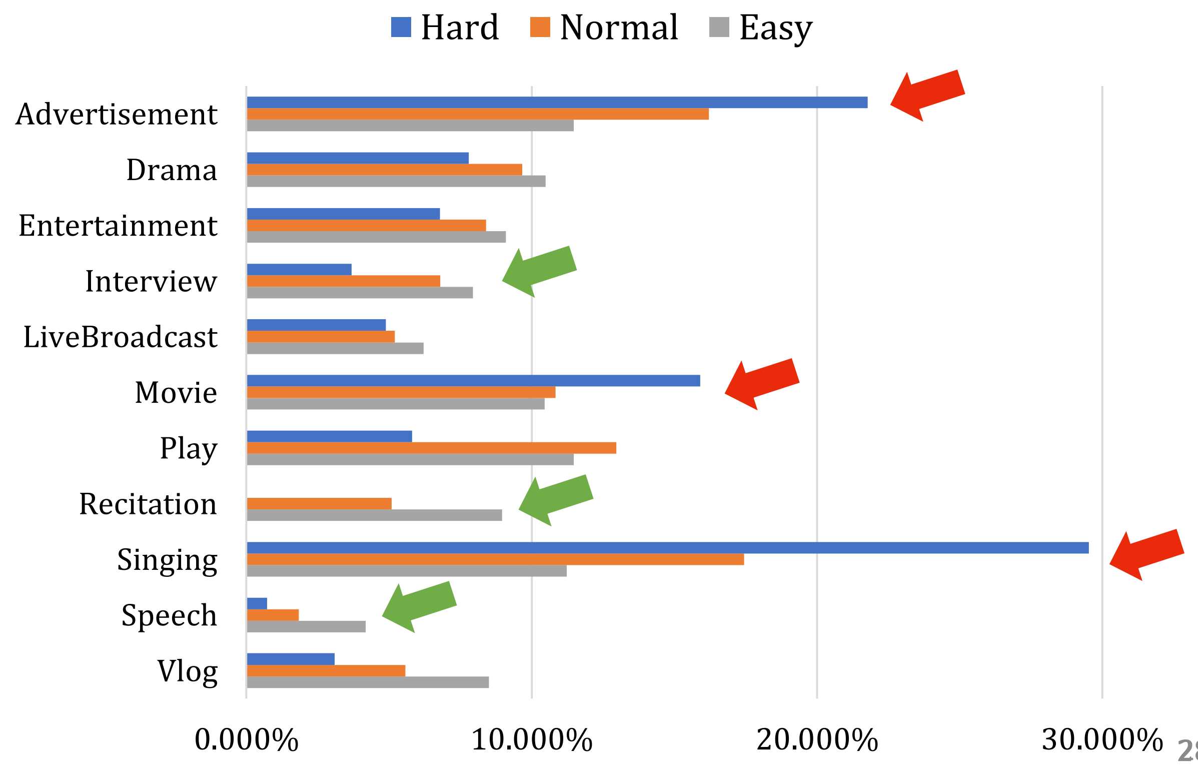


Baseline C-P Map

Genre distributions of three trial configs (Target trials)



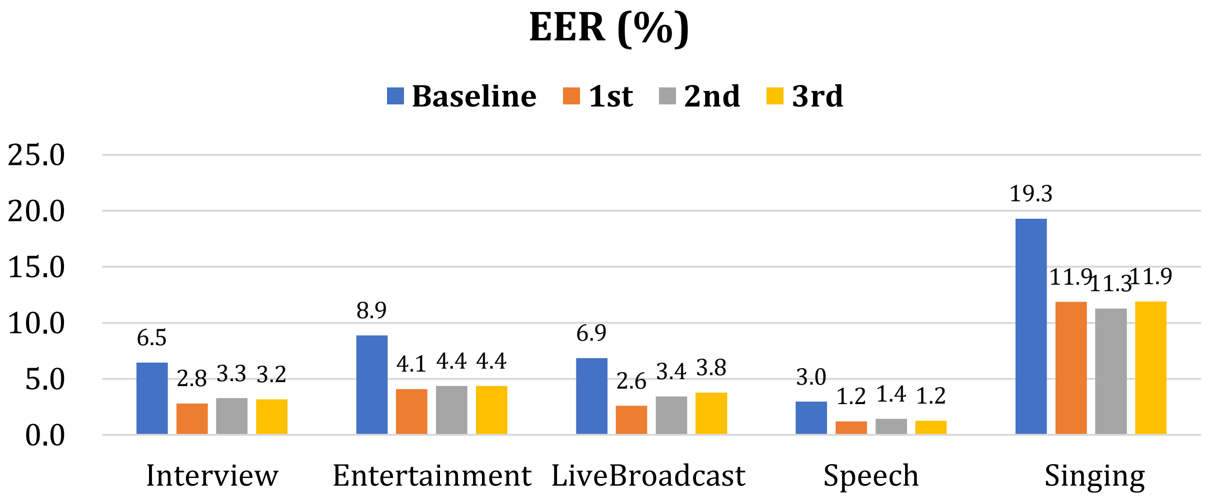
Genre distributions of three trial configs (Target trials) with Normalization

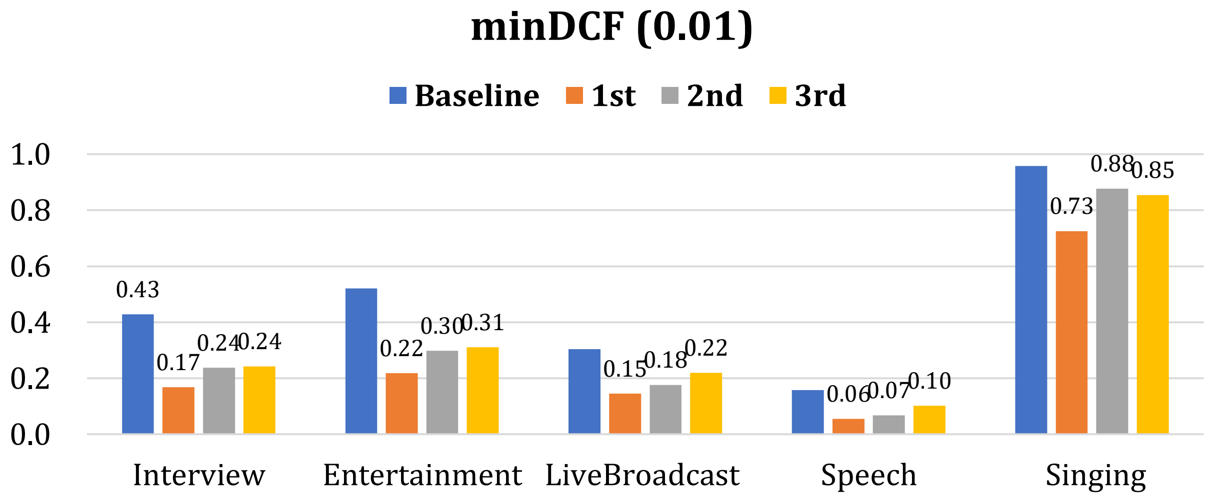


Task 1 SV: Open Track

Multi-genre analysis

- Select 5 genres which have the most number of test trials.
- Make comparison amongst Baseline and Top-3 systems.



