

PLAYING A PART: SPEAKER VERIFICATION AT THE MOVIES

Andrew Brown^{1*}, Jaesung Huh^{1*}, Arsha Nagrani^{1,†*}, Joon Son Chung², Andrew Zisserman¹

¹Visual Geometry Group, University of Oxford, UK

²Naver Corporation, South Korea

{abrown, jaesung, arsha, joon, az}@robots.ox.ac.uk

ABSTRACT

The goal of this work is to investigate the performance of popular speaker recognition models on speech segments from movies, where often actors intentionally disguise their voice to play a character. We make the following three contributions: (i) We collect a novel, challenging speaker recognition dataset called *VoxMovies*, with speech for 856 identities from almost 4000 movie clips. *VoxMovies* contains utterances with varying emotion, accents and background noise, and therefore comprises an entirely different domain to the interview-style, emotionally calm utterances in current speaker recognition datasets such as VoxCeleb; (ii) We provide a number of domain adaptation evaluation sets, and benchmark the performance of state-of-the-art speaker recognition models on these evaluation pairs. We demonstrate that both speaker verification and identification performance drops steeply on this new data, showing the challenge in transferring models across domains; and finally (iii) We show that simple domain adaptation paradigms improve performance, but there is still large room for improvement.

Index Terms: speaker recognition, speaker verification, domain adaptation.

1. INTRODUCTION

All the world's a stage, and all the men and women merely players; They have their exits and their entrances, and one man in his time plays many parts.
- W. Shakespeare, As You Like It

After hearing the actor Steve Martin's smooth American accent on the Ellen show, is it possible to recognise that he is the voice behind 'Inspector Jacques Clouseau', the comical French character in the movie 'Pink Panther', without access to his visual appearance? Or recognise Anne Hathaway's voice in the movie 'Les Misérables', where she plays 'Fantine', singing through tears about her sadness and desperation? In this paper we investigate the challenging task of speaker recognition for actors from two different domains – the first being when they are speaking '*naturally*' in interviews, and second while *playing a part* in a movie, where they may be intentionally modifying their voice in order to play the role of a character or show emotion.

While recent years have shown great successes in speaker recognition [11, 19, 42, 44], these successes have been reliant on the collection of large, labelled datasets such as VoxCeleb [12, 35, 36] and others [16, 30]. The VoxCeleb datasets, while valuable, have been collected *entirely from interviews* of celebrities in YouTube videos and are limited in terms of linguistic content (celebrities mostly speak about their professions [33]), emotion, and background noise. In contrast, movies contain speech covering emotions such as anger, sadness, assertiveness, and fright, and varied background conditions – imagine the shouting in a violent scene from an action movie, or a romantic scene of reconciliation in a romcom. As we show in this paper, models trained on VoxCeleb, when applied to a novel domain such as speech in movies, suffer from significant degradation in

performance. In order to accurately measure this, there is a compelling need for real-world datasets and evaluation sets across these domains. Collecting and annotating datasets for every new domain encountered in the real-world, however can be an extremely expensive and time-consuming process. We introduce a scalable method to automatically generate data in a new domain (movies), and investigate the performance of state-of-the-art speaker recognition models on this data, where actors are intentionally disguising their voice. Being able to detect human identity under such conditions of spoofing is valuable for security and authentication [6, 9, 14], and as shown by psychology studies [23, 40], is a challenging task even for humans.

In order to encourage research in domain adaptation for speaker recognition, we make the following three contributions: (i) We collect a novel speaker recognition dataset called *VoxMovies*, from 3,792 popular movie clips uploaded to YouTube. Our dataset consists of almost 9,000 utterances from 856 identities that appear in the VoxCeleb datasets, and contains challenging emotional, linguistic and channel variation (Fig. 1); (ii) We provide a number of domain adaptation evaluation sets, and benchmark the performance of state-of-the-art speaker recognition models on these evaluation pairs. We demonstrate that performance drops steeply on this new data for both speaker verification *and* identification, showing the challenge in transferring models across domains (from interviews to movies). We also investigate performance on positive pairs sampled across different movies, and reveal further performance drops; and finally (iii) We demonstrate that domain adaptation approaches added on top of already trained models improve performance, but there is still a severe degradation.

