Speech Communication 137 (2022) 77–91

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

CN-Celeb: Multi-genre speaker recognition ☆

Lantian Li [a,1], Ruiqi Liu [a,b,1], Jiawen Kang [a,c,1], Yue Fan [f], Hao Cui [f], Yunqi Cai [d], Ravichander Vipperla [e], Thomas Fang Zheng [a], Dong Wang [a,*]

[a] Center for Speech and Language Technologies (CSLT), BNRist at Tsinghua University, Beijing, China
[b] China University of Mining and Technology-Beijing, China
[c] Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong
[d] Department of Computer Science and Technology, Tsinghua University, China
[e] Samsung AI Center, Cambridge, UK
[f] Key Laboratory of Transient Physics, Nanjing University of Science and Technology, China

ARTICLE INFO

Keywords:
Speaker recognition
Multi-genre
Speech corpus

ABSTRACT

Research on speaker recognition is extending to address the vulnerability in the wild conditions, among which genre mismatch is perhaps the most challenging, for instance, enrollment with reading speech while testing with conversational or singing audio. This mismatch leads to complex and composite inter-session variations, both intrinsic (i.e., speaking style, physiological status) and extrinsic (i.e., recording device, background noise). Unfortunately, the few existing multi-genre corpora are not only limited in size but are also recorded under controlled conditions, which cannot support conclusive research on the multi-genre problem. In this work, we firstly publish CN-Celeb, a large-scale multi-genre corpus that includes in-the-wild speech utterances of 3000 speakers in 11 different genres. Secondly, using this dataset, we conduct a comprehensive study on the multi-genre phenomenon, in particular the impact of the multi-genre challenge on speaker recognition and the performance gain when the new dataset is used to conduct multi-genre training.

1. Introduction

Speaker recognition aims to verify the claimed identity of a person using her/his spoken utterance as input modality. With several decades of research, the performance of speaker recognition systems has been remarkably improved, and commercial usage has been made feasible in certain conditions (Campbell, 1997; Reynolds, 2002; Hansen and Hasan, 2015).

A long-standing theme in speaker recognition research has been the problem of tackling various speaker-independent variations in the speech signal. These variations could be extrinsic or intrinsic. The most significant extrinsic variations include diversity in recording device, ambient acoustics, background noise, transmission channel, and distortions introduced in pre-processing algorithms. Intrinsic variations refer to both the minor and universal randomness in the movement of pronunciation apparatus as well as more explicit diversity in speaking style (e.g., reading or spontaneous), speaking rate, emotion, and physical status. These variations pose the major challenge for speaker recognition systems.

The history of speaker recognition research can be seen as a pursuit towards solving the impact of these variations on the recognition accuracy. For instance, initial research was constrained to text-dependent tasks and focused on solving the variation caused by pronunciation randomness, in which the Hidden Markov Model (HMM) was the most popular (Parthasarathy and Rosenberg, 1996). Later research attempted to solve text-independent tasks and had to deal with phonetic variation, which boomed the Gaussian Mixture Model with Universal Background Model (GMM-UBM) architecture (Reynolds et al., 2000). Further research tried to address inter-session variation caused by channels and speaking styles, for which the i-vector/PLDA architecture was the most successful (Dehak et al., 2011). Recently, the research focus has been targeted towards dealing with complex variations in the *wild* scenarios, for which deep learning methods have been demonstrated to be the most powerful (Variani et al., 2014; Li et al., 2017; Snyder et al., 2018; Okabe et al., 2018).

Multi-genre scenario is perhaps the most challenging scenario for speaker recognition, as it involves nearly all the complex variations one can imagine. For example, a speaker aged 20 may be registered with

☆ This document was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61633013 and No. 62171250.
* Corresponding author.
E-mail address: wangdong99@mails.tsinghua.edu.cn (D. Wang).
[1] Lantian Li, Ruiqi Liu and Jiawen Kang are joint first authors.

https://doi.org/10.1016/j.specom.2022.01.002
Received 23 November 2020; Received in revised form 28 October 2021; Accepted 3 January 2022
Available online 22 January 2022
0167-6393/

the system using an interview speech, recorded in a quiet environment, with a relatively formal speaking style, and using a far-field table microphone; while the test may be with a singing speech of the same speaker at age 40, in a live show under music background, with a close-talk hand-held microphone. From another perspective, multi-genre is not an artificial challenge, it is indeed encountered in many real-life applications. For instance, a good speaker recognition system must be able to accept a user several months after the registration, even if the user uses a different cellphone and speaks in a different style. In summary, we argue that good performance on multi-genre scenarios is a sufficient and necessary condition for practical success of speaker recognition research.

Unfortunately, the existing state-of-the-art techniques perform poorly in multi-genre conditions. Our recent experimental results show that a system trained with the most popular recipe and using the largest corpus publicly available (VoxCeleb) performed quite poorly on CN-Celeb1 (Fan et al., 2020), a multi-genre corpus that we have recently published. Specifically, the results in terms of Equal Error Rate (EER) was 3.75% on SITW, the accompanying test data of VoxCeleb. On CN-Celeb.E, a subset of CN-Celeb1 consisting of 200 speakers, the EER result was 15.52%, an increase of 300% compared to the SITW result.

The importance of multi-genre scenario and the below-par performance of existing techniques in this scenario necessitates a concentrated research effort in this direction. However, the present CN-Celeb1 corpus (Fan et al., 2020) is insufficient to derive comprehensive research conclusions due to its limited data size. CN-Celeb1 has only 1000 speakers and 270 h of speech signals in total. The limited data size makes it hard to be used as a standalone training set, especially when the models are based on deep learning methods. This has been clearly demonstrated in our previous experiments (Fan et al., 2020) where the performance of i-vector system was better than that of x-vector system (14.24% vs. 14.78%), when the two systems were trained with 800 speakers (CN-Celeb1.T) and tested on the rest 200 speakers (CN-Celeb.E). Moreover, the performance of x-vector system trained with CN-Celeb1.T was even worse than the one trained with VoxCeleb, even though the former is based on multi-genre data and so is under matched condition. The data insufficiency is an obstacle for researchers for a deep investigation into the multi-genre challenge.

In this work, we firstly publish a new large-scale multi-genre corpus, called CN-Celeb2. CN-Celeb2 shares the same 11 genres as CN-Celeb1, but the data size is much larger. It contains over 520,000 utterances from 2000 Chinese celebrities. The two multi-genre corpora make up the overall CN-Celeb corpus,[2] which can be used to perform fully multi-genre training and test. Secondly, based on the new CN-Celeb database, we conduct a comprehensive study on multi-genre speaker recognition. In particular, we employ multi-genre training to improve model robustness in cross-genre scenarios, and also investigate the efficacy of a meta-learning approach to improve model generalization for novel genres.

2. Speaker recognition: challenge, technique and data

We firstly present a historical review of speaker recognition research. Different from the previous overviews that concentrate on details of speaker recognition techniques (Campbell, 1997; Reynolds, 2002; Hansen and Hasan, 2015), this review focuses on the interaction amongst scientific challenge, technical development and data accumulation. With this outlook, we categorize the development of speaker recognition research into 4 phases, as outlined in Table 1.

[2] http://cnceleb.org.

2.1. Phase 1: Randomness in pronunciation

The initial foray into speaker recognition focused on the randomness in pronunciation. One cannot produce the same words/utterances in exactly the same way. Early speaker recognition research focused on solving this type of variation, by using either non-parametric methods such as Dynamic Time Warping (DTW) (Rosenberg, 1976) or with parametric models such as Hidden Markov Model (HMM) (Parthasarathy and Rosenberg, 1996; Matsui and Furui, 1993).

Researchers in this period often used small self-collected datasets. For example, Doddington recorded 123 male speakers in a sound booth using a dynamic microphone, where 63 males were used for target trials and the rest 60 males for imposter trials (Rosenberg, 1976). Similarly, Parthasarathy et al. recorded 51 males and 49 females over long distance telephony channel, and each speaker made 26 calls uttering the same phrase (Parthasarathy and Rosenberg, 1996).

2.2. Phase 2: Phonetic variation

Further investigations attempted to deal with the phonetic variation — the main obstacle towards text-independent speaker recognition. GMM (Reynolds, 1995) and its adapted version, GMM-UBM (Reynolds et al., 2000) were demonstrated to be the most effective towards this end.

A large dataset is necessary to train the Gaussian components, and so data requirement during this phase was more demanding than Phase 1. Moreover, the data used by researchers began to be more standardized, partly due to the NIST Speaker Recognition Evaluation (SRE) started in 1996 (Alvin and Martin, 2004). One of the most popular corpora during this phase was Switchboard collected by the Linguistic Data Consortium (LDC).[3] This corpus incorporated several collections, each of which includes hundreds of speakers and thousands of conversations. It was extensively used in the NIST SRE series from 1996–2003, and also formed an important part of the training set in the later NIST SREs.

2.3. Phase 3: Session variation

Session variation refers to the systematic change in speaking style or acoustic condition when the same speaker speaks in different sessions. Kenny's important work on Joint Factor Analysis (JFA) (Kenny et al., 2003, 2007; Kenny, 2005) paved the way for solving general variations between sessions. The i-vector model, a successor of JFA, made a further leap (Dehak et al., 2011) in producing session-based vectors that involve all types of long-term variations, and leave the task of discriminating different types of variations to a back-end model. Numerous results have demonstrated that the i-vector model could achieve very good performance with accompanying probabilistic linear discriminant analysis (PLDA) (Ioffe, 2006; Prince and Elder, 2007) as its back-end model.

The data requirement to solve session variation is much more demanding than the previous two phases. Especially, differentiating session variation requires single-speaker multiple-condition (SSMC) data. This type of data is much harder to collect as it requires the same speaker providing utterances under different conditions. Fortunately, NIST SRE, after year by year evaluation, offered a large amount of SSMC data, and the research conducted in this phase mostly used the NIST SRE data, which was primarily collected by LDC under the Mixer protocol (Cieri et al., 2007).

[3] https://www.ldc.upenn.edu/.

Table 1
History of speaker recognition research.

Phases	Approximate period	Variations	Techniques	Typical Corpora
I	1970–2000	Pronunciation	DTW, HMM	Private data
II	1995–2010	Phone	GMM-UBM	Switchboard, NIST SRE (−03)
III	2005–2016	Session	JFA, i-vector/PLDA	Mixer, NIST SRE (04–16)
IV	2017–	Complex	DNN	VoxCeleb, NIST SRE (18–19)

2.4. Phase 4: Complex variation

The session variation, especially covered by the NIST SRE, varies in channels, background noises and even languages. However, these variations are largely under control. The speech data collected by the Mixer project, be it telephonic conversations or interviews, was constrained by the collection process, e.g., the participants were fully cooperative.

Recently, researchers are attempting to solve a more challenging task: recognizing speakers *in the wild*. A key feature of this task is that the speakers do not cooperate with or are even aware of being recorded, and the recording conditions are fully unconstrained, leading to more complex variation. For example, the audio/video posted on YouTube may be fully spontaneous and recorded by diverse devices.