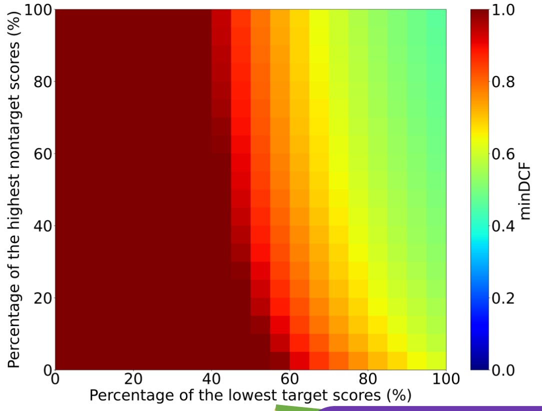


- Different genres present obviously different performance.
- Under different genres, the Top-3 systems show different advantage.

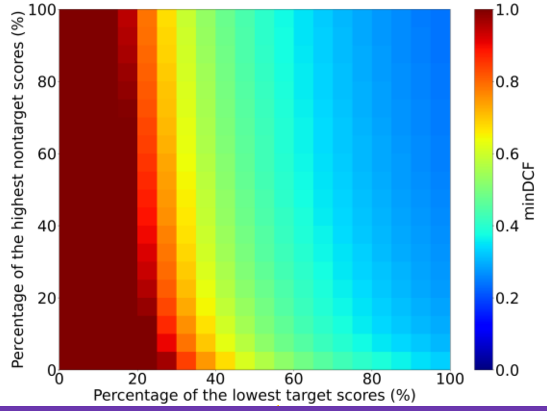
Task 1 SV: Open Track

- More fine-grained comparison with minDCF (0.01) via *C-P map*

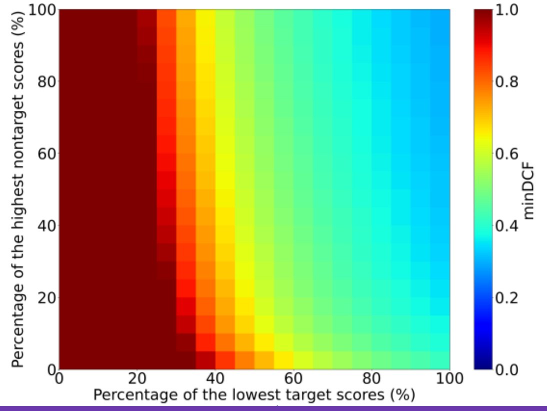
Baseline



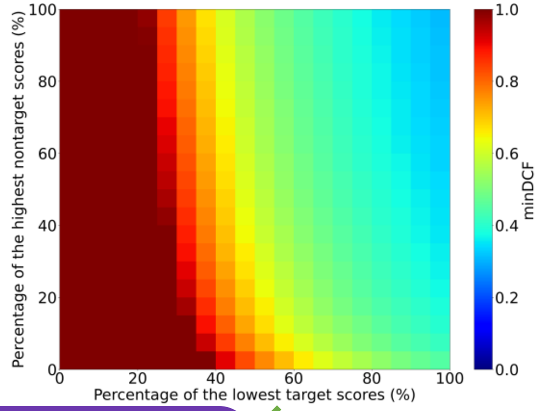
1st



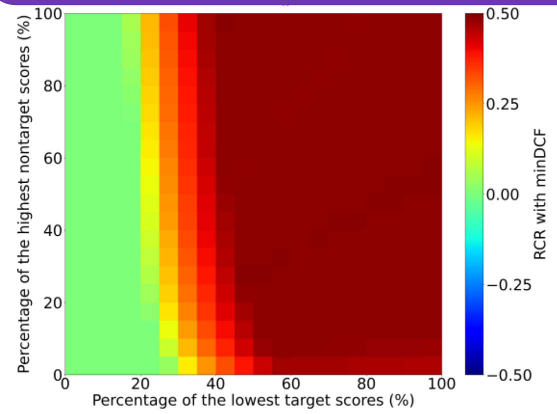
2nd

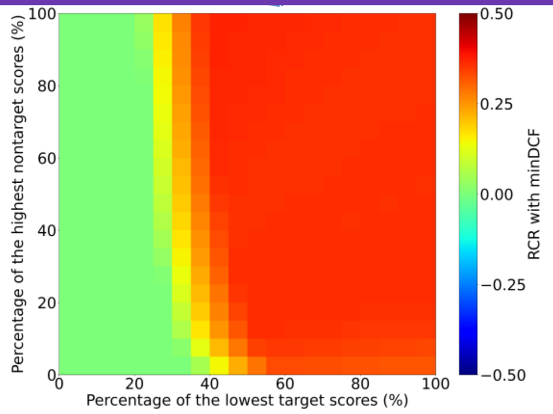


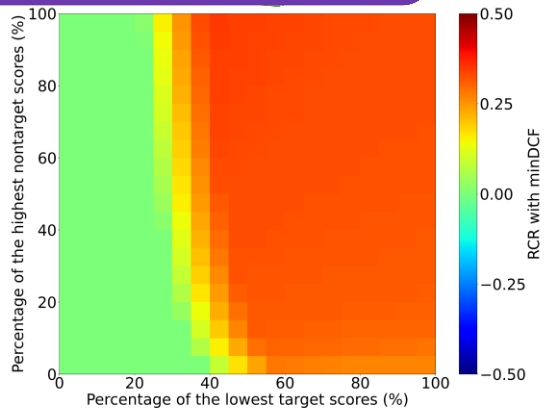
3rd



The Top-3 systems clearly outperform the Baseline system under *Easy/Normal* trial configs. However, this advantage becomes marginal under *Hard* trial configs.