VoxMovies has been used to create a challenging test set for the VoxCeleb Speaker Recognition Challenge [34] (VoxSRC2020)¹. Data: <https://www.robots.ox.ac.uk/~vgg/data/voxmovies/>.

2. RELATED WORK

Due to the reliance of machine learning on large, labelled datasets, domain adaptation has become popular in fields such as computer vision [24, 43], text classification [27, 46], speech enhancement [26, 32] and speaker verification [17, 18, 20]. Recent interest in domain adaptation for speaker recognition has largely focused on boosting the performance on datasets such as NIST-SRE16 or Speakers in the Wild (SITW) [31], which does not have a large training set. In this case, most methods train on VoxCeleb [12, 36] and evaluate on these evaluation sets.

Recent methods focus on adversarial training techniques, with [45] introducing domain adversarial training that exploits domain prediction, and [28] proposing a channel adversarial training method by adding a channel discriminator to mitigate the domain mismatch problem by using the video labels in VoxCeleb, with both works [28, 45] using gradient reversal layers. [41] proposes an end-to-end domain adversarial training method adopting the architecture of Wasserstein Generative Adversarial Network (GAN) [2] with adversarial domain loss and cross-entropy loss. It significantly improves the

* These authors contributed equally to this work.

† Now at Google Research.

¹<http://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html>

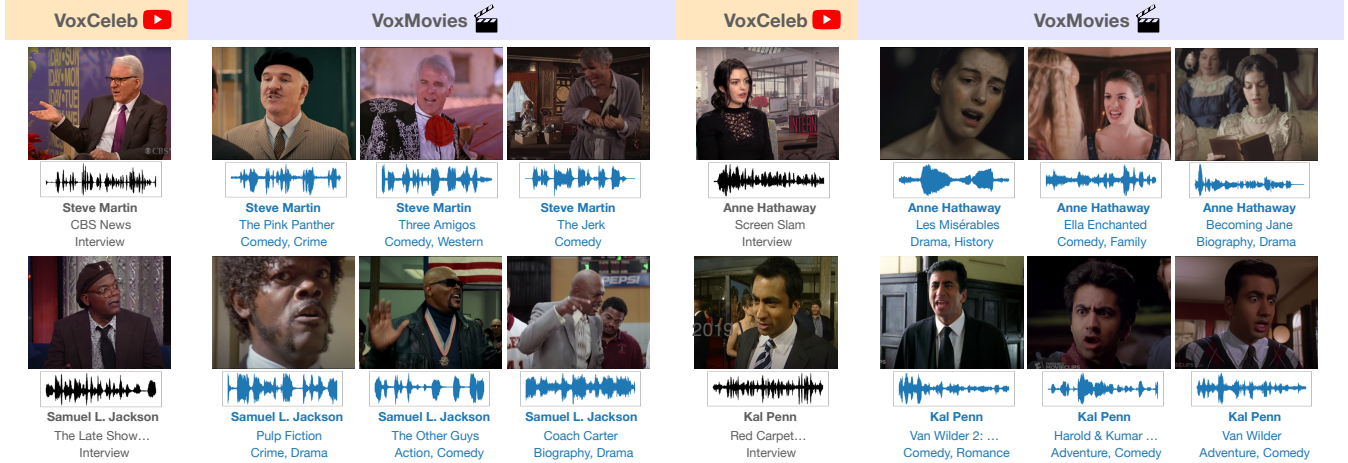


Fig. 1: Domain Gap between VoxCeleb and VoxMovies: Unlike VoxCeleb (left in each panel), which consists of utterances from interviews, VoxMovies is sourced from movies of different genres (right in each panel). While VoxCeleb features utterances largely with calm, unvaried emotions, VoxMovies has challenging emotion and background noise. We show only a single frame for each utterance, below which is the name and genre of the film/video. VoxMovies contains 24 utterances from 5 movies for Anne Hathaway, including singing in the musical, ‘Les Misérables’, reading in an English accent in ‘Becoming Jane’, and arguing heatedly in an American accent in ‘Ella Enchanted’. Samuel L. Jackson has 59 utterances from 14 movies, including his famous, assertive speech in ‘Pulp Fiction’ and bragging as an arrogant policeman in ‘The Other Guys’.

