LEVERAGING SPEAKER ATTRIBUTE INFORMATION USING MULTI TASK LEARNING FOR SPEAKER VERIFICATION AND DIARIZATION

Chau Luu, Peter Bell, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, UK
{chau.luu, peter.bell, s.renals}@ed.ac.uk

ABSTRACT

Deep speaker embeddings have become the leading method for encoding speaker identity in speaker recognition tasks. The embedding space should ideally capture the variations between all possible speakers, encoding the multiple aspects that make up speaker identity. In this work, utilizing speaker age as an auxiliary variable in US Supreme Court recordings and speaker nationality with VoxCeleb, we show that by leveraging additional speaker attribute information in a multi task learning setting, deep speaker embedding performance can be increased for verification and diarization tasks, achieving a relative improvement of 17.8% in DER and 8.9% in EER for Supreme Court audio compared to omitting the auxiliary task. Experimental code has been made publicly available.

Index Terms— Speaker verification, speaker diarization, multitask learning, deep neural network, deep learning

1. INTRODUCTION

Obtaining speaker discriminative features is an important step for many speaker recognition tasks, such as speaker verification and speaker diarization. In recent years, extracting speaker embeddings from the intermediate layer of a neural network has become the state-of-the-art method for both tasks [1, 2], outperforming the historically successful i-vector method [3].

The ideal speaker embedding space should discriminatively capture the variations in speakers, encoding the multiple contributing factors that make up speaker identity. For example, attributes such as gender and accent intuitively should contribute to where an utterance of a given speaker should be placed in the embedding space. Indeed, previous work has shown that both deep speaker embeddings and i-vectors encode a wide variety of information and meta-information about speakers and utterances, such as speaker emotion [4, 5], accent and language [6] or speaker gender, channel and transcription information [7].

For example, In the field of forensic phonetics and acoustics [8, 9], speaker classification is performed by human experts for suspects in criminal cases. These experts typically profile speakers according to a set of characteristics that allow their voices to be qualitatively compared, such as gender, age, dialect, accent, medical conditions and sociolect (education level, which affects things like lexicon, syntax and stylistics) [8]. This raises the question as to whether these attributes that contribute to speaker identity can be leveraged explicitly to improve the descriptiveness of a deep embedding space.

This work proposes using multi-task learning (MTL) [10] to leverage speaker attribute information when training deep speaker embeddings. Specifically, by web scraping nationality labels for VoxCeleb [11, 12] and age information of speakers in US Supreme Court recordings, we show that training on nationality and age classification tasks in conjunction with the standard speaker classification can improve deep speaker embedding performance on both verification and diarization tasks. Experimental and web scraping code has been made available[1].

The concept of MTL revolves around the idea that machine learning models that may be used to solve different problems using the same data can benefit from sharing a common representation. The work of [13] found improvement in increasing the robustness of a hybrid RNN/HMM ASR system by performing speech enhancement as an additional task to classification, with both tasks relying on the same hidden representation, while the work of [14] found that training to simultaneously predict both context-dependent and context-independent targets regularly improved performance in an ASR setting.

Previous works implementing MTL for speaker recognition specifically have explored using the word spoken in the utterance as an additional task in training deep speaker embeddings to increase verification performance [15]. Likewise, [16] found that learning to additionally classify the phonetic information also improved performance. Instead of utilizing transcription information, this work proposes utilizing information that is explicitly connected to the speaker identity to establish a more descriptive speaker embedding space.

There may be several explanations for why MTL may improve performance for speaker recognition. Ideally, the embedding space provided by the extractor should be able to de-

Supported by an EPSRC iCASE studentship collaboration with the BBC.

[1] https://github.com/cvqluu/MTL-Speaker-Embeddings

scribe multiple properties of the data. By additionally and explicitly training towards an embedding space that can describe these speaker attributes, it follows that this may lead to a more descriptive and discriminative embedding space.

2. MULTITASK LEARNING

The concept of multi-task learning (MTL) is that machine learning models may benefit from sharing the same representations when solving different tasks on the same data. In the context of deep learning, this typically means initial layers of a neural network are shared between tasks, after which task specific layers act upon these shared representation of the input data.

