Towards A Fault-tolerant Speaker Verification System: A Regularization Approach To Reduce The Condition Number

Siqi Zheng, Gang Liu, Hongbin Suo, Yun Lei

Machine Intelligence Technology, Alibaba Group

{zsq174630,g.liu, gaia.shb, yun.lei}@alibaba-inc.com

Abstract

Large-scale deployment of speech interaction devices makes it possible to harvest tremendous data quickly, which also introduces the problem of wrong labeling during data mining. Mislabeled training data has a substantial negative effect on the performance of speaker verification system. This study aims to enhance the generalization ability and robustness of the model when the training data is contaminated by wrong labels. Several regularization approaches are proposed to reduce the condition number of the speaker verification problem, making the model less sensitive to errors in the inputs. They are validated on both NIST SRE corpus and far-field smart speaker data. The results suggest that the performance deterioration caused by mislabeled training data can be significantly ameliorated by proper regularization.

Index Terms: Speaker Verification, Mislabeled Data, Entropy Minimization, Loss Regularization, Co-Training, Condition Number

1. Introduction

Recently we witnessed an increase in a variety of application scenarios of speaker verification, ranging from call centers to smart speakers [1][2]. The problem of wrong labeling data is less of a concern in the traditional telephone setting, in which there is usually a one-to-one mapping between the speaker and the recording device (such as a personal cell phone). However, in more complicated scenarios such as smart speakers, whose popularity are raising rapidly recently, the labeling of online data is a lot more challenging. These home virtual-assistants are designed to be shared among a group of users. Hence a simple data cleaning pipeline no longer meets the need of harvesting large-scale online speakers data for training.

In industry, a traditional approach to collect training corpus for smart speakers is to recruit subjects and record their voices in a simulation studio, which is expensive and highly time-consuming. An alternative solution is to cluster users' voices based on each device Serial ID using a pre-trained model learned from other data sets. Due to problems such as domain mismatch, it is expected to have performance degradation [3] and will unavoidably assign wrong labels to unneglectable proportions of training samples.

Despite the fact that some unlabeled and automatically labeled data are included in NIST SRE 16 and 18 corpus[4][5], little attention has been drawn to the study of training with potentially mislabeled speakers data. This is because the majority of NIST SRE corpus is clearly labeled and is sufficient to train a well-behaved speaker verification model. The potential deterioration incurred by mistakes in automatically labeled data is marginal. In the complex real-world settings, however, there is a stronger motivation to focus on the correctness of the labels. As the complications of different types of data increase, we often find ourselves in situations with limited clean, labeled speakers and abundant of inexpensive, unlabeled utterances. Thus we are interested in investigating to what extent mislabeled data can harm the performance of the speaker verification system and how we can minimize such harm.

Conditions	Durations	EER(male)	EER(female)
No mislabeled	full-full	**2.67%**	**2.81%**
10% mislabeled	full-full	12.81%	11.58%
20% mislabeled	full-full	14.86%	15.44%
No mislabeled	10s-10s	**14.86%**	**15.06%**
10% mislabeled	10s-10s	24.47%	25.00%
20% mislabeled	10s-10s	27.27%	29.93%

Table 1: *The mislabeling effect of speaker verification system on the NIST SRE corpus*

To get a taste of the impact of the mislabeled data on the speaker verification system, a pilot experiment is performed on the training corpus of NIST SRE 04-10. 10% or 20% of the total utterances in the corpus are randomly selected and assigned to labels of other speakers. As shown in Table 1, when the 10% labels are mistaken-ed, the equal error rates increase absolutely around 10%, for both male and female. When the proportion increases to 20%, further deterioration is observed. Note that the "full-full" duration condition means the full utterance is used for both enrollment and test, while "10s-10s" means only the first 10s of an utterance is used for both enrollment and test.

Let $\mathcal{F} : \mathbf{x} \mapsto \mathbf{y}$ be the mapping of training data $\mathbf{x}$ to a learned model $\mathbf{y}$. The condition number of $\mathcal{F}$ is a measure of sensitivity of the function to the changes or errors in the inputs and is formally defined as the asymptotic supremum of the relative derivative of the function with respect to the inputs[6][7],

$$\lim_{\epsilon \to 0} \sup_{\|\delta x\| \leq \epsilon} \frac{\|\delta \mathcal{F}\|}{\|\delta x\|}.$$

When facing a corpus whose samples may be partially mislabeled, it is of our interest to reduce the condition number of the problem so that small errors in the inputs do not result in large fallacy in the learned model.