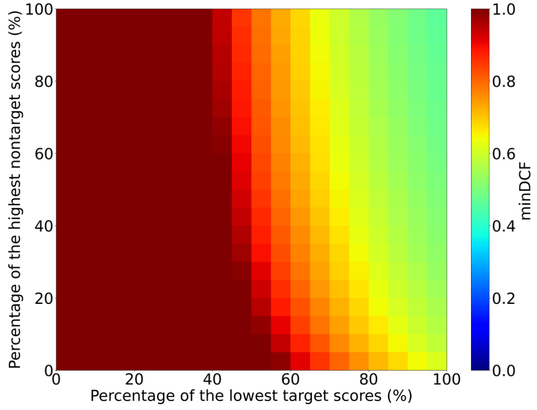




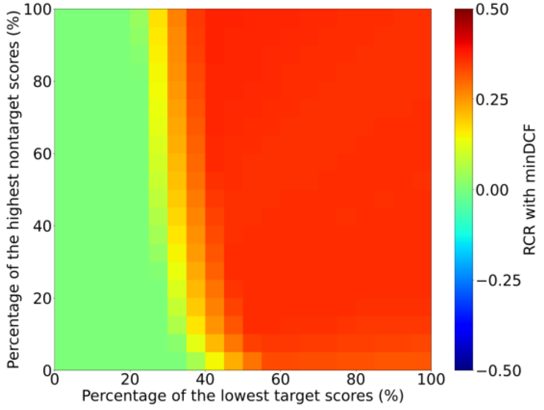
Task 1 SV: Fixed Track vs. Open Track

- Fixed/Open Comparison with minDCF (0.01) via *delta C-P map*

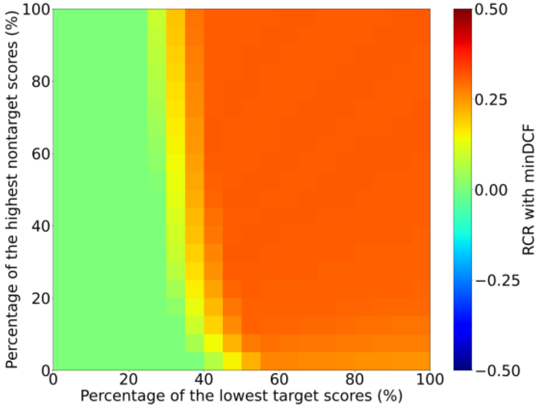
Baseline C-P Map



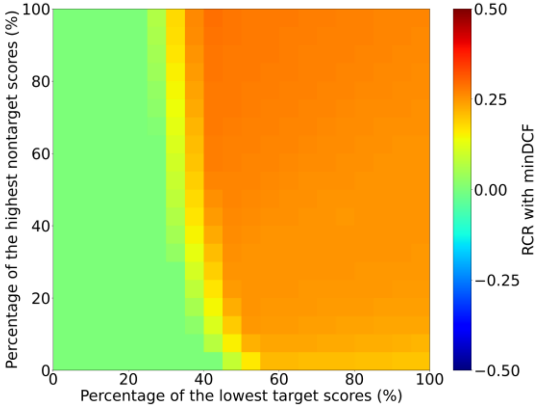
Fixed Track: 1st



Fixed Track: 2nd



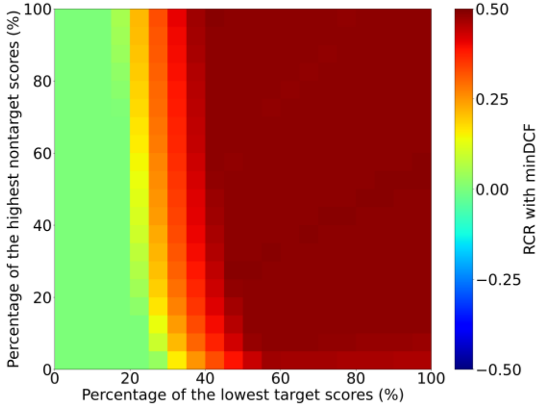
Fixed Track: 3rd



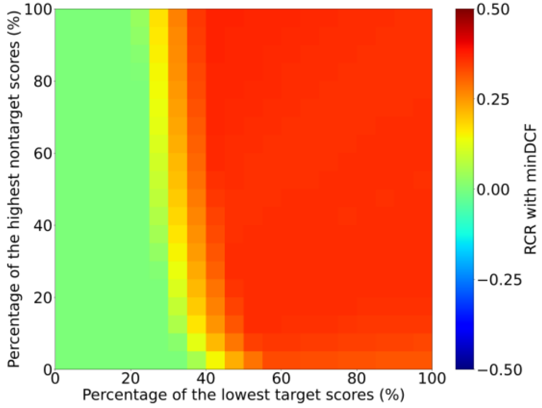
Systems in the Open track is superior to systems in Fixed track.

The more data, the better performance, especially on *E*asy/*N*ormal trial configs.

