Experimental Report of Further Investigation on CN-Celeb

1 Model evaluation on CN-Celeb

In this expriment, we evaluate the performances of current best methods on more real-world condition. These states of the art methods including different network structures, feature pooling methods and loss functions. All of these model are trained on *VoxCeleb* and evaluated on *CN-Celeb*.

1.1 Dataset

VoxCeleb: Data involves two parts: *VoxCeleb1* and *VoxCeleb2*. We used *SITW*, a subset of *VoxCeleb1* as the evaluation set. The rest of *VoxCeleb1* was merged with *VoxCeleb2* to form the training set (simply denoted by *Vox- Celeb*). The training set involves 1, 236, 567 utterances from 7, 185 speakers, and the evaluation set involves 6, 445 utter- ances from 299 speakers (precisely, this is the *Eval. Core* set within *SITW*).

CN-Celeb: The entire 1, 000 speakers were used to do evaluation, including 130, 106 of 130, 108 (the other 2 utterances are deceted as silence). In detail, we randomly selected 5 utterances for each speaker to enroll, and all of the rest utterances are put into test set. Note that the enrollment set only contains 975 of 1, 000 speakers, since the other 25 speakers don't have at least 6 utterances (5 for enrollment and 1 for test).

1.2 Experiments setting:

Data preprocessing 40-demensional FBANK are used as input feature. Date augmentation are processed as same as Kaldi SITW recipe. And the final input segments are randomly selected between 200-400 frames.

Network structure We choose TDNN based x-vector system and ResNet based r-vector system as network structure. Both of these models use the second last dense layer as embedding layer.

- The **x-vector** system is similar to Kaldi TDNN recipe, while it uses BN+ReLU rather than ReLU+BN, and there is no dilation used. The TDNN topology is shown in Table 1 (a).
- The **r-vector** system follows ResNet-34 network of *VoxCeleb Speaker Recognition Challenge 2019* champion model, while there are also some differences. As is stated in tf-kaldi toolkit, the original ResNet uses max/average pooling after the last block to reduce feature dimension. To avoid eliminating the time resolution, it uses conv layer plus dense layer to replace max/average pooling. The ResNet topology is shown in Table 1 (b).

Layer	Layer context	Output
Frame1	[t-2, t-1, t, t+1, t+2]	512
Frame2	[t-2, t-1, t, t+1, t+2}	512
Frame3	[t-3, t-2, t-1, t, t+1, t+2, t+3]	512
Frame4	[t]	512
Frame5	[t]	512
Pooling	[0, T]	-
Segment6	[0, T]	512
Segment7	[0, T]	512
Classification	[0, T]	N

(a)

Layer	Structure	Output
Conv2D-1	3 × 3, Stride 1	512
ResNetBlock-1	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$, Stride 1	40 × L × 32
ResNetBlock-2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$, Stride 2	20 × L/2 × 32
ResNetBlock-3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 3$, Stride 2	10 × L/4 × 64
ResNetBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$, Stride 2	5 × L/8 × 256
Pooling	-	-
Segment6	-	512
Segment7	-	512
Classification	-	N

(b)

Table 1. (a) TDNN topology (b) ResNet-34 topology. N is the number of speaker, which is 64 in our experiments. L is the input segment frames.

Pooling method Traditional statistic pooling and attentive pooling method are implemented to evaluate. To have a further understanding of attentive method. We have conducted a sub-experiment to see how attention-head number affect the performance, whose result are shown in Table 2. For quick verification, this sub-experiment only uses CN-Celeb to do training and testing. From the result, we though multi-head attention cannot improve the performance. In hence, we finally only applied 1-head attention pooling on VoxCeleb experiments. (But now we think this experiment may can't reflect same trend on VoxCeleb)

Attention-head number	1	2	3	4	5
EER on CN-Celeb200	12.06%	12.22%	12.52%	12.13%	12.62%

Table 2. Multi-head attention experiment on CN-Celeb

Loss Function We applied traditional Softmax and ArcSoftmax loss function in experiments. ArcSoftmax also called Additive Angular Margin Softmax, which introduced angular margin, weight normalization and feature normalization to original Softmax. Although Softmax has many improved versions, ArcSoftmax has proved its performance in *VoxCeleb Speaker Recognition Challenge 2019*. However, we mention that ArcSoftmax is hard to optimize, it always applied with some tricks (egs. partly finetuning). For ArcSoftmax parameter, we set λ_b = 1000, γ = 0.00001, α = 5. Besides, relevant paper has revealed that the optimal margin factor *m* is 0.20 ~ 0.25 for VoxCeleb set, so we choose m=0.25 in our experiments.