Addressing such complex variations is highly challenging. Fortunately, the DNN-based methods have shown great potential in dealing with this problem (Variani et al., 2014; Li et al., 2017; Snyder et al., 2018; Okabe et al., 2018). So far, the most popular deep learning architecture is based on the concept 'deep embedding', which converts speech segments of variable lengths to fixed-length continuous vectors. Accompanied by a back-end scoring model (e.g., PLDA), the deep embedding approach has gained the state-of-the-art performance.

The most successful deep embedding model is the x-vector model, proposed by Snyder et al. (2018). Recent progress on the deep speaker embedding approach includes more comprehensive architectures (Chung et al., 2018; weon Jung et al., 2019), improved pooling methods (Okabe et al., 2018; Cai et al., 2018; Chen et al., 2019; Xie et al., 2019), better training criteria (Li et al., 2016; Ding and He, 2018; Wang et al., 2019b; Bai et al., 2020; Gao et al., 2019; Zhou et al., 2019), and better training schemes (Li et al., 2019; Wang et al., 2019a; Stafylakis et al., 2019). Another popular deep learning approach is the end-to-end modeling, which discriminates speakers of two speech segments directly (Heigold et al., 2016; Zhang et al., 2016; Rahman Chowdhury et al., 2018). A key advantage of the end-to-end approach is that the training and test are based on the same criterion, which ensures the test is optimal if the data is sufficient and the training can be well conducted. However, the training process is often tricky (Wang et al., 2017).

Due to the data-driven nature, DNN-based methods are data-hungry. Ideally, the training data must comprise all the potential variations and their combinations that may appear in real applications. To meet this demand, researchers from SRI released SITW, the first dataset in unconstrained conditions (McLaren et al., 2016). This dataset contains audio from 299 speakers, with an average of 8 different sessions per speaker. Although very valuable, SITW is too small to be used as a training set. Oxford released a large in-the-wild corpus VoxCeleb1 in 2017 (Nagrani et al., 2017), and an even larger one VoxCeleb2 in 2018 (Chung et al., 2018). The total number of speakers in the two corpora exceeds 7000, which is fairly large for speaker recognition research. More importantly, both SITW and VoxCeleb are free, which significantly promoted the recent research on complex and unconstrained conditions.

2.5. Data is still insufficient

With decades of research, complex variation can be partially addressed, thanks to the SITW and VoxCeleb corpora. However, the present research is yet to solve the truly complex (and difficult) variations. In fact, most of SITW/VoxCeleb data are from interviews, so the variation on speaking styles has been largely constrained. Most importantly, the recording condition and speaking style of each speaker in these corpora do not change much, which means that speakers and conditions may be heavily coupled (Shon and Glass, 2020).

As we have mentioned, the multi-genre scenario involves the truly complex variation, making it one of the most difficult conditions for speaker recognition research.[4] Unfortunately, none of existing datasets is really multi-genre, including SITW and VoxCeleb, which makes the research on multi-genre speaker recognition nearly impossible. The recently published CN-Celeb1 corpus, allows some preliminary studies in this direction (Kang et al., 2020; Kataria et al., 2020; Mo and Xu, 2020; Chen et al., 2021). However, the lack of multi-genre *training* data severely precludes further study on this important subject. We therefore start by building a large-scale multi-genre training set, and then study some simple techniques to address the multi-genre challenge.

3. CN-Celeb2: features and collection pipeline

3.1. Revisit CN-Celeb1

The CN-Celeb1 corpus is a free and public speaker recognition dataset released by the research group of the authors (Fan et al., 2020). The speech data were collected from Bilibili, a public media source, using an automated pipeline similar to the one used to collect VoxCeleb (Nagrani et al., 2017). Human check was arranged to ensure the quality of the collected data. Especially, CN-Celeb1 was intentionally designed to cover multiple genres, in particular cross-genre situations. The entire dataset contains more than 130,000 utterances from 1000 Chinese celebrities, and covers 11 different genres in real world. Readers can refer to the original publication (Fan et al., 2020) for more details of the data profile and the collection pipeline.

We note that choosing celebrities as the target speakers is important for CN-Celeb1 to achieve its goal. Celebrities naturally appear in multiple situations and speak in multiple genres, which perfectly matches our research on multi-genre phenomenon. However, we do not expect that the speaking style of celebrities would be fully spontaneous, and we can neither guarantee that speech of celebrities can perfectly represent that of the general public. A 100% spontaneous speech dataset covering a large population and diverse genres is certainly valuable, but constructing such a dataset will be very difficult, considering the constraints in terms of legislation and technical possibility. At present, although not fully spontaneous and representative, speech of celebrities is sufficient for our research purpose.

Recently, CN-Celeb1 has attracted increasing attention and several multi-genre studies have been carried out using this dataset (Kang et al., 2020; Kataria et al., 2020; Mo and Xu, 2020; Chen et al., 2021). Although very valuable, CN-Celeb1 is not large enough to be used as a standalone training set. We therefore present a new large-scale multi-genre corpus called CN-Celeb2 to meet the requirements for multi-genre training. This section summarizes the data collection pipeline and presents the data profile of this new corpus. CN-Celeb1 and CN-Celeb2 make up the overall CN-Celeb corpus. It has been published in OpenSLR[5] and is freely available for researchers.

[4] Others may include recognition with disguised speech or non-speech signals such as laugh and cough (Zhang et al., 2018).

[5] https://openslr.org/82/.

Table 2
Comparison between *CN-Celeb1* and *CN-Celeb2*.

	CN-Celeb1	CN-Celeb2
Language	Chinese	Chinese
Genre	11	11
# of Sources	1	5
# of Spks	1000	2000
# of Utters	130,109	529,485
# of Hours	274	1090
# of SSMC Spks	745	658
Human check	Yes	Yes

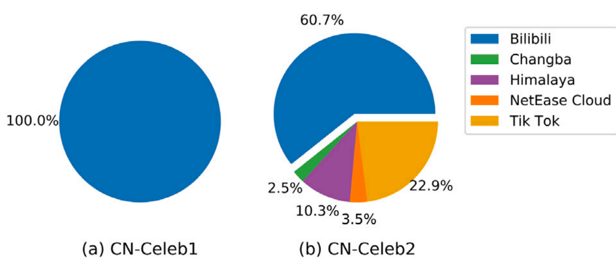


Fig. 1. The media source distribution of (a) *CN-Celeb1* and (b) *CN-Celeb2*.

Table 3
The audio duration distribution of *CN-Celeb1* and *CN-Celeb2*.

Duration (s)	CN-Celeb1		CN-Celeb2	
	# of Utters	Proportion	# of Utters	Proportion
<2	41,658	32.02%	36,505	6.89%
2–5	38,629	29.69%	57,215	10.81%
5–10	23,497	18.06%	266,799	50.39%
10–15	10,687	8.21%	154,120	29.11%
>15	15,638	12.02%	14,846	2.80%

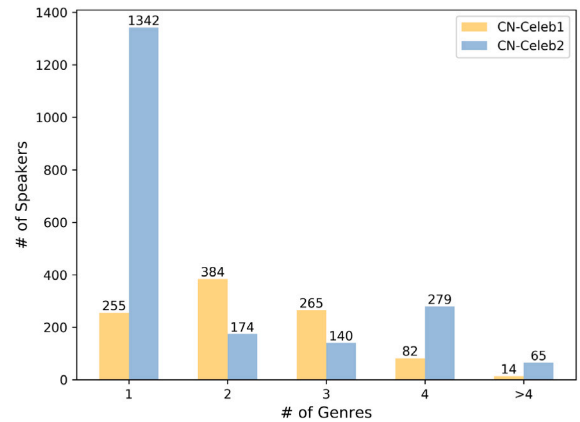


Fig. 2. The distribution of multi-genre speakers in *CN-Celeb1* and *CN-Celeb2*.

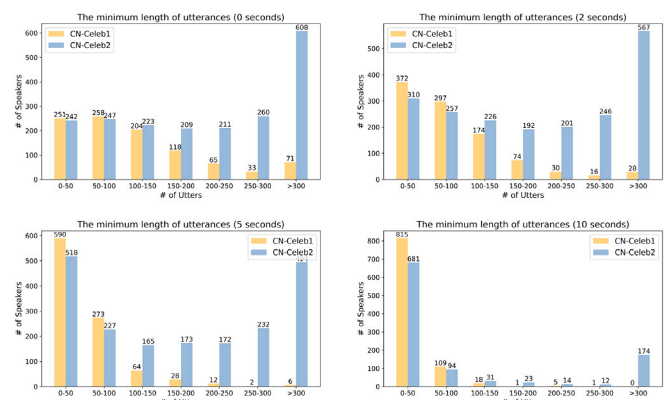


Fig. 3. The distribution of speakers that have different numbers of utterances in *CN-Celeb1* and *CN-Celeb2*. For the four plots, the minimum length of the utterances is set to be 0 s, 2 s, 5 s, 10 s respectively, representing that utterances shorter than the minimum length are ignored when computing the statistics.

3.2. Data description

CN-Celeb2 shares the same features as CN-Celeb1. Both the corpora were collected from Chinese open media, and all the constituent speakers are Chinese celebrities. The overall statistics are shown in Table 2.

- The data volume of CN-Celeb2 is larger than CN-Celeb1. CN-Celeb2 contains 529,485 utterances from 2000 Chinese celebrities, the total speech duration is 1090 h, which is around 4 times the volume of CN-Celeb1.
- CN-Celeb2 was collected from more media sources compared to CN-Celeb1. All the data of CN-Celeb1 were collected from Bilibili.[6] For CN-Celeb2, we collected singing data from NetEase Cloud[7] and Changba,[8] recitation data from Himalaya,[9] and vlog data from Tik Tok.[10] Fig. 1 shows the source distribution of the two datasets.
- The audio duration distributions of CN-Celeb1 and CN-Celeb2 are shown in Table 3. It can be seen that short utterances account for a larger proportion in both CN-Celeb1 and CN-Celeb2, which reflects the scenario of most real-life applications but leads to a more complex challenge for speaker recognition research.
- CN-Celeb2 includes more data for the genres that were not well covered by CN-Celeb1, such as vlog, live broadcast. The genre distributions of CN-Celeb1 and CN-Celeb2 are shown in Table 4.
- CN-Celeb2 contains less multi-genre speakers than CN-Celeb1. Fig. 2 shows the distribution of multi-genre speakers in the two datasets. The reason is that the number of celebrities who are active in multiple domains is limited, making it very difficult to collect multi-genre data. Compared to CN-Celeb1 where 75% speakers are multi-genre, there are only 33% multi-genre speakers in CN-Celeb2. Note that although a large proportion of the speakers are single-genre, the speech utterances were still collected from multiple sessions in diverse conditions. Such multi-session data is also very valuable and can be used to develop techniques that can address the multi-genre challenge with limited multi-genre data.
- The averaged number of utterances per speaker is 265 in CN-Celeb2 and 130 in CN-Celeb1. Fig. 3 shows the number of speakers that have different numbers of utterances in the two datasets, where the minimum length of the utterances counted in the statistics is set to be 0 s, 2 s, 5 s, 10 s respectively in the four plots. In other words, utterances shorter than the minimum length are ignored when computing the statistics. It reflects the proportion of speakers that have different numbers of utterances.
- The averaged number of sessions per speaker is 17 in CN-Celeb2 and 6 in CN-Celeb1. Note that we treat each video as a single session, even though some videos may involve multiple sessions (e.g., in a movie or play). Fig. 4 shows the number of speakers that have different numbers of sessions in the two datasets, where the minimum length of the utterances counted in the statistics is set to be 0 s, 2 s, 5 s, 10 s respectively in the four plots. It reflects the proportion of speakers that have different numbers of sessions.