performance on language adaptation. [4] also introduces a domain adaptation technique to new language or recording conditions by using a Domain Adversarial Neural Speaker Embeddings (DANSE) model which contains a 1-dimensional self-attentive residual block. [5] explores various GANs to make the domain discriminator unable to distinguish whether embeddings are from the source or target domain. [25] exploits Model-Agnostic Meta-Learning (MAML), projecting speaker representations to a generalized embedding space and achieves better results on CNCeleb dataset [16]. Other research methods train channel-invariant, noise-robust speaker recognition models which can help improve the overall performance of the model in diverse domains. [47] shows a multi-task learning method that trains both a speaker recognition network and a discriminator that distinguishes types of noise in speaker representations. [10] also tackles a similar problem by using an environment network and a KL-divergence based confusion loss to learn speaker-discriminative, environment-invariant representations.

Unlike existing works, we provide novel domain data for the *same identities as in VoxCeleb*, allowing us to investigate both cross-domain speaker verification, where pairs have one segment from either domain, and cross-domain identification, where speaker identification models are tested on a new, previously unseen domain. Both have not been previously possible. We focus on the domain adaptation from interviews to movies for male and female actors (explored for faces in [37]). We propose and benchmark on such evaluations conditions, and additionally on several within-domain verification tasks.

3. CROSS-DOMAIN DATA

Our goal in this work is to investigate the effects of cross-domain speaker verification in movies. The domain change we focus on here is from YouTube interviews (domain *D-I*), to speech in different genres of movies (domain *D-M*). The data for the two different domains are sourced as follows:

D-I: Interviews from VoxCeleb [12, 36] These datasets are sourced solely from interviews uploaded to YouTube. These are mainly in studio, outdoor or red-carpet locations. In turn, and due to the often *professional* context, voices are mainly calm and rarely show any strong emotion. These utterances are degraded with real world noise that would be expected from these environments, such as background chatter or laughter.

D-M: Movies from CMD [3] (**VoxMovies**) For the second domain, we curate a dataset of speech from movies, called VoxMovies, which

consists of 8,905 utterances for 856 different identities, sourced from 3,792 video clips from 1,452 movies. These movies cover a range of genres (see Figure 2). The utterances in VoxMovies are sourced from the Condensed Movies dataset (CMD) [3], which covers the *key scenes* from movies. The distinctive change of domain can be seen in the following characteristics of VoxMovies:

(1) Emotion: In line with different movie genres, the utterances cover emotions such as anger, sadness, assertiveness, and fright. Furthermore, the videos in CMD represent scenes that are integral to the story-line and the different character developments, such as a fight between two main characters, or when they make up later in the film. Hence the utterances in the dataset often capture the most emotional parts of each movie.

(2) Background noise: Each key scene in the CMD dataset covers many different settings, from a loud basketball stadium in *Coach Carter*, to an 18th century gathering in *Becoming Jane* (see Figure 1). This represents a far more varied set of degradation for each of the utterances. Also, there is often background music.

Importantly, in VoxMovies this variety of emotion and background noise is seen both within and across different identities. Firstly this is because on average each identity has utterances from 2.7 different movies (see Table 1), and these movies are likely to be of different genres. Secondly, the videos in this dataset show the important character arcs of these identities within each movie, where they show different emotions at different points in the story-line. Examples and further details can be found in Figure 1.

Note that the utterances in VoxMovies are all from identities that are represented in VoxCeleb1 and VoxCeleb2. We create an evaluation set, VoxMovies-(Test), featuring identities from VoxCeleb1. We also provide a small amount of VoxMovies data for training domain adaptation methods, VoxMovies-(Train), using a subset of the identities in the VoxCeleb2 dev set. There is no identity overlap between these partitions. Statistics are shown in Table 1.

4. DATASET COLLECTION PIPELINE

Our dataset collection pipeline is similar to the one used to collect the VoxCeleb datasets, albeit applied to YouTube clips of movie scenes from the Condensed Movies Dataset (CMD) [3], described below.

