ACCELERATED UNSUPERVISED CLUSTERING IN ACOUSTIC SENSOR NETWORKS USING FEDERATED LEARNING AND A VARIATIONAL AUTOENCODER

Luca Becker, Alexandru Nelus, Rene Glitza, and Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
{firstname.lastname}@rub.de

2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914753

ABSTRACT

In this paper we present an accelerated algorithm for clustering source-dominated microphones in acoustic sensor networks. Predicated on privacy-preserving unsupervised clustered federated learning that groups microphones by evaluating the similarity of model weight updates, we introduce a light-weight variational autoencoder and equip the algorithm with supplementary control criteria for faster convergence. We validate the quality, degree of acceleration and utility of our method using clustering-based and classification-based tasks. Compared to the previously employed deterministic autoencoder, we observe a significantly lower number of client-server communication rounds at the price of a minor reduction in clustering performance.

Index Terms— clustered federated learning, DNN, autoencoder, acoustic sensor networks, distributed classification

1. INTRODUCTION

Acoustic sensor networks (ASNs) enjoy increasing attention thanks to their capability of exploiting information from various wide-spread acoustic sensors. They provide the means for numerous use cases [1], such as source localization [2, 3], speech enhancement [4], and activity monitoring [5]. Further augmentation of ASN-based applications can be achieved by grouping acoustic sensors into clusters which are dominated by a specific source ("source-dominated clusters") [6, 7].

In order to implement the aforementioned augmentation in a privacy-aware fashion and without the need of prior knowledge regarding the number of active sources, we have proposed to use unsupervised clustered federated learning (UCFL) [8, 9]. The latter is derived from clustered federated learning (CFL) [10, 11, 12] and has been successfully used for clustering sensor nodes in smart-home environments [9]. Generally, CFL trains a neural network (NN) model at node-level and re-groups nodes predicated on the similarity of their weight updates. CFL has been mainly designed for supervised classification scenarios with abundant training data [12]. Therefore, it is not trivial to implement CFL in ASN scenarios that require clustering results based on short segments of audio and that lack training labels. For this reason, UCFL introduces a light-weight autoencoder, where only a designated bottleneck layer is trained by the ASN nodes, circumventing thus the need for vast amounts of labeled data. In this work we further improve the UCFL procedure by significantly reducing the amount of algorithm iterations needed for convergence. This is achieved by replacing the deterministic autoencoder model by a variational autoencoder [13], where weight updates carry a stochastic component that is further used to accelerate the method.

For evaluating our method, we make use of a complex simulated acoustic scenario introduced in [9]. The UCFL steering criteria calibrated for the high acoustic variability encountered in [9] are further amended here in order to comply with the updated algorithm. The evaluation of clustering contains distance-based metrics, measures for convergence speed and a scenario-wide gender recognition task, accounting for clustering quality, degree of acceleration and utility, respectively.

This paper is structured as follows: in Sections 2, 3, and 4, we provide an overview of related publications, summarize UCFL, discuss our intuition on the VAE in UCFL, and motivate a novel convergence criterion. In Section 5, we briefly outline the experimental setup, the utilized evaluation metrics, and underlying data sets. Finally, we present experimental results and conclusions in Sections 6 and 7, respectively.

2. RELATION TO PRIOR WORK

Previous investigations on microphone clustering are based on e.g., the coherence function [14, 15], eigenvalue decomposition [6], power spectral densities [16] or cepstral features [7, 17]. Although being efficient, they unfortunately lack privacy considerations. As an alternative, Nelus et al. [8, 9] successfully extended CFL [10, 11, 12] towards an unsupervised, privacy-aware clustering procedure.

Nonetheless, the method described in [8, 9] suffers from a relatively large number of communication and re-training rounds which are costly on nodes with limited computational power and consequently discourage potential applications in realistic smart-home systems.

This work has been supported by the German Research Foundation (DFG) - Project Number 282835863.

In addition to that, variational autoencoders have been successfully incorporated into *federated learning* (FL) for recommendation systems [18] and anomaly detection [19]. However, these approaches are limited to FL and are not evaluated in the context of CFL.

