Study on acoustic model personalization in a context of collaborative learning constrained by privacy preservation

Salima Mdhaffar[1][0000−0002−8472−6890], Marc Tommasi[2][0000−0003−2838−4408], and Yannick Estève[1][0000−0002−3656−8883]

[1] LIA - Avignon Université, France
[2] Université de Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, France
firstname.lastname@univ-avignon.fr

Abstract. This paper investigates different approaches in order to improve the performance of a speech recognition system for a given speaker by using no more than 5 minutes of speech from this speaker, and without exchanging data from other users/speakers. Inspired by the federated learning paradigm, we consider speakers that have access to a personalized database of their own speech, learn an acoustic model and collaborate with other speakers in a network to improve their model. Several local personalizations are explored depending on how aggregation mechanisms are performed. We study the impact of selecting, in an adaptive way, a subset of speakers's models based on a notion of similarity. We also investigate the effect of weighted averaging of fine-tuned and global models. In our approach, only neural acoustic model parameters are exchanged and no audio data is exchanged. By avoiding communicating their personal data, the proposed approach tends to preserve the privacy of speakers.
Experiments conducted on the TEDLIUM 3 dataset show that the best improvement is given by averaging a subset of different acoustic models fine-tuned on several user datasets. Our approach applied to HMM/TDNN acoustic models improves quickly and significantly the ASR performance in terms of WER (for instance in one of our two evaluation datasets, from 14.84% to 13.45% with less than 5 minutes of speech per speaker)

Keywords: automatic speech recognition · privacy-protection · collaborative learning · acoustic models · personalization

1 Introduction

User interface of modern electronic and personal devices are more and more based on voice interaction and this tendency would probably continue increasing during the next years. Automatic speech recognition (ASR) is the technology at the core of voice interaction systems. To attain a satisfactory level of usability, ASR models need to be trained with a huge amount of training data costly to be collected and annotated. But an important issue with the data collection is

privacy preservation. Indeed, some users are now very reluctant to use software solutions that do not preserve their privacy. Efforts towards privacy have been made by European states and the General Data Protection Regulation (GDPR) for instance constraints the way to realize data collection. Speech signals can be considered as sensitive information because in addition to the linguistic content, speech also brings information about the speaker: identity, gender, age, health, emotion...[8]

Different levels of privacy preservation can be defined according to the private information to preserve. Two main approaches have been proposed depending on the information to hide. In the Interspeech VoicePrivacy Challenge [13], the aim of privacy preservation consisted in modifying the speech representation features, trying to remove the speaker identity without removing the linguistic content. In this scenario, the data is first anonymized, and then collected. Even if very promising results have been reached with these contributions, data anonymization is still imperfect and negatively impacts the performance of ASR systems. Another approach consists in avoiding to share data: data is only used locally, on the user device, to personalize the model to this user. Then the models are exchanged, assuming that adapted models contain less sensitive information than the data itself. Such approaches have been used in different works for speech recognition [7], mainly through the use of distributed learning to speech the acoustic model train process [16].

Instead of targeting the improvement of a single general model by sharing anonymized data or applying a distributed learning approach, we propose in this paper to focus on the personalization of an initial model to each user. Inspired by widespread of powerful personal devices, we consider speakers that have access to a personalized database of their own speech. In this scenario, closely related to personalized federated learning [5,14], it is possible to both locally fine-tune an acoustic model and collaborate with other speakers in a network to improve their own model.

The paper is organised along the following lines. Section 2 presents related works. Section 3 details the model adaptation. The experimental setup are described in Section 4. The experimental results are presented in Section 5 before concluding and giving some perspectives in Section 6.

2 Related Work

In this study the term '*personalization*' can be interpreted as '*speaker adaptation*', more used in the speech community. A nice overview of speaker adaptation techniques for neural acoustic models has been presented in [1], that classifies adaptation techniques into three categories: embedding-based approaches that relies to the use of auxiliary speaker-dependent features like i-vectors [3], model-based approaches that relies to speaker data to update the neural weights, and data augmentation approaches '*which attempt to synthetically generate additional training data with a close match to the target speaker, by transforming the existing training data*'. No approach based on the use of collaborative train-

ing was mentioned in this paper for speaker adaptation, but collaborative training has already been investigated for acoustic model training. In [2,7], federated learning was applied to improve a general shared acoustic model with the goal of privacy preservation, but no speaker adaptation was targeted. Federated learning was also experimented in [4] to speed up the training process and improve the shared general acoustic model performance.