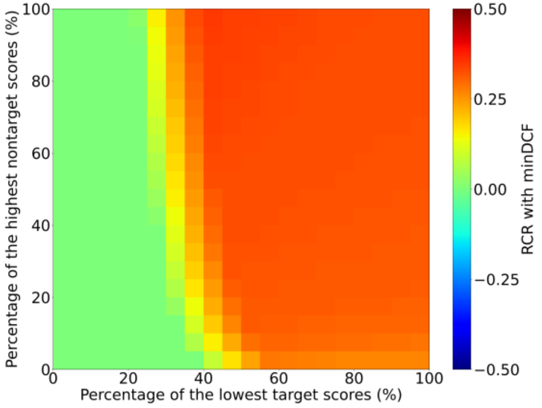
Open Track: 1st



Open Track: 2nd

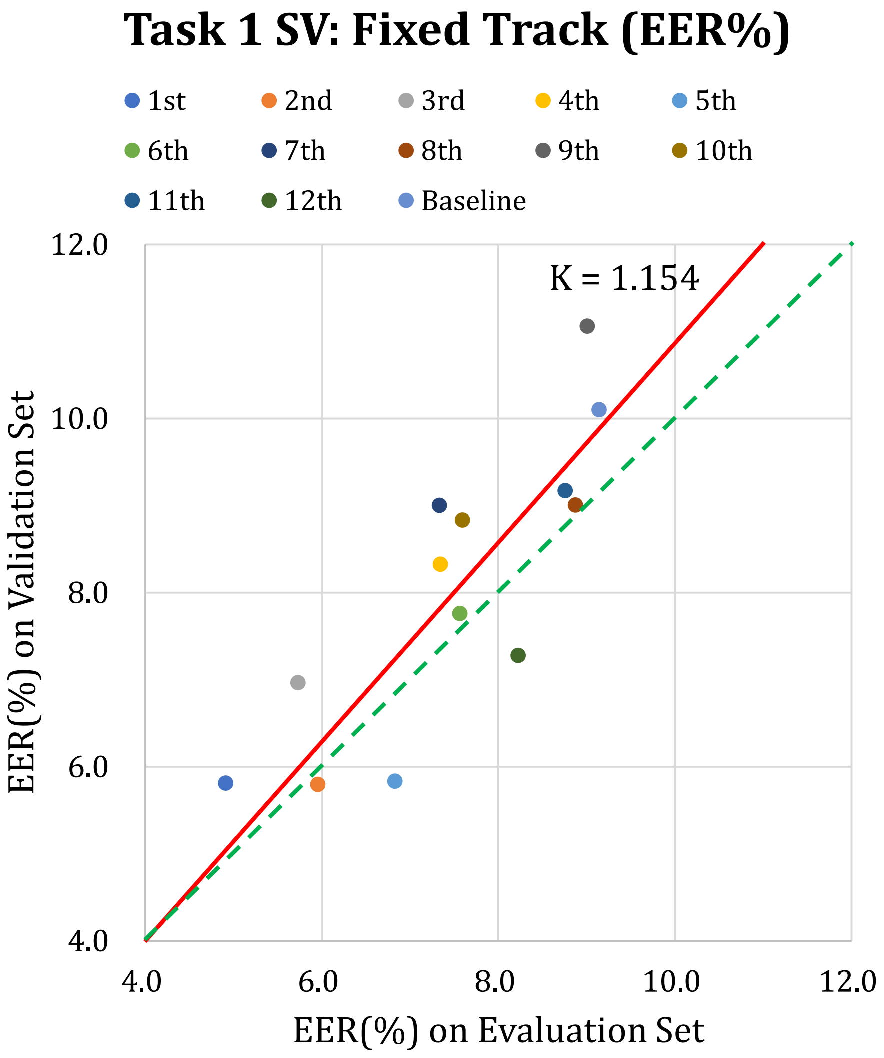


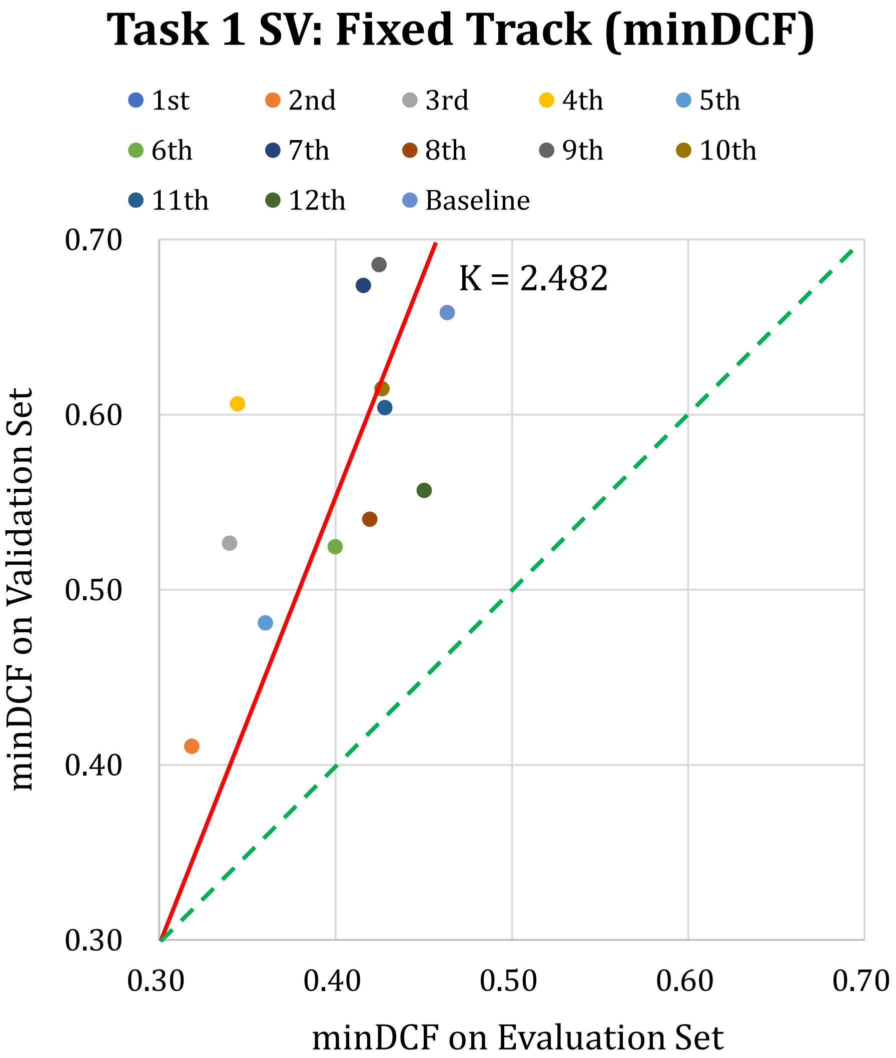
Open Track: 3rd



Task 1 SV: Evaluation vs. Validation

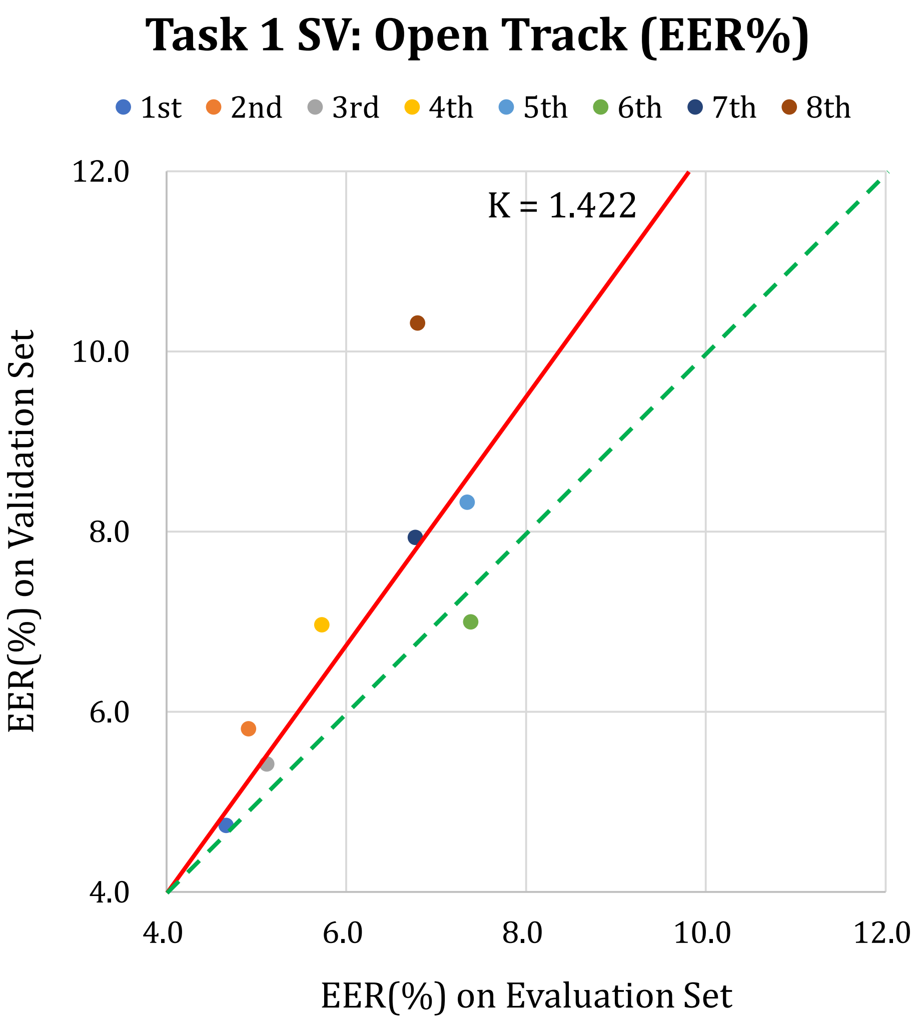
- Visible vs. Blind ➡ Innovativeness vs. Practicability