[6] https://bilibili.com.
[7] https://music.163.com/.
[8] https://changba.com/.
[9] https://www.ximalaya.com/.
[10] https://www.douyin.com/.

Table 4
The genre distribution of *CN-Celeb1* and *CN-Celeb2*.

Genres	CN-Celeb1			CN-Celeb2		
	# of Spks	# of Utters	# of Hours	# of Spks	# of Utters	# of Hours
Advertisement	17	120	0.18	66	1542	3.86
Drama	160	7247	6.43	268	13,116	16.32
Entertainment	483	22,064	33.67	616	31,982	60.84
Interview	780	59,317	135.77	519	34,024	81.28
Live Broadcast	129	8747	16.35	388	167,019	439.95
Movie	62	2749	2.20	133	4449	5.77
Play	69	4245	4.95	127	14,992	22.04
Recitation	41	2747	4.98	218	58,231	129.18
Singing	318	12,551	28.83	394	42,157	75.19
Speech	122	8401	36.22	394	36,680	82.58
Vlog	41	1894	4.15	488	125,293	177.00
Overall	1000	130,109	273.73	2000	529,485	1090.01

Table 5
Comparison of existing speaker recognition datasets.

Name	Collection environment	Language	Data source	# of Spks	# of Utters	Free
Forensic comparison (Morrison et al., 2015)	Clean	Australian English	Mobile	552	1264	Yes
Free ST Chinese Mandarin[a]	Clean	Chinese	Mobile	855	102,600	Yes
TIMIT (Fisher, 1986; Zue and Seneff, 1996)	Clean	English	Telephone	630	6300	No
SWB (Godfrey et al., 1992)	Clean	English	Telephone	3114	33,039	No
CSLU[b]	Mostly clean	English	Telephone	500	6000	No
NIST SRE (Gonzalez-Rodriguez, 2014; Greenberg et al., 2020)	Clean, noisy	Multilingual	Telephone, microphone	–	–	No
Aishell-1 (Bu et al., 2017)	Clean	Chinese	Mobile	400	140,000	Yes
Aishell-2 (Du et al., 2018)	Clean	Chinese	Mobile	1991	1,000,000	Yes
RSR2015 (Larcher et al., 2014)	Clean	English	Mobile, tablet	300	190,000	No
RedDots (Lee et al., 2015)[c]	Clean	Multilingual	Mobile	62	13,500	Yes
HI-MIA (Qin et al., 2020)	Near/far-field	Chinese, English	Microphone, mobile	340	3,940,000	Yes
BookTubeSpeech (Pham et al., 2020)	Multi-media	English	BookTube	8450	38,707	Yes
SITW (McLaren et al., 2016)	Interview	English	Open-source media	299	2800	Yes
VoxCeleb1 (Nagrani et al., 2017)	Mostly interview	Mostly English	YouTube	1251	153,516	Yes
VoxCeleb2 (Chung et al., 2018)	Mostly interview	Multilingual	YouTube	6112	1,128,246	Yes
CN-Celeb1 (Fan et al., 2020)	Multi-genre	Chinese	Bilibili	1000	130,109	Yes
CN-Celeb2	**Multi-genre**	**Chinese**	**Multi-media sources**	**2000**	**529,485**	Yes

[a] http://www.openslr.org/38/.
[b] https://catalog.ldc.upenn.edu/LDC2006S26.
[c] Here presents RedDots_r2015q4_v1 released up to August 17th 2015.

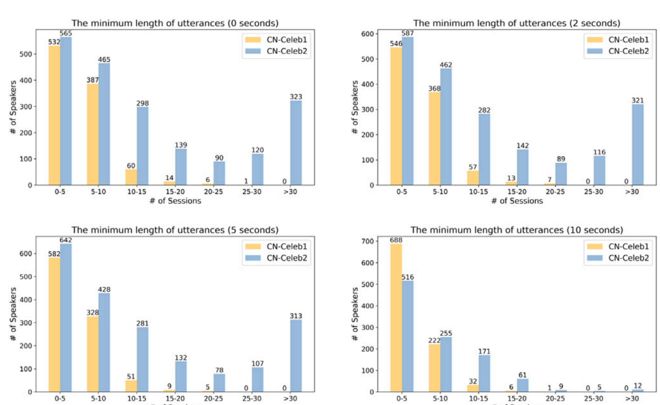


Fig. 4. The distribution of speakers that have different number of sessions in *CN-Celeb1* and *CN-Celeb2*. For the four plots, the minimum length of the utterances is set to be 0 s, 2 s, 5 s, 10 s respectively, representing that utterances shorter than the minimum length are ignored when computing the statistics.

We also compare CN-Celeb1 and CN-Celeb2 with some existing datasets in Table 5. Note that NIST SRE dataset has not been enumerated as it is not a standalone corpus and changes in composition with each release.

3.2.1. *Collection pipeline*

CN-Celeb2 was collected following a similar pipeline as CN-Celeb1. The source code has been published online to help readers reproduce our work and collect their own data.[11]

Broadly, the collection process comprises two stages: in the first stage, potential segments of the Person of Interest (POI) were extracted from a large amount of raw videos with an automatic tool, and then in the second stage, human check was employed to remove incorrect segments. This process is much faster than purely human-based segmentation, and also avoids potential errors caused by a purely automated process, as VoxCeleb has employed. We highlight that the human check is important in our case: because the multi-genre data are very complex in both video and audio, the purely automatic process cannot deal with such complexity. Although the human check makes the process more costly, it results in a more valuable dataset.

Our automatic pipeline is largely borrowed from the one used for collecting VoxCeleb1 (Nagrani et al., 2017) and VoxCeleb2 (Chung et al., 2018), with some modifications to increase the efficiency and precision. In particular, we introduced an additional relaxation & recovery step that employs both image and speech information to validate the extracted segments. The detailed steps of the collection process are summarized as follows.

[11] https://github.com/celebrity-audio-collection/videoprocess.

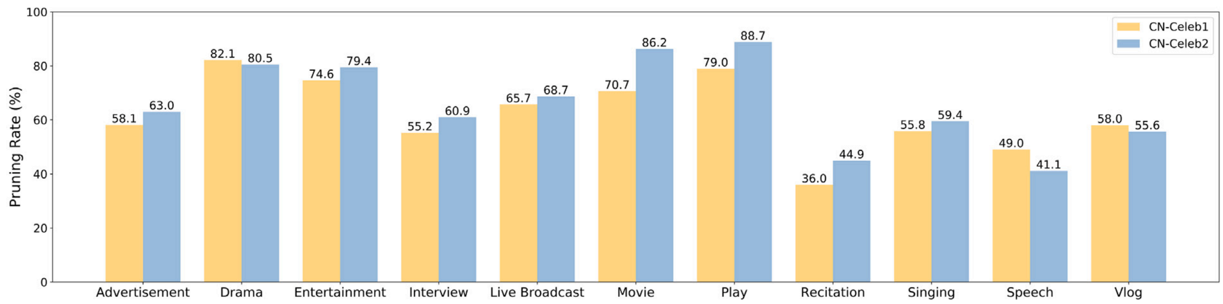


Fig. 5. Pruning rate with human check when collecting *CN-Celeb1* and *CN-Celeb2*.

- **STEP 1. POI list design**. We manually selected 2000 Chinese celebrities as our target speakers. These speakers were mostly from the entertainment sector, including singers, drama actors /actresses, news reporters and interviewers. Regional diversity was also taken into account so that variations in accent were covered.
- **STEP 2. Pictures and videos download**. Pictures and videos of the 2000 POIs were downloaded from several media sources by searching for the names of the persons.
  For pictures, we developed a crawler to download pictures of POIs using the search engine Baidu.[12] For each POI, 120 pictures were downloaded and 10 clear pictures were selected by a human examiner. Since the POIs are well-known, the selection was very easy and the errors were rare. We then arranged a double check process, in which the selected 10 pictures were rechecked by another examiner.
  For videos, we firstly searched the POI name in the source media. In order to specify that we were searching for POI names, the word 'human' was appended to the search queries. Secondly, for each POI, at most 10 videos were manually selected and downloaded for each genre, depending on how many videos can be found in that genre. One examiner was responsible for one POI, and a double check process was arranged to ensure the quality.
- **STEP 3. Face detection and tracking**. For each POI, we firstly obtained the portrait of the person by detecting and clipping the face images from all pictures of that person. The RetinaFace algorithm was used for the detection and clipping (Deng et al., 2020). Thereafter, video segments that contained the target person were extracted. This was achieved via the following three-step process: (1) For each frame, detect all the faces appearing in the frame using RetinaFace; (2) Determine if the target person appears by comparing the POI portrait and the faces detected in the frame using the ArcFace face recognition system (Deng et al., 2019). (3) Apply the MOSSE face tracking system (Bolme et al., 2010) supported by OpenCV Tracker[13] to produce face streams.
- **STEP 4. POI speaking verification by SyncNet**. As in Nagrani et al. (2017), we employed a mouth-speech synchronization detection system (Chung and Zisserman, 2016) to verify that the target person (POI) is speaking, by testing if the mouth movement of the target person is synchronized with the speech signal. This is necessary especially in genres such as movie, drama and entertainment where the target person appears in the video but the speech is from other persons. A pre-trained SyncNet model[14] was used in our implementation.
- **STEP 5. Relaxation & recovery by speaker diarization**. Although SyncNet worked well in most cases, it failed for videos of complex genres such as advertisement, movie and vlog. In these genres, scenes may change abruptly in time, leading to a large number of small POI segments. To solve this problem, we employed a relaxation & recheck process: Firstly relaxes the result of SyncNet by merging the adjacent POI segments if their distance is less than 10 frames, and secondly recovers the true POI part from the merged segments, by using a speaker diarization system. The details of the recovery step are as follows: we used an off-the-shelf speaker diarization system[15] to split the input speech into speaker-homogeneous trunks. The diarization system was based on the UIS-RNN model (Zhang et al., 2019), which firstly split the entire speech into small pieces of 1 s length with 0.6 s overlap, and then extracted the speaker embeddings of all the pieces using a pre-trained VGG speaker recognition model (Xie et al., 2019). The UIS-RNN model then clustered the embeddings of all the pieces into several clusters, each corresponding to a single speaker. According to the clustering result, adjacent pieces were merged together if they were from the same cluster, resulting into speaker homogeneous trunks.
  In order to utilize the diarization result, the *POI cluster* was firstly identified as the one that overlapped with the SyncNet output most. Then the overlap between the speaker-homogeneous trunks of the POI cluster (output from the diarization) and the merged POI segments from SyncNet (output from the relaxation) was output as the final POI speech segments.
- **STEP 6. Human check**. The POI segments produced with the above automated pipeline were finally checked by humans. To ensure the quality, we designed an iterative process: For each POI, the extracted POI segments were firstly assigned to an examiner for the first-round full check, and then 20 segments were sampled and assigned to another examiner for spotting check. If the accuracy returned by the spotting check was lower than 90%, the task would be bounced back to the first examiner for the second-round full check. This process repeated until the spotting check returned an accuracy higher than 90%.
  According to our experience, this human check is rather efficient: one could check 1 h of speech in 1 h. As a comparison, if we do not apply the automated pre-selection, checking 1 h of speech requires about 4 h.