3 Model adaptation

Our objective is to locally improve the acoustic model for a target speaker by taking advantage from both local pre-existing data and from pre-existing models specific to other users. In our scenario, a global acoustic model is available, trained on the initial corpus. This global model is distributed to all the devices, on which it is possible to fine-tune a local instance of the global model by exploiting locally the user data. These local models can be shared in order to indirectly take benefit from the local data used to their adaptation, through a model averaging.

Since the number of speakers (*i.e* devices) can be very high, and so the number of adapted models, it seems relevant to propose a strategy to better select these local models that could be use to adapt the model of a target speaker. In a classical hybrid HMM/DNN speech recognition acoustic features like MFCC (Mel-Frequency Cepstral Coefficients) are generally augmented with additional speaker-specific features like i-vectors [6] or x-vectors [12] that can capture information about the speaker. In this work, we assume that this kind of information cannot be exchanged between the different devices since we want to avoid to share explicit knowledge about the speaker and the linguistic content present in the data. To select the best candidate models from the other speakers, we suggest to consider the euclidean distance between candidate models and the model fine-tuned on the target user data.

Our study explores different ways to adapt a HMM/DNN acoustic model through the use of model averaging, local fine-tuning, or a combination of these approaches. The aim of this adaptation is to modify the parameters of the (generic) neural network involved in the HMM/DNN architecture. Fine-tuning consists in continuing the training process of the generic acoustic model on a small dataset of the target speaker, by taking care on avoiding overfitting. Model averaging consists on computing a model whose each weight is the average of the weights extracted from a set of models that share the same neural topology.

Figure 1 illustrates the adaptation approach explored in this work. In this framework, no user data is shared: the fine-tuning is made locally, only adapted models can be exchanged.

4 Experimental setup

This section describes the ASR system, the experimental methodology and the datasets used for the experiments on speaker adaptation through fine-tuning and

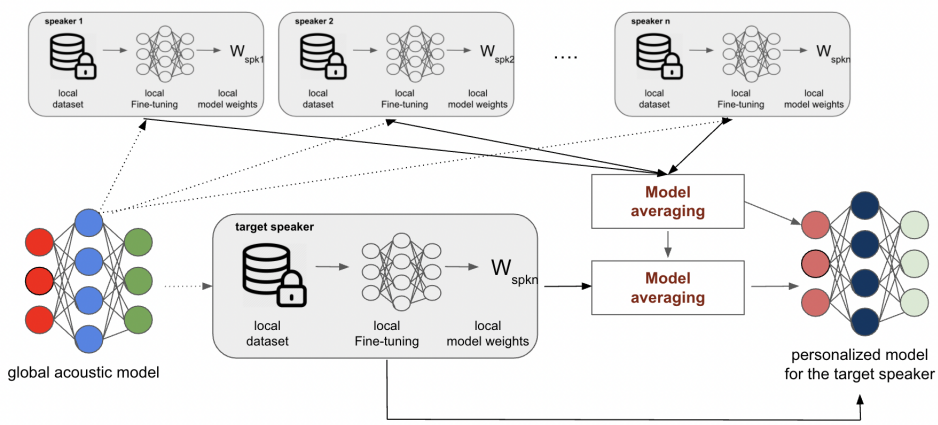

Fig. 1. Model personnalization for a target speaker

model averaging, applied in addition to the classical use of i-vectors as auxiliary input features for neural network speaker adaptation.

4.1 ASR system

The ASR system is based on the Kaldi toolkit [9]. Acoustic models are based on a chain-TDNN approach [10]. The chain-TDNN setup is based on 13 layers with dimension 512 and is trained on cepstral mean and variance normalized 40-dimensional MFCC features concatenated with 9 left and 9 right neighbor frames. We also incorporates i-vectors as an additional input features. The acoustic model has about 14 million parameters. The initial and final learning rates were equal to 0.00025 and 0.000025 respectively. Training audio samples were randomly perturbed in speed and volume during the training process. This approach is commonly called audio augmentation.