[8] offers an informative read for the general entropy regularization framework. It discusses when and how entropy regularization can benefit a semi-supervised learning problem. It provides theoretical reasoning that the conditional entropy can be substituted by an empirical distribution. Motivated by the idea, we propose a regularized entropy loss for speaker verification as a replacement of the loss function in x-vector system[9].

A common practice to improve the generalization ability of the learned model is to simply increase the size of training corpus. In most cases, however, acquiring large-scale correctly-labeled speakers data is impractical. Co-Training is an effective

regularization approach to reduce the condition number of the training problem, allowing us to train with large-scale unlabeled speakers data using inferred labels[10].

Another effective approach to improve model robustness without having substantial quantities of labeled data for training is proposed in [11]. Snyder et al. demonstrate that augmenting labeled data with random music, noise, and reverberation can improve the performance of the model. Motivated by the idea of data augmentation, we introduce a segment reshuffling process for potentially mislabeled data. Instead of augmenting training data with music and reverberations collected from other domain, we find that mislabeled data themselves can be a good resource for data augmentation.

2. Systems

2.1. Baseline

The x-vector speaker verification system described in [9] is used as the baseline in this study. Five layers of frame-level time-delay neural network architecture is implemented before a global average pooling layer is used to aggregate frame-level information. Several segment-level layers are followed before the softmax output layer. A cross entropy loss is used to train the time-delay neural network:

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} d_{nk} \ln\left(P(spkr_k|x_{1:T}^{(n)})\right), \quad (1)$$

when there are N utterances from K speakers, $d_{nk} = 1$ if the speaker label for segment n is k and 0 otherwise. $P(spkr_k|x_{1:T}^{(n)})$ measures the probability that utterance n with input frames $\{x_1^{(n)}, x_2^{(n)}, \ldots, x_T^{(n)}\}$ belongs to speaker k.

2.2. Entropy Loss Regularization

The identity function d_{nk} in Equation (1) takes the value of only 0 and 1. When dealing with labels that come with uncertainty, we relax the constraint to take any values between 0 and 1. The identity function d_{nk} is replaced by a probability function that measures the likelihood utterance n belongs to speaker k.

Following the notations in Equation (1), we propose a regularized form of Entropy Loss from (1), as below:

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(I_{lab}(n,k)\cdot F(n,k) + I_{inf}(n,k)\cdot G(n,k)\Big), \quad (2)$$

where

$$F(n,k) = d_{nk}\ln\left(P(spkr_k|x_{1:T}^{(n)})\right),$$

$$G(n,k) = P(spkr_k|x_{1:T}^{(n)})\ln\left(P(spkr_k|x_{1:T}^{(n)})\right),$$

and $I_{lab}(n,k)$ is the indicator function that equals 1 if the utterance n from speaker k is manually labeled and 0 otherwise. On the contrary, $I_{inf}(n,k)$ is the indicator function that equals 1 if the label of utterance n from speaker k is inferred by the classifier and 0 otherwise.

When a mislabeled sample is encountered in training, the regularized function would multiply the entropy loss by a deduced probability, resulting in a loss strictly less than the original loss used for backpropagation. This suggests that the training of the model is less vulnerable to an error in input, or, in other words, smaller condition number.

To minimize the impact of the poor samples or wrongly labeled samples, a quality-based thresholding mechanism is adopted to do sample selection for LDA and PLDA training. For each of the speaker embedding n that is inferred to have label k, mean similarity score between n and all other embeddings in k is calculated. Embeddings with the least scores are dropped before training LDA and PLDA. A good overall estimate of the quality of training data is helpful for determining a threshold for dropping utterances. If the threshold is too loose, the system may include dirty samples in training while too strict might force the system to reject samples that are instrumental for training a robust model. Lastly, a self-attention pooling layer is also applied to replace the global average temporal pooling layer in the baseline system.

2.3. Regularization by Segments Reshuffling

The training data can be "regularized" by segments reshuffling process. Equipped with a sufficiently good pre-trained model, speaker verification system can estimate the class confidence of an utterance. A mean similarity score is calculated between each utterance and all other utterances that are marked with the same class (or label). Less confident utterances are segmented and reshuffled for data augmentation. The pre-trained model can be trained with a small set of clean labeled data, or out-of-domain data. This segments reshuffling approach is especially promising on smart speakers data, where users accumulate lots of short utterances during the course of daily interaction.

Let $S = S_C \cup S_D$ be the set of all training utterances inferred to be from speaker A, where S_C is the set of utterances that the system has more confidence in the labeling, and S_D is set of utterances whose labels the system is less certain. Let M be the number of utterances in S and s_i be the mean similarity score between utterance i, denoted as u_i, and all other utterances $u_j \in S, j \neq i, S = \{u_1, \ldots, u_M\}$. Now let $S_C = \{u_i | \forall i \leq M \textit{ where } s_i > \alpha\}$ for some constant α, and $S_D = \{u_j | \forall j \leq M \textit{ where } s_j \leq \alpha\}$.