Condensed Movies Dataset. CMD [3] consists of key scenes from over 3K movies, totalling 1,270 hours. Provided alongside the dataset are cast

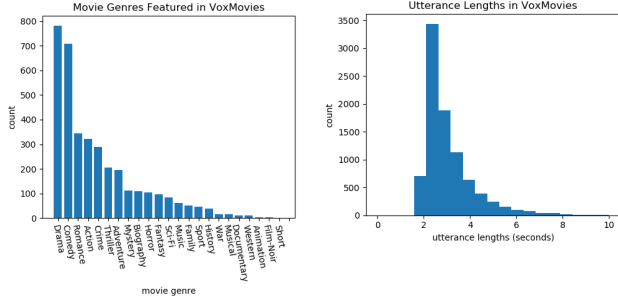


Fig. 2: Statistics of the VoxMovies dataset. (a) The different movie genres that the utterances in VoxMovies are sourced from, and (b) the distribution of utterance lengths. The minimum length is 2 seconds by design choice.

Partition	ID Source	IDs	# Utter.	Clips / ID	Movies / ID
VoxMovies-(Test)	VoxCeleb1	485	4,943	5.1	2.7
VoxMovies-(Train)	VoxCeleb2	371	3,962	5.0	2.5

Table 1: Dataset Statistics for VoxMovies. The utterances are sourced from the Condensed Movies Dataset [3], which contains clips of key scenes from movies. Identities (IDs) overlap with the VoxCeleb1 (test) and VoxCeleb2 (dev) datasets.

lists for each of the featured movies, face-tracks for each of the clips, and face embeddings for discriminating identity for each of these face-tracks. The cast lists give the names of people who are likely to appear in the clip. Our three stage method for collecting VoxMovies is as follows:

Stage 1. Sourcing candidate names. We compute the intersection between the VoxCeleb1 and VoxCeleb2 identities and the CMD cast lists.

Stage 2. Face verification. In this stage, face-tracks from the CMD dataset are classified as to whether they depict any of the candidate names from the previous stage using a three step process. (1) Example face images for each of the candidate names for the train and test set are sourced from the VGGFace2 [8] and VGGFace1 [38] datasets (these face datasets contain the same identities as the VoxCeleb2 and VoxCeleb1 datasets respectively). (2) 256D Embeddings are taken from the final hidden layer of a SE-Net50 architecture CNN for each of the face images, trained for discriminating face-identity using the VGGFace2 dataset [8]. For each identity, these embeddings are average pooled across all example face-images and L2 normalised, leaving one embedding per identity. (3) Verification is performed by computing the cosine similarity between the identity embeddings and the face-track embeddings from CMD. Any face-track with a similarity to an identity embedding above a high threshold is assigned that name.

Stage 3. Active speaker verification. Here we identify which face-tracks from the previous stage are speaking. This is performed using an audio-visual synchronisation network [13], which predicts the correlation between the audio track and the motion of the mouth, and outputs a synchronisation confidence score for each 5 frame window. A window with confidence above a high threshold is classified as speaking. Face-tracks with a minimum of 2 seconds of consecutive speech are kept, and the audio track is clipped to the speaking period. See Figure 2 for a histogram of segment lengths.

Discussion. A manual check of VoxMovies-(Test) reveals that the automated process has a precision of > 99%. The < 1% false positive face verifications, or false positive active speaker verifications are manually removed. The high thresholds for stages 2 and 3 are chosen as 0.55 and 0.13, respectively. These values (both cosine similarity) achieved high precision on a manually labelled validation set. As shown in Table 1, VoxMovies-(Test) has slightly more utterances than VoxMovies-(Train). This is due to the fact that a larger proportion of the identities in VoxCeleb1 are

Eval. set	Positives Source	Negatives Source	# Utter.	# Positive Pairs	# Negative Pairs
E-1	<i>D-M (same)</i>	<i>D-M</i>	20,572	10,286	10,286
E-2	<i>D-I, D-M</i>	<i>D-I, D-M</i>	46,578	23,289	23,289
E-3	<i>D-I, D-M</i>	<i>D-I</i>	46,804	23,402	23,402
E-4	<i>D-I, D-M</i>	<i>D-M</i>	46,866	23,433	23,433
E-5	<i>D-M (diff)</i>	<i>D-M</i>	41,090	20,545	20,545