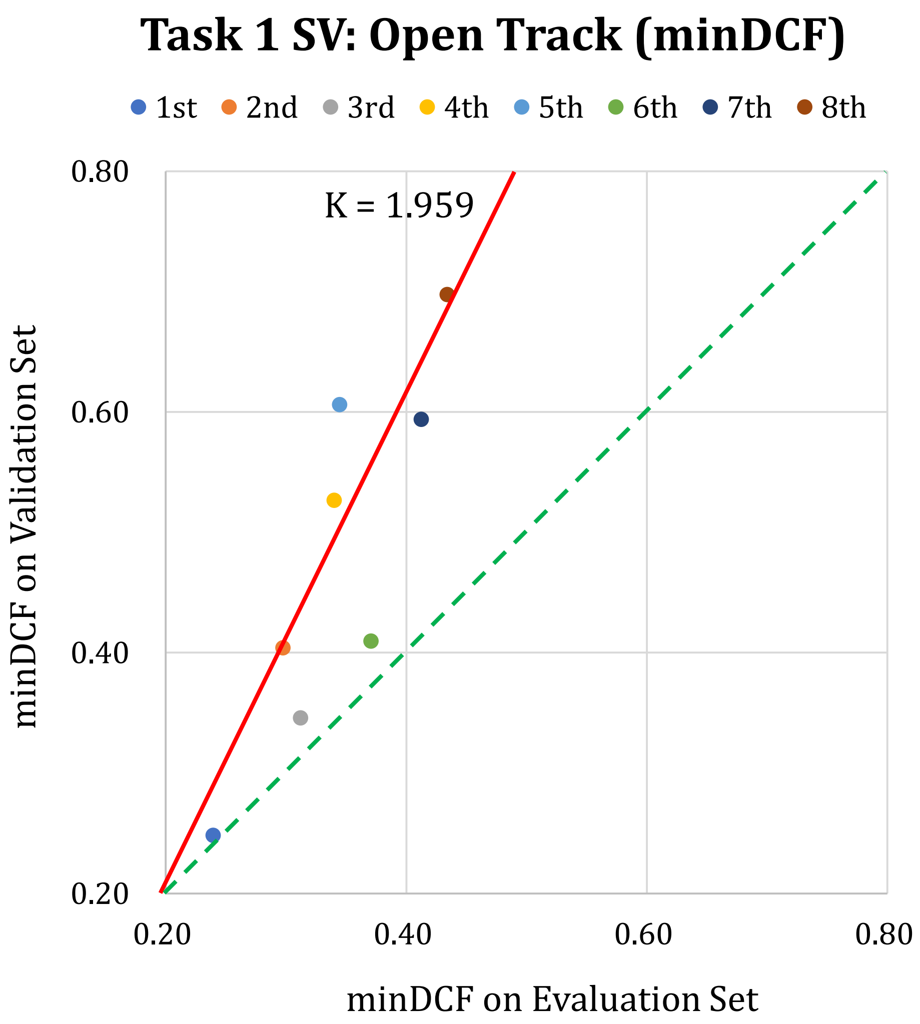




Task 1 SV: Evaluation vs. Validation

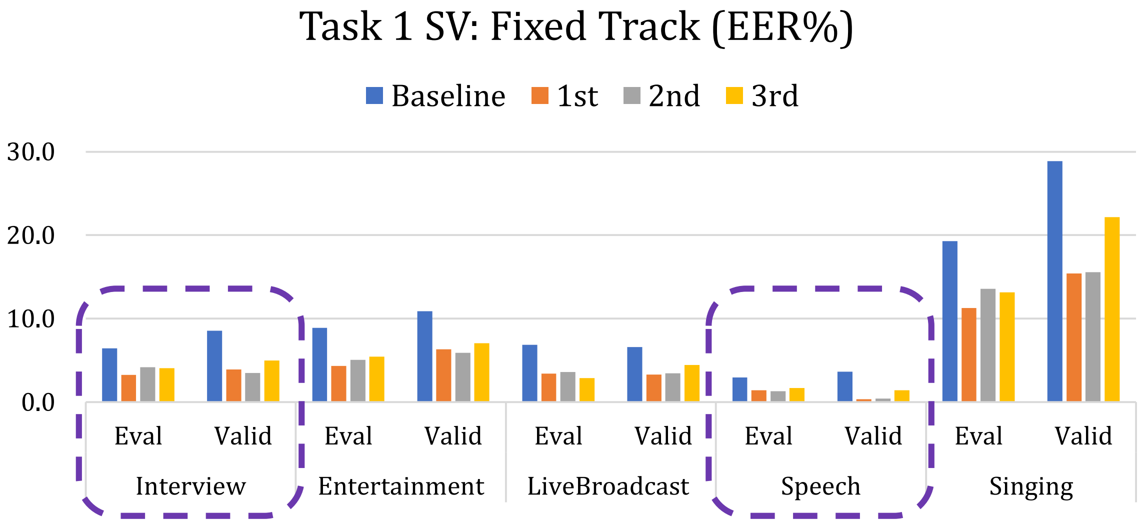
- Visible vs. Blind ➡ Innovativeness vs. Practicability