Besides, we also conducted an experiment to test the optimal *m* for CN-Celeb dataset. As is shown in Table 3, smaller *m* like 0.10 is more suitable for CN-Celeb, it may because CN-Celeb is much more complex, there is not big margin between speaker classes.

Margin factor	m=0.05	m=0.10	m=0.15	m=0.20	m=0.25	m=0.30
EER on CN-Celeb	12.64%	**12.49%**	12.77%	12.63%	12.78%	13.15%

Table 3. ArcSoftmax experiment on CN-Celeb

Back-End The 512-dimensional embedding vectors extracted from the second last dense layer are projected to 128-dimensional vectors by LDA, and finally the PLDA model was employed to score the trials.

1.3 Experiments result:

Based on above model and method, we evaluated the performance of 8 different models both on SITW and CN-Celeb. The results refer to Table 4 and Figure 1.

Network	Pooling	Loss	EER on SITW	EER on CN-Celeb
TDNN	Stat.	Softmax	2.433%	16.07%
TDNN	Att.	Softmax	2.406%	15.97%
TDNN	Stat.	ArcSoftmax	2.488%	15.82%
TDNN	Att.	ArcSoftmax	2.57%	15.96%
ResNet-34	Stat.	Softmax	2.406%	15.37%
ResNet-34	Att.	Softmax	2.16%	15.56%
ResNet-34	Stat.	ArcSoftmax	**1.958%**	15.56%
ResNet-34	Att.	ArcSoftmax	2.296%	15.49%

Table 4. SOTA Models on SITW and CN-Celeb

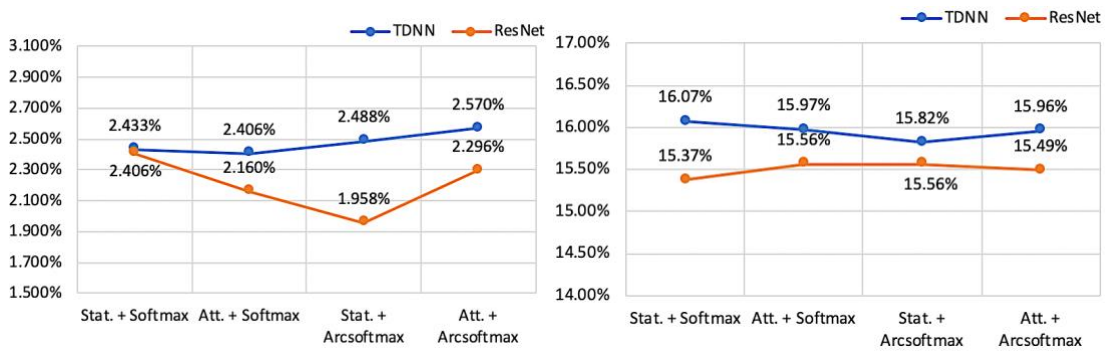


Figure 1. SOTA Models on SITW (left) and CN-Celeb (right).

It is obvious that the performances of current models are much worse on complex condition than that of constrained condition. Besides, we also find that the results do not show similar trend between VoxCeleb and CN-Celeb. On the one hand, it is because the EERs of CN-Celeb is not stable. On the other hand, it implys that there is a bias bottleneck in current models on unconstrained condition, it is a bias because these improved methods do not have much attributes to this stubborn bottleneck, sicne it has never been considered when designing model.

The feature vector distribution of 10 speakers are drawed in the following Figuire 2. These vectors are extracted form ResNet + ArcSoftmax model (the 7th row in Table 4). From this figure we can find out that after applying ArcSoftmax, the margin between different classes is big enough to distinguish from each other, however most class appears domain shift problem. In conclusion, the limitation of current models is not discriminant ability, but the ability of handing different uncertainties in real-world condition.

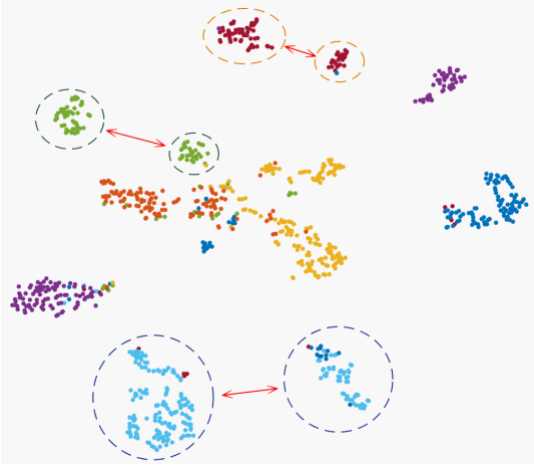


Figure 2. Feature vector distribution of 10 speaker. Extractor from ResNet + ArcSoftmax model (the 7th row in Table 4).