When fine-tuning the generic model on target speaker data, we modify only the value of learning rate (the initial and final learning rates were equal to 0.000025 and 0.000015 respectively) and all hyperparameters (i.e. learning rate and local epochs number) are assumed to be homogeneous among all workers.

We make available complete recipes for building the generic acoustic model and the fine-tuned models[3].

As described below, the TEDLIUM 3 dataset was used to train the acoustic models. Data used to train the model is not a part of the TEDLIUM 3 data and is described in [11]. The language model used in our experiment is a 4-gram model, which was pruned to 10 million n-grams.

[3] https://github.com/mdhaffar/Acoustic_model_personalisation

4.2 Experimental Methodology

The experiments are conducted in the following way. We start by building a generic ASR model using a large set of utterances from many speakers. Then, every speaker is associated with a worker that fine-tunes the generic model with a fresh set of utterances of the given speaker. We obtain a set of fine-tuned ASR models. Then we try to collaboratively improve these fine-tuned models in different ways.

Let us consider in the following a set of n different speakers. Let us denote by G the generic model, P_s the fine-tuned model of speaker s. We call an average model of a given set of models, the model defined by the average of model parameters component wise. We denote by $\bar{P}_s$ the per-speaker average of all fine-tuned models $P_{s'}$ except P_s: i.e $\bar{P}_s = \frac{1}{n-1} \sum_{s' \neq s} P_{s'}$. For a given speaker s and an integer k, we denote by $\bar{B}_s^{\leq k}$ the average of the k best models, measured by the WER on the set of utterances of s used to fine-tune P_s.

We also performed a hierarchical clustering on personal models represented by the vector of their weights of the first layer only. This choice of the first layer as a representative layer comes from a preliminary study we made. This study showed us that the word error rate obtained by using an acoustic model fine tuned on a non-target speaker and the Euclidean distance between the first layer of this model and the first layer of the model fine-tuned on the target speaker data is the most correlated, in comparison to the use of other layers. This is illustrated by Figure 2.

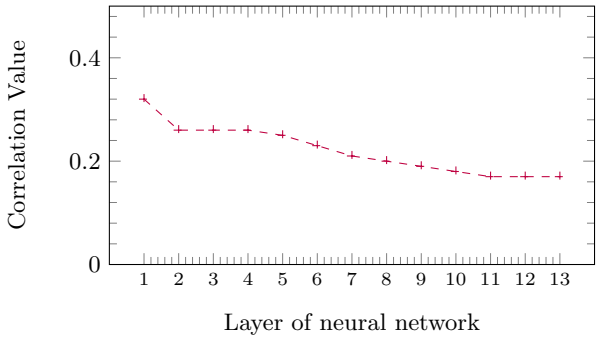

Fig. 2. Pearson correlation between WER and Euclidian distance in function of the layer order on the *perso*1 dataset, described in Section 4.3

For the hierarchical clustering, we use the Numpy library[4] with the *ward* linkage function. Let us recall that the principle of the hierarchical clustering is to build a hierarchy of clusters in bottom-up fashion. The Euclidean distance

[4] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html`

between weight vectors is used to compute the distance between neural network models. In an iterative process, the two closest clusters are successively merged until only one remains. The output can be represented by a dendrogram. The Ward linkage function [15] is used to evaluate the distance between clusters. It is based on minimum variance method and allows to minimize the total within-cluster variance.

Using this dendrogram, for a given speaker s and an integer k, we compute $\bar{D}_s^{\leq k}$ the average of k closest models to P_s (in terms of distance within the dendrogram). [5]

We perform different kinds of aggregation of the fine-tuned models using a weighted average combined to:

1. the generic model G,
2. or the per-speaker average of all fine-tuned models $\bar{P}_s$
3. or the k-best fine-tuned models $\bar{B}_s^{\leq k}$ (models that have the lowest WER on data from speaker s)
4. or the k-nearest neighbours models $\bar{D}_s^{\leq k}$ where the similarity is given by a dendrogram of Euclidean distance between model weights
5. or the average of k models taken at random among all fine-tuned models $\bar{R}_s^k$.

