

FEDERATED MARGINAL PERSONALIZATION FOR ASR RESCORING

Zhe Liu, Fuchun Peng

Facebook AI, Menlo Park, CA, USA

ABSTRACT

We introduce federated marginal personalization (FMP), a novel method for continuously updating personalized neural network language models (NNLMs) on private devices using federated learning (FL). Instead of fine-tuning the parameters of NNLMs on personal data, FMP regularly estimates global and personalized marginal distributions of words, and adjusts the probabilities from NNLMs by an adaptation factor that is specific to each word. Our presented approach can overcome the limitations of federated fine-tuning and efficiently learn personalized NNLMs on devices. We study the application of FMP on second-pass ASR rescoring tasks. Experiments on two speech evaluation datasets show modest word error rate (WER) reductions. We also demonstrate that FMP could offer reasonable privacy with only a negligible cost in speech recognition accuracy.

Index Terms— Federated learning, language modeling, automatic speech recognition, second-pass rescoring, personalization

1. INTRODUCTION

In recent years, there has been a rise in the popularity of a distributed learning technique called *federated learning* (FL) [1, 2, 3, 4]. It protects the privacy of data by training a shared machine learning model in a decentralized manner on users’ local devices, so that raw data never leaves physical devices. Each client model takes a global model from the central server for parameters initialization, and trains its own private local model using personal data. FL has been applied in many fields including recommendation [5], keyboard suggestion [6], keyword spotting [7], phenotyping [8] and health care [9].

Among these applications, *language modeling* is one of the most common tasks and serves as an important module in automatic speech recognition (ASR) [10]. In particular, neural network language models (NNLMs) typically outperform traditional n -gram language models in better keeping track of long range dependency [11, 12], and are widely used in the second-pass decoding via N -best or lattice rescoring [13].

A common issue arising after deploying an ASR model on user device is the discrepancy between training data and actual data received on local devices. In the case of language modeling, the language and style of real users’ utterances can

be very different from those of generic training corpus. The most general method to address this challenge is integrating a separate *personalized* language model trained on device in the FL framework. Particularly, the method of *federated fine-tuning* on private data has been explored in recent literature [14, 15, 16, 17], where we start with a general language model downloaded from the central server and have it continuously updated on devices using distributed parameter fine-tuning.

However, there exists several limitations confronting this federated fine-tuning method: (1) If each user only generates a very small number of utterances, such data by itself cannot be used for updating the general language model; (2) Fine-tuning a large pre-trained general language model on personal data tends to suffer from overfitting, or *catastrophic forgetting* [18]; (3) Training neural models on user devices typically has resource constraints and could be computationally expensive.

In this paper, we introduce *federated marginal personalization* (FMP), a novel approach for continuously updating personalized NNLMs on private devices using FL. Instead of fine-tuning the parameters of NNLMs on user personal data from first-pass ASR decoded words, FMP regularly estimates and updates global (server side) and personalized (client side) unigram distributions, and multiplies the probabilities from on-device NNLMs by a factor specific to each word. Then the resulting adapted language models are utilized in the second-pass ASR rescoring. Our proposed method can overcome the limitations of federated fine-tuning framework and efficiently learn personalized NNLMs. We also demonstrate that FMP satisfies utterance-level *differential privacy* (DP) [19, 20] with only a negligible cost in speech recognition accuracy.

The idea of leveraging word frequencies to bias language model probabilities was originally presented in [21], further studied in [22], and authors in [23] adopt such fast marginal adaptation framework to adapt recurrent neural network language models. In our work, we continue this line of research and take advantage of FL to interpolate global and personal marginals of word distributions for on-device NNLM personalization. To the best of our knowledge, our paper is the first one that leverages FL to explore fast marginal language model personalization with its application in ASR rescoring.

The rest of the paper is organized as follows. In Section 2, we introduce the FMP approach for ASR rescoring tasks. We evaluate the proposed method in Section 3 and demonstrate privacy analysis in Section 4. We conclude in Section 5.