This concept can be applied to a speaker embedding extractor, such as the x-vector network [1], which takes in as input some acoustic features and performs speaker classification. The speaker embeddings are extracted from an intermediate layer in this network. If the layers up until the extracted embedding are considered to be the embedding extractor, one can consider the remaining layers to be a task specific 'head'. The standard x-vector network has a single task specific head, a feed forward network performing speaker classification. Multi task learning can be applied by adding additional, separate task-specific heads with their own loss functions which also act on the embedding.

Starting from the assumption that a speaker classification loss will always be applied, for M additional tasks, their losses $\mathcal{L}_m$ can be combined in the following manner,

$$\mathcal{L}_{\text{multi-task}} = \mathcal{L}_{\text{speaker}} + \sum_{i=1}^{M} \lambda_m \mathcal{L}_{\text{m}} \tag{1}$$

where each additional task loss $\mathcal{L}_{\text{m}}$ is weighted by some chosen loss weighting λ_m, relative to the speaker loss $\mathcal{L}_{\text{speaker}}$.

3. SPEAKER ATTRIBUTES

In order to explore leveraging speaker attributes for MTL, this requires the availability of data which has this meta-information. The following sections will detail the datasets and additional attributes to be explored: The Supreme Court of the United States corpus, for which we obtained information on speaker age, and VoxCeleb, which has obtainable information about speaker nationality.

3.1. Age: SCOTUS Corpus

The Supreme Court of the United States (SCOTUS) is the highest level court in the United States. Audio recordings and transcriptions of SCOTUS oral arguments have been made available via the Oyez project[2]. This corpus is rare in that

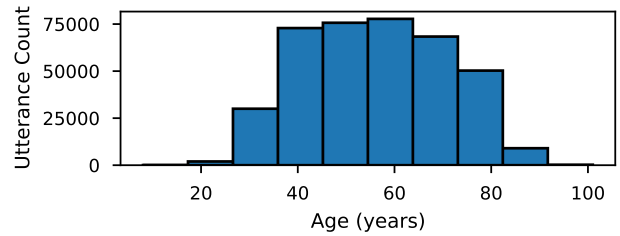

Fig. 1: The age distribution of utterances in the SCOTUS corpus, split into 10 bins.

it features certain speakers across many years. These speakers are the justices (judges) of the Supreme Court, who serve indefinitely on the court until their retirement. The average length of a supreme court justice's tenure at time of writing is approximately 17 years. The phenomenon of how a human voice can change due to the effects of age is fairly well known [17], with degradation in the vocal folds sometimes leading to changes in the fundamental frequency of the voice, and other qualities in vocal delivery, such as jitter, shimmer and volume. Severe cases can lead to diagnoses of vocal disorders [18]. These effects on a speaker's voice makes age a fascinating aspect to explore of what contributes to speaker identity.

For this work, 1022 digital recordings with 913 unique speakers across 1032 hours of audio were considered. The vast majority of speakers however are not the justices and thus do not have easily obtainable dates of birth. However, many of these speakers are law professionals and thus have publicly available dates for when they were registered to start practising law or graduated their final law degrees. The date of birth can then be inferred by subtracting 25, the usual time taken to complete these qualifications. The interested reader can refer to the code repository for more information. One can see the distribution of ages of utterances in Figure 1, with an average age of around 56.

3.2. Nationality: VoxCeleb

VoxCeleb 1 and 2 [11, 12] are speaker recognition datasets featuring celebrity speakers, meaning certain speaker attributes can be found in the public domain. Speaker nationality labels can be web scraped from sources such as Wikipedia and are a proxy for speaker accent, an attribute which is clearly informative of speaker identity.

4. EXPERIMENTAL SETUP

Experiments were performed using the SCOTUS and VoxCeleb datasets, evaluating them both on verification and diarization tasks for both SCOTUS and VoxCeleb.