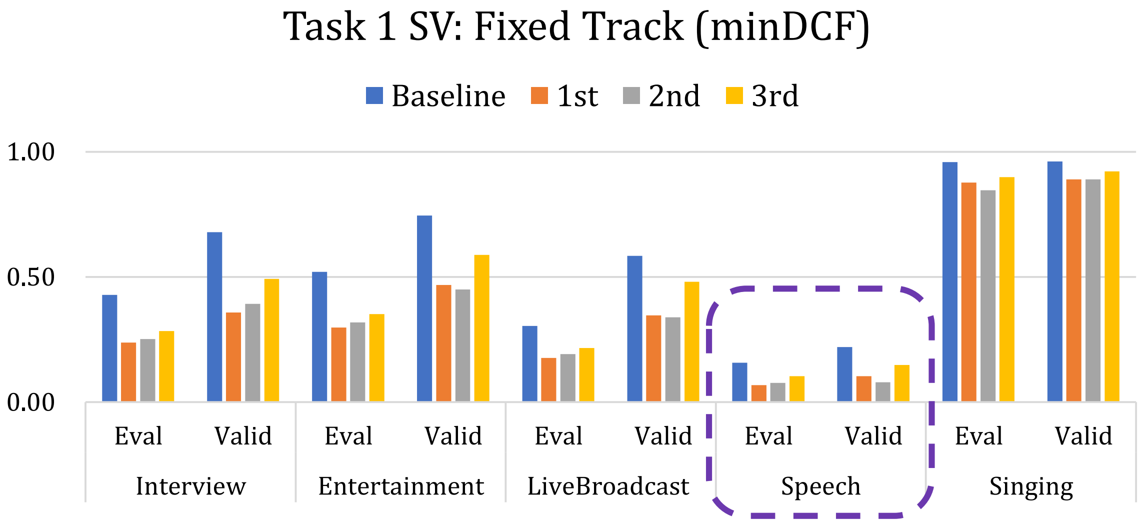


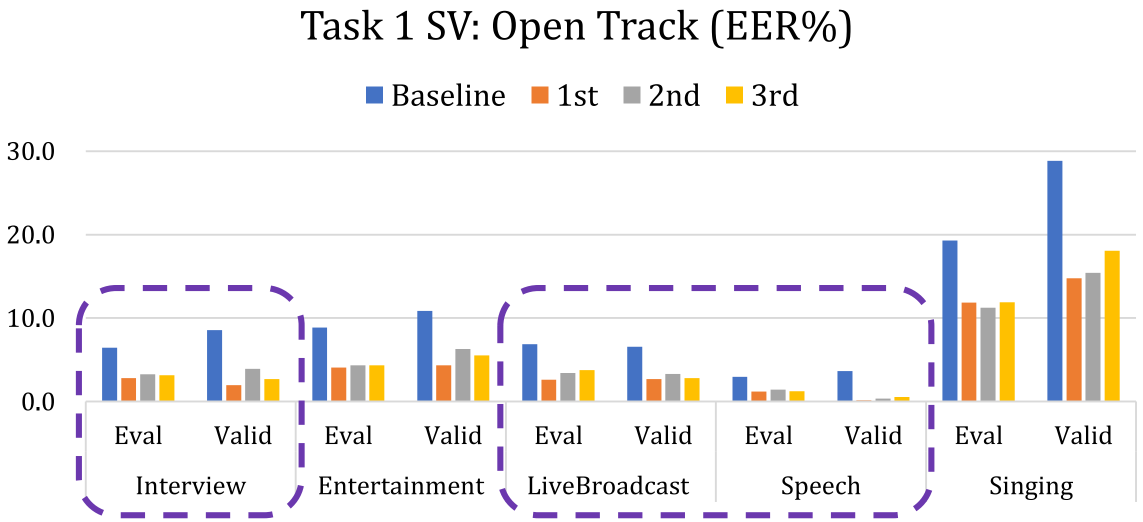


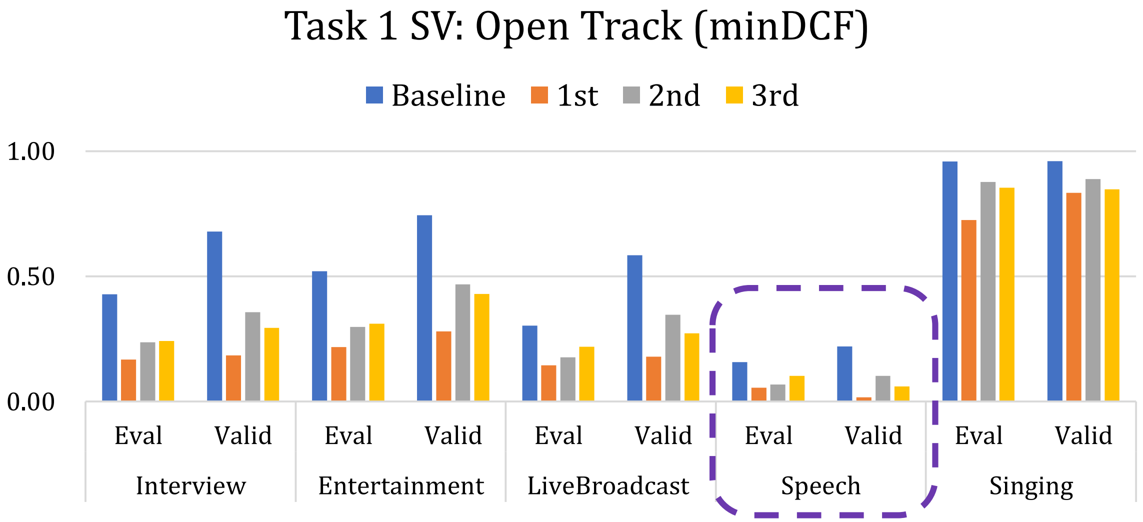
Task 1 SV: Evaluation vs. Validation

Evaluation vs. Validation on Different Genres





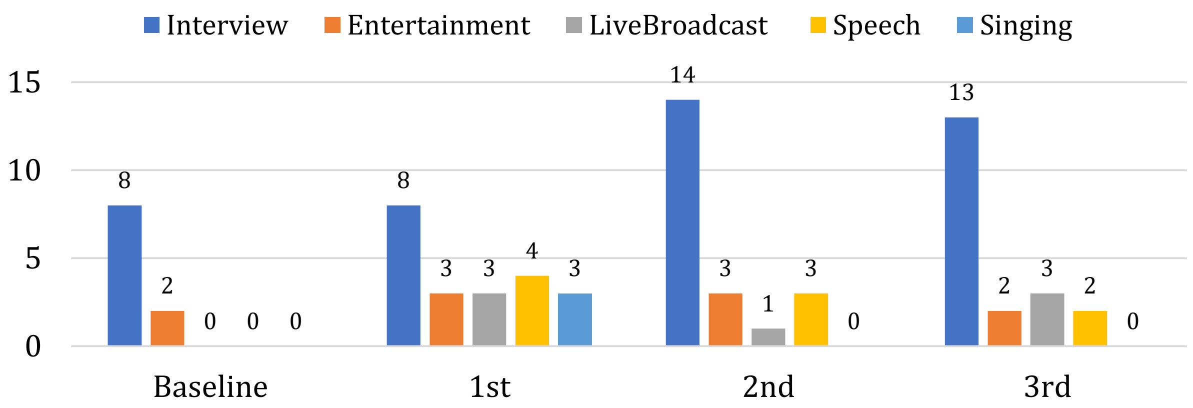




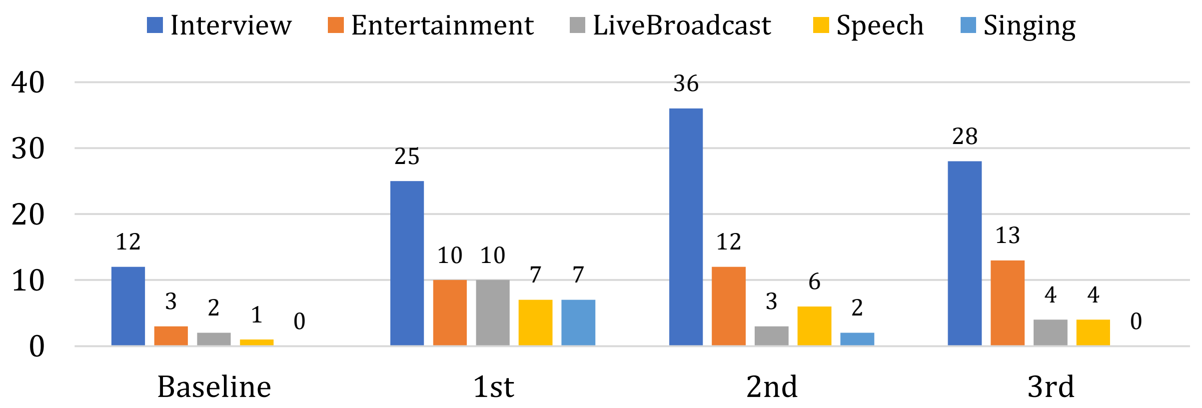
Task 2 SR: Open Track

System comparison

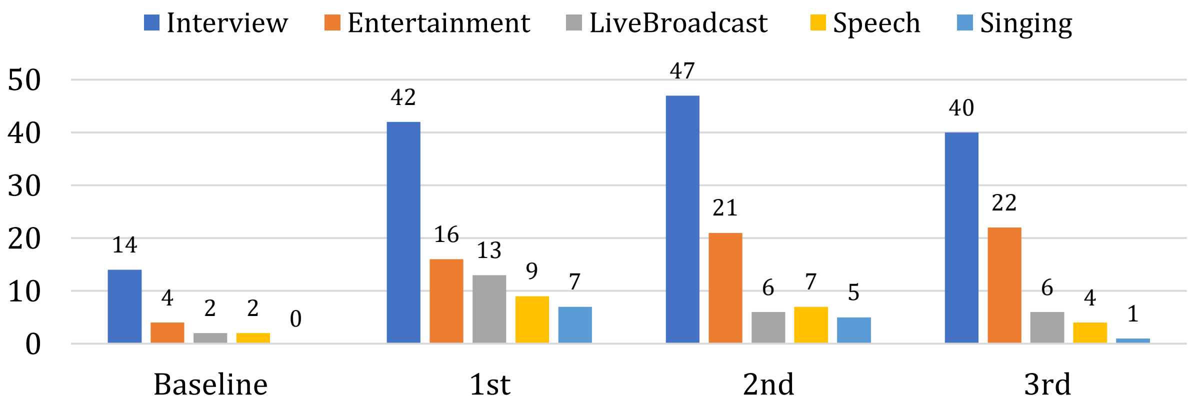
Top-1 retrieval counts



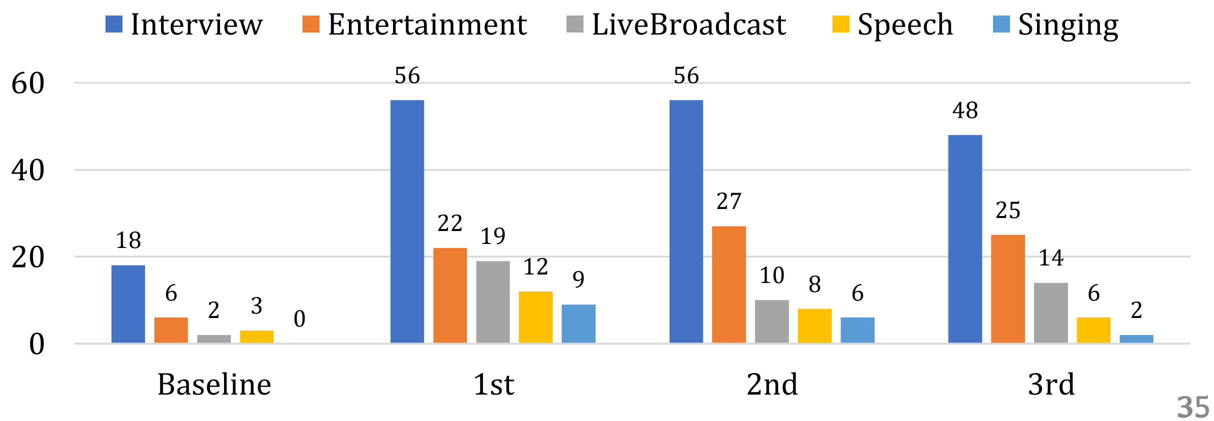
Top-3 retrieval counts