Table 2: Statistics for the different evaluation sets: Our evaluation sets are sourced from two different domains, interview material (*D-I*) and movie material (*D-M*). Key: **Utter.**: Utterances., *D-M (same)*: Segments in a pair are sourced from the same movie, *D-M (diff)*: Segments sourced from different movies.

well-known actors than in VoxCeleb2, and hence they appeared in more of the movies in CMD. Impressively, on average, each identity has utterances from 2.7 different movies, with Robert DeNiro appearing in 25 movies.

5. EXPERIMENTS

5.1. Evaluation Tasks

Our goal is to determine the performance of state-of-the-art models trained on the VoxCeleb2 dev set for speaker recognition performance on data from the movie domain (*D-M*).

Verification: We use the VoxCeleb1 and VoxMovies-(Test) datasets for evaluation, as these sets have no overlap with the identities in the VoxCeleb2 dev set (which is used to train all baselines). Given the two domains (*D-I* and *D-M*), we note that there are three different ways that pairs can be sourced for evaluation: pairs can be sourced entirely from *D-I*, entirely from *D-M*, or from both, where one utterance is from *D-I* and one is from *D-M*. For positive pairs sourced from *D-M*, we can add the further constraint that both utterances must come from the same movie (*D-M (same)*) or from different movies (*D-M (diff)*). We use these options to create 5 challenging evaluation sets E1-E5 (Table 2), of increasing difficulty. More details are provided in Section 6.

Identification: We also demonstrate domain mismatch from VoxMovies and VoxCeleb, with speaker identification. Here we add a single linear layer on baseline models trained on VoxCeleb2 dev set for verification. The task is to fine-tune this layer with the cross-entropy loss using utterances from 485 speakers in the VoxCeleb1 dataset (using only the training data from the identification split, see Table 5 in [36]). We then test performance using the VoxCeleb1 identification test data (in-domain test set) and the utterances from VoxMovies (out of domain test set). **Evaluation Metrics:** For verification, we report equal error rate (EER-%) and minimum detection cost (minDCF) with $C_{miss} = 1$, $C_{fa} = 1$ and $P_{target} = 0.05$. The formula of minimum detection cost function is the same as the one used in NIST SRE [1] and the VoxSRC2020 evaluations. For identification, we report top 1 and top 5 % accuracy.

5.2. Baseline models

We compare four baseline models to investigate the performance on our dataset. All models are trained on the VoxCeleb2 dev set.

1. I-vector [15]: Following the Kaldi [39] VoxCeleb recipe v1, we extract 400D features, followed by Probabilistic LDA (PLDA) scoring.

2. X-vector [42]: Following the Kaldi [39] VoxCeleb recipe v2, we train a x-vector model with a PLDA back-end to extract 512D features.

3. Thin ResNet-34 [7]: Consists of 34 residual blocks (one-fourth of the channel dimensions of original ResNet-34 [21]), with self-attentive pooling. [11] trains this network with an angular variant of the prototypical loss. We use the pre-trained model which is publicly available [11].

4. Thick ResNet-34 [22]: This has double the number of channels in Thin ResNet-34 and uses attentive statistical pooling to model higher order

Eval set	I-vec. [15]		X-vec. [42]		Thin-R34 [7]		Thick-R34 [22]	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
VoxCbl†	5.53	0.336	3.30	0.220	2.05	0.166	1.05	0.084
VoxCbl-H†	9.13	0.467	5.82	0.352	4.37	0.283	2.39	0.154
VoxCbl-E†	5.55	0.338	3.35	0.221	2.27	0.164	1.22	0.086
VoxSRC20*	11.07	0.566	8.22	0.450	6.35	0.374	3.79	0.213
E-1	16.6	0.727	12.92	0.665	9.72	0.562	6.09	0.365
E-2	17.91	0.822	14.75	0.712	10.58	0.610	7.40	0.423
E-3	18.91	0.917	13.58	0.806	10.58	0.666	7.50	0.484
E-4	19.83	0.872	21.56	0.845	12.52	0.737	9.23	0.579
E-5	21.5	0.913	17.97	0.861	14.11	0.760	10.47	0.574

Table 3: Baseline results on various VoxMovies test evaluation pairs. We show the performance of 4 popular state-of-the-art speaker recognition models. † Cleaned versions of these evaluation pairs from the VoxCeleb1 dataset. *VoxSRC2020 validation set¹.