The weighted average of P_s with one these models $M \in \{G, \bar{P}_s, \bar{B}_s^{\leq k}, \bar{D}_s^{\leq k}, R_s^k\}$ is computed by $\alpha P_s + (1-\alpha)M$, i.e. in a component-wise convex combination of the weights.

4.3 Datasets

All experiments to train acoustic models were conducted with the TEDLIUM 3 dataset, a large corpus of 452 hours of TED talks given by 2,295 speakers. The dataset is ready for training ASR systems but also dedicated to speaker adaptation tasks. We processed the dataset in an original way for this set of experiments. We split it into three parts so that the sets of speakers in each part are pairwise disjoints. Characteristics of the three parts are reported in Table 1. The first part is called *generic* and has been used to train an initial acoustic model for ASR. The two other parts called *perso*1 and *perso*2 are used for 2 distinct trials of model personalization and evaluation. In each part $p \in \{perso1, perso2\}$, for each speaker s, we consider a small subset of 5 minutes of speech data called $train_p^s$ to fine-tune a per-speaker model and the remaining is called $test_p^s$ and used for evaluation. These datasets are never shared (or merged) with other data. We consider them as personal and private datasets belonging to speakers. The average duration of $test_p^s$ data is presented in the third line in Table 1. For the reproductibility of experimental results by research community, we give the list of the new division of the dataset[6].

[5] Note that this set may not be unique and we build it iteratively, starting from the closest cluster and choosing models uniformly at random in the last iteration.

[6] https://github.com/mdhaffar/Acoustic_model_personalisation

	generic	*perso*1	*perso*2
Duration (hours)	200	150	170
Duration of speech (hours)	170	125	150
Average duration per speaker (minutes)	-	8.5	8.1
Number of speakers	880	650	765

Table 1. TEDLIUM3 dataset

5 Experimental results

In our experiments, we take k equals to 50, except for $\bar{B}_s^{\leq k}$, where $k = 10$. We measure the average of WER of different models on the $test_p^s$ data. In a more formal way, if we denote by $\mathrm{WER}(M, S)$ the word error rate of model M on the dataset S, then we compute averages in the following way. For each part $p \in \{perso1, perso2\}$, and for a base model M_s in $\{G, \bar{P}_s, \bar{B}_s^{\leq k}, \bar{D}_s^{\leq k}, R_s^k\}$, we compute the average WER on part p as $\frac{1}{n}\sum_{s=1}^{n} \mathrm{WER}(\alpha P_s + (1-\alpha)\bar{M}_s, test_p^s)$. The word error rate of the generic model G and the fine-tuned model P_s are given in Table 2.

	*perso*1	*perso*2
Generic model G	15.43	14.84
Speaker model P_s	15.04	14.63

Table 2. Word Error Rate of the generic and the fine-tuned models.

The values of average WER for the different base models computed as explained above are reported in Table 3 for the values of $\alpha = 0$, when the private data of the given speaker has not been used and $\alpha = 1/2$ when the fine-tuned model and the other model equally contribute to the averaged model. A graphical representation of the results is given in Figure 3 for α varying between 0 and 1.

	$\alpha = 0$		$\alpha = 0.5$	
Base model	*perso*1	*perso*2	*perso*1	*perso*2
G	15.43	14.84	14.43	13.92
$\bar{P}_s$	15.21	14.62	14.38	13.82
$\bar{D}_s^{\leq 50}$	15.49	14.81	14.59	14.01
$\bar{R}_s^{50}$	15.25	14.66	14.35	13.84
$\bar{B}_s^{\leq 10}$	14.72	13.8	14.13	13.45

Table 3. Results of collaborative learning with ($\alpha = 0.5$) and without ($\alpha = 0$) fine-tuned model of the target speaker.