For the SCOTUS experiments, to train the speaker embedding extractors, all utterances which belonged to a speaker with unknown age were discarded, and utterances longer than

[2] `https://www.oyez.org/about`

14 seconds were split up into non-overlapping 10 second utterances, with minimum length 4 seconds. 30 dimensional MFCCs were extracted from these utterances, also removing silence frames and determined by an energy based voice activity detection. The utterances were split into train and test by recording, at an approximate 80% train proportion, and ensuring that the train and test distributions for age using 10 uniformly spaced bins were approximately similar. In order to evaluate the SCOTUS data as a verification task, 15 positive and negative trials were selected for each speaker in the test recordings, excluding any speakers seen in the training set from these trials. Cosine similarity was used to score all embeddings.

The speaker embedding extractors for both VoxCeleb and SCOTUS followed the original x-vector architecture, up until the embedding layer which had 256 hidden units. For classification heads that utilized the standard cross entropy loss, these had a similar architecture to the x-vector network, having two hidden layers also of dimension 256 before being projected to the number of classes. The non-linearity used throughout was Leaky ReLU. For classification heads that used an angular penalty loss, these were simply a single affine matrix on top of the embedding layer that projected into the number of classes, using the CosFace [19] loss.

Embedding extractors were trained with different configurations of classification heads, along with different weightings for each loss, and then embedding performance was assessed based on verification and diarization performance. An adversarial recording classification head with a negative coefficient λ predicting the recording, using a gradient reversal layer was also employed [20, 21]. Control experiments were performed by randomly shuffling the labels of the additional tasks, such as age, so that the distribution of these labels was maintained despite them being converted into noise.

For all embedding extractor training setups for SCOTUS, regardless of the number of classification heads, networks were trained for 50,000 iterations with batch size 500.

For diarization of the SCOTUS corpus, embeddings were extracted for every 1.5s with 0.75s overlap using reference speaker activity detection segmentation, using cosine similarity for scoring and using agglomerative hierarchical clustering to the oracle number of speakers. Due to the supreme court justices appearing in both train and test recordings, the diarization error rate was evaluated for two scenarios: The first evaluation scenario was standard in that all the speech segments were scored, including the speakers which appeared in the training set. The second evaluation scenario was to only score portions of the speech in which unseen speakers were talking.

Speaker embedding extractors for VoxCeleb were trained on the VoxCeleb 2 training set, with 5994 speakers, augmented in the standard Kaldi [22] fashion. Speakers who were the only members of their nationality in the training set were grouped into the same class as the speakers who could not have their nationality scraped. Networks were trained for 100,000 iterations with batch size 500.

5. RESULTS AND DISCUSSIONS

The results for adding a 10-class age classification task to speaker embedding extractors for the SCOTUS corpus can be seen in Table 1, separated into the models trained with cross entropy loss on the speaker classification head, and the models trained with CosFace loss on the speaker classification head. The Diarization Error Rate (DER) results are split into two scoring scenarios, 'All' indicates that all speech was scored and 'Unseen' indicates that only speech segments from unseen speakers was scored. Adding gender classification was not found to have any positive effect, and thus these results have been omitted for brevity.

As we can see for the standard cross entropy loss, for configurations of the age loss with $\lambda \leq 0.5$, the verification and diarization performance was improved over both the baseline and the adversarial recording task. In comparison, both the control experiment of randomly shuffled labels and the control experiment featuring completely random speaker labels yielded no improvement.

Table 1 also features an experiment in which the only classification head was the Age classification head, and this performs surprisingly well on the speaker verification and diarization tasks, despite only being trained to distinguish between 10 age categories. The age accuracy of the networks trained with both speaker and age outperformed that of training on age alone, suggesting that performing tasks in combination was able to improve both tasks. This is supported by previous MTL literature, which indicates training on multiple tasks may be helpful to each individually.

For SCOTUS utterances, it is clear from Table 1 that CosFace improves the speaker verification and diarization performance over the standard cross entropy loss, with the change in the scale of the weightings λ to account for the change in the relative scale of $\mathcal{L}_{\text{speaker}}$. The addition of the age classification head similarly improves over the baseline, producing the best results for verification and diarization for all the configurations shown, making a relative improvement of 17.8% in DER and 8.9% in EER over the 'Only Speaker' CosFace baseline. However, the performance on VoxCeleb suggests the benefit of the CosFace loss does not carry over across domains, which is supported by the findings of [23] with regards to angular penalty losses.