3.3. Pruning rate

Human check is the most costly step in the CN-Celeb pipeline. An interesting question is that if this check is necessary, especially for the

12 https://image.baidu.com/.
13 https://learnopencv.com/object-tracking-using-opencv-cpp-python/.
14 https://github.com/joonson/syncnet_python.
15 https://github.com/taylorlu/Speaker-Diarization.

genres that are relatively easy to deal with. For example, for interview, the accuracy of the automatic process might be sufficiently high — at least VoxCeleb1 and VoxCeleb2 were collected using a similar pipeline, without any human check.

To investigate how necessary the human check is, we compute the *pruning rate* for each genre, i.e., the proportion of frames that were pruned by human check. The results are shown in Fig. 5. Firstly, it can be seen that for some genres (e.g., speech, recitation, interview), the pruning rate is relatively small, indicating that the automatic process can be regarded as reliable; however for other genres (e.g., play, movie), the pruning rate is very high, which means there is a big proportion of frames produced by the automated process that are incorrect. These results confirm the necessity of the human check in complex genres. As we will see in the next section (Table 9), speaker recognition performance is often worse on genres with a high pruning rate. This observation indicates that more complex the genre is, more necessary the human check is.

4. Experiment I: Multi-genre challenge

In this section, we will study the basic performance of speaker recognition systems on multi-genre conditions. Our goal is to investigate the behavior of the state-of-the-art systems when the enrollment/test genre is different from that of the training, and when the enrollment and test are in different genres.

4.1. Basic results

We built two speaker recognition systems, one is based on the i-vector model and one is based on the x-vector model. The two systems are used as baseline systems to test the single-genre performance and multi-genre performance on SITW and CN-Celeb.E, respectively.

4.1.1. Data

VoxCeleb[16]: This is used as the training data. It comprises VoxCeleb1 and VoxCeleb2, amounting to 2000+ hours of speech signals from 7000+ speakers. Data augmentation was applied to improve robustness, with the MUSAN corpus (Snyder et al., 2015) to generate noisy utterances, and the room impulse responses (RIRS) corpus (Ko et al., 2017) to generate reverberant utterances.

SITW: This dataset is used for testing single-genre performance. It comprises 6445 utterances from 299 speakers (precisely, this is the Eval.Core set within SITW). Note that this dataset is similar to VoxCeleb in terms of data properties, and so there is no genre mismatch when used as the test set. The test protocol (trials for test) follows the Kaldi SITW recipe.[17]

CN-Celeb.E: This dataset is a subset of CN-Celeb1, containing 18,224 utterances from 200 speakers. All the speakers are multi-genre. Note that data of the interview genre is similar to VoxCeleb and SITW, although they are from different media sources. Therefore, there is a domain mismatch between training and enrollment/test, while the genres are the same. During the test, speakers enroll once and are tested against multiple utterances. The enrollment speech (might be split into multiple utterances) for each speaker is 28 s on average. The average length of the test utterances is 8 s, and there are 18,024 test utterances in total, 90 utterances per speaker in average. The trials are produced by cross pairing the enrollment speech and the test utterances, amounting to 3,604,800 gender-independent test trials. More details of the test data and trials are shown in Table 6.

SITW(S): This is an auxiliary dataset for performance analysis. As the average length of SITW is much longer than that of CN-Celeb.E, the results on these two datasets are not directly comparable. For a more reasonable comparison, we trim the utterances of SITW to match the average length of the utterances in CN-Celeb.E, which is 28 s and 8 s for enrollment and test, respectively. This new dataset is called SITW(S), and the test protocol on this dataset is the same as in SITW.

Table 6
Data profile of CN-Celeb.E.

Enroll data	Avg. Length per Utt	28 s
	# of Utters per Spk	5
Test data	Avg. Length per Utt	8 s
	# of Utters per Spk	90
Gender info	# of Female	84
	# of Male	116
Trials	# of Target	18,024
	# of Nontarget	3,586,776

4.1.2. Baseline systems

In this study, we firstly use the SITW recipe of the Kaldi toolkit (Povey et al., 2011) to build our i-vector and x-vector baselines. This basic recipe may not achieve the best performance on a particular dataset, but has been demonstrated to be highly competitive and generalizable by many researchers with their own data and model settings. We therefore consider that this recipe can represent a stable state-of-the-art technique. Moreover, using this recipe allows others to reproduce our results easily.

i-vector system: The i-vector model was built following the Kaldi SITW/v1 recipe. The acoustic features comprise 24-dimensional MFCCs plus the log energy, augmented by first- and second-order derivatives, resulting in a 75-dimensional feature vector. Moreover, cepstral mean normalization (CMN) is employed to normalize the channel effect, and an energy-based voice active detection (VAD) is used to remove silence segments. The universal background model (UBM) consists of 2048 Gaussian components, and the dimensionality of the i-vector is set to be 400. For the back-end model, LDA is firstly employed to reduce the dimensionality to 150, and then PLDA is used to score the trials.

x-vector system: The x-vector model was created following the Kaldi SITW/v2 recipe. The acoustic features are 30-dimensional MFCCs. The DNN architecture involves 5 time-delay (TD) layers to learn frame-level deep speaker features, and a temporal statistic pooling (TSP) layer is used to accumulate the frame-level features to utterance-level statistics, including the mean and standard deviation. After the pooling layer, 2 fully-connection (FC) layers are used as the classifier, for which the outputs correspond to the number of speakers in the training set. Once trained, the 512-dimensional activations of the penultimate layer are read out as an x-vector. The back-end model is the same as in the i-vector system, which includes LDA for dimensional reduction, and PLDA to score the trials.

4.1.3. Baseline results

The overall results in terms of equal error rate (EER) are shown in Table 7. It can be seen that both the i-vector and x-vector systems obtain reasonable performance on SITW, and the results are similar to the official results released with Kaldi recipes. The results on SITW(S) are slightly worse than those on SITW, which is expected as the enrollment and test utterances are shorter in this dataset. The performance on CN-Celeb.E is much worse. For example, compared to the x-vector results on CN-Celeb.E and SITW(S), the EER on CN-Celeb.E increases by more than 300%. These results clearly indicate that the state-of-the-art speaker recognition systems cannot inherently deal with the complexity introduced by multiple genres.

[16] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/.
[17] https://github.com/kaldi-asr/kaldi/egs/sitw.

Table 7
EER (%) results with the i-vector and x-vector baseline systems.

System	Training set		Test set		
	Front-end	Back-end	SITW	SITW(S)	CN-Celeb.E
i-vector	VoxCeleb	VoxCeleb	5.66	7.41	18.37
x-vector	VoxCeleb	VoxCeleb	3.48	4.62	16.59

Table 8
EER (%) results of more powerful x-vector systems. 'TSP' represents temporal statistic pooling. 'SAP' represents self-attentive pooling. 'AAM' represents additive angular margin.

Topology	Pooling	Loss	SITW	CN-Celeb.E
TDNN	TSP	Softmax	2.43	16.87
TDNN	TSP	AAM-Softmax	2.49	16.65
TDNN	SAP	Softmax	2.41	17.11
TDNN	SAP	AAM-Softmax	2.57	16.96
ResNet-34	TSP	Softmax	2.41	16.74
ResNet-34	TSP	AAM-Softmax	1.96	16.51
ResNet-34	SAP	Softmax	2.16	17.33
ResNet-34	SAP	AAM-Softmax	2.30	16.52

Table 9
EER (%) results of the baseline systems in different genres on CN-Celeb.

Genres	# of Spks	# of Utters	i-vector	x-vector
Advertisement	75	781	12.43	9.37
Drama	377	4521	14.66	11.70
Entertainment	1020	18,931	9.48	7.31
Interview	1253	41,586	9.06	6.98
Live broadcast	496	154,249	6.79	5.42
Movie	165	1495	14.17	11.47
Play	170	5476	13.87	11.56
Recitation	259	58,839	19.21	16.55
Singing	683	38,879	23.37	20.86
Speech	331	39,792	4.19	3.21
Vlog	524	120,812	7.92	5.31
Overall	3000	485,361	8.75	7.43

4.1.4. More powerful x-vector systems

We have also implemented more powerful x-vector systems by using arguably more advanced techniques. In this work, we used an open-source code[18] and tested a bunch of state-of-the-art architectures/techniques, such as ResNet (Chung et al., 2018; Zeinali et al., 2019), self-attentive pooling (Zhu et al., 2018) and additive angular margin loss (Deng et al., 2019), on SITW and CN-Celeb.E. The results are shown in Table 8. It can be found that although these more advanced systems obtain obvious performance improvements over the basic x-vector system on SITW (1.96% vs. 3.48%), they did not show clear superiority on CN-Celeb.E (16.51% vs. 16.59%). It indicates that these advanced techniques may simply overfit to the training condition and so are of little help in solving the multi-genre challenge. This is another reason why we use the basic x-vector architecture as our baseline.

4.2. Within-genre results

In this section, we break down the multi-genre tests and compare the performance in different genres. We focus on the case where the enrollment and test utterances are from the same genre. To ensure the confidence of experimental results, we use the overall CN-Celeb dataset (3000 speakers in total) for test and also filter away short utterances with less than 5 s duration. Table 9 presents the EER results, and Fig. 6 shows the DET curves. In the test for each genre, 5 utterances of each speaker are randomly selected for enrollment, and the remaining utterances are used for test.

It can be observed that with both the i-vector and the x-vector systems, performance on different genres is substantially different. For genres such as speech, live broadcast, vlog and interview, the EER results are less than 8%, and the performance is relatively acceptable. While for genres such as singing, recitation, drama and movie, the EER results are more than 12%, and the performance is quite unacceptable. The performance discrepancy on different genres could be attributed to two reasons: Firstly, the speaker-independent variation is naturally much more significant for some genres compared to others. For example, the channel, background and speaking style in speech and interview tend to be more controlled than those in singing and drama. The more complex the variation, the more difficult it is for the speaker traits to be identified. Secondly, the i-vector and x-vector models are trained with VoxCeleb that mainly consists of interview speech. This perhaps makes the model biased to interview and similar genres such as speech and live broadcast.