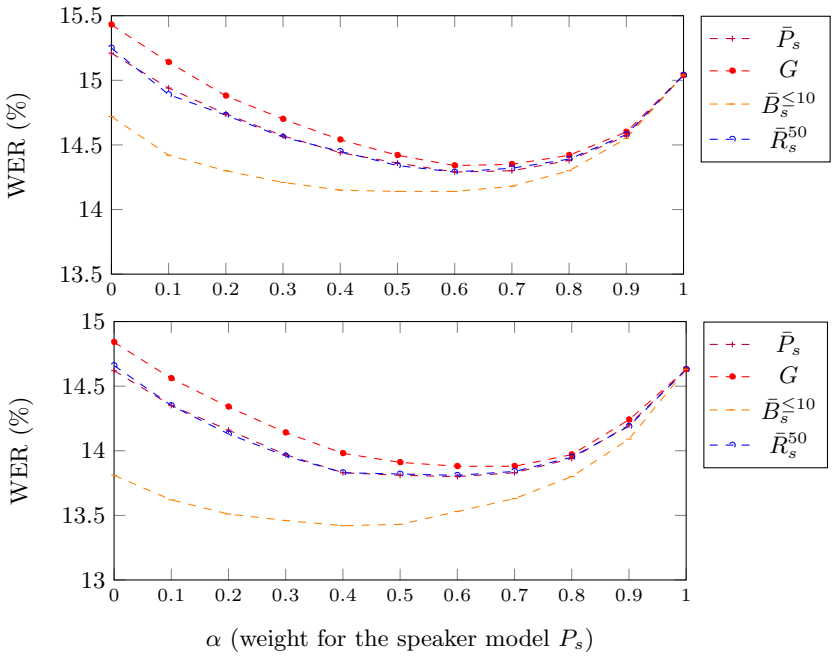

Fig. 3. WER according to different weighted average of P_s with four different aggregated models: all fine-tuned models $\bar{P}_s$, generic model G, 10-best (WER) fine-tuned models $\bar{B}_s^{\leq 10}$, 50 random models $\bar{R}_s^{\leq 50}$ on *perso*1 (top) and *perso*2 (bottom).

The significance of our results is measured using WER and using the 95% confidence interval. Confidence interval for the *perso*1 is 0.08 and 0.07 for *perso*2. Confidence interval means that if the improvement in WER exceed this value, we can consider it as significant improvement. The generic model G is considered as a good model as it is trained by mixing all the training data such that training is carried out on public dataset. WER of G evaluated on the test part of the two datasets *perso*1 and *perso*2 are presented in line 1 in Table 2. We observe that the values are significantly different. The second line in Table 2 presents results for the speaker model (the local model trained by fine-tuning using only 5 minutes of speech from the target speaker). Compared to the generic model in Table 2, speaker model improves the WERs for the two sets *perso*1 and *perso*2. It is also treated as a good model since it is trained using data from the target speaker.

Table 3 shows results obtained under various aggregation of acoustic model. Results are given for two configurations: with and without the local fine-tuned model of the target speaker.

Collaborative learning without the local fine-tuned model of the target speaker ($\alpha = 0$): The second line presents results of aggregation of all fine-tuned models except the target speaker. An improvement is shown compared to the generic model (from 15.43% to 15.21% for *perso*1 and from 14.84% to 14.62% for *perso*2). This improvement is significant since it exceeds the confidence interval value (the absolute gain for *perso*1 is 0.22 (0.22>0.08) and the same for *perso*2).

The third line in Table 3 presents results of aggregation of k-nearest neighbours models. The selection of k-nearest neighbours models does not improve results compared to the generic model and the model of all speakers (15.49% WER for *perso*1 and 14.81% WER for *perso*2). This selection is compared to a random selection. Results are presented in the line 4 in Table 3. Results of random selection are better than the results of dendrogram-based selection. Surprisingly, the benefit of using close models is not empirically demonstrated. This may be due to several factors. Either the distance is not reflecting a notion of usefulness or some amount of diversity is necessary to obtain models that behave well on new data. It should be noted that the first layer has maybe a too large number of parameters to compute a meaningful distance. We also tried to reduce the dimension of this vector, but the impact on the correlations with the WER (computed in a similar way than in Fig 2) was not observable.