2. METHODS

In this section, we describe the FMP approach on second-pass ASR rescoring tasks. To start with, we train a general NNLM $p_{\text{nnlm}}(w|h)$ using background corpus on the server side, and deploy this model to each physical device for second-pass N -best rescoring. Here, w stands for any word and h represents the context history. Let $u(w)$ be the discrete unigram distribution estimated from the background corpus. It is delivered to each local device as well along with the initial deployment of ASR model and second-pass NNLM rescorer.

Our approach can be outlined in Algorithm 1, with details provided in Sections 2.1, 2.2 and 2.3.

Algorithm 1: FMP approach for ASR rescoring.

```

hyper-parameters  $\lambda, \alpha, \beta, \sigma$ ;
initialize  $p_{\text{nnlm}}(w|h), u(w)$ ;
for each FL round  $t = 0, 1, 2, \dots$  do
  for each client  $i = 1, 2, \dots, n$  do
    if  $t = 0$  then
      |  $p_{i,\text{nnlm}}^t(w|h) \leftarrow p_{\text{nnlm}}(w|h)$ ;
    else
      |  $p_{i,\text{nnlm}}^t(w|h) \leftarrow$  according to Eq. (1);
    end
    conduct ASR rescoring using  $p_{i,\text{nnlm}}^t(w|h)$ ;
     $q_i^t(w) \leftarrow$  according to Eq. (4);
     $c_i^t \leftarrow$  according to Eq. (5);
  end
   $\bar{q}^t(w) \leftarrow$  according to Eq. (6);
end

```

2.1. Client-side model update

Each client i receives the global unigram distribution $\bar{q}^t(w)$ from server by the end of round t and performs the following update on client-side NNLM in round $t + 1$

$$p_{i,\text{nnlm}}^{t+1}(w|h) = \frac{1}{Z_i^t(h)} \cdot \left(\frac{g_i^t(w)}{u(w)} \right)^\lambda \cdot p_{\text{nnlm}}(w|h), \quad (1)$$

where

$$g_i^t(w) := (1 - \alpha - \beta) \cdot u(w) + \alpha \cdot \bar{q}^t(w) + \beta \cdot q_i^t(w). \quad (2)$$

Here, $q_i^t(w)$ is the personalized unigram distribution for user i estimated from ASR decoded text in an unsupervised manner, hyper-parameter $\lambda \geq 0$ controls the scaling factor of marginal adaptation, and $Z_i^t(h)$ is a normalization constant. We defer the estimation method of $q_i^t(w)$ to Section 2.2.

Notice that the numerator of the scaling factor, $g_i^t(w)$, is a linear interpolation of background unigram estimates $u(w)$, global unigram estimates $\bar{q}^t(w)$, and personalized unigram estimates $q_i^t(w)$, with interpolation weights of $1 - \alpha - \beta \geq 0$,

$\alpha \geq 0$ and $\beta \geq 0$, respectively. Here, our intuition is that the updated NNLM should respect the general words, *in-domain* or globally trending words, as well as personalized words that are particularly uttered by user i . For example, in voice search applications, once initial models are shipped to local devices, we use global unigram distribution to adapt live traffic, and personalized unigram distribution to account for user-level language and style. Also, we still require general background unigram distribution in the interpolation to prevent NNLMs from overfitting on the decoded text received on the devices.

For efficiency purposes in ASR rescoring tasks, we use an unnormalized version of Equation (1) to adjust the NNLM output word probabilities during second-pass rescoring.

2.2. Estimation of personalized unigram distribution

After a word is spoken, there is more chances that it is spoken again by the same user. For each user i , personalized unigram distribution $q_i^t(w)$ exploits the word distribution of historical context up to FL round t . In particular, we maintain a running “cache” that keeps track of the word counts from historical ASR decoded text, and estimate the corresponding unigram distribution by counts normalization and smoothing.