3. UNSUPERVISED CLUSTERED FEDERATED LEARNING

As introduced in [8] and further refined for more complex multi-source environments in [9], unsupervised clustered federated learning is an extension of clustered federated learning (CFL) [10, 11, 12] that may be used to cluster ASN nodes when labeled data is limited. The goal of CFL, and consequently UCFL, is to organize nodes $\mathrm{n}_1, \ldots, \mathrm{n}_M$ with *incongruent* data distributions into separate clusters. For this purpose, Sattler et al. [10] propose computing node weight update vectors $\Delta\boldsymbol{\theta}_i$ using *stochastic gradient descent* (SGD) and assemble them in the *cosine similarity matrix* $\boldsymbol{A} = \Delta\boldsymbol{\theta}^T \Delta\boldsymbol{\theta}$, where $\Delta\boldsymbol{\theta} = (\Delta\boldsymbol{\theta}_1/\|\Delta\boldsymbol{\theta}_1\|, \ldots, \Delta\boldsymbol{\theta}_M/\|\Delta\boldsymbol{\theta}_M\|)$ contains the L_2-normalized weight update vectors [9, 10]. In the case of UCFL, the weight updates $\Delta\boldsymbol{\theta}_i$ of each node n_i are extracted from a bottleneck layer of an autoencoder model (see Section 4.1) that is pre-trained on log-mel band energy features (LMBE). Afterwards, the nodes transmit their weight updates to the *server* which then computes the cosine similarity matrix. Predicated on the cosine similarity matrix, the server bi-partitions the nodes into two emerging clusters c_1 and c_2 with corresponding cosine similarity matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, such that the minimum intra-cluster similarity is larger than the maximum inter-cluster similarity.

In order to evaluate whether a cluster c_k contains congruent nodes, the mean and maximum L_2-norms of the nodes' weight update vectors are computed as [10]

$$\Delta\bar{\theta}_{c_k} = \left\| \frac{1}{|c_k|} \sum_{\mathrm{n}_i \in c_k} \Delta\boldsymbol{\theta}_i \right\| \text{ and } \Delta\hat{\theta}_{c_k} = \max_{\mathrm{n}_i \in c_k} (\|\Delta\boldsymbol{\theta}_i\|) \,. \quad (1)$$

Bi-partitioning of the cluster c_k is triggered when the conditions $\Delta\bar{\theta}_{c_k} \leq \varepsilon_1$ and $\Delta\bar{\theta}_{c_k}/\Delta\hat{\theta}_{c_k} \geq \varepsilon_2$ are met, implying that the cluster is close to a stationary solution while a subset of nodes still has large individual weight updates and thus, the cluster is *incongruent*. After assigning nodes to clusters $\mathrm{n}_i \rightarrow c_k$, the server determines average cluster-wise weight updates $\Delta\boldsymbol{\theta}_{c_k} = 1/|c_k| \sum_{\mathrm{n}_i \in c_k} \Delta\boldsymbol{\theta}_i$, where $|c_k|$ denotes the cardinality of set c_k. These cluster-wise weight updates are applied on the server-side and the new weights $\boldsymbol{\theta}_{c_k}$ are transmitted to the nodes $\mathrm{n}_i \in c_k$ which replace their old weights by $\boldsymbol{\theta}_{c_k}$. We provide a visualization of data transmission in UCFL in Fig. 1.

This procedure is repeated until $\tau = \tau_{\max}$ rounds are reached or a convergence criterion is verified. After the first round, the first threshold is dynamically initialized as $\varepsilon_1 = \Delta\bar{\theta}_{c_1} + 0.05\Delta\hat{\theta}_{c_1}$.

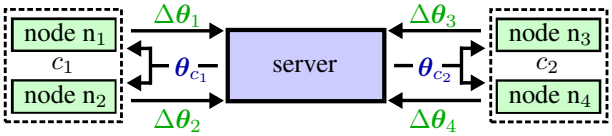

Fig. 1: Data transmission in UCFL for $M = 4$ nodes and two potential clusters c_1 and c_2.

The above considerations result in a set of clusters $C = \{c_1, \ldots, c_K\}$. In addition, Nelus et al. provide membership values (MVs) that express a node's confidence towards the assigned cluster. MVs can have a beneficial impact on further clustering-based applications [17] (see Sec. 6.2).

4. ACCELERATED UCFL

4.1. Variational autoencoder in UCFL

Previous work on UCFL [8, 9] investigated the case where the weight updates $\Delta\boldsymbol{\theta}_i$ are extracted from the bottleneck of a deterministic autoencoder (AE). This work, on the other hand, employs a *variational* autoencoder (VAE) [13] instead and exploits its stochastic properties in order to accelerate the clustering process.