Top-5 retrieval counts



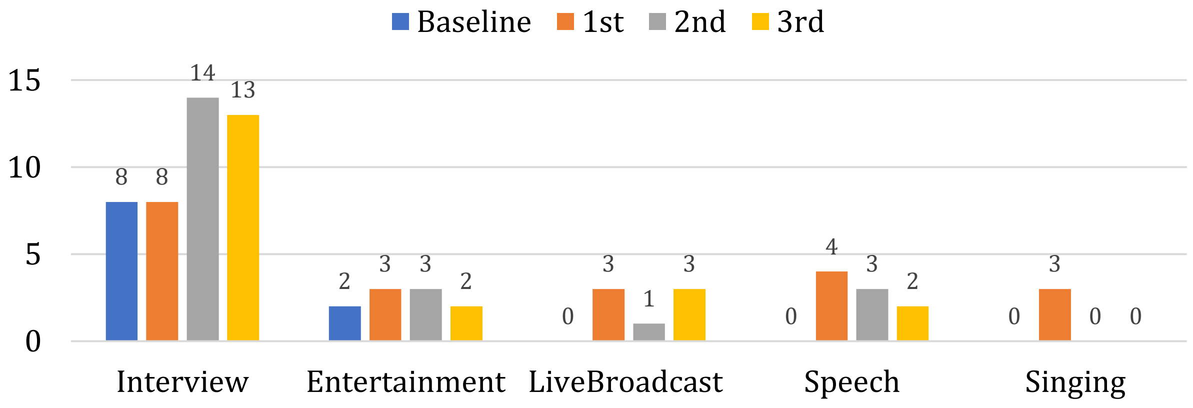
Top-10 retrieval counts



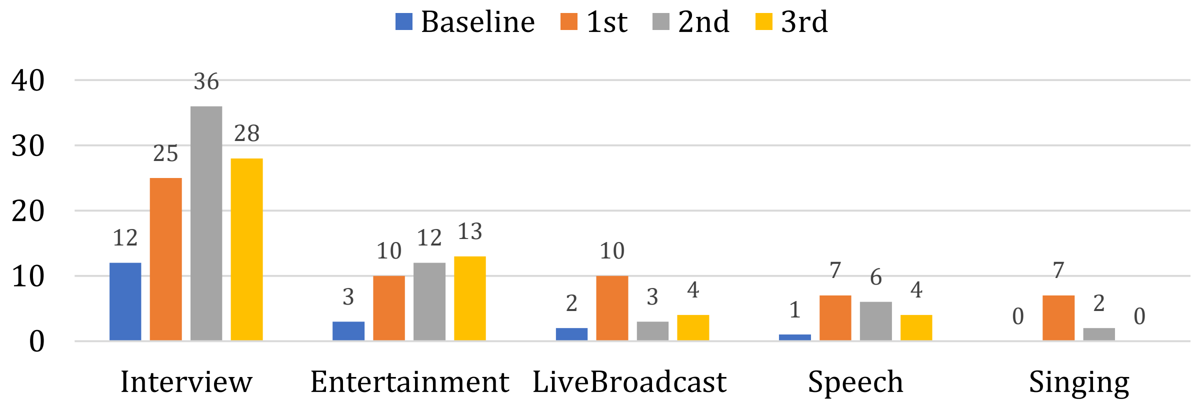
Task 2 SR: Open Track

Genre comparison

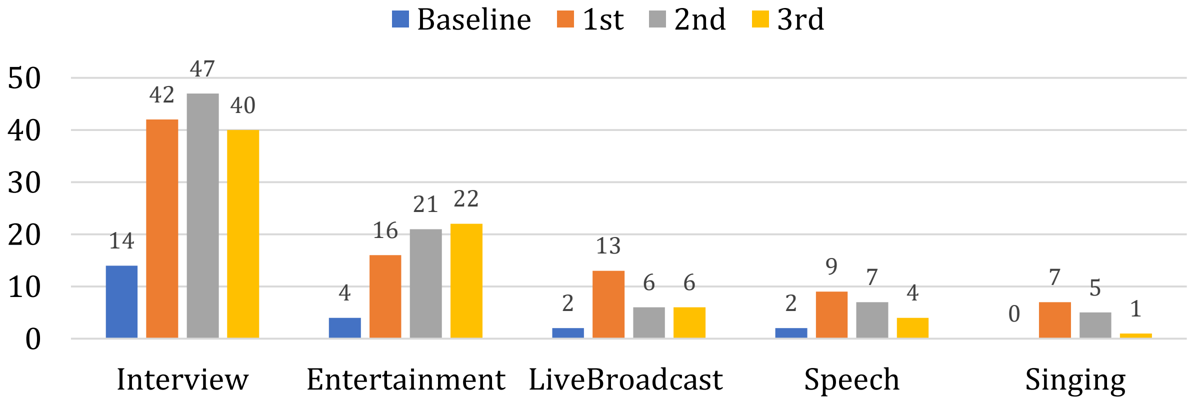
Top-1 retrieval counts



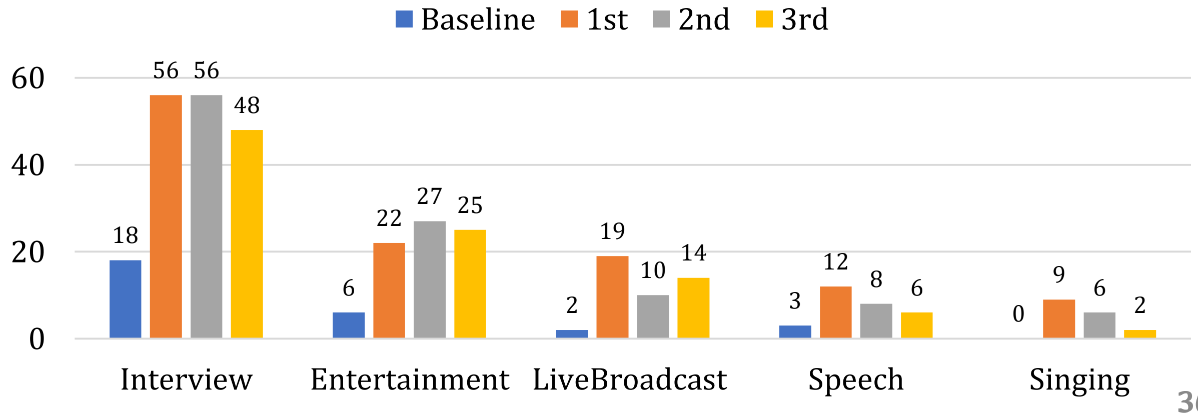
Top-3 retrieval counts



Top-5 retrieval counts



Top-10 retrieval counts



OUTLINE

- Data, Tasks and Baselines
- Technical Summary
- System Analysis
- **The Next CNSRC**

The Next CNSRC

Evaluation Protocol

- Advocate and standardize the ‘**Pre-Evaluation + Post-Validation**’ mode

SV Task

- Employ more convincing and all-round evaluation measurement (e.g., C-P map).
- Design various types of trials (e.g., easy/normal/hard trials) to evaluate system.

SR Task

- Consider the computation complexity (e.g., time and memory)

New Tasks

- For multi-speaker genres, such as interview or entertainment, to design a speaker localization and recognition task.

Many Thanks !

http://cnceleb.org/workshop