2 Affective factors in CN-Celeb

With the conclusion of weakness on concertainties. Further experiments have conducted to work out What the concertainties are? And how they affect models? This part focus on genres, speech length, gender and age factors to do analysis. All of the experiments in this part are based on ResNet + ArcSoftmax model (the 7th row in Table 4).

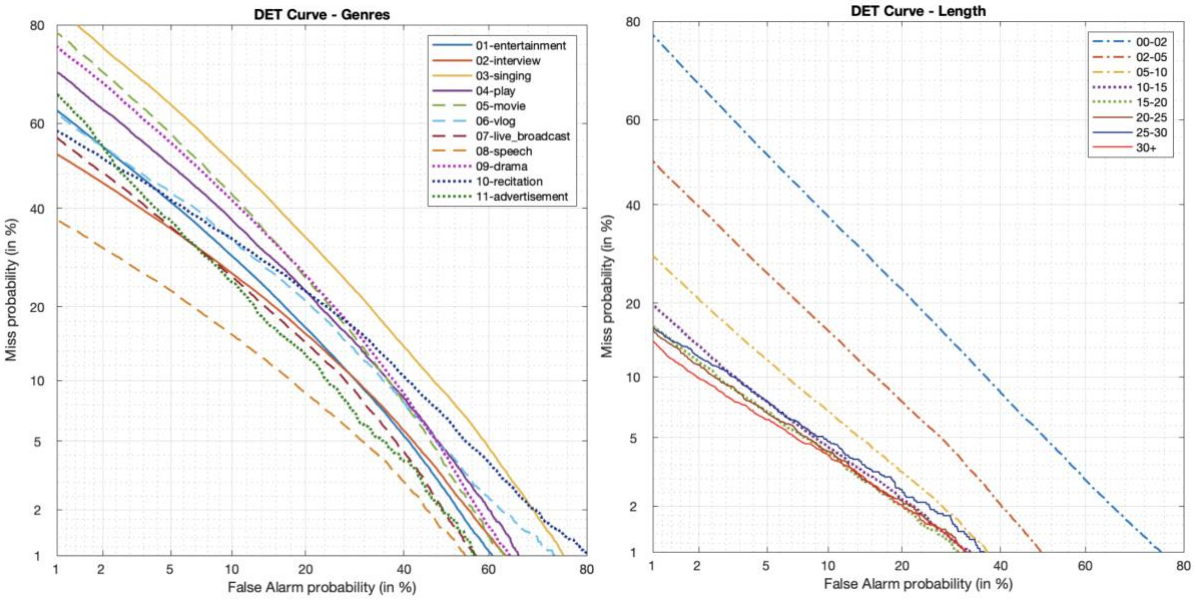
2.1 Genres

All genres in CN-Celeb set are seperately enroll and test to do full-cross experiment. For gener *g*, randomly select 5 utterances for each speaker that has more than 5 *g* gener utterances. For total trial, all the rest utterances other than enroll set are put into trial list. The result refers to Table 5.

test / enroll	01-entertain	02-interview	03-singing	04-play	05-movie	06-vlog	07-live_broa	08-speech	09-drama	10-recitation	11-advertise	total
01-entertain	16.15%	16.70%	26.89%	24.62%	20.57%	22.30%	18.71%	14.75%	21.24%	16.33%	24.68%	18.30%
02-interview	19.44%	13.27%	28.46%	20.11%	28.05%	19.13%	18.18%	14.47%	25.93%	16.41%	17.00%	17.70%
03-singing	29.04%	27.63%	25.68%	28.12%	37.17%	29.16%	25.86%	21.63%	35.05%	23.44%	33.33%	26.10%
04-play	24.63%	21.49%	27.12%	20.09%	22.78%	11.67%	24.59%	40.30%	26.27%	20.69%	18.75%	21.32%
05-movie	21.85%	20.26%	33.06%	27.02%	22.19%	20.93%	16.13%	27.50%	23.64%		33.33%	22.42%
06-vlog	29.40%	18.05%	36.06%	12.22%	31.58%	17.29%	18.54%	16.67%	22.45%			20.52%
07-live_broa	21.57%	17.55%	27.80%	23.33%	25.71%	23.38%	13.38%	17.36%	18.34%	25.00%	10.00%	17.00%
08-speech	18.32%	14.72%	23.82%	30.10%	16.67%	25.49%	20.45%	8.61%	16.67%	16.14%	16.36%	13.03%
09-drama	23.17%	20.85%	35.52%	23.23%	33.63%	23.13%	17.13%	11.85%	21.79%	43.08%	20.00%	22.61%
10-recitation	27.16%	16.79%	22.67%	26.33%		33.33%	28.01%	14.84%	46.67%	13.23%	25.00%	21.51%
11-advertise	21.16%	16.46%	50%	13.95%	20.90%		19.05%	9.90%	33.33%		13.33%	15.96%

Table 5. Cross genres experiments.