Another observation is that even for the interview genre, the performance is much worse than that obtained on SITW (6.98% vs. 3.48% on the x-vector system). This is clearly caused by the discrepancy between channels and languages of the two sources of CN-Celeb and SITW (Bilibili, etc., for CN-Celeb while YouTube for SITW). It indicates that the true performance of the present state-of-the-art speaker recognition system is not as good as one though from the results reported on SITW, even without genre mismatch.

Taking the x-vector system as an example, if we set the overall EER (7.43%) as the threshold for an *acceptable* system, the present speaker recognition system can only obtain reasonable performance in a few genres. These genres are speech, vlog, live broadcast, interview and entertainment. This clearly demonstrated how challenging the multi-genre problem is.

4.3. Cross-genre results

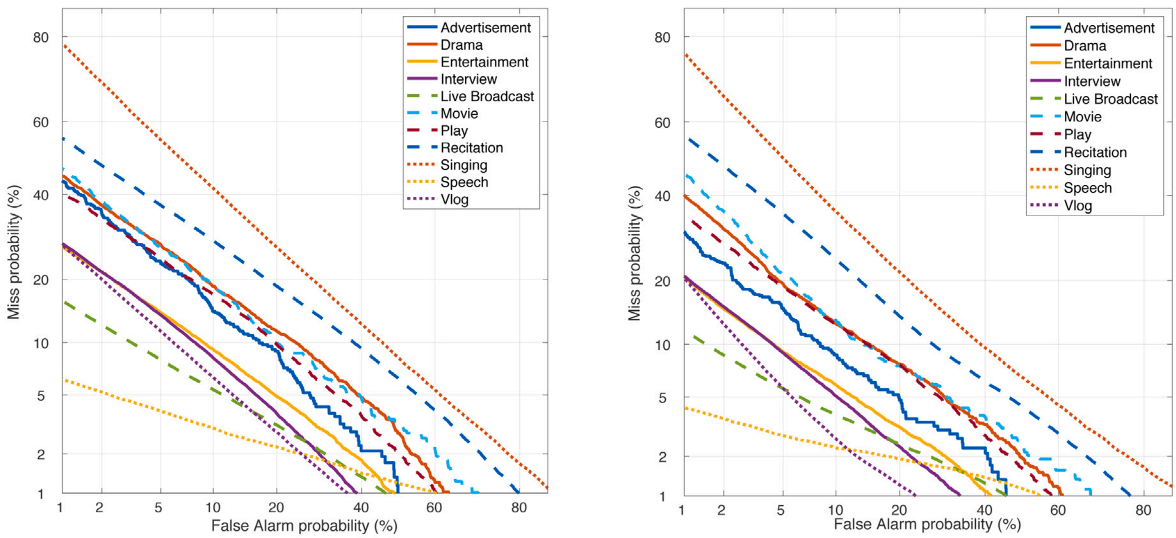
In this section, we focus on cross-genre test and compute a genre-to-genre performance matrix. Figs. 7 and 8 show the EER results with the i-vector system and the x-vector system, respectively. The numerical values shown in the blocks are the EER results under the enrollment genre corresponding to its row and the test genre corresponding to its column. Note that the diagonal results show the in-genre results in Table 9. The last column shows the overall results that the enrollment is based on one genre and test is on all the genres. Note that there are two blank cells (recitation-advertisement and recitation-advertisement). This is because there is only 1 recitation-advertisement speaker, which makes the EER result unreliable.

Firstly, paying attention to the overall results (the last column) enrolled with each genre, it can be found that the best performance is obtained when the enrollment is with the speech genre, and the worst performance is obtained when the enrollment is with the singing genre. Roughly stating, the simpler the enrollment condition (e.g., speech, interview, etc.), the better is the average performance obtained. This is good news as one tends to enroll in the silent environments with careful pronunciation.

Secondly, the results with the same enroll-test pair (e.g., singing-speech and speech-signing) are roughly the same. This phenomenon was also observed in some previous studies on multiple speaking styles, e.g., Park et al. (2017), Shriberg et al. (2009), Park et al. (2016) and Shriberg et al. (2008). This indicates that the variation in the enrollment genre is similar to the variation in the test genre.

Thirdly, the cross-genre performance is determined by two factors: (1) the complexity of the enrollment/test genre, (2) the degree of match between the two genres. For example, when the enrollment is movie, the EER is 14.17% when the test genre is movie, which is not very bad due to the matched genre; however the EER is 11.67% when the test genre is speech, which is even better, due to the simpler condition of the speech genre. In general, worse performance is obtained when the enrollment and test genres are less matched.

[18] https://github.com/kjw11/tf-kaldi-speaker.



(a) i-vector (b) x-vector

Fig. 6. DET curves of different genres with the i-vector and x-vector systems.

Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement	12.43	22.69	21.66	12.43	12.33	36.14	18.52	-	24.36	19.62	10.89	17.89
Drama	24.71	14.66	24.12	23.62	18.54	26.20	11.71	37.88	34.17	30.00	16.70	21.81
Entertainment	17.14	23.57	9.48	12.55	15.73	23.26	17.96	9.15	28.89	13.22	13.36	14.12
Interview	21.17	22.81	14.30	9.06	13.69	20.69	15.60	12.75	31.66	11.86	14.45	14.24
Live Broadcast	11.11	20.93	17.05	15.51	6.79	23.15	14.01	21.84	24.81	23.08	12.88	7.29
Movie	29.21	26.72	17.97	22.61	15.41	14.17	10.61	14.29	29.57	11.67	29.73	20.23
Play	9.09	20.53	16.64	19.43	14.29	26.67	13.87	5.88	29.66	14.78	18.18	14.55
Recitation	-	24.35	29.11	12.02	10.50	10.00	6.64	19.21	31.94	10.18	33.33	13.48
Singing	24.91	32.53	32.65	31.95	26.41	32.00	17.58	23.81	23.37	18.45	25.88	20.33
Speech	31.18	23.08	14.86	13.00	17.86	25.00	12.02	6.55	22.15	4.19	33.65	6.19
Vlog	10.53	22.45	27.36	13.30	12.70	17.81	16.13	20.00	26.10	33.66	7.92	7.83

EER (%) 38.0 21.0 4.0

Fig. 7. Cross-genre tests with the i-vector system. The lightness of the color corresponds to the numerical value of the EER (%).

Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement	9.37	20.17	15.63	10.18	6.61	31.33	18.52	-	25.26	25.16	8.42	17.10
Drama	22.35	11.70	21.37	22.37	14.35	25.88	8.46	31.82	32.78	29.31	16.48	20.22
Entertainment	14.11	19.61	7.31	9.70	12.45	18.40	14.09	9.78	27.53	12.07	10.87	11.90
Interview	20.15	21.72	11.46	6.98	11.04	18.77	12.62	7.84	28.92	10.72	11.91	12.40
Live Broadcast	8.50	17.38	14.40	13.50	5.42	18.52	11.48	22.99	22.34	16.99	9.76	6.00
Movie	29.21	25.00	15.07	19.13	12.57	11.47	12.46	14.29	29.42	14.51	25.38	18.29
Play	9.09	16.84	14.91	19.29	15.09	24.67	11.56	5.88	28.25	11.82	18.18	12.65
Recitation	-	20.00	23.58	8.61	9.24	20.00	3.69	16.55	34.72	4.79	33.33	9.71
Singing	24.91	27.56	31.49	28.87	21.11	29.00	18.83	36.51	20.86	21.86	24.07	18.76
Speech	29.03	25.27	12.37	10.41	12.66	20.83	10.66	2.91	24.95	3.21	36.54	5.18
Vlog	5.26	19.73	26.75	10.45	10.38	15.07	9.68	20.00	25.40	36.63	5.31	5.32

EER (%) 37.0 20.0 3.0

Fig. 8. Cross-genre tests with the x-vector system. The lightness of the color corresponds to the numerical value of the EER (%).

In summary, the cross-genre phenomenon is highly complex, and in most of the conditions, the performance is not acceptable. Taking the x-vector system as an example, if we set the overall EER (7.43%) as the threshold for acceptance, there are only several conditions can be deemed acceptable, as shown in Fig. 9. These observations indicate once again that cross-genre is a very challenging problem.

For a deep analysis, we also report results in two metrics related to score calibration: the cost of log likelihood ratio (LLR) denoted by C_{llr} and the minimum cost of LLR denoted by C_{llr}^{min} (Ramos and Gonzalez-Rodriguez, 2008; Brümmer and Du Preez, 2006). Compared to EER, C_{llr} evaluates the averaged performance over all the possible settings on the priors and costs of the target and non-target trials, by assuming that the PLDA scores are LLRs. In the cross-genre situation, the enrollment and test conditions are drastically different, and so the PLDA scores may significant deviated from the true LLRs. In this case, a large C_{llr} could be attributed to the lack of both discrimination and regularization of the PLDA scores, where discrimination refers to how well the scores can distinguish target and non-target trails, while regularization refers to how well the scores represent LLRs . A score calibration can be designed to map the scores to LLRs. This map is monotonic and so does not change the discrimination capacity of the scores but improves regularization. When the calibration is perfect (which can be obtained with a finite evaluation set), the resultant C_{llr} is C_{llr}^{min}. Therefore the difference between C_{llr} and C_{llr}^{min} (sometimes called C_{loss}) reflects how

Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement					6.61							
Drama												
Entertainment			7.31									
Interview				6.98								
Live Broadcast					5.42							6.00
Movie												
Play								5.88				
Recitation							3.69			4.79		
Singing												
Speech								2.91		3.21		5.18
Vlog	5.26										5.31	5.32

Fig. 9. Cross-genre results (EER %) for acceptable conditions (EER threshold = 7.43%) with the x-vector baseline system.

Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement	8.41	16.73	16.30	8.71	10.59	25.62	12.11	-	21.50	20.67	7.14	13.79
Drama	20.65	7.18	17.38	16.95	12.81	18.58	7.92	23.44	29.31	25.78	12.52	14.69
Entertainment	14.05	18.76	5.56	9.03	13.00	18.71	14.87	7.13	25.31	8.90	11.04	10.18
Interview	18.65	19.03	10.40	4.93	11.65	16.55	13.30	9.73	30.92	9.70	13.40	10.24
Live Broadcast	10.71	15.41	13.30	11.88	2.91	18.76	9.64	21.65	23.15	15.77	8.52	3.61
Movie	23.05	17.11	14.15	17.27	14.24	8.71	8.22	1.31	27.98	10.45	21.38	15.53
Play	4.63	15.61	13.87	16.18	10.21	17.30	6.64	1.31	24.54	9.77	18.90	10.31
Recitation	-	17.05	21.02	9.75	7.83	5.78	3.63	7.32	22.88	6.83	27.32	7.45
Singing	23.08	25.85	26.60	30.52	25.90	27.31	18.28	21.29	12.03	24.33	23.61	17.36
Speech	28.51	22.47	9.87	8.86	14.08	17.18	11.69	4.63	24.68	1.29	33.84	2.62
Vlog	10.76	15.92	21.19	11.06	6.75	15.92	12.51	12.75	26.49	34.06	4.20	4.57

C_{llr}
35.0
18.0
1.0

(a) C_{llr} results with the i-vector system under cross-genre tests

Enroll \ Test	Advertisement	Drama	Entertainment	Interview	Live Broadcast	Movie	Play	Recitation	Singing	Speech	Vlog	Total
Advertisement	0.397	0.636	0.571	0.399	0.398	0.801	0.438	-	0.692	0.567	0.324	0.540
Drama	0.679	0.464	0.670	0.642	0.544	0.719	0.372	0.809	0.877	0.720	0.506	0.622
Entertainment	0.511	0.667	0.321	0.405	0.494	0.639	0.546	0.320	0.777	0.390	0.448	0.449
Interview	0.607	0.662	0.449	0.302	0.460	0.596	0.480	0.367	0.822	0.377	0.472	0.447
Live Broadcast	0.385	0.594	0.516	0.468	0.245	0.624	0.423	0.649	0.702	0.518	0.412	0.258
Movie	0.719	0.682	0.544	0.654	0.513	0.460	0.353	0.040	0.822	0.372	0.713	0.610
Play	0.187	0.616	0.552	0.603	0.447	0.675	0.432	0.077	0.784	0.477	0.576	0.472
Recitation	-	0.704	0.723	0.387	0.314	0.270	0.208	0.582	0.746	0.304	0.542	0.433
Singing	0.712	0.828	0.831	0.831	0.731	0.854	0.549	0.663	0.699	0.567	0.714	0.618
Speech	0.693	0.694	0.447	0.399	0.491	0.525	0.420	0.212	0.634	0.159	0.878	0.215
Vlog	0.323	0.582	0.679	0.422	0.387	0.556	0.362	0.397	0.757	0.823	0.282	0.282

C_{llr}^{min}
0.88
0.46
0.04

(b) C_{llr}^{min} results with the i-vector system under cross-genre tests

Fig. 10. Cross-genre tests with the i-vector system. The lightness of the color corresponds to the numerical value of the C_{llr}/C_{llr}^{min}.

the PLDA scores biased from LLRs and how much the score calibration may contribute.

It should be noted that score calibration does not change EERs with each cross-genre test as the calibration is simply a monotonic score mapping. However, it may improve system performance on the overall test, by applying different calibration models for different cross-genre tests. This is because after the test-dependent calibration, the scores of different tests become comparable and a cross-test threshold is applicable.

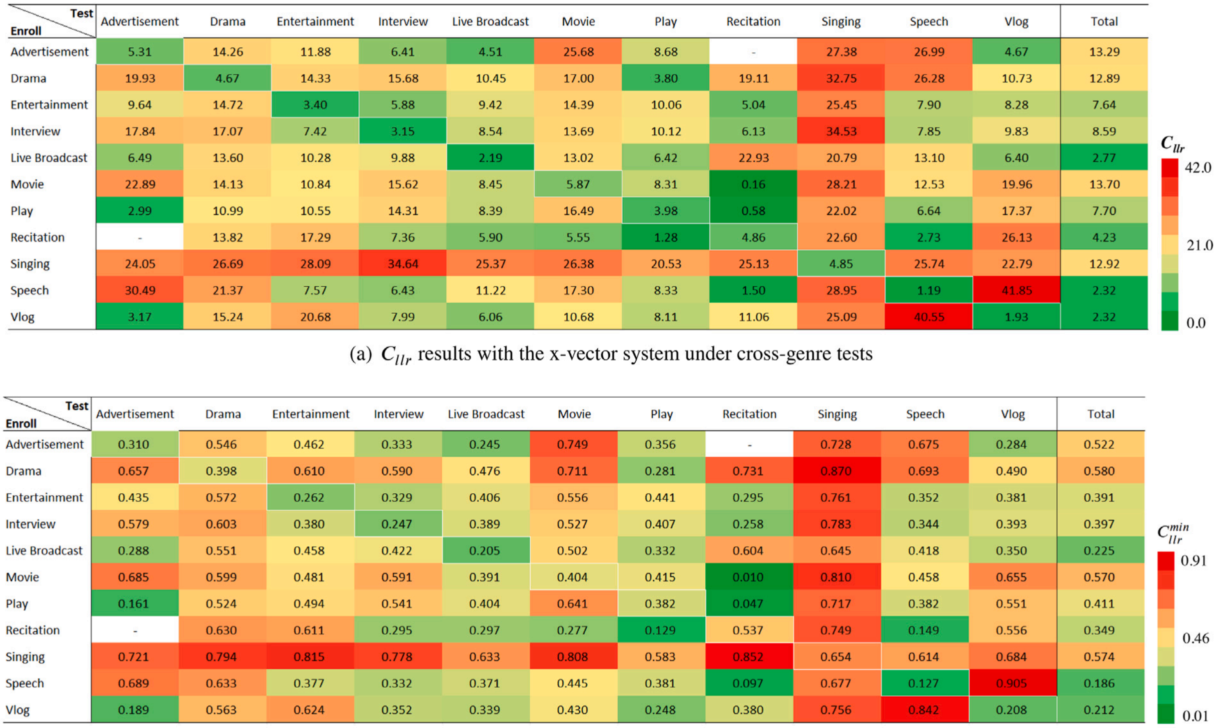
The C_{llr}/C_{llr}^{min} results are shown in Figs. 10 and 11 for the i-vector and x-vector systems, respectively. Note that the last column 'Total' reports the results with all the test trials pooled of each row.

We firstly observe that the performance tendency of the C_{llr}^{min} results are similar to that of the EER results. This is not very surprising as both evaluate the discrimination power of the scores, though C_{llr}^{min} reflects the expected error rate while EER is the error rate at the equilibrium point of false acceptance and false rejection. Due to the similar trend, we will keep use EER as the main metric and report the EER results only when discussing the relative performance.

Moreover, we found that there is a large gap between C_{llr} and C_{llr}^{min}, and this gap is more significant for the test scenarios where the enroll-test mismatch is more obvious (specified by results in EER and C_{llr}^{min}). This indicates that the PLDA scores are far from LLRs, and score calibration is important in real applications where a threshold is required for decision making.

4.4. Statistical analysis

In this section, we analyze the performance degradation under within-genre and cross-genre conditions. A key insight is that if the distribution of the speaker vectors remains the same as in the training condition, then the performance with the PLDA scoring will be optimal assuming the PLDA model is well trained with the training data (Wang, 2020). Therefore, the performance degradation we observe in the multi-genre test (either the within-genre test or the cross-genre test) can be understood by the change in the statistics of the distribution. Since PLDA is the back-end scoring model, we compute the statistics related to PLDA, including the inter-speaker variance, intra-speaker variance



(a) C_{llr} results with the x-vector system under cross-genre tests

(b) C_{llr}^{min} results with the x-vector system under cross-genre tests

Fig. 11. Cross-genre tests with the x-vector system. The lightness of the color corresponds to the numerical value of the C_{llr}/C_{llr}^{min}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 10
Inter-speaker variances (S_b) and intra-speaker variances (S_w) of i-vectors and x-vectors derived from VoxCeleb and different genres of CN-Celeb.

Genres	i-vector		x-vector	
	S_b	S_w	S_b	S_w
VoxCeleb	0.920	1.042	1.053	0.894
Advertisement	0.443	1.020	1.162	3.167
Drama	0.297	1.227	0.982	4.835
Entertainment	0.300	1.176	0.889	4.179
Interview	0.297	1.114	0.755	3.967
Live broadcast	0.357	1.052	0.822	2.023
Movie	0.404	1.210	1.201	5.289
Play	0.250	1.232	0.868	4.740
Recitation	0.360	1.481	0.816	2.379
Singing	0.204	1.226	0.618	2.884
Speech	0.542	1.096	1.038	2.225
Vlog	0.349	1.307	0.830	2.726

Table 11
Mean shifts of i-vectors and x-vectors on different genres of CN-Celeb. *Euc.* represents Euclidean distance and *1-Cos.* represents Cosine similarity.

Genres	i-vector		x-vector	
	Euc.	*1-Cos.*	*Euc.*	*1-Cos.*
Advertisement	1.616	1.305	1.554	1.208
Drama	1.575	1.240	1.515	1.148
Entertainment	1.590	1.263	1.551	1.203
Interview	1.562	1.220	1.532	1.174
Live broadcast	1.498	1.122	1.533	1.174
Movie	1.587	1.259	1.519	1.153
Play	1.482	1.097	1.449	1.049
Recitation	1.491	1.111	1.498	1.123
Singing	1.653	1.366	1.526	1.164
Speech	1.410	0.994	1.426	1.017
Vlog	1.516	1.150	1.543	1.190

and the global mean shift. We will compute these statistics of each genre and observe the statistics change amongst different genres.

We firstly compute the inter-speaker variance and intra-speaker variance of each genre in Table 10. Besides, we also compute the mean shift between VoxCeleb and different genres of CN-Celeb. The mean vector of VoxCeleb is regarded as a reference vector, and the mean vectors of different genres in CN-Celeb are regarded as genre vectors. The mean shifts can be computed between the reference vector and different genre vectors based on the Euclidean distance and Cosine similarity. Results are shown in Table 11. Note that all these results are computed in the PLDA transformed space.

Firstly, it can be found that the statistics of VoxCeleb are more similar to matched genres (e.g., speech and interview) compared to unmatched genres (e.g., singing and recitation), and the mean shift is less significant in the case of matched genres. As the PLDA is trained on VoxCeleb, if the statistics change and the mean shift are significant, the performance will be impacted. Referring to the results in Table 9, it can be observed that the genre incurring the most significant statistics change and mean shift suffers from the most performance reduction.

The statistics change and the mean shift cause more severe problems in the cross-genre scenario, as the enroll data and test data in this scenario possess different statistical properties but they have to be represented in a single PLDA model. We presented a deep analysis on this enroll-test mismatch problem in our recent study (Li et al., 2020), but mismatch caused by the cross-genres challenge is yet to be thoroughly studied.

4.5. Qualitative analysis

In this section, we analyze the distribution of the speaker vectors by visualization. Data from 10 speakers of 11 genres are selected to generate speaker vectors. The t-SNE toolkit (Maaten and Hinton, 2008) is applied to project the speaker vectors to a 2-dimensional space. Figs. 12 and 13 present the distribution of i-vectors and x-vectors, respectively.