Evaluation set	Top1 accuracy	Top5 accuracy
VoxCeleb1-test†	89.47%	97.38%
VoxMovies-(Test)	52.23%	73.31%

Table 4: Identification results for 485 identities on the VoxMovies-(Test) set (out of domain) and on the VoxCeleb1 test set (same domain). Note how performance drops steeply on the out of domain test set. † VoxCeleb1 test set for identification.

Eval set	Baseline	FT	S-norm	FT + S-norm
E-1	6.09	5.76	5.89	5.66
E-2	7.40	7.10	7.18	7.03
E-3	7.50	8.38	8.16	8.48
E-4	9.23	7.37	8.03	7.19
E-5	10.47	9.55	10.15	9.35
E-3a	1.15	1.53	1.29	1.58
E-3b	0.87	0.97	0.68	0.98
E-3c	7.72	9.90	9.64	10.23

Table 5: Domain transfer results for Thick ResNet-34. EER(%) is reported. **FT:** Fine-tuning on the VoxMovies-(Train) set. **S-norm:** Score-normalisation.

statistics such as standard deviation. The model is trained with both angular prototypical loss and vanilla softmax loss to improve performance. We use the pre-trained model which is publicly available [11]. This model currently represents the state-of-the-art on the VoxCeleb1 test sets. 512D features are extracted for both thick and thin ResNet-34 architectures.

5.3. Domain Transfer

In this section, we implement two common domain transfer methods using the *small* amount of data provided in the VoxMovies-(Train) set.

Fine-tuning on a small amount of target domain data. We fine-tune the pretrained Thick-ResNet34 with data from the VoxMovies-(Train) set and the VoxCeleb2 dev set (overlapping speakers only). To decrease the domain gap between the datasets, we always pick one utterance from VoxMovies and another from VoxCeleb2 to form positive pairs in each mini-batch. This forces the model to decrease the distance between embeddings from the same speaker’s utterances, hence reducing the domain gap during training. The network is trained with angular prototypical loss [11], for 500 epochs using Adam with learning rate of 1e-5. Only the last fully-connected layer is fine-tuned while weights of other layers are fixed.

Score Normalisation. [29] introduces various score normalisation techniques for test conditions with diverse domains. A *cohort* is used to estimate the amount of shift and scale for normalisation, allowing robust threshold setting. We experiment with the Z-norm, T-norm and S-norm, and find the

best performance to be using the S-norm (Sec.2 in [29]). We use the VoxMovies-(Train) set as the cohort (speakers in the VoxMovies-(Train) and (Test) sets are disjoint, which fits the assumption of cohort selection).

6. RESULTS

Verification using Baseline Models. The results for the baseline models on the different verification tasks are given in Table 3. The change of domain in the VoxMovies evaluation sets offers a significant challenge – the Thick ResNet-34 which achieves an impressive 1.05 EER on the VoxCeleb1 test set can only achieve 6.09 EER on the least challenging set, E-1. When comparing eval sets that share a positives or negatives source (see Table 2), several conclusions are made: (1) Verifying the same speaker with cross domain utterances (*D-I*, *D-M* - E-4) is harder than with utterances from the same movie (*D-M* (*same*) - E-1). Interestingly, this shows that the change in an actor’s voice from a calm interview setting in *D-I* to strong emotions, accents and different background noise in *D-M*, is more challenging for speaker verification than the differing emotions in an actor’s voice within the same movie at different points in the story-line. Furthermore, the same speaker from different movies (*D-M* (*diff*) - E-5) is harder still, showing that an actor’s voice changes most between different movies. (2) Negative pair verification is hardest when the negatives are both taken from the unseen domain, *D-M* (E-4). This is more difficult than when they are taken from *D-I* (E-3) or *D-I*, *D-M* (E-2), which are of roughly equal difficulty to the Thick ResNet-34. This is largely because the baseline models were trained on *D-I*, so any source of negatives or positives exclusively from that domain will be less challenging.