Note that the blank blocks in Table 5 mean there are no "target" pairs in relative trials list, so can't calaulate EER. And for some experiments, there are not enough "target" pairs in trial list to get a statistical EER. For example, "11-advertisement_to_03-singsing" trials only have 2 "target" pairs, hence the result is relatively abnormal.
Also, we use total test result to draw the DET curves of different geners. As it shown in Figure 3 (a), diferent genres show a layered distribution. The speech gener gets the best EER performance while singing genre get sthe worst. Based on the general trend, we can find that the genres that always accompanied by background voice (singing, movie, drama) generally have worse result than that of steady conversation (speech, interview, live broadcast).



(a) (b)

Figure 3. DET curve of genres experiments (a) and speech length experiments (b).

2.2 Speech Length

To estimate the affect of speech length, we divide the CN-Celeb test set into 8 subset with different speech length. Table 6 shows corresponding EER and Figure 3 (b) shows the DET curve. Experimental results presente obvious positive correlation between EER and speech length. Moreover, when speech length reachs 15s, this positive correlation appears saturation. It means for speaker recognition task, speech of 15s contain all the information we need. On the other hand, this experiment shows the short utterances largely affect the model performance on CN-Celeb.

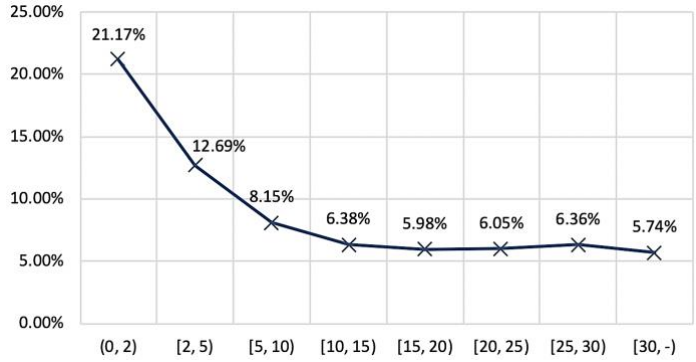


Figure 4. EER of different speech length.

Length(s)	(0, 2)	[2, 5)	[5, 10)	[10, 15)	[15, 20)	[20, 25)	[25, 30)	[30, -)
Proportion	32%	30%	18%	8%	4%	2.5%	1.5%	4%
EER	21.17%	12.69%	8.149%	6.381%	5.976%	6.046%	6.364%	**5.739%**

Table 6. EER of different speech length.

2.3 Gender

The total CN-Celeb dataset includes 423 female speakers and 577 male speakers. The total set and every genre are divided according to gender to evaluate, results are as follows. Nearly in every genre, male speakers has better result than female speaker. This may because approximately 60% of training speakers in VoxCeleb are male. Besides, the paly, vlog, speech, advertisement genres have abnormal, this mostly due to gender unbalance under these genres.