Let K be the number of utterances in S_C. Each of the utterances $u_j \in S_D$ are split into N segments, $N \leq K \leq M$. N utterances are randomly selected from S_C. Each of the chunked segments in turn are attached to the N utterance of S_C, as shown in Figure 1. For each of the utterance u_j in the uncertain set S_D, we reshuffle and attach the segments in the same manner and create a new set S_j. Finally, we obtain a new set of training utterances for speaker A: $S' = S_C \cup S_1 \cup \ldots \cup S_{M-K}$. If $N = K = M$, then S_D is empty and no reshuffling is performed, in other words, $S' = S$.

To see how segments reshuffling helps to reduce the condition number and improve generalizability of the deep neural network, consider the case when the segmented utterance is indeed a misclassified speaker. Instead of treating the entire utterance as a sample data used for updating the deep neural network, we now see it as segments of noises attached to other utterances from the specified speaker. In the setting of smart speakers in a family, for example, the mislabeled speakers are usually other family members whose voices may sometimes occur as background noise in the recordings. This is more helpful than using random music or noise from completely different channel domain. Therefore, segments reshuffling turns a harmful sample into a useful resource for data augmentation.

It may raise concerns that there are situations when we mistakenly reshuffle segments from the correctly labeled recordings and hence lose some useful information that the deep neural network can learn from. First it needs to be noted that the infor-

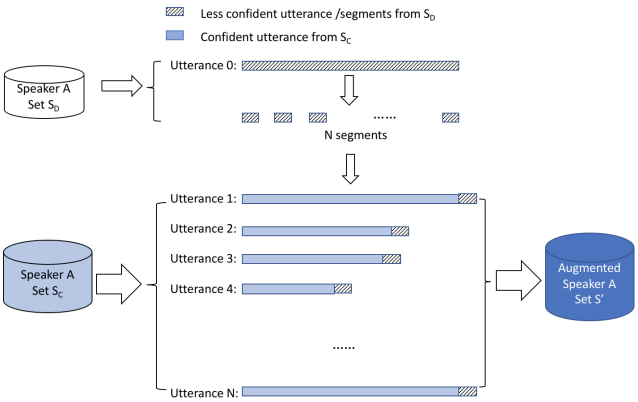

Figure 1: *Illustration of segments reshuffling and augmentation on a single less confident utterance. Utterance 0 is selected, without replacement, from the set of less confident utterances. Utterance 1 to N are randomly selected, with replacement, from the set of confident utterances.*

mation from the segments are never lost - they are attached to other recordings and are still being used repeatedly in the model training. Secondly, since we only segment least confident utterances from a bad quality data set, a good data selection criteria can ensure the majority of the segmented recordings are the mislabeled ones. Statistically speaking, the gains from regularizing mislabeled speaker segments is expected to outweigh the potential loss resulted from missing a few informative samples.

2.4. Co-Training

Regularization can also be achieved at the model level via co-training. Two speakers classifiers are trained simultaneously, each serves as a regularizer for the other. Under certain conditions, each of the model plays the role as a rein that helps prevent the other from being misled too much by bad labels.

The labeled data set is split into two sets to independently train the two classifiers, A and B. Through the classifiers each of the unlabeled utterance is projected onto two independent speaker feature spaces to produce, as [10] put it, two different "views". Two conditions have to be satisfied in order to achieve a good performance boost using co-training. First, each of the two classifiers, when working by themselves, are sufficiently good. Second, the training data are conditionally independent.

Figure 2 illustrates an example of the same set of samples projected onto two speaker feature spaces. If the two spaces are conditionally independent given speaker class, we expect utterances with highest confidence from feature space A to have a relatively random distribution when viewing from feature space B.

This provides the ground for training with unlabeled data. The randomness in distribution in speaker feature space B indicates that sometimes data with highest confidence in speaker feature space A can provide additional helpful information to correct mislabeled data from classifier B.

Similarly, results that classifier B finds most confident with can also be informative to update classifier A if these data are classified differently in speaker feature space A. Therefore, iteratively training both models simultaneously leads to a quick expansion in training data set and therefore better robustness.