Identification. We use the Thin ResNet-34 model for identification. Table 4 shows identification accuracy on both domains. As expected, the identification accuracy drops significantly from VoxCeleb1-test to VoxMovies-(Test) by 37.24% (top1% acc.). Anne Hathaway is one of the hardest to identify (top1 acc. 29.17% - down from 90% in VoxCeleb1-test), whereas Samuel L. Jackson is easier (acc. 100% drops to 72.88% in movies). This may be due to Hathaway’s multiple accents in her movies (Fig. 1).

Domain Transfer. We report the results in Table 5. Fine-tuning with VoxMovies-(Train) reduces the EER on most of the evaluation sets. The largest improvement is seen in evaluation sets that showed the worst performance in baseline experiments, namely by 2.04% in E-4 and 1.12% in E-5. Both E-4 and E-5 source negatives from *D-M*, showing that fine-tuning and score normalisation work well when transferring existing models from *D-I* to *D-M*. E-3 on the other hand, which sources negatives from *D-I* shows performance degradation. We conclude that fine-tuning on *D-M* degrades the performance on negative pairs only from *D-I*. To verify this conclusion, we introduce three new evaluation sets, E-3a (positives: *D-I*, negatives: *D-I*), E-3b (positives: *D-I*, negatives: *D-I*, *D-M*), E-3c (positives: *D-M*, negatives: *D-I*) in Table 5. The fine-tuned model becomes worse at verification in *D-I* (as shown by the degradation of E-3a). Performance on negatives from *D-I* contribute most to this (as shown by the degradation of E-3c, relative to E-3b).

7. CONCLUSION

In this paper, we provide a novel speaker recognition dataset from movies called VoxMovies which contains almost 9,000 utterances with diverse emotion, accents and background conditions. We demonstrate that state-of-the-art models trained on interview data from VoxCeleb degrade significantly on cross-domain evaluation sets from VoxMovies and while simple domain adaptation techniques boost performance, there is still large room for improvement. We hence encourage the research community to develop new methods and systems for this challenging new domain.

Acknowledgements. We are grateful to Max Bain for his help with the CMD dataset. AB is funded by an EPSRC DTA Studentship, JH by a Global Korea Scholarship, and AN by a Google PhD Fellowship. This work is supported by the EPSRC Programme Grant Seebibyte EP/M013774/1.