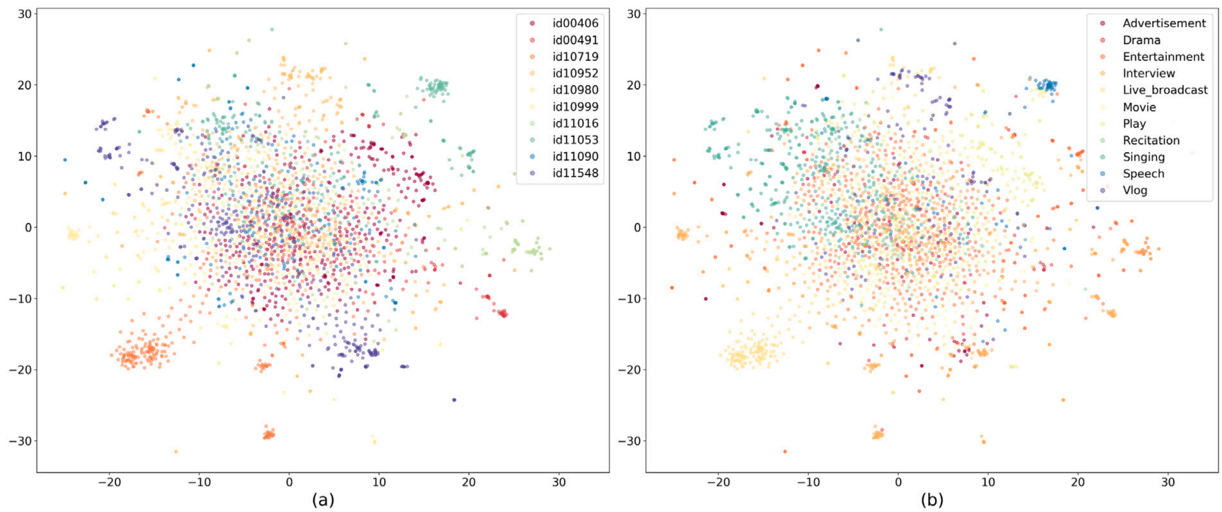


Fig. 12. The i-vector distribution plotted by t-SNE, where (a) each color represents a speaker, (b) each color represents a genre. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

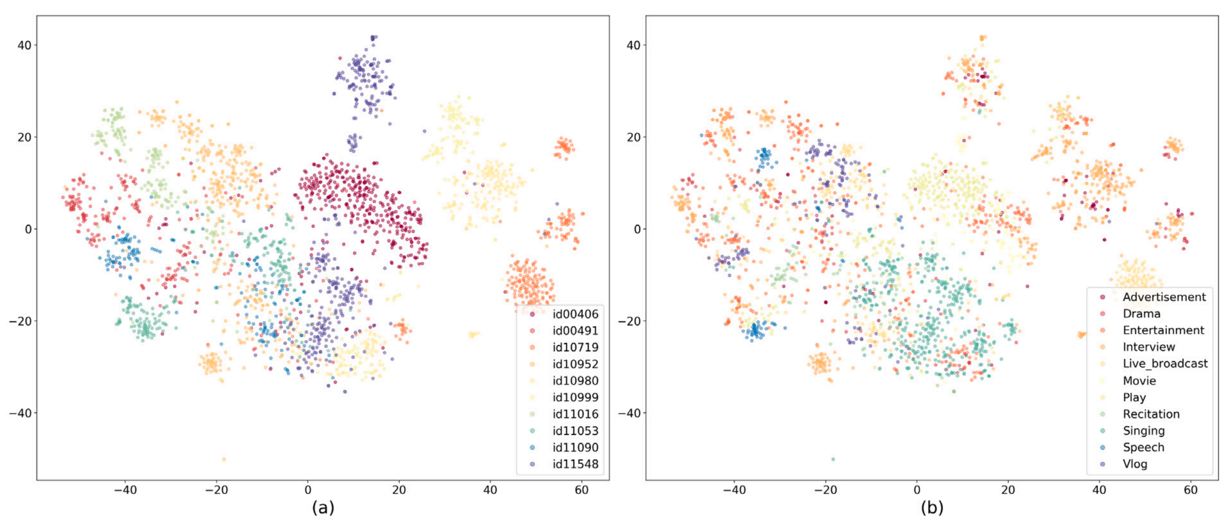


Fig. 13. The x-vector distribution plotted by t-SNE, where (a) each color represents a speaker, (b) each color represents a genre. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 12, it can be seen that with i-vectors, speakers are largely intermingled with each other. This is not surprising as the i-vector model is purely unsupervised and reflects variations of both speaker traits and acoustic conditions. Therefore, it is naturally hard to discriminate amongst speakers in multi-genre conditions.

For x-vectors presented in Fig. 13, one can observe larger inter-speaker distance and smaller intra-speaker distance compared to i-vectors. This indicates that the x-vector model has its advantage to tackle the acoustic complexity associated with multiple genres. Nevertheless, the genre complexity still leads to complicated intra-speaker distributions and overlap among different speakers. This demonstrates that multi-genre speaker recognition is quite challenging.

5. Experiment II: Multi-genre training

A straightforward approach to improve the performance under multi-genre conditions is to train the speaker recognition system using multi-genre data, called **multi-genre (MG) training**. Correspondingly, training using single-genre data (e.g., VoxCeleb) is called **single-genre (SG) training**. In the following experiments, we use CN-Celeb.T to denote the speech data in CN-Celeb but not in CN-Celeb.E. We use VoxCeleb for SG training, and CN-Celeb.T (2800 speakers in total) for MG training.

Besides, as shown in Fig. 2, there is a large proportion of multi-genre speakers in CN-Celeb (also in CN-Celeb.T). These multi-genre speakers are the most important for MG training, as their data can inform the model what variations are caused by genres. In order to investigate the contribution of the multi-genre speakers (which to some extent are truly multi-genre data), we relabel CN-Celeb.T such that the data from the same speaker for different genres are treated as originating from different speakers. This relabeled dataset is denoted by CN-Celeb.T/SI, where SI means speaker isolation. We call CN-Celeb.T/SI as **partial multi-genre data**, and the training on CN-Celeb.T/SI as **partial MG training**.

In this experiment, we compare the overall performance on CN-Celeb.E with different training schemes. Since the baseline system consists of two components: the i-vector/x-vector front-end model and the PLDA back-end model, we investigate the impact of MG training on the two components respectively. The results are shown in Table 12.

Table 12
Overall EER (%) results with single-genre, multi-genre and partial multi-genre training.

System	Scheme	Front-end	Back-end	CN-Celeb.E	
				Cosine	PLDA
i-vector	(a)	VoxCeleb	VoxCeleb	20.88	18.37
	(b)	VoxCeleb	CN-Celeb.T	20.88	15.30
	(c)	VoxCeleb	CN-Celeb.T/SI	20.88	16.31
	(d)	CN-Celeb.T	CN-Celeb.T	19.29	14.01
	(e)	CN-Celeb.T/SI	CN-Celeb.T/SI	19.25	14.81
x-vector	(a)	VoxCeleb	VoxCeleb	20.13	16.59
	(b)	VoxCeleb	CN-Celeb.T	20.13	13.44
	(c)	VoxCeleb	CN-Celeb.T/SI	20.13	14.76
	(d)	CN-Celeb.T	CN-Celeb.T	20.35	12.52
	(e)	CN-Celeb.T/SI	CN-Celeb.T/SI	20.83	13.65

5.1. Front-end model training

For the front-end models, we firstly compare the performance between SG training (a) and MG training (d). To eliminate the impact of the back-end model, we just look at the results with cosine scoring. It can be observed that MG training does not give a clear advantage over SG training, especially with the x-vector model (20.35% vs. 20.13%). This may be attributed to the bias in speaker numbers (2800 in CN-Celeb.T vs. 7000+ in VoxCeleb). Note that with the i-vector model, the MG training leads to slightly better performance than the SG training (19.29% vs. 20.88%) although the MG training used much less data. This better performance could be explained by the fact that the training data (CN-Celeb.T) and the test data (CN-Celeb.E) are coherent in the MG training, in both languages and acoustic conditions. This coherence is important for the i-vector model that is generative and descriptive.

Secondly, we compare the performance between MG training (d) and partial MG training (e). We again focus on the cosine scoring. Due to the unsupervised training strategy of the i-vector model, MG training and partial MG training obtain the same EER results. For the x-vector model, partial MG training is a bit inferior to MG training (20.83% vs. 20.35%). This is expected as the speaker labels of the partial multi-genre data lose the cross-genre information after speaker isolation.

5.2. Back-end model training

To investigate the impact of MG training on the back-end PLDA model, we compare the performance between SG training (a) and MG training (b). It can be seen that performance with PLDA scoring improves with the MG training, for both the i-vector and x-vector systems.

Secondly, when comparing the training scheme (c) to (a) and (b), it can be seen that although the partial MG training is worse than the MG training (14.76% vs. 13.44% for the x-vector system), it greatly outperforms the SG training (14.76% vs. 16.59% for the x-vector system). This indicates that even without cross-genre speakers (true multi-genre data), data collected from multiple genres are still very useful. This is good news, as collecting partial multi-genre data is much cheaper than collecting true multi-genre data.

In summary, the multi-genre data is important for MG training and can improve the performance on multi-genre test. The MG training can be employed to either the front-end model or the back-end PLDA; though the best performance is obtained when both are MG trained. Partial MG training is not as effective as the true MG training, but it can provide reasonable gains with a low cost.

6. Conclusion

In this paper, we presented a comprehensive study for multi-genre speaker recognition. To make the study feasible, we firstly collected and published a large-scale multi-genre corpus, CN-Celeb2. Combined with the previously published CN-Celeb1, we have sufficient data to train and test speaker recognition systems in multi-genre conditions.

Based on the new dataset, we firstly evaluated the performance of the state-of-art speaker recognition systems in the multi-genre scenario, through which we identified the most difficult genres, and demonstrated that the major challenge of multi-genre speaker recognition lies in both genre complexity and genre mismatch. In the second experiment, we employed multi-genre training to tackle the multi-genre difficulties. Significant performance improvement was obtained, and importance of multi-genre speakers was identified.

Multi-genre speaker recognition is very important but is also very challenging. The research presented in this paper should be regarded as an initial and preliminary study in this direction. Lots of work remains to be done on this topic; to mention a few: (1) collection of more multi-genre data to support the research; (2) development of more powerful front-end models in order to produce genre-independent vectors; (3) discovery of more powerful back-end models to handle the changed statistics from one genre to another; (4) exploration of physiological models that can describe the intrinsic change of human pronunciation in different genres. We anticipate that the multi-genre challenge will be one of the prime obstacles that needs to be tackled before the speaker recognition techniques find ubiquitous applicability in practice.