The performance of VoxCeleb models can be seen in Table 2. For both the CosFace and standard models, adding a nationality classification task produced improvements in both EER and DER over their respective baselines, with the CosFace models showing a 3% relative improvement in EER and 12.9% in DER.

While both adding both age and nationality tasks improved SCOTUS and VoxCeleb models respectively, the

	SCOTUS				VoxCeleb
			DER		
	EER	Age Acc.	All	Unseen	EER
Only Speaker (Baseline)	3.14%	-	27.58%	19.75%	15.51%
+Random labels ($\lambda_{\text{rand}} = 0.2$)	3.78%	-	27.87%	19.14%	17.02%
+Age ($\lambda_{\text{age}} = 0.8$)	3.36%	**78.1%**	33.38%	23.12%	15.85%
+Age ($\lambda_{\text{age}} = 0.5$)	2.89%	**78.1%**	26.72%	18.74%	15.30%
+Age ($\lambda_{\text{age}} = 0.2$)	**2.68%**	77.9%	26.14%	18.02%	**14.95%**
+Age ($\lambda_{\text{age}} = 0.1$)	2.99%	77.2%	**25.99%**	**17.68%**	15.03%
+Recording ($\lambda_{\text{rec}} = -0.1$)	4.83%	-	26.18%	17.71%	15.39%
Only Speaker (CosFace Baseline)	2.71%	-	26.51%	19.75%	16.52%
+Age ($\lambda_{\text{age}} = 0.1$)	**2.47%**	76.3%	25.34%	15.88%	16.72%
+Age ($\lambda_{\text{age}} = 0.05$)	2.52%	**75.4%**	23.10%	15.94%	**15.94%**
+Age ($\lambda_{\text{age}} = 0.01$)	2.62%	70.9%	**21.80%**	**14.08%**	16.32%
+Age ($\lambda_{\text{age}} = 0.001$)	2.62%	57.3%	23.41%	15.35%	16.41%
+Recording ($\lambda_{\text{rec}} = -1.0$)	3.67%	-	22.19%	15.56%	15.97%
Only Age	3.99%	77.0%	37.08%	26.18%	17.76%
Only Gender	19.42%	-	58.02%	44.78%	27.89%
Only Random	23.31%	-	69.13%	48.97%	30.18%

Table 1: Verification and diarization performance for various models trained on SCOTUS utterances, with some models trained on additional age and recording classification tasks.

	EER	Nat.Acc.	DER
Only Speaker (Baseline)	6.19%	-	26.07%
+Random ($\lambda_{\text{rand}} = 0.2$)	8.31%	-	38.41%
+Nationality ($\lambda_{\text{nat}} = 0.5$)	6.37%	**72.2%**	29.32%
+Nationality ($\lambda_{\text{nat}} = 0.2$)	**5.89%**	72.0%	25.42%
+Nationality ($\lambda_{\text{nat}} = 0.1$)	5.90%	71.5%	**25.28%**
Only Sp. (CosFace BL)	3.04%	-	28.39%
+Nationality ($\lambda_{\text{nat}} = 0.1$)	2.99%	71.5%	26.28%
+Nationality ($\lambda_{\text{nat}} = 0.05$)	**2.95%**	72.3%	**24.73%**
+Nationality ($\lambda_{\text{nat}} = 0.01$)	3.00%	**72.9%**	26.61%
Only Nationality	13.38%	**73.8%**	57.97%

Table 2: Verification and diarization performance of models trained on VoxCeleb 2. The EER is on VoxCeleb. The DER is evaluated on SCOTUS, as in Table 1, scoring all speech regions.

age task had a larger overall impact. We suggest that the reason for this, along with the gender task not improving performance on SCOTUS, is that both gender and nationality tasks are ultimately coarse groupings of the existing speaker classification task, and thus not introducing as much new information. Despite this, the gains that can be found on VoxCeleb models with adding nationality as a task may suggest that providing the network with additional prior information about the underlying structure of the data can still improve performance on unseen speakers.