A VAE consists of a probabilistic encoder $q_\psi(\boldsymbol{z}|\boldsymbol{x})$ and a probabilistic decoder $p_\psi(\boldsymbol{x}|\boldsymbol{z})$ that are realized using neural networks with encoder and decoder parameters ψ, input representation $\boldsymbol{x}$ and latent representation $\boldsymbol{z}$ that is determined via stochastic sampling from the encoder. For training, Kingma et al. [13] suggest to maximize the *evidence lower bound* (ELBO) which is given by

$$\arg\max_{\psi} \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x})}[\log(p_\psi(\boldsymbol{x}|\boldsymbol{z}))] - D_{\mathrm{KL}}[q_\psi(\boldsymbol{z}|\boldsymbol{x}) \| p_\psi(\boldsymbol{z})], \quad (2)$$

where D_{KL} is the Kullback-Leibler-Divergence and $p_\psi(\boldsymbol{z})$ is the prior of the latent representation $\boldsymbol{z}$. In this work, we assume that the latent space can be modeled via multi-variate Gaussians so that $\boldsymbol{z}$ can be expressed as

$$\boldsymbol{z} = \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\sigma}_{\boldsymbol{x}} \odot \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad (3)$$

where $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\boldsymbol{\sigma}_{\boldsymbol{x}}$ denote the latent mean and standard deviation vector, respectively. As in [8] and [9], we compute the nodes' weight updates $\Delta\boldsymbol{\theta}_i$ by applying stochastic gradient descent (SGD) onto the parameters of the bottleneck of a pre-trained VAE model h while the remaining layers are frozen (application of weight updates is suppressed). Although both layers generating $\boldsymbol{\sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\mu}_{\boldsymbol{x}}$ can be adapted, we utilize only the layer that produces $\boldsymbol{\sigma}_{\boldsymbol{x}}$ as the bottleneck for the weights $\boldsymbol{\theta}_i$ because using the means did not give better results in preliminary experiments. Thus, the layer generating $\boldsymbol{\mu}_{\boldsymbol{x}}$ is frozen as well. Due to the fact that SGD applied to $\boldsymbol{\theta}_i$ partially relies on the latent representation $\boldsymbol{z}$, our intuition is that this stochasticity further influences the corresponding weight updates $\Delta\boldsymbol{\theta}_i$ and thus, weight updates tend to be more dissimilar, prompting more frequent bi-partitioning as compared to [8, 9].

4.2. Suppressing re-training

Considering that calculating $\Delta\boldsymbol{\theta}_i$ in each communication round can be computationally expensive for the nodes, we now introduce a transition criterion that determines when the nodes no longer re-compute and transmit $\Delta\boldsymbol{\theta}_i$. This way, the server exclusively performs cosine similarity evaluation and bi-partitioning on the most recent weight updates it received from the nodes. For this purpose, we make use of the Frobenius norm of the difference between the (global) cosine similarity matrix at round τ and its predecessor as

$$\|\boldsymbol{A}^{(\tau)} - \boldsymbol{A}^{(\tau-1)}\|_{\mathrm{F}} < \varepsilon_3, \tag{4}$$

indicating that the weight updates $\Delta\boldsymbol{\theta}_i$ are no longer depending on τ and thus, further re-computations and transmissions, as depicted in Fig. 1, are not necessary.

4.3. Intra- and cross-cluster cosine similarity

In order to formulate a dynamic convergence criterion for UCFL, we make use of the short-time average intra-cluster similarity $\bar{\mu}^{(\tau)}_{\mathrm{intra}}$ and short-time average cross-cluster similarity $\bar{\mu}^{(\tau)}_{\mathrm{cross}}$ defined as

$$\bar{\mu}^{(\tau)}_{\mathrm{intra}} = \frac{1}{|C|}\sum_{c_k\in C}\frac{1}{|c_k|(|c_k|-1)}\sum_{\mathrm{n}_i\in c_k}\sum_{\mathrm{n}_j\in c_k\setminus\{\mathrm{n}_i\}} A^{(\tau)}_{i,j}, \text{ and} \tag{5}$$

$$\bar{\mu}^{(\tau)}_{\mathrm{cross}} = \frac{1}{|C|(|C|-1)}\sum_{c_k\in C}\sum_{c_l\in\{C\setminus c_k\}}\frac{1}{|c_k||c_l|}\sum_{\mathrm{n}_i\in c_k,\mathrm{n}_j\in c_l} A^{(\tau)}_{i,j}. \tag{6}$$