CRediT authorship contribution statement

Lantian Li: Conceptualization, Methodology, Software. **Ruiqi Liu:** Methodology, Software. **Jiawen Kang:** Methodology, Software. **Yue Fan:** Resources, Data curation. **Hao Cui:** Resources, Data curation. **Yunqi Cai:** Writing – review & editing. **Ravichander Vipperla:** Writing – review & editing. **Thomas Fang Zheng:** Funding acquisition. **Dong Wang:** Conceptualization, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alvin, M.P., Martin, A., 2004. NIST speaker recognition evaluation chronicles. In: Proceedings of Odyssey: The Speaker and Language Recognition Workshop. Citeseer.
Bai, Z., Zhang, X.-L., Chen, J., 2020. Partial AUC optimization based deep speaker embeddings with class-center learning for text-independent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6819–6823.
Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M., 2010. Visual object tracking using adaptive correlation filters. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2544–2550.
Brümmer, N., Du Preez, J., 2006. Application-independent evaluation of speaker detection. Comput. Speech Lang. 20 (2–3), 230–275.
Bu, H., Du, J., Na, X., Wu, B., Zheng, H., 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). IEEE, pp. 1–5.
Cai, W., Chen, J., Li, M., 2018. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In: Proceedings of Odyssey: The Speaker and Language Recognition Workshop. pp. 74–81.
Campbell, J.P., 1997. Speaker recognition: A tutorial. Proc. IEEE 85 (9), 1437–1462.
Chen, N., Villalba, J., Dehak, N., 2019. Tied mixture of factor analyzers layer to combine frame level representations in neural speaker embeddings. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 2948–2952.

Chen, Z., Wang, S., Qian, Y., 2021. Self-supervised learning based domain adaptation for robust speaker verification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5834–5838.

Chung, J.S., Nagrani, A., Zisserman, A., 2018. VoxCeleb2: Deep speaker recognition. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 1086–1090.

Chung, J.S., Zisserman, A., 2016. Out of time: automated lip sync in the wild. In: Asian Conference on Computer Vision. Springer, pp. 251–263.

Cieri, C., Corson, L., Graff, D., Walker, K., 2007. Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH).

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S., 2020. RetinaFace: Single-stage dense face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5203–5212.

Ding, W., He, L., 2018. MTGAN: Speaker verification through multitasking triplet generative adversarial networks. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 3633–3637.

Du, J., Na, X., Liu, X., Bu, H., 2018. AISHELL-2: transforming mandarin ASR research into industrial scale. arXiv preprint arXiv:1808.10583.

Fan, Y., Kang, J., Li, L., Li, K., Chen, H., Cheng, S., Zhang, P., Zhou, Z., Cai, Y., Wang, D., 2020. CN-CELEB: a challenging Chinese speaker recognition dataset. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7604–7608.

Fisher, W.M., 1986. Ther DARPA speech recognition research database: specifications and status. In: Proc. DARPA Workshop on Speech Recognition, Feb. 1986. pp. 93–99.

Gao, Z., Song, Y., McLoughlin, I., Li, P., Jiang, Y., Dai, L.-R., 2019. Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 361–365.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1. IEEE Computer Society, pp. 517–520.

Gonzalez-Rodriguez, J., 2014. Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996–2014). Loquens.

Greenberg, C.S., Mason, L.P., Sadjadi, S.O., Reynolds, D.A., 2020. Two decades of speaker recognition evaluation at the national institute of standards and technology. Comput. Speech Lang. 60, 101032.

Hansen, J.H., Hasan, T., 2015. Speaker recognition by machines and humans: A tutorial review. IEEE Signal Process. Mag. 32 (6), 74–99.

Heigold, G., Moreno, I., Bengio, S., Shazeer, N., 2016. End-to-end text-dependent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5115–5119.

Ioffe, S., 2006. Probabilistic linear discriminant analysis. In: European Conference on Computer Vision (ECCV). Springer, pp. 531–542.

Kang, J., Liu, R., Li, L., Cai, Y., Wang, D., Zheng, T.F., 2020. Domain-invariant speaker vector projection by model-agnostic meta-learning. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH).

Kataria, S., Nidadavolu, P.S., Villalba, J., Dehak, N., 2020. Analysis of deep feature loss based enhancement for speaker verification. In: Proceedings of Odyssey: The Speaker and Language Recognition Workshop. pp. 459–466.

Kenny, P., 2005. Joint factor analysis of speaker and session variability: Theory and algorithms. In: CRIM, Montreal,(Report) CRIM-06/08-13, Vol. 14. pp. 28–29.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15 (4), 1435–1447.

Kenny, P., Mihoubi, M., Dumouchel, P., 2003. New MAP estimators for speaker recognition. In: Eighth European Conference on Speech Communication and Technology.

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5220–5224.

Larcher, A., Lee, K.A., Ma, B., Li, H., 2014. Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Commun. 60, 56–77.

Lee, K.A., Larcher, A., Wang, G., Kenny, P., Brümmer, N., Leeuwen, D.v., Aronowitz, H., Kockmann, M., Vaquero, C., Ma, B., et al., 2015. The RedDots data collection for speaker recognition. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH).

Li, L., Chen, Y., Shi, Y., Tang, Z., Wang, D., 2017. Deep speaker feature learning for text-independent speaker verification. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 1542–1546.

Li, R., Li, N., Tuo, D., Yu, M., Su, D., Yu, D., 2019. Boundary discriminative large margin cosine loss for text-independent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6321–6325.

Li, L., Wang, D., Kang, J., Wang, R., Wu, J., Gao, Z., Chen, X., 2020. A principle solution for enroll-test mismatch in speaker recognition. arXiv preprint arXiv:2012.12471.

Li, L., Wang, D., Xing, C., Zheng, T.F., 2016. Max-margin metric learning for speaker recognition. In: 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 1–4.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Matsui, T., Furui, S., 1993. Concatenated phoneme models for text-variable speaker recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2. IEEE, pp. 391–394.

McLaren, M., Ferrer, L., Castan, D., Lawson, A., 2016. The speakers in the wild (SITW) speaker recognition database. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 818–822.

Mo, J., Xu, L., 2020. Weighted cluster-range loss and criticality-enhancement loss for speaker recognition. Appl. Sci. 10 (24), 9004.

Morrison, G., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B., De Souza, S., Cummins, N., Chow, D., 2015. Forensic database of voice recordings of 500+ Australian English speakers.

Nagrani, A., Chung, J.S., Zisserman, A., 2017. VoxCeleb: A large-scale speaker identification dataset. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 2616–2620.

Okabe, K., Koshinaka, T., Shinoda, K., 2018. Attentive statistics pooling for deep speaker embedding. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 2252–2256.

Park, S.J., Sigouin, C., Kreiman, J., Keating, P.A., Guo, J., Yeung, G., Kuo, F.-Y., Alwan, A., 2016. Speaker identity and voice quality: Modeling human responses and automatic speaker recognition. In: Proceedings of The Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 1044–1048.

Park, S.J., Yeung, G., Kreiman, J., Keating, P.A., Alwan, A., 2017. Using voice quality features to improve short-utterance, text-independent speaker verification systems. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 1522–1526.

Parthasarathy, S., Rosenberg, A.E., 1996. General phrase speaker verification using sub-word background models and likelihood-ratio scoring. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, Vol. 4. IEEE, pp. 2403–2406.

Pham, M., Li, Z., Whitehill, J., 2020. Toward better speaker embeddings: Automated collection of speech samples from unknown distinct speakers. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7089–7093.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, no. EPFL-CONF-192584. IEEE Signal Processing Society.

Prince, S.J., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: 2007 IEEE 11th International Conference on Computer Vision. IEEE, pp. 1–8.

Qin, X., Bu, H., Li, M., 2020. Hi-mia: A far-field text-dependent speaker verification database and the baselines. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, pp. 7609–7613.

Rahman Chowdhury, F.R., Wang, Q., Moreno, I.L., Wan, L., 2018. Attention-based models for text-dependent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5359–5363.

Ramos, D., Gonzalez-Rodriguez, J., 2008. Cross-entropy analysis of the information in forensic speaker recognition. In: Proceedings of Odyssey: The Speaker and Language Recognition Workshop. International Speech Communication Association.

Reynolds, D.A., 1995. Automatic speaker recognition using Gaussian mixture speaker models. Linc. Lab. J. Citeseer.

Reynolds, D.A., 2002. An overview of automatic speaker recognition technology. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 4. IEEE, pp. IV–4072.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. 10 (1–3), 19–41.

Rosenberg, A.E., 1976. Automatic speaker verification: A review. Proc. IEEE 64 (4), 475–487.

Shon, S., Glass, J., 2020. Multimodal association for speaker verification. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 2247–2251.

Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S.S., Jameel, H., Richey, C., Goodman, F., 2008. Effects of vocal effort and speaking style on text-independent speaker verification. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH).

Shriberg, E., Kajarekar, S., Scheffer, N., 2009. Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions? In: Tenth Annual Conference of The International Speech Communication Association.

Snyder, D., Chen, G., Povey, D., 2015. MUSAN: A music, speech, and noise corpus. arXiv:1510.08484.
Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-vectors: Robust DNN embeddings for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5329–5333.
Stafylakis, T., Rohdin, J., Plchot, O., Mizera, P., Burget, L., 2019. Self-supervised speaker embeddings. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 2863–2867.
Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J., 2014. Deep neural networks for small footprint text-dependent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4052–4056.
Wang, D., 2020. Remarks on optimal scores for speaker recognition. arXiv preprint arXiv:2010.04862.
Wang, D., Li, L., Tang, Z., Zheng, T.F., 2017. Deep speaker verification: Do we need end to end? In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp. 177–181.
Wang, S., Rohdin, J., Burget, L., Plchot, O., Qian, Y., Yu, K., Cernocky, J., 2019a. On the usage of phonetic information for text-independent speaker embedding extraction. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 1148–1152.
Wang, J., Wang, K.-C., Law, M.T., Rudzicz, F., Brudno, M., 2019b. Centroid-based deep metric learning for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3652–3656.
weon Jung, J., Heo, H.-S., ho Kim, J., jin Shim, H., Yu, H.-J., 2019. RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 1268–1272.
Xie, W., Nagrani, A., Chung, J.S., Zisserman, A., 2019. Utterance-level aggregation for speaker recognition in the wild. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5791–5795.
Zeinali, H., Wang, S., Silnova, A., Matějka, P., Plchot, O., 2019. BUT system description to voxceleb speaker recognition challenge 2019. arXiv preprint arXiv:1910.12592.
Zhang, S.-X., Chen, Z., Zhao, Y., Li, J., Gong, Y., 2016. End-to-end attention based text-dependent speaker verification. In: Spoken Language Technology Workshop (SLT). IEEE, pp. 171–178.
Zhang, M., Kang, X., Wang, Y., Li, L., Tang, Z., Dai, H., Wang, D., 2018. Human and machine speaker recognition based on short trivial events. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5009–5013.
Zhang, A., Wang, Q., Zhu, Z., Paisley, J., Wang, C., 2019. Fully supervised speaker diarization. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6301–6305.
Zhou, J., Jiang, T., Li, Z., Li, L., Hong, Q., 2019. Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 2883–2887.
Zhu, Y., Ko, T., Snyder, D., Mak, B., Povey, D., 2018. Self-attentive speaker embeddings for text-independent speaker verification. In: Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH). pp. 3573–3577.
Zue, V.W., Seneff, S., 1996. Transcription and alignment of the TIMIT database. In: Recent Research Towards Advanced Man-Machine Interface Through Spoken Language. Elsevier, pp. 515–525.