The age labels in comparison do not always correspond with the speaker label. As such, they provide a source of information to the network of how variations within the same speaker should be structured with regards to age. This distinction of having intra-speaker variation captured according to some attribute is what may separate the descriptive power of age labels from nationality or gender.

For both age and nationality tasks, adding them also increased performance when evaluating on out of domain data compared to the baseline, with SCOTUS models performing better on VoxCeleb, and vice versa, suggesting that encouraging the embedding space to describe these additional attributes may improve overall robustness also, even if these extra task labels are not available in the target domain.

Overall, this work demonstrated that training on auxiliary speaker attribute tasks in addition to speaker classification can yield strong performance gains for both verification and diarization.

6. CONCLUSIONS

In this work we showed that training on additional tasks relating to speaker attributes, specifically age and nationality, alongside the standard speaker classification can improve the performance of deep speaker embeddings for both verification and diarization. Experimental results on SCOTUS and VoxCeleb showed that explicitly encouraging the embedding space to describe the auxiliary attributes improved performance, even across domain. This improvement was found even with potentially noisy web scraped labels about speakers.

7. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: robust DNN embeddings for speaker recognition," in *IEEE ICASSP*, 2018, pp. 5329–5333.

[2] Gregory Sell, David Snyder, Alan Mccree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur, "Diarization is Hard: some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Interspeech*, 2018, pp. 2808–2812.

[3] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[4] Jennifer Williams and Simon King, "Disentangling style factors from speaker representations," in *Interspeech*, 2019, vol. 2019-Septe, pp. 3945–3949.

[5] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak, "X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition," apr 2020, pp. 7169–7173, Institute of Electrical and Electronics Engineers (IEEE).

[6] Soumi Maiti, Erik Marchi, and Alistair Conkie, "Generating Multilingual Voices Using Speaker Space Translation Based on Bilingual Speaker Data," in *IEEE ICASSP*, 2020, vol. 2020-May, pp. 7624–7628.

[7] Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur, "Probing the Information Encoded in X-Vectors," in *IEEE ASRU*, 2019, pp. 726–733.

[8] Michael Jessen, "Speaker classification in forensic phonetics and acoustics," *Lecture Notes in Computer Science*, vol. 4343 LNAI, pp. 180–204, 2007.

[9] John H.L. Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[10] Rich Caruana, *Multitask Learning*, pp. 95–133, Springer US, Boston, MA, 1998.

[11] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Interspeech*, vol. 2017-August, pp. 2616–2620, 2017.

[12] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech*, vol. 2018-Septe, no. ii, pp. 1086–1090, 2018.

[13] Shahla Parveen and Phil Green, "Multitask learning in connectionist robust ASR using recurrent neural networks," in *Eurospeech*, 2003, pp. 1813–1816.

[14] Peter Bell, Pawel Swietojanski, and Steve Renals, "Multitask learning of context-dependent targets in deep neural network acoustic models," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 238–247, 2017.

[15] Subhadeep Dey, Takafumi Koshinaka, Petr Motlicek, and Srikanth Madikeri, "DNN Based Speaker Embedding Using Content Information for Text-Dependent Speaker Verification," in *IEEE ICASSP*, 2018, vol. 2018-April, pp. 5344–5348.

[16] Yi Liu, Liang He, Jia Liu, and Michael T. Johnson, "Introducing phonetic information to speaker embedding for speaker verification," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–17, dec 2019.

[17] Peter B. Mueller, "The aging voice," *Seminars in Speech and Language*, vol. 18, no. 2, pp. 159–168, 1997.

[18] Sarah K. Rapoport, Jayme Meiner, and Nazaneen Grant, "Voice Changes in the Elderly," aug 2018.

[19] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," 2018.

[20] Yaroslav Ganin and Victor Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," 9 2014.

[21] Chau Luu, Peter Bell, and Steve Renals, "Channel adversarial training for speaker verification and diarization," in *IEEE ICASSP*, 2020, pp. 7094–7098.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE ASRU*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[23] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, Pedro A. Torres-Carrasquillo, and Najim Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech and Language*, vol. 60, 2020.