References

- [1] *NIST 2018 Speaker Recognition Evaluation Plan*, 2018 (accessed 31 July 2020), https://www.nist.gov/system/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf, See Section 3.1.
- [2] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan,” *arXiv preprint arXiv:1701.07875*, 2017.
- [3] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, “Condensed movies: Story based retrieval with contextual embeddings,” in *Asain Conference on Computer Vision (ACCV)*, 2020.
- [4] G. Bhattacharya, J. Alam, and P. Kenny, “Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training,” in *Proc. ICASSP*, 2019.
- [5] G. Bhattacharya, J. Monteiro, J. Alam, and P. Kenny, “Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification,” in *Proc. ICASSP*, 2019.
- [6] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, “Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion,” in *INTERSPEECH*, 2017.
- [7] W. Cai, J. Chen, and M. Li, “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system,” *Speaker Odyssey*, 2018.
- [8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.
- [9] Z. Chen, Z. Xie, W. Zhang, and X. Xu, “Resnet and model fusion for automatic spoofing detection,” in *INTERSPEECH*, 2017.
- [10] J. S. Chung, J. Huh, and S. Mun, “Delving into voxceleb: environment invariant speaker recognition,” *Speaker Odyssey Workshop*, 2020.
- [11] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, “In defence of metric learning for speaker recognition,” in *INTERSPEECH*, 2020.
- [12] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in *INTERSPEECH*, 2018.
- [13] S.-W. Chung, J. S. Chung, and H.-G. Kang, “Perfect match: Improved cross-modal embeddings for audio-visual synchronisation,” in *Proc. ICASSP*, 2019.
- [14] R. K. Das, J. Yang, and H. Li, “Long range acoustic features for spoofed speech detection,” in *INTERSPEECH*, 2019.
- [15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.
- [16] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, “Cn-celeb: a challenging chinese speaker recognition dataset,” in *Proc. ICASSP*, 2020.
- [17] D. Garcia-Romero and A. McCree, “Supervised domain adaptation for i-vector based speaker recognition,” in *Proc. ICASSP*, 2014.
- [18] D. Garcia-Romero, A. McCree, S. Shum, and C. Vaquero, “Unsupervised domain adaptation for i-vector speaker recognition.”
- [19] D. Garcia-Romero, A. McCree, D. Snyder, and G. Sell, “Jhu-hltcoe system for the voxsrc speaker recognition challenge,” in *Proc. ICASSP*, 2020.
- [20] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, “Improving speaker recognition performance in the domain adaptation challenge using deep neural networks,” in *SLT*, 2014.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016.
- [22] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, “Clova baseline system for the voxceleb speaker recognition challenge 2020,” *arXiv preprint arXiv:2009.14153*, 2020.
- [23] A. Hirson and M. Duckworth, “Glottal fry and voice disguise: a case study in forensic phonetics,” *Journal of biomedical engineering*, 1993.
- [24] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, “Cycada: Cycle-consistent adversarial domain adaptation,” in *Proc. ICML*, 2018.
- [25] J. Kang, R. Liu, L. Li, Y. Cai, D. Wang, and T. F. Zheng, “Domain-invariant speaker vector projection by model-agnostic meta-learning,” *arXiv preprint arXiv:2005.11900*, 2020.
- [26] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, “Noise adaptive speech enhancement using domain adversarial training,” *arXiv preprint arXiv:1807.07501*, 2018.
- [27] P. Liu, X. Qiu, and X. Huang, “Adversarial multi-task learning for text classification,” *arXiv preprint arXiv:1704.05742*, 2017.
- [28] C. Luu, P. Bell, and S. Renals, “Channel adversarial training for speaker verification and diarization,” in *Proc. ICASSP*, 2020.
- [29] P. Matejka, O. Novotný, O. Plchot, and L. Burget, “Analysis of score normalization in multilingual speaker recognition,” 2017.
- [30] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, “The speakers in the wild (sitw) speaker recognition database,” 2016.
- [31] M. McLaren, A. Lawson, L. Ferrer, D. Castan, and M. Graciarena, “The speakers in the wild speaker recognition challenge plan,” *Interspeech 2016 Special Session, San Francisco*, 2015.
- [32] Z. Meng, J. Li, Y. Gong *et al.*, “Adversarial feature-mapping for speech enhancement,” in *INTERSPEECH*, 2018.
- [33] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, “Disentangled speech embeddings using cross-modal self-supervision,” in *Proc. ICASSP*, 2020.
- [34] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, “VoxSRC 2020: The second voxceleb speaker recognition challenge,” 2020.
- [35] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech & Language*, vol. 60, p. 101027, 2020.
- [36] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: a large-scale speaker identification dataset,” in *INTERSPEECH*, 2017.
- [37] A. Nagrani and A. Zisserman, “From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script,” *BMVC*, 2018.
- [38] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proc. BMVC.*, 2015.
- [39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *IEEE workshop on automatic speech recognition and understanding*, 2011.
- [40] A. R. Reich and J. E. Duke, “Effects of selected vocal disguises upon speaker identification by listening,” *The Journal of the ASA*, 1979.
- [41] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, “Speaker verification using end-to-end adversarial language adaptation,” in *Proc. ICASSP*, 2019.
- [42] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *Proc. ICASSP*, 2018.
- [43] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proc. CVPR*, 2017.
- [44] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in *Proc. ICASSP*, 2018.
- [45] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, “Unsupervised domain adaptation via domain adversarial training for speaker recognition,” in *Proc. ICASSP*, 2018.
- [46] Y. Zhang, R. Barzilay, and T. Jaakkola, “Aspect-augmented adversarial networks for domain adaptation,” *Transactions of the Association for Computational Linguistics*, 2017.
- [47] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, “Training multi-task adversarial network for extracting noise-robust speaker embedding,” in *Proc. ICASSP*, 2019.