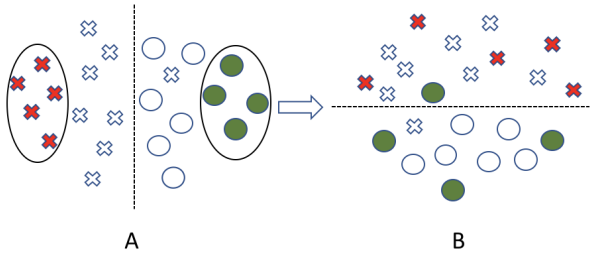

Figure 2: *Multiple views of the co-training. A: Utterances with the highest confidence from speaker feature space A; B: Distribution of the selected utterances from A in speaker feature space B. The disagreement between A and B about the green dot lies above the line in B suggests it might be mislabeled.*

3. Corpora

The proposed methods are validated on two drastically different domains of data - English dominated data from NIST SRE corpus; and mandarin far-field utterances collected from online home-based smart speakers named Tmall Genie.The sampling frequency for these two are 8kHz and 16kHz, respectively. Even though NIST SRE data is not a probable scenario with large proportion of mislabeled data, it is chosen for two reasons, 1) it is a data source of high labeling quality to easily imitate some controlled cases; and 2) unlike almost all industrial proprietary data, it is publicly available for community to reproduce the results in this study. We are interested to see to what extent the discussed regularization approaches can mitigate the performance deterioration shown in Table 1.

SRE training data comes from NIST SRE04-08, a total of 57,517 utterances from 5,767 speakers. Evaluation data set consists of 5,251 males and 6,732 females from NIST SRE10 evaluation set.

Training		unlabeled	
# of speakers	# of utterances	# female	# of male
5767	57517	6732	5251

Table 2: *Summary of data from NIST SRE corpus used in this study.*

On the other hand, 22 millions unlabeled recordings and 0.22 million labeled recordings are collected from Tmall Genie smart speakers. All of the labeled utterances are recorded in a simulation environment, including both far-field and near-field recordings. 5,000 speakers are asked to make recordings in simulation studios, averagingly 44 utterances per speaker. The actual number of speakers from 22 millions unlabeled data is unknown. The inferred number of speakers obtained through unsupervised clustering is around 180,000.

All of the recordings are combinations of wake-up keywords and command messages users speak to the smart speakers. The average length of the recordings are under 2s. No more information can be acquired other than the masked IDs of devices that generated the recordings, and we have no knowledge of how much of the 22 millions data are actually mislabeled.

	labeled	unlabeled
# of speakers	5,000	180,000
# of utterances	0.22M	22M

Table 3: *Summary of data from smart speakers corpus. Note that the number of unlabeled speakers are inferred through unsupervised clustering.*

4. Experiment and Results

The effects of regularization approaches on NIST SRE10 evaluation set are listed in Table 4 and 5. Under the condition of 10% mislabeled data and "full-full" durations, the EER degrades from 2.67% to 12.81% for male. After regularization, the EER is improved to 6.95%. In other words, the regularization approaches ameliorate the negative effects caused by ten percent mislabeled data by 58%. To sum up, regularization approaches can help avoid more than half performance degradation resulted from wrong labels across all scenarios.

Conditions	Durations	EER(baseline)	EER(Regularized)
10% mislabeled	full-full	12.81%	6.95%
20% mislabeled	full-full	14.86%	9.46%
10% mislabeled	10s-10s	24.47%	18.74%
20% mislabeled	10s-10s	27.27%	19.91%

Table 4: *The effects of regularization approaches to the training of mislabeled data on the NIST SRE corpus - male*

Conditions	Durations	EER(baseline)	EER(Regularized)
10% mislabeled	full-full	11.58%	6.51%
20% mislabeled	full-full	15.06%	9.93%
10% mislabeled	10s-10s	25.00%	19.68%
20% mislabeled	10s-10s	29.93%	21.47%

Table 5: *The effects of regularization approaches to the training of mislabeled data on the NIST SRE corpus - female*

Above-mentioned results suggest that regularization approaches are effective on NIST SRE corpus, where wrong labels are artificially created in a controlled manner. It is worthwhile to explore its validity on the real-world smart speakers data, in which wrong labels exist by nature.

First we are interested in discovering the effects of simply expanding the size of training data by a large magnitude, without applying any regularization techniques. Table 6 shows that the EER of the model learned from 0.2 millions of labeled utterances is 10.86%. The learned model is then used as a simple classifier to automatically cluster each of the 20 millions unlabeled utterances, based on the device serial ID. It is expected that this will present a training corpus with a considerable proportion of bad labels. However, we can still observe a slight improvement using tremendous unlabeled data with EER reducing from 10.86% to 9.80%. The demographic structure of the data from online smart speakers provides a possible explanation for this improvement. A preliminary investigation on a random sample of 100 online devices suggests at least 70% of the data has high speaker labeling quality[12].