For unigram counting, we can leverage the decoded words from all the N -best hypotheses obtained from first-pass ASR decoding, with the use of *Gaussian kernel* weighting

$$K(s) = \exp(-(\text{rank}(s) - 1)^2 / (2\sigma^2)). \quad (3)$$

Here s represents an ASR decoded hypothesis, $\text{rank}(s)$ stands for the rank of s among the N -best list, and the bandwidth hyper-parameter $\sigma > 0$ controls the weighting scale. Notice that as σ approaches infinity, we approximately have uniform weights over N -best list; when σ is close to zero, we put zero weights on all hypotheses except the 1-best hypothesis.

Personalized unigram distribution can be estimated by

$$q_i^t(w) = \text{Smoothing} \left(\sum_{s \in S_i^t} K(s) \cdot c_s(w) / c_i^t \right), \quad (4)$$

$$c_i^t := \sum_w \sum_{s \in S_i^t} K(s) \cdot c_s(w), \quad (5)$$

where S_i^t contains all the hypotheses generated from user i by the end of round t , and $c_s(w)$ computes the count of word w among the hypothesis s . It is worth noting that in practice the personalized unigram distribution $q_i^t(w)$ can be estimated continuously to allow finer personalization, for example, it can be refreshed every time that the ASR model transcribes a new utterance on device, instead of being updated only once per each FL round.

In our application of second-pass ASR rescoring, NNLMs adapt the text data that is labeled by first-pass ASR decoding in an unsupervised manner. An alternative approach is to take advantage of the *soft labels* predicted by ASR models [24] and estimate unigram distributions accordingly. This method is beyond the scope of this study.

2.3. Server-side model update

After on-device estimation for round t , these locally updated unigram distributions $q_i^t(w)$ are sent to the central server for global aggregation. We adopt the following global update rule of *federated averaging* [3]

$$\bar{q}^t(w) = \frac{\sum_{i=1}^n c_i^t q_i^t(w)}{\sum_{i=1}^n c_i^t}, \quad (6)$$

where $q_i^t(w)$ is the estimated unigram distribution for user i and round t , and c_i^t is the corresponding sum of word pseudo-counts which serves as the weight for averaging. Notice that after updating, the global unigram distribution $\bar{q}^t(w)$ will be sent back to local devices and utilized in round $t + 1$.

In practical implementation, instead of sending each local personalized unigram distribution to the server, user devices can send distribution deltas [25], i.e., the difference between current personalized distribution and the global distribution before updating. Moreover, we typically sample only a subset of users before performing federated averaging.

The proposed FMP method is efficient and computationally cheap. Also, it only requires clients to exchange unigram frequencies with the server, which significantly reduces the data communication cost comparing to federated fine-tuning.

3. EXPERIMENTS

3.1. Datasets

In our experiments, the first-pass ASR model is trained using the in-house video ASR datasets (14K hours), which are sampled from public social media videos and de-identified before transcription; both transcribers and researchers do not have access to any user-identifiable information (UII). For second-pass rescoring, the general background text that we use to train NNLM is a corpus of public Facebook posts and comments, which contains around 30M English sentences.

We evaluate the proposed method on two speech datasets. The first is a *curated* set of carefully select very clean videos. Each video is segmented into multiple chunks of utterances. The second dataset is the *Augmented Multi-Party Interaction* (AMI) Meeting data [26]. It includes scenario meetings (with roles assigned for participants) and non-scenario meetings (where participants were free to choose topics). The sizes of these datasets are summarized in Table 1.

Table 1. Summary of speech evaluation datasets.

Feature	Evaluation Dataset	
	Curated Video	AMI Meeting
Num. of videos/meetings	78	63
Num. of utterances	1,015	12,643
Num. of words	74,248	89,666

3.2. Setups

For the first-pass ASR model, we use connectionist temporal classification (CTC) [27] criterion to learn an encoder-only model and is further composed with a 5-gram language model in a standard weighted finite-state transducers (WFST) framework. Here we adopt a latency-control bi-directional LSTM (LC-BLSTM) encoder with 6 layers of 1000 hidden units. For second-pass rescoring, we utilize a Transformer [28] based language model with word embeddings dimension 256, feed-forward network (FFN) dimension of 1024, 3 decoder blocks, 4 attention heads, and dropout of 0.10.