These are re-computed in each round τ. We argue that a large $\bar{\mu}^{(\tau)}_{\mathrm{intra}}$, along with a small $\bar{\mu}^{(\tau)}_{\mathrm{cross}}$ indicate good clustering. Consequently, we define a criterion $|\bar{\mu}^{(\tau)}_{\mathrm{intra}} - \bar{\mu}^{(\tau)}_{\mathrm{cross}}| < \varepsilon_4$ that terminates UCFL when the absolute difference between $\bar{\mu}^{(\tau)}_{\mathrm{intra}}$ and $\bar{\mu}^{(\tau)}_{\mathrm{cross}}$ falls below a threshold ε_4.

5. EXPERIMENTAL SETUP

5.1. Scenario description

We evaluate our proposed method under the same experimental conditions as in [9]. Thereby, a virtual apartment model (43.5 m^2, five rooms) with $Z = 4$ randomly positioned acoustic sources and 41 microphone nodes is deployed. Under these circumstances, we model the microphone signal $x_i(t)$ of the i-th node,

$$x_i(t) = \sum_{z=1}^{Z} h_{z,i}(t) * s_z(t)\,, \tag{7}$$

as a superposition of all source signals $s_z(t)$, each of which convolved $(*)$ with the corresponding room impulse response (RIR) $h_{z,i}(t)$ from source z to node n_i. These microphone signals are then further transformed into log-mel band energy features (LMBE) with the same configuration as in [8, 9].

In [9], Nelus et al. proposed a set of RIRs based on ten distinct source-node constellations. These are used in conjunction with 20 gender-balanced speaker groups from the dataset described in Section 5.2 to create 200 unique simulated scenarios. For UCFL we empirically set the parameters $\varepsilon_1 = 0.045$, $\varepsilon_2 = 0.84$, $\varepsilon_3 = 3$, $\varepsilon_4 = 0.55$ and $\tau_{\max} = 9$. The structure of the VAE h used for UCFL is based on the AE model described in [9] (Table I). The dense bottleneck layer of the AE is replaced in the VAE by two dense layers that produce vectors $\boldsymbol{\sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\mu}_{\boldsymbol{x}}$. These are required for the re-parameterization described in (3) and each consists of 29 neurons with ReLU activation. Consequently, the weight update vector $\Delta\boldsymbol{\theta}_i$ of the layer generating $\boldsymbol{\sigma}_{\boldsymbol{x}}$ has 870 elements which is equal to the bottleneck size in [9].

5.2. Database

For comparability, the database used in this work is identical to the database used in [8] and [9]. It is derived from the *train-clean-100* subset that is part of the LibriSpeech corpus [20] and sampled at $f_s = 16$ kHz. It contains audiobooks of 125 female and 126 male speakers and is further processed into 25006 utterances of 10s duration using voice activity detection (VAD). This database is subdivided into *Libri-server* and *Libri-client*, whereas the first contains 78 female and 79 male speakers for pre-training the VAE h and a gender recognition model (see Section 5.3.3), and the latter contains the remaining speakers for UCFL and gender recognition inference.

5.3. Evaluation measures

5.3.1. Cluster-to-source distance

As in [17], we make use of the normalized cluster-to-source (CTS) distance $\tilde{d}^{s_z}_{c_k}$ that is generalized for multi-source environments [8, 9] as $\tilde{d}^{s_z}_{c_k} = \|\boldsymbol{\rho}_{s_z} - \boldsymbol{\rho}_{c_k}\|/\bar{d}_S$, where $\boldsymbol{\rho}_{s_z}$ is the position of source s_z, $\boldsymbol{\rho}_{c_k}$ denotes the geometrical mean of nodes in the cluster c_k, weighted by MVs and $\bar{d}_S$ represents the average of all unique source-pair distances.

5.3.2. Convergence speed

To quantify the impact of our method with respect to the acceleration of UCFL, we measure the number of rounds τ for each simulated scenario until convergence. Additionally, for each scenario, we mark the round after which (4) holds true, implying that no more federated training is performed for the remaining rounds.