Collaborative learning with the local fine-tuned model of the target speaker ($\alpha = 0.5$): As shown in Table 2, the fine-tuned model using a small local dataset improves WER compared to the generic model. So, we decide to take benefit from this improvement by aggregating models obtained with $\alpha = 0$ with the fine-tuned model of the target speaker (note that this model is trained with a very small dataset). Results in Table 3 shows a significant improvement in WER for all kinds of aggregation.

Acoustic models are prone to overfitting when the training dataset is limited. This could explain why the speaker model P_s cannot get very high performance,

since speakers' models are trained using only 5 minutes. Averaging speakers' models with the target speaker model allows us to produce a more accurate model than the target speaker model. The weight value used to combine the target speaker model P_s and an aggregated model has an influence on the resulting model combination. This is illustrated in Figure 3. Results in Figure 3 show also that with a good aggregated model, there is less need of the fine-tuned model P_s to get better results. This is particularly visible when combining P_s to the $\bar{B}_s^{\leq 10}$ aggregated model, made by averaging the ten models fine-tuned on other speakers that got the lowest WERs when applied to the target speaker data. This final combined model outperforms all the other ones, but its usage seems unrealistic since a local decoding process on the target user data is necessary for all the available non-target speaker models P'_s. However, these results provide good indications to continue this work on acoustic model collaborative personalization with privacy preservation constraints.

6 Conclusion

In this paper, we investigate a collaborative learning algorithm to locally improve the performance of an automatic speech recognition system, without sharing data (but models). In this purpose, we suggest to take benefit from acoustic models that have been separately fine-tuned for each user, in addition to the local fine-tuning on the target speaker data. Two kinds of local personalizations are explored, based on a fine-tuning processed on local data, and model averaging. Significant improvements are observed when these two local personalizations are combined through a weighted average. We also observed that a random selection of non-target speaker models gives better results than a non-naive approach. In a scenario where computations are not limited, a selection of non-target speaker models based on their performances – in terms of WER – on the target speaker data gives the best results.

7 Acknowledgements

This work was supported by the French National Research Agency under project DEEP-PRIVACY (ANR-18-CE23-0018).

References

1. Bell, P., Fainberg, J., Klejch, O., Li, J., Renals, S., Swietojanski, P.: Adaptation algorithms for neural network-based speech recognition: An overview. arXiv preprint arXiv:2008.06580 (2020)
2. Cui, X., Lu, S., Kingsbury, B.: Federated acoustic modeling for automatic speech recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6748–6752. IEEE (2021)

3. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798 (2010)
4. Dimitriadis, D., Kumatani, K., Gmyr, R., Gaur, Y., Eskimez, S.E.: A federated approach in training acoustic models. In: Proc. Interspeech (2020)
5. Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 3557–3568. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/24389bfe4fe2eba8bf9aa9203a44cdad-Paper.pdf
6. Gupta, V., Kenny, P., Ouellet, P., Stafylakis, T.: I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 6334–6338. IEEE (2014)
7. Leroy, D., Coucke, A., Lavril, T., Gisselbrecht, T., Dureau, J.: Federated learning for keyword spotting. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6341–6345. IEEE (2019)
8. Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M.A., Mtibaa, A., et al.: Preserving privacy in speaker and speech characterisation. Computer Speech & Language **58**, 441–480 (2019)
9. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF, IEEE Signal Processing Society (2011)
10. Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S.: Purely sequence-trained neural networks for asr based on lattice-free mmi. In: Interspeech. pp. 2751–2755 (2016)
11. Rousseau, A., Deléglise, P., Esteve, Y.: Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In: LREC. pp. 3935–3939 (2014)
12. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333. IEEE (2018)
13. Tomashenko, N., Srivastava, B.M.L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.F., Noé, P.G., et al.: Introducing the voiceprivacy initiative. arXiv preprint arXiv:2005.01387 (2020)
14. Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., Ramage, D.: Federated evaluation of on-device personalization (2019)
15. Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association **58**(301), 236–244 (1963)
16. Zhang, W., Cui, X., Finkler, U., Kingsbury, B., Saon, G., Kung, D., Picheny, M.: Distributed deep learning strategies for automatic speech recognition. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5706–5710. IEEE (2019)