As it turns out, regularization approaches discussed in Sec.2 further boost the performance when training with 20 millions of unlabeled utterances. Entropy Minimization, Segments Reshuffling, and Co-Training together improve the EER from 9.80% to 6.33%. Since there is no way of knowing the truth labels of all 22 millions utterances, we are not able to quantitatively measure the exact effect of regularization, as we did for NIST SRE corpus. However, a relative reduction of 35.4% in EER demonstrates that regularization has effectively reduced the condition number of mislabeled smart speakers training problem.

Re-visiting the definition in Sec. 1, let $\mathcal{F} : \mathbf{x} \mapsto \mathbf{y}$ be the original training problem and $\mathcal{G} : \mathbf{x} \mapsto \mathbf{y}$ be the regularized one. The three regularization approaches make the output $\mathbf{y}$ less sensitive to a unit of change in input $\mathbf{x}$. In other words,

$$\lim_{\epsilon \to 0} \sup_{\|\delta x\| \leq \epsilon} \frac{\|\delta \mathcal{G}\|}{\|\delta x\|} < \lim_{\epsilon \to 0} \sup_{\|\delta x\| \leq \epsilon} \frac{\|\delta \mathcal{F}\|}{\|\delta x\|}.$$

Therefore, regularization is able to better handle errors in automatically labeled data and hence allows deep neural network to effectively leverage the benefit of large training data.

# of utterances	Corpus Quality	EER
0.2M	No mislabeled	10.86%
20M	With mislabeled	9.80%

Table 6: *Effects of increasing the size of corpus without any regularization methods on far-field smart speakers corpus*

Methods	EER
Baseline - no regularization	9.80%
Entropy Minimization	7.54%
Ent. Min. + Segments Reshuffling	7.07%
Ent. Min. + Seg. Res. + Co-Training	6.33%

Table 7: *Performance comparison of regularization methods on far-field smart speakers corpus - trained on 20M utterances*

5. Conclusion

In this paper the impact of a non-negligible proportion of mislabeled training data on a speaker verification system is explored, which has rarely been addressed before. Experiments on NIST SRE corpus show that mislabeled data can severely damage the performance of speaker verification systems. Then three regularization approaches are proposed to remediate the damage. First, a modified version of entropy loss in the x-vector system is proposed, which relaxes the constraints of speaker identity function to a probabilistic one. Secondly, the segments of least confident speaker utterances are reshuffled and attached to randomly selected utterances from the same speaker. This regularizes the effects of bad labels and serves as a means of data augmentation. Lastly, co-training is employed to boost performance by tremendously increasing the training data size.

All three regularization methods successfully reduce the condition number of mislabeled speakers training problem. Significant performance improvements are observed on both NIST SRE and far-field smart speakers corpus. The results are quite meaningful, especially in the industrial setting, as the increasing complications of real-world application scenarios often leave us with contaminated online data for training. The proposed methods can help the speaker verification system leverage the bonanza of the ever-growing online data while tolerate its inherent missed-information. This study serves as a pilot step towards a fault-tolerant speaker verification system. As a next step, more accurate labeling quality measure can be explored to strike a balance between sample variety and wrong automatic labeling.

6. References

[1] F. A. R. E. Chowdhury, Q. Wang, I. Lopez-Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20*, 2018, pp. 5359–5363. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461587

[2] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20*, 2018, pp. 4879–4883. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462665

[3] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9*, 2014, pp. 4002–4006. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6854353

[4] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24*, 2017, pp. 1353–1357. [Online]. Available: http://www.isca-speech.org/archive/Interspeech/2017/abstracts/0458.html

[5] NIST, "The 2018 NIST speaker recognition evaluation." [Online]. Available: https://www.nist.gov/sites/default/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf

[6] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley, 1980.

[7] D. A. Belsley, *Conditioning diagnostics: collinearity and weak data in regression*. Wiley, 1991.

[8] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, 2004, pp. 529–536. [Online]. Available: http://papers.nips.cc/paper/2740-semi-supervised-learning-by-entropy-minimization

[9] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden*, 2017, pp. 999–1003. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0620.html

[10] A. Blum and T. M. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26,*, 1998, pp. 92–100. [Online]. Available: https://doi.org/10.1145/279943.279962

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20*, 2018, pp. 5329–5333. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461375

[12] S. Zheng, G. Liu, H. Suo, and Y. Lei, "Autoencoder-based semi-supervised curriculum learning for out-of-domain speaker verification," in *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings*, 2019.