Table 1: Normalized cluster-to-source distance $\tilde{d}_{c_k}^{s_z}$ from cluster c_k to source s_z, averaged over 200 simulated scenarios (clusters c_{13} - c_{16} are neglected).

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
s_1	**0.19**	0.76	0.79	0.96	0.32	0.76	0.73	0.84	0.38	0.59	0.71	0.63
s_2	0.73	**0.25**	1.2	1.0	0.78	0.35	1.0	0.92	0.76	0.46	0.93	0.84
s_3	0.78	1.0	**0.25**	0.92	0.77	0.95	0.44	0.82	0.86	0.9	0.57	0.76
s_4	1.0	0.98	0.93	**0.25**	1.0	0.94	0.8	0.46	1.0	0.98	0.74	0.78

5.3.3. Gender recognition

Aiming to compare the performance of our method with prior work [9], we perform gender recognition on every node's microphone signal $x_i(t)$. The ground-truth label for $x_i(t)$ is assigned by the gender label of the source signal whose direct component has the shortest delay in the acoustic transmission path towards that node. Cluster-wise predictions are given by the MV-weighted average of all node predictions and ground-truth labels are determined by the majority vote of all nodes inside the cluster. For evaluation purposes, we compute the F_1-score (F_1) and Accuracy (A_{cc}) for the four clusters with the highest classification confidence. Gender recognition is performed for all simulated scenarios and the NN-classifier is the same as in [9], except that the model is trained for 27 epochs and the dropout factor is increased to 0.7 with test metrics $(A_{\text{cc}}, F_1) = (99.1\%, 98.2\%)$ on convolved but not superposed signals.

6. EXPERIMENTAL RESULTS

6.1. Clustering results

For each simulated scenario, UCFL produces a variable amount of up to 16 clusters. Based on this observation, we compute the CTS-distance for each cluster-source pair and sort the clusters c_k depending on their distances to sources $s_1 - s_4$. The CTS-distances for the first 12 clusters are presented in Table 1. A small value for $\tilde{d}_{c_k}^{s_z}$ indicates that a cluster c_k is close to a source s_z, whereas $\tilde{d}_{c_k}^{s_z} > 1$ implies that the distance between the cluster and source is larger than the average distance between sources [8, 9]. Especially the first four clusters underline this notion: while having a relatively small distance to one source, each of these clusters exhibits a large distance to the other sources. Hence, these clusters are dominated by one source and can further be used for clustering-based applications. Comparing these results with [9], we find that the proposed VAE approach gives slightly larger distances between the geometrical means of nodes (centroids) and their corresponding sources.

Nevertheless, since our method aims at accelerating UCFL, Figure 2 (left) provides a comparison of effective rounds τ_{eff}, after which UCFL converges for both the proposed method and the procedure described in [9]. It can be observed that the majority of simulations using the baseline method requires up to 11 rounds, while the proposed approach does not require more than nine rounds. We also

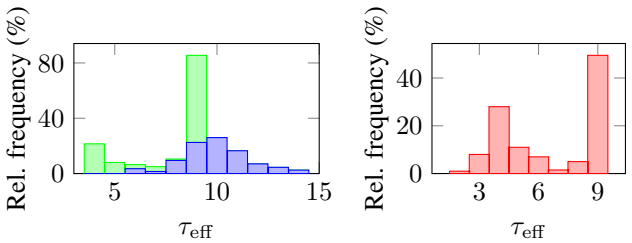

Fig. 2: Left: relative frequencies of effective rounds τ_{eff} required for UCFL to converge for the baseline method ([9], blue, 6% outliers with $\tau_{\text{eff}} > 15$ are neglected) and our accelerated method (green, for clarity: bars are stacked). Right: relative frequencies of τ_{eff} after which node weight updates are kept fixed. Graphs summarize results from 200 scenarios.