In our experiments, the baseline method is the first-pass ASR decoding with second-pass N -best rescoring using the general background NNLM. For each utterance, we generate 100-best hypotheses for *Curated Video* and 20-best hypotheses for *AMI Meeting* datasets. For each hypothesis, its NNLM score is linearly combined with the score from first-pass 5-gram language model using interpolation weight 0.50.

To simulate the server-client environments for evaluating FL based approaches, we treat each video in *Curated Video* data (or each meeting in *AMI Meeting* data) as a client, and any utterances that belong to the same video (or meeting) are thus considered as being received on devices and transcribed by client-side first-pass ASR model with second-pass NNLM rescoring. Utterances from the same video (or meeting) are ranked based on the starting timestamp of recording. For the proposed FMP method with total number of FL rounds being T (varied in our experiments), we evenly partition utterances from the same video (or meeting) into $T + 1$ groups, and assume all utterances from group t are received and processed in round $t - 1$ of FL, where $t = 1, \dots, T + 1$.

3.3. Evaluation results

We evaluate the proposed FMP method by ASR rescoring task on *Curated Video* and *AMI Meeting* datasets. We set hyper-parameters $\alpha = 0.5, \beta = 0.25$, and $\sigma = 5.0$; λ is tuned on a small validation set and kept as the same across all our experiments. We measure the impact of different choices of hyper-parameters in Section 3.4. Table 2 presents the word error rate (WER) results of FMP with various numbers of FL rounds T comparing to the baseline approach. We can see that FMP improves WERs consistently on both datasets (relatively 2.4% gain on *Curated Video* and 4.8% gain on *AMI Meeting* datasets). In particular, the improvement becomes slightly larger as T increases, which is expected since we can better leverage global marginals by more frequent updates.

3.4. Comparison of different hyper-parameters

Table 3 shows the WERs comparison results among various hyper-parameters of the proposed FMP method ($T = 2$ for *Curated Video* data and $T = 10$ for *AMI Meeting* data). For the interpolation weights α and β in marginal adaptation, we

Table 2. WERs from the baseline and FMP methods.

Method	Evaluation Dataset	
	Curated Video	AMI Meeting
Baseline	7.85	32.83
FMP w/ FL rounds $T = 1$	7.67 (-2.3%)	31.44 (-4.2%)
FMP w/ FL rounds $T = 2$	7.66 (-2.4%)	31.34 (-4.5%)
FMP w/ FL rounds $T = 5$	-*	31.27 (-4.8%)
FMP w/ FL rounds $T = 10$	-	31.24 (-4.8%)

* Not evaluated due to small number of utterances in each FL round.

can see that both of them play their crucial roles and perform better than only using either of them. Regarding the Gaussian kernel bandwidth σ , we can see that $\sigma = 1$ achieves the best results on both datasets, better than only using the 1-best hypothesis ($\sigma = 0.1$) or all N -best hypotheses with equal weights ($\sigma = 100$) to estimate personalized distributions.

Table 3. WERs comparison among various hyper-parameters (interpolation weights α, β ; bandwidth σ) of FMP method.

Hyper-Parameters of FMP		Evaluation Dataset	
		Curated Video	AMI Meeting
$\alpha = 0.5, \beta = 0.25$	$\sigma = 5$	7.66	31.24
	$\sigma = 0.1$	7.67	31.02
	$\sigma = 1$	7.65	30.96
	$\sigma = 10$	7.66	31.38
	$\sigma = 100$	7.67	31.42
$\alpha = 0.0, \beta = 0.75$	$\sigma = 5$	7.70	31.56
$\alpha = 0.75, \beta = 0.0$	$\sigma = 5$	7.72	31.31

4. PRIVACY ANALYSIS

A differentially private mechanism enables the public release of model parameters with a strong privacy protection [19, 20].

Definition 4.1 (DP) A randomized mechanism \mathcal{M} with a domain \mathcal{D} and range \mathcal{S} satisfies (ϵ, δ) -DP if for any two adjacent datasets $d, d' \in \mathcal{D}$ and for any subset $S \subseteq \mathcal{S}$, it holds that

$$P(\mathcal{M}(d) \in S) \leq e^\epsilon P(\mathcal{M}(d') \in S) + \delta. \quad (7)$$

Here d and d' are defined to be *adjacent* if d' can be formed by adding or removing a single training example from d .

In order to achieve DP, some randomness must be introduced to the algorithm. We use the *Laplace* mechanism [29] which adds a Laplace noise to the output of a query function

$$\bar{q}_{\epsilon, \text{DP}}^t(w) = \frac{\sum_{i=1}^n c_i^t q_i^t(w) + r_\epsilon^t(w)}{\sum_{i=1}^n c_i^t + \sum_w r_\epsilon^t(w)}, \quad (8)$$

$$r_\epsilon^t(w) \stackrel{\text{i.i.d.}}{\sim} \text{Laplace}(1/\epsilon), \quad (9)$$

where $\text{Laplace}(b)$ stands for a Laplace distribution with mean 0 and variance $2b^2$, and parameter ϵ controls the strength of privacy protection. Intuitively, a larger ϵ leads to stronger privacy protection, but can degrade the model accuracy.

Note that count has *sensitivity* 1, i.e. maximum difference in the query function output from adjacent datasets. For simplicity, we assume that any sensitive word is uttered at most once in any utterance. It is straightforward to show (see [29]) that randomized FMP with the server-side update provided in Equation (8) satisfies $(\epsilon, 0)$ -DP at utterance level.

It is worth noting that the definition of *adjacent datasets* in Definition 4.1 depends on the application. Most prior work on DP deals with example level (or utterance level in our case). For ASR tasks, a better definition is user-level adjacency for protecting whole user histories in the training set [14], since a sensitive word may be uttered several times by an individual user. In such case, we need counting capping to give upper bounds of sensitivity. We leave this for future work.

We evaluate our approach using ϵ values from 0.1 to 2.0. Here, we use the 1-best hypothesis for personalized unigram estimation ($\sigma = 0.1$), and set $T = 2$ for *Curated Video* data and $T = 10$ for *AMI Meeting* data. Table 4 shows the WERs comparison results, where we can see that randomized FMP can offer reasonable utterance-level privacy protection with a relatively small cost in speech recognition accuracy. Particularly, the WER degradation is less than 1% with $\epsilon \geq 0.5$.

Table 4. WERs comparison among various randomized FMP methods ($\alpha = 0.5, \beta = 0.25, \sigma = 0.1$) satisfying $(\epsilon, 0)$ -DP.

Method	Evaluation Dataset	
	Curated Video	AMI Meeting
FMP w/ DP not satisfied	7.67	31.02
FMP w/ (2.0, 0)-DP satisfied	7.67 (+0.0%)	31.09 (+0.2%)
FMP w/ (1.0, 0)-DP satisfied	7.67 (+0.0%)	31.17 (+0.5%)
FMP w/ (0.5, 0)-DP satisfied	7.68 (+0.1%)	31.26 (+0.8%)
FMP w/ (0.1, 0)-DP satisfied	7.74 (+0.9%)	31.59 (+1.8%)

5. CONCLUSION

In this work, we leverage FL to study fast marginal language model personalization with its application in ASR rescoring. Our approach could overcome the cold start and catastrophic forgetting issues confronting traditional federated fine-tuning, and efficiently learn personalized NNLMs on local devices. Experiments on two speech evaluation datasets show modest WER reductions. In the future, we plan to explore higher order of n -gram distribution for deeper personalization.

6. ACKNOWLEDGEMENTS

Thank Kshitiz Malik and Yutong Pang for discussions.

7. REFERENCES

- [1] Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik, “Federated optimization: Distributed machine learning for on-device intelligence,” *arXiv preprint arXiv:1610.02527*, 2016.
- [2] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, “Federated learning: Strategies for improving communication efficiency,” *NeurIPS Workshop on Private Multi-Party Machine Learning*, 2016.
- [3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *AISTATS*, 2017.
- [4] Robin C Geyer, Tassilo Klein, and Moin Nabi, “Differentially private federated learning: A client level perspective,” *arXiv preprint arXiv:1712.07557*, 2017.
- [5] Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He, “Federated meta-learning with fast convergence and efficient communication,” *arXiv preprint arXiv:1802.07876*, 2018.
- [6] Kenneth C Arnold, Krzysztof Z Gajos, and Adam T Kalai, “On suggesting phrases vs. predicting words for mobile text composition,” in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016.
- [7] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau, “Federated learning for keyword spotting,” in *Proc. ICASSP*, 2019.
- [8] Yejin Kim, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang, “Federated tensor factorization for computational phenotyping,” in *Proc. SIGKDD*, 2017.
- [9] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang, “Federated learning for healthcare informatics,” *Journal of Healthcare Informatics Research*, pp. 1–19, 2020.
- [10] Joshua Goodman, “A bit of progress in language modeling,” *Computer Speech and Language*, vol. 15, pp. 403–434, 2001.
- [11] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur, “Recurrent neural network based language model,” in *Proc. Interspeech*, 2010.
- [12] Xie Chen, Xunying Liu, Mark JF Gales, and Philip C Woodland, “Improving the training and evaluation efficiency of recurrent neural network language models,” in *Proc. ICASSP*, 2015.
- [13] Hainan Xu, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur, “A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition,” in *Proc. ICASSP*, 2018.
- [14] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang, “Learning differentially private recurrent language models,” in *Proc. ICLR*, 2018.
- [15] Vadim Popov, Mikhail Kudinov, Irina Piontkovskaya, Petr Vytovtov, and Alex Nevidomsky, “Distributed fine-tuning of language models on private data,” in *Proc. ICLR*, 2018.
- [16] Shaoxiong Ji, Shirui Pan, Guodong Long, Xue Li, Jing Jiang, and Zi Huang, “Learning private neural language modeling with attentive aggregation,” in *IJCNN*. IEEE, 2019, pp. 1–8.
- [17] Mingqing Chen, Ananda Theertha Suresh, Rajiv Mathews, Adeline Wong, Cyril Allauzen, Françoise Beaufays, and Michael Riley, “Federated learning of n-gram language models,” *arXiv preprint arXiv:1910.03432*, 2019.
- [18] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio, “An empirical investigation of catastrophic forgetting in gradient-based neural networks,” *arXiv preprint arXiv:1312.6211*, 2013.
- [19] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.
- [20] Cynthia Dwork and Aaron Roth, “The algorithmic foundations of differential privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.
- [21] Reinhard Kneser, Jochen Peters, and Dietrich Klakow, “Language model adaptation using dynamic marginals,” in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [22] Mittul Singh, Youssef Oualil, and Dietrich Klakow, “Approximated and domain-adapted LSTM language models for first-pass decoding in speech recognition,” in *Proc. Interspeech*, 2017.
- [23] Ke Li, Hainan Xu, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, “Recurrent neural network language model adaptation for conversational speech recognition,” in *Proc. Interspeech*, 2018.
- [24] Sungho Shin, Kyuyeon Hwang, and Wonyong Sung, “Generative knowledge transfer for neural language models,” *arXiv preprint arXiv:1608.04077*, 2016.
- [25] Duc Bui, Kshitiz Malik, Jack Goetz, Honglei Liu, Seungwhan Moon, Anuj Kumar, and Kang G Shin, “Federated user representation learning,” *arXiv preprint arXiv:1909.12535*, 2019.
- [26] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., “The ami meeting corpus: A pre-announcement,” in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.
- [27] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proc. ICML*, 2006.
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Proc. NeurIPS*, 2017.
- [29] Cynthia Dwork, “Differential privacy: A survey of results,” in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.