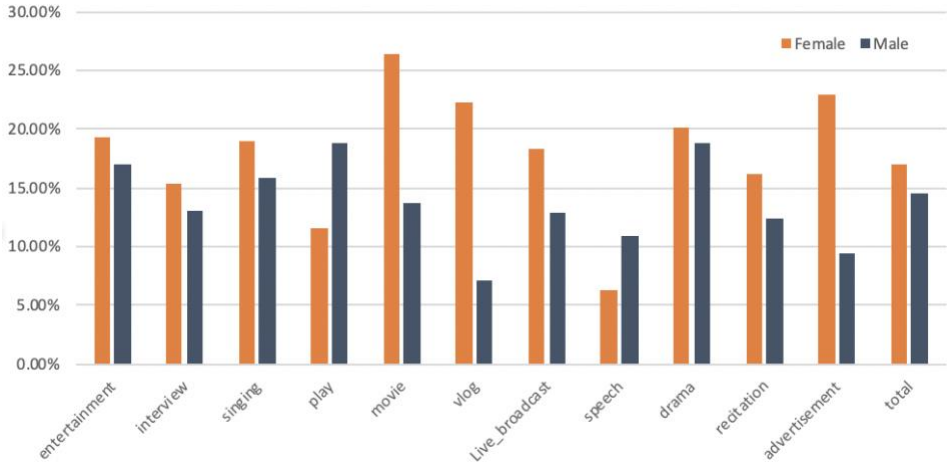
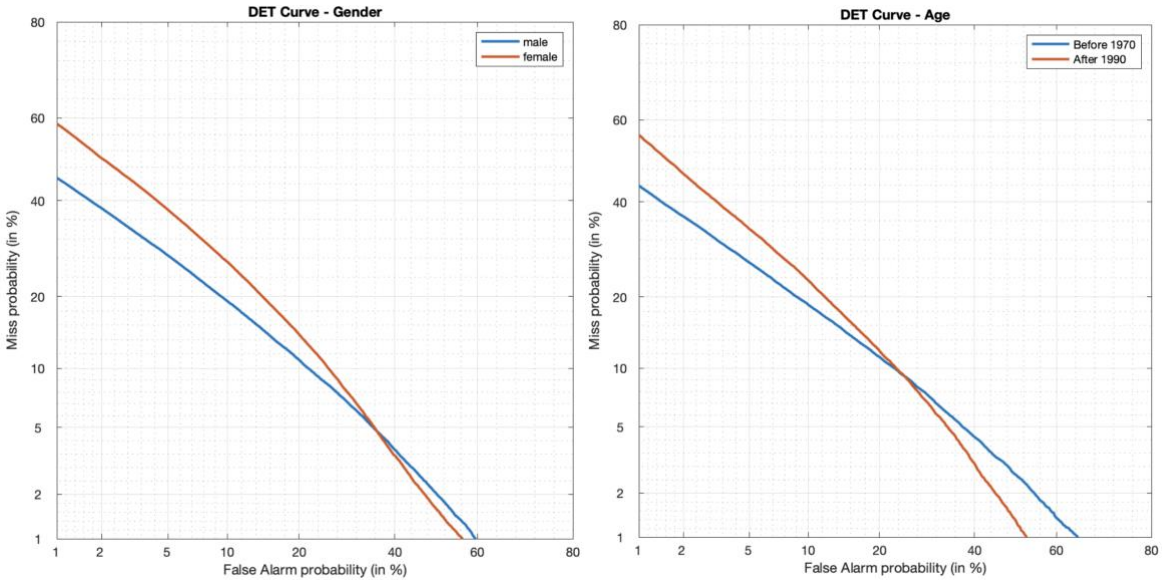


Figure 5. EER of female and male in each genre.

Gener	Ent.	Inter.	Sin.	play	Mov.	vlog	Live	Spe.	drama	Reci.	Ad.	total
Female (%)	19.36	15.32	19.03	11.56	26.44	22.22	18.26	6.26	20.17	16.21	22.89	**17.00**
Male (%)	16.97	13.12	15.89	18.73	13.66	7.19	12.91	10.96	18.73	12.43	9.38	**14.60**

Table 7. EER of female and male.



(a) (b)

Figure 6. DET curve of gender experiments (a) and speech age experiments (b).

2.4 Age

For age experiments, considering 80% of speakers in CN-Celeb are middle-age, we choose an extrame comparison of speakers borned before 1970 ang speakers borned after 1990. The EER results and propertries of 1970 set and 1990 set are show in Table 8, DET curve are shown in Figure 6 (b). From the statistic information, we can see the

1970 set and 1990 set are comparable. Both of the final EERs and DET curves show there are not obvious accuracy differece across age.

Set		1970	1990
EER		**14.58%**	**15.67%**
Speaker number		210	153
Utterance number		27837	18841
Average length		8.37s	6.61s
Genres	Ent.	15%	13%
	Reci.	3%	1%
	Ad.	0%	0%
	Int.	44%	53%
	Sin.	9%	10%
	Play	7%	1%
	Mov.	1%	0%
	Vlog	0%	3%
	Live	0%	13%
	Spe.	12%	0%
	Dra.	4%	0%

Table 8. Properties of 1970 set and 1990 set.

3 Conclusion

Current speaker recognition model do not perform well on more complex condition, in feature vector space, domain shift problem is common in most cross-genre speakers. Further experiments have revealed some affective factors for such a unsatisfactory performance: (1) For different speech genres, the recognition accuracy of complex genres (like singing) can be twice lower than the steady genres (like speech). (2) In terms of speech length, the EER increase rapidly when speech get longer, and finally converges in about 15s. Generally, the recognition accuracy of long utterances (longer than 15s) can be three times higher than short utterances (shorter than 2s). (3) As for gender, the EER of male speaker is 3% lower than that of female speaker, and the same trend has occurred in neatly all genres. (4) The age factor doesn't show obvious contribution, there are only 1% EER differece between 2 extrame age comparison set.