Table 2: Gender recognition Accuracy (A_{cc}) and F_1-score (F_1) of estimated clusters, without and with membership value (MV) weighting using threshold v, averaged over 200 scenarios. Average node-based metrics without clustering are provided as a reference (no clust.).

	no clust.	no MV	$v = 0$	$v = 0.3$	$v = 0.4$	$v = 0.6$	$v = 0.8$	$v = 0.9$
VAE-A_{cc} (%)	84.9	87.9	93.7	93.6	93.8	94.5	**95.9**	95.2
VAE-F_1 (%)	83.2	87.7	93.2	93.1	93.4	94.1	**95.6**	94.9
AE-A_{cc} (%)	82.9	84.7	95.3	95.7	95.9	96.2	**96.4**	**96.4**
AE-F_1 (%)	81.3	84.0	95.1	95.5	95.7	96.1	**96.3**	**96.3**

evaluated $\tau_{\max} = 9$ for the AE which, however, did not yield convincing results. Moreover, Fig. 2 (right) reveals that in roughly fifty percent of simulated scenarios, costly training on nodes is suppressed as described in (4) before the last round. Consequently, UCFL is further accelerated.

6.2. Gender recognition results

As in [8, 9, 17], we utilize gender recognition as an objective measure to evaluate the utility of our method. A detailed description of the metrics is given in Section 5.3.3. In Table 2, we showcase gender recognition results for the proposed and the baseline methods with a supplementary MV-threshold v, i.e. nodes with MVs smaller than v are discarded. As seen in Table 2, using MVs provides a significant advantage for gender recognition performance. Nonetheless, the evaluation metrics for the baseline method are around 0.5 % higher, when compared to the proposed method, indicating along with the larger distances of sources to cluster centroids, that the acceleration of UCFL comes with a noticeable, yet small trade-off in quality and utility.

7. CONCLUSIONS

We proposed and evaluated an accelerated adaption of UCFL that leads to a large reduction of iterations required for the algorithm to converge. Even though a small decrease in performance is observed, we regard it as an acceptable trade-off w.r.t the reduction of clustering costs. Consequently, this indicates a strong potential towards more realistic and real-time applications of our method on low-cost devices.

8. REFERENCES

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011, pp. 1–6.

[2] Y. Yan, X. Shen, F. Hua, and X. Zhong, "On the semidefinite programming algorithm for energy-based acoustic source localization in sensor networks," *IEEE Sensors Journal*, vol. 18, no. 21, pp. 8835–8846, 2018.

[3] A. Brendel and W. Kellermann, "Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 61–75, 2019.

[4] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[5] J. Ebbers, M. Keyser, and R. Haeb-Umbach, "Adapting sound recognition to a new environment via self-training," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1135–1139.

[6] M. Bahari, L. Hamaidi, M. Muma, J. Plata-Chaves, M. Moonen, A. Zoubir, and A. Bertrand, "Distributed multi-speaker voice activity detection for wireless acoustic sensor networks," *arXiv:1703.05782*, 2017.

[7] S. Gergen, R. Martin, and N. Madhu, "Source separation by fuzzy-membership value aware beamforming and masking in ad hoc arrays," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[8] A. Nelus, R. Glitza, and R. Martin, "Estimation of microphone clusters in acoustic sensor networks using unsupervised federated learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 761–765.

[9] A. Nelus, R. Glitza, and R. Martin, "Unsupervised clustered federated learning in complex multi-source acoustic environments," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1115–1119.

[10] R. Sattler, K. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3710–3722, 2021.

[11] F. Sattler, S. Wiedemann, K. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, pp. 1–8, 2019.

[12] F. Sattler, K. Müller, T. Wiegand, and W. Samek, "On the byzantine robustness of clustered federated learning," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. 2020, pp. 8861–8865, IEEE.

[13] D. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[14] I. Himawan, I. McCowan, and S. Sridharan, "Clustering of ad-hoc microphone arrays for robust blind beamforming," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 2814–2817.

[15] S. Pasha, J. Donley, and C. Ritz, "Blind speaker counting in highly reverberant environments by clustering coherence features," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1684–1687.

[16] Y. Zhao, J. Nielsen, J. Chen, and M. Christensen, "Model-based distributed node clustering and multi-speaker speech presence probability estimation in wireless acoustic sensor networks," *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 4189–4201, 2020.

[17] S. Gergen, A. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Process.*, vol. 107, pp. 21–32, 2015.

[18] M. Polato, "Federated variational autoencoder for collaborative filtering," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.

[19] Z. Gu and Y. Yang, "Detecting malicious model updates from federated learning on conditional variational autoencoder," in *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2021, pp